Towards Reliable Causal Machine Learning for Macroeconomics

David Bruns-Smith

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-168
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-168.html

August 9, 2024

Towards Reliable Causal Machine Learning for Macroeconomics

By

David A Bruns-Smith

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Avi Feller, Co-chair
Professor Jacob Steinhardt, Co-chair
Professor Emi Nakamura

Summer 2024

Towards Reliable Causal Machine Learning for Macroeconomics

Abstract

Towards Reliable Causal Machine Learning for Macroeconomics

By

David A Bruns-Smith

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Avi Feller, Co-chair

Professor Jacob Steinhardt, Co-chair

The 21st century has seen an explosion in the availability of economic data, and machine learning tools for making predictions from that data. Motivated by these developments, in this dissertation, I consider the broad question of: what, if anything, can machine learning contribute to macroeconomic policy-making? In Chapter 2, I begin with a case study of a pure prediction problem in Icelandic tax data, and show that machine learning is quantitatively and qualitatively useful for this problem. But in economic policy settings, we want to predict the effect of an intervention, a much more challenging problem than the standard supervised learning task. Therefore, the rest of my dissertation focuses on using machine learning for observational causal inference. In Part 2, I consider the "no unobserved confounders" case, where we assume that we observe all of the relevant covariates. In this setting, the causal inference problem reduces to a prediction task under covariate shift, and we can debias causal effect estimates using the density ratio - the object that measures how the covariate distributions shift. In high dimensions, density ratios are typically not well behaved, and I help make progress on this front in Chapter 3 by drawing connections between density ratio estimation and so-called "balancing weights" estimators via a duality argument. Then in Chapter 4, I apply these results to obtain a broad set of numerical equivalence results for debiased machine learning estimators, which results in a number of implications for undersmoothing and hyperparameter tuning in practice. In Part 3, I turn to the setting where we do potentially have unobserved confounders, making unbiased recovery of the causal effect impossible. Instead, we use "sensitivity analysis", which measures how quickly the estimated causal relationship degrades with hypothetical confounding. Of particular relevance to macroeconomics, I develop algorithms for the dynamic setting where causal effects unroll over time, adopting the Reinforcement Learning framework. Chapter 5 considers the tabular setting, and Chapter 6 extends these results to function approximation with machine learning.

Contents

List of Figures

List of Tables

Acknowledgments

I took an unusual path through my Computer Science PhD. I began by building hardware accelerators for gene sequencing applications, and ultimately found a home studying causal inference and macroeconomics. Pulling this off required a tremendous amount of support — both in terms of research, and administratively — and so there are number of people I have to thank.

I will begin with my two wonderful advisors; I managed to win the academic jackpot not once but twice. Avi Feller supported me from the very beginning of my transition into social science research, ever since I reached out to him after a causal inference reading group meeting. This began administratively ("you're a very unusual case"), but after many casual causal conversations, Avi became my main methodological mentor. I learned a lot from his breadth of methodological and applied experience, but also his focus on getting a precisely scoped research question, and his relentless focus on the main takeaway when writing papers. I really took advantage of Avi's care and availability, and I enjoyed our near daily conversations on Zoom, over Slack, and in person. I hope to continue to enjoy them!

My second advisor, Emi Nakamura, taught me economics essentially from scratch. When I set out to do economic policy research, I emailed a dozen or so economics professors to see if they were interested in collaborating with a computer science PhD student. Emi was the only one to respond, and we had a meeting on the calendar within days. She took a very uncertain bet on working with me, including funding me after only a semester of getting to know me, and I'm forever grateful for that. I appreciate her guidance over a variety of theoretical and empirical projects, and her mentorship has fundamentally (and maybe even idiosyncratically) defined how I conceptualize the practice of economics. I am deeply thankful for that.

I had a number of other mentors and collaborators along the way. My undergraduate advisor and first boss, Rich Lethin, encouraged me to go to grad school and pursue academia. It was probably through him that I was admitted to Berkeley in the first place. Krste Asanovic and Lisa Wu Wills were my computer architecture mentors upon arriving at Berkeley, and they taught me the importance of always building the real thing. Jon Steinsson, aside from ongoing mentorship on a number of empirical projects, emphasized for me why one would ever want to do economic theory, and was also the first person to lead me through teaching undergraduates. Angela Zhou is a wonderful collaborator, but also one of my biggest inspirations: I would like to match her unusual combination of deep technical work with actually meaningful applied projects. Betsy Ogburn and Oliver Dukes taught me most of what I know about semiparametric statistics, and were always wonderful to chat with during our long group meetings. Vira Semenova showed me that sometimes theory can help you solve what look like purely computational problems. I'm thankful for all of the time everyone in the list above spent to help me become a better researcher.

Administratively, none of this would have gotten off the ground without the help of Hilary Hoynes. When I dropped out of my computer architecture lab, there was no one there to fund me. I had taken Hilary's public policy class on the safety net, and fittingly she generously secured funding for me for a year via an Opportunity Lab fellowship. This then allowed

me to focus on working with Emi for a semester, which in turn led to Emi becoming my full time advisor. So a heartfelt thank you to Hilary! Similarly, I thank Jesse Rothstein for important early conversations, where he suggested I not lock myself narrowly into a particular conception of labor economics. This advise ultimately landed me in Emi's office. From the EECS department, I must thank Ben Recht, Jacob Steinhardt, Jean Nguyen, and Naomi Yamasaki for their invaluable help with navigating the administrative realities of having one statistics advisor and one economics advisor while a student in computer science.

I met many wonderful peers and colleagues during my time at Berkeley. This includes the students in the ASPIRE and ADEPT labs, especially David Biancolin, Orianna DeMasi, Adam Izraelevitz, Kyle Kovacs, Eric Love, Albert Magyar, Albert Ou, Nathan Pemberton, and Colin Schmidt. Once I left ADEPT, one unexpected benefit was that I got to know many different cohorts of students. In statistics, Amanda Glazer and Jake Spertus. In machine learning theory, Frances Ding, Robbie Netzorg, and Ricardo Sandoval. In economics, Raheem Chaudhry, Nicole Gandre, and Jessie Harney. With all of these people I have many fond memories. Many of them helped me through rough spots of the PhD at various times.

Thanks also to my good friends who are mostly on the other side of the United States: Damian Krebs, Tim Laciano, Maya Major, Mark Moore, and Caleb Small. Thank you to my housemate of three years Branden Ghena, who ultimately introduced me to Meghan. And to my neighbors and found-family members Anne Mayoral and Remi Myers. I love you all very much.

Finally thank you to my mom and dad, Pammy and Kenny. I owe everything to you. To my sister, Alex, who is an inspiration to me in her drive and passion for doing good in the world. And to my wife and partner Meghan. You are the love of my life and I draw endless strength and joy from your support and from supporting you in turn. Thank you!

Part I

Prediction in Macroeconomics

Chapter 1

Introduction

1.1 Causal Inference and Machine Learning in Macroeconomics

In order to achieve our various social and political goals, we would like to predict the effect of economic policy interventions before rolling them out. Due to the tremendous complexity of the economy (in the United States, over 300 million people and over 30 million businesses, for example), it's virtually impossible to reason out policy consequences from only first principles, and so, increasingly, we turn to the empirical data on historical interventions, randomized and otherwise, to further build our understanding of policy impacts.

The 21st century has brought a rapid expansion in the size, quality, and availability of economic data, including both survey and large administrative datasets. There have concurrently been two major developments in empirical economics. The first is the "Credibility Revolution" - an emphasis on experimental design and the credibility of drawing causal conclusions from historical data. This development highlights the fundamental gap between predicting trends in historical data versus predicting the impact of an intervention. Causal inference has seen wide adoption and development in applied microeconomics and labor economics in particular. While the adoption in macroeconomics has been considered uneven [8], there has been considerable progress on this front [251, 242, 286, 213]. Interestingly, in macroeconomics, research has not typically involved using explicit causal inference machinery like potential outcomes — although there are notable exceptions [241].

The second development is the incorporation of "machine learning" — a set of statistical tools that leverage large amounts of data and computing power to predict an outcome given various inputs without the practitioner having to precisely specify the (potentially very complicated) relationship between the inputs and outcomes beforehand. As with causal inference, machine learning is popular in applied microeconomics and labor economics — especially in the burgeoning intersection of "causal ML." One advantage of machine learning methodologies in this context is the ability to explore heterogeneity by flexibly modelling how treatment effects depend on covariates [16]. By contrast, in macroeconomics, machine

learning has been nearly exclusively restricted to (1) fitting large-scale macroeconomic DSGE models [93], and (2) macroeconomic forecasting [112].

1.2 Overview of this Dissertation

The content in this dissertation includes parts of previously-published papers that I co-authored with Oliver Dukes, Avi Feller, Emi Nakamura, Betsy Ogburn, and Angela Zhou [41, 45, 43, 44, 42].

Predictions and Interventions

In this dissertation, I focus on developing the machine learning and causal ML toolkit for macroeconomics applications (very broadly construed). In Chapter 2, before turning to causal inference, I begin with a simple case study of a prediction problem in macroeconomics where machine learning works out-of-the-box. In this application, we use an Icelandic administrative tax dataset to predict future household labor income given current household information. Crucially, the goal here is not to take this predictor and claim that we can forecast the future income of any new household with high accuracy. This might be a fundamentally spurious endeavor [263]. Instead, we are inspired by Milton Friedman's work on the permanent income hypothesis, i.e. that a core component of household behavior is expectations about future income. Friedman explicitly conceptualized the core concept as a conditional expectations of future income — see Figure 2 of the chapter "The Permanent Income Hypothesis" from "A Theory of the Consumption Function" [99]. But when Friedman wrote Theory of the Consumption Function, the statistical machinery of the time made estimation of the curve of expected income over future periods very difficult; indeed, Friedman mostly sets aside statistical questions, and remarks only that conditioning on coarse groups like age-year buckets is undesirable. However, estimating the conditional expectation of future income is precisely a prediction problem, which the machine learning toolbox can estimate flexibly even while conditioning on a very large and granular information set. By solving this prediction problem, we are able to give a descriptive quantitative analysis of income risk and its persistence over time across an entire population.

However, our ultimate goal is to predict the effect of economic policies, and machine learning prediction tools are not well-suited to this task when applied naively. Let's say that we are considering rolling out a jobs training program in Philadelphia, and we would like to predict what impact it will have on participants' future earnings. We have some data from a jobs training program that happened recently in a very similar city — let's say Baltimore. In this dataset, we observe some people who participated in job training and some who did not. It will not suffice to simply fit a machine learning model that predicts future earnings given participation in job training plus other baseline characteristics. If in our dataset, those who receive training have systematically less education and lower previous earnings, then it is likely that their future earnings will be systematically lower than those who do not receive

training, *even if* job training does in fact boost their future income. Machine learning is very good at finding and exploiting complicated sources of correlation in data if they improve prediction, and so a machine learning algorithm fit on this dataset will (accurately) infer that those who received job training are more likely to have lower future earnings. But these predictions do not reflect the underlying causal effect of the job training program.

The previous paragraph discussed the problem very informally. To formally reason about causal effects, we will need to introduce some notion of counterfactuals. We will use the "potential outcomes" framework [138], which posits that for each individual, there exist two possible outcomes: future earnings with job training, $Y(1)$, and future earnings without job training, $Y(0)$. The fundamental problem is that in our dataset, we necessarily only observe one of these two outcomes for each individual; for those who do receive job training, we cannot observe what *would have happened* in the hypothetical possible world in which they never received job training, i.e. $Y(0)$. In general, we need some assumptions to make it possible to use the observed potential outcomes to infer something about the unobserved potential outcomes.

Causal Inference with No Unobserved Confounders

One such assumption is to claim that we observe all relevant variables — sometimes called "no unobserved confounding", "conditional ignorability", "conditional exogeneity", among other things. In the job training example, our problem was that job training participation might be correlated with education or previous earnings, which we might erroneously attribute to a causal effect of the training program if we fit a predictive model. However, if we observe education and previous earnings, then we could correct for those differences. This is the setting for Part 2 of my dissertation.

My contribution to the literature on causal inference with no unobserved confounding begins by recognizing that this setting is formally equivalent to the problem of "covariate shift" in machine learning — I am not the first to make this connection, but this viewpoint leads to a number of useful insights. Consider our job training example, and assume (for the moment) that the only important confounding factors are education and previous income. If we fit a machine learning model to predict the observed $Y(0)$ amongst those without job training, we could then apply that model to predict the unobserved $Y(0)$ for those who did receive job training without having to worry about bias from confounding. Unfortunately, the distribution of education and previous income are still systematically different between the two groups, and the predictive capabilities of machine learning model depend on test samples being drawn from the same distribution as the data used to train the model. In the extreme case where everyone in the dataset who receives job training has very little education and everyone in the dataset who does not receive job training has a extensive education, it is simply not possible to transfer a useful predictive model from one population to the other.

It turns out that many fields, including survey sampling, econometrics, statistics, and machine learning, have been studying this covariate shift problem for decades, but under different names. The key object goes by many names but is always the same: the density

ratio, the likelihood ratio, survey weights, importance weights, the inverse propensity score weights, or the inverse probability weights. Different disciplines have developed various ways of estimating the density ratio, and in Chapter 3, I demonstrate some surprising numerical equivalences between seemingly unrelated algorithms. In particular, I show that a class of modern methods in statistics called "balancing weights" and recent methods from econometrics called "automatic estimators of the Riesz representer" are numerically equivalent. But these methods are in turn exactly equivalent to an older estimator in computer science called "direct density ratio estimation" (also "least squares density ratio estimation" or the "convex surrogate loss for the likelihood ratio"). And these methods are a generalization of an even older technique from survey sampling called "calibration weighting." I develop these equivalences in Chapter 3 using a fairly general duality argument. While doing so, we produce some new insights: while we can view these different methods as estimating the density ratio over some function class, we can alternatively view the estimand as a "tailored" density ratio that is only sensitive to certain functions. The very surprising result is that these two viewpoints arrive at *the same answer*.

Inspired by the results in Chapter 3, in Chapter 4 we consider a popular machine learning framework for estimating causal effects with no unobserved confounders called "automatic debiased machine learning" (AutoDML) — known in other literatures as "GREG", "augmented balancing weights", "augmented minimax linear estimation", or "approximate residual balancing". AutoDML uses the estimators for the density ratio from Chapter 3 and uses them to correct for the covariate shift problem when using machine learning in causal inference. Recall that our goal is to predict what would have happened to those who received job training in the counterfactual world where they instead had not received training. We do this by fitting a machine learning predictor of $Y(0)$ amongst those who never received training and then applying it to those who did. The issue is that the predictor faces a potentially different distribution of education and prior income which may threaten the validity of the predictions. In AutoDML, we estimate the density ratio via balancing weights, which measures how education and previous income has shifted, and then add a correction term to the machine learning estimate using the density ratio estimate.

Our main result is that for linear models in some (potentially infinite-dimensional) basis, this procedure of combining the machine learning predictive estimate with the density ratio estimate, can be rewritten as a single predictive model that is a mixture of the original model and the unregularized least squares estimate. That is, we shift our model toward one that overfits in the training data, but in a principled way according to how the control and treated populations differ. In the special case where both the predictive model and density ratio model use ℓ_2-norm regularization, we show that the result is *exactly* equivalent to making the predictive model overfit more, and in particular this procedure is equivalent to boosting with ridge regression. We connect these results to the broad literature on "undersmoothing" — another word for overfitting — and develop some practical considerations for hyperparameter tuning in AutoDML.

Causal Inference with Unobserved Confounders

In our earlier job training example, we worried that the effect of factors like education or previous income would be mistakenly attributed to the effect of the job training program. The solution was to measure education and previous income and adjust for how they differ between the control and treated populations. But what about other unobserved factors? If education and prior income were systematically different, we might actively expect other factors to differ as well. And while education and prior income might already be available, either collected as part of the job training program or fused in from another administrative dataset, by contrast consider childhood nutrition or childhood exposure to heightened cortisol levels. These factors impact future earnings even conditional on education, are likely to be different between treated and control populations if those who receive job training are systematically more disadvantaged, and are very unlikely to be measured in an available dataset. In this case, it is impossible to exactly recover the causal effect of job training. I consider this setting in Part 3 of the dissertation.

As in Part 2, this setting is also a case of distribution shift. Importantly, however, it's not just the covariate distributions that shift, but also the relationship between the covariates and the potential outcomes ("conditional shift"). The problem is severe — we train a predictor between A and $B1$ and are then asked to use A to predict $B2$. If we don't know anything about $B2$ it's relationship to A could be arbitrarily different and our predictor would be useless. Without any further restrictions, it's impossible to get a single unbiased point estimate. Instead, we will assess the *sensitivity* of our results to potential differences between $B1$ and $B2$. That is, let's say we have a single measure of how different $B1$ is from $B2$ given A, call it Λ. In Part 3, we will give a particular definition of Λ, but for now just assume that when $\Lambda = 1$, then $B1$ and $B2$ have the same relationship given A. As Λ grows, $B1$ and $B2$ become more and more dissimilar. Even though we don't know exactly how $B1$ and $B2$ differ, given a predictor of $B1$ given A, we can still build upper and lower bounds on the best predictor of $B2$ given A for each value of Λ. If these upper and lower bounds are not too far apart for a large range of Λ's, then we have some evidence that our results are not very sensitive to hypothetical unobserved factors — such factors would have to be extremely strong to overturn our analysis. On the other hand, if even for very small values, like $\Lambda = 1.1$, the upper and lower bounds become far apart, then our analysis is very sensitive to confounding. Even a small violation of the "no unobserved confounding" assumption would threaten the results, and it is necessary to go collect new data.

In Chapters 5 and 6, we show how to compute these upper and lower bounds on the prediction function of interest for the popular "marginal sensitivity model" in the challenging *dynamic* setting: i.e. we causal effects that unroll over time. Formally, we adopt a Markov decision process (MDP) framework, which immediately connects our results to the large literature on Reinforcement Learning. In Chapter 5 we consider discrete covariates (called the "tabular" setting) and in Chapter 6, we generalize to continuous covariates with arbitrary machine learning function approximation. In Chapter 6, we also extend our results to learning a robust optimal policy under confounding.

Finally, while this dissertation is largely technical and methodological in nature, much of the work has had direct application in my empirical macroeconomics research. I will largely leave these results for other publications, but in a concluding Chapter 7, I will discuss where some of my results have been most useful, and also some future promising directions for further development of causal inference and machine learning in macroeconomics.

Chapter 2

Case Study: Income Prediction

2.1 Introduction

In this chapter, we introduce an application of pure prediction in economics. As motivation, economic hardship and the dynamics of socioeconomic inequality depend crucially on the *income shocks* that households face. Such shocks come in a variety of forms: positive shocks include promotions and stimulus checks; negative shocks include job loss, illness, and lack of available working hours. A large body of research shows that unexpected income shocks pass through to changes in household spending and saving [262, 201, 233], with wide heterogeneity in responses across household characteristics [164, 20, 178, 90, 104]. Households' responses to shocks are also key drivers of financial fragility [208, 2, 222], the effectiveness of fiscal policy [163, 141, 18], and the evolution of wealth inequality [235, 74, 14].

To study the impacts of shocks on households and the corresponding implications for subsidy allocation or macroeconomic policy, we first need to be able to measure them. While we observe changes in income from one period to the next in many economic datasets, we rarely observe what proportion of those changes were unexpected shocks, or whether those shocks were temporary or will persist long into the future. Consider a household that makes $60,000 dollars annually. In the next year, the household might simultaneously experience a promotion to site manager — a persistent increase of $5,000 a year — and an especially-snowy spring construction season — a temporary decrease of $20,000 a year — for a total observed shock of -$15,000. Does this hypothetical household spend more because their expected income will be higher into the future? Do they (or can they) spend less to weather the larger but temporary negative shock? Do they have a savings buffer to draw upon, or does the unexpected temporary loss of income cause them to miss their mortgage payments? These various considerations are difficult to tease apart based on the observed, overall shock alone.

In the economics literature, the workhorse statistical model for analyzing shocks and their persistence is a panel model where current income is the sum of a random *transient* shock and unobserved *permanent* income that evolves according to an autoregressive process [3, 202, 37]. Statistical estimands of interest, such as the size of transient and persistent

shocks, are then defined with respect to the parameters of this model. Importantly, however, this model embeds a series of assumptions about the income process that impose strong homogeneity across households and over time, such as assuming that households across the income distribution face shocks of the same size, severely limiting our ability to understand critical sources of variation.

In this paper, we instead propose to directly estimate income shocks and their persistence from income data. We first propose a non-parametric estimand for income shocks — defined outside of any particular statistical model — in terms of the conditional expectation of future income given the information known in the present. Estimating these conditional expectations for a particular population requires finding the best mean-squared error predictors for data drawn from that population, a task we can perform using off-the-shelf supervised learning tools with strong uniform convergence guarantees. Our procedure outputs estimates of income shocks associated with each income observation along with the persistence of those shocks at several horizons into the future. These shocks can then be used in downstream tasks like estimating households' consumption/savings response, calibrating models for the evolution of wealth inequality, or as real-world datasets for studying algorithmic fairness.

Contributions:

- We provide a nonparametric definition of income shocks that relaxes strong functional form assumptions, and allows researchers to assess heterogeneity in the size and persistence of income shocks across observed features.
- As a real-world application, we estimate income shocks in Iceland by predicting labor income at various horizons into the future using a large administrative tax dataset.
- We document several features of the estimated shocks that are not captured by standard economic parametric income models, including: a much larger magnitude of income risk faced by individuals at the bottom of the income distribution; an exponential decay in the persistence of shocks on average over time; wide hetereogeneity in the persistence of shocks across household circumstances; and substantial asymmetry between positive and negative shocks.

We hope to draw attention to an under-utilized role for prediction in the social sciences, where supervised learning models are used as an approximation of a conditional expectation, rather than used to predict future outcomes for new, potentially out-of-distribution observations. We also hope to further connect the parallel research on income shocks in economics and computer science. The large body of research on income uncertainty and household responses in economics can bring valuable insight to the recent literature on algorithmic fairness and inequality. Likewise, the powerful non-parametric modelling and optimization toolboxes from computer science can shed new light on the dynamics of income.

2.2 Related Work

Economics

A large literature on economic theory studies household responses to income shocks. The permanent income hypothesis [99, 67] suggests that households will smooth consumption over the lifecycle, and predicts very small responses to temporary shocks but large responses to permanent shocks. A literature on precautionary savings explores why consumption seems to track income so closely in the data with an emphasis on uninsurable idiosyncratic income risk [50, 49, 134, 113, 163].

Most closely related to the current work is the literature on transient-persistent income process models. Linear transient-persistent autoregressive models have been widely used [3, 202, 37], and we will compare these parametric models to our non-parametric estimands in Section 2.3.

Several recent papers have critiqued these models for failing to match key stylized facts documented in real income data. For example, Guvenen et al [118] emphasize the substantially higher skewness and kurtosis exhibited by US income data that linear panel models have difficulty reproducing. Other papers address the assumptions about the persistence of shocks embedded in standard models. For example, De Nardi et al [75] discretize income into buckets, and then non-parametrically estimates a first-order Markov chain for transitions between these discrete states. They find evidence for heterogeneity in the persistence of shocks and substantial deviations from the AR(1) process in [37]. Arellano et al [10, 11] propose a Bayesian approach for estimating the posterior of permanent income (defined as a latent variable) using expectation-maximization. Straub [288] creates proxies for permanent income in real data by averaging together several past and future income observations — a procedure that could be seen as a very simple version of predicting income at several future horizons.

Computer Science and Machine Learning

Recently, a growing literature in computer science and algorithmic fairness has also emphasized the role of income shocks. This includes algorithm and mechanism design research for subsidy allocation where the level of household income and wealth as well as their susceptibility to shocks play central roles [2, 222, 231]. Other work studies the dynamics of income inequality over time [127, 244], including their implications for policy interventions. In many ways, this literature has close connections to the macroeconomic consumption and inequality literature. In fact, Nokhiz et al [222] solves and simulates from a macroeconomic consumption model with incomplete markets and precautionary savings motives in the style of Hubbard et al [134] or Gourinchas and Parker [113].

Most of the related work in computer science has used simulated income data. For example, Abebe et al [2] simulate shocks as arriving via a Poisson process. Nokhiz et al simulate income with a first-order Markov chain over discretized income states. On the other hand, D'Amour et al [71] have an *implicit* model for income shocks in their simulated model

for loan repayment. In their model, the probability of repayment is a deterministic function of credit score, embedding a number of assumptions about credit score calculations and the income risk faced by households that necessarily partially determines their ability to repay. Our work is complementary to the work above and provides an alternative to simulation, measuring the degree of labor income risk directly in real-world data.

Reader et al [244] suggest modelling the evolution of income inequality as a linear dynamical system, with policy interventions and feedback loops modelled as a PID controller. Our method similarly has connections to the controls literature. As we will discuss, the transient-persistent models for income can be formulated as partially-observed dynamical systems, finite-sample estimation of which has featured in recent research on system identification [276, 182, 204, 183].

An adjacent literature studies the dynamics of income *between* generations, mostly focused on interventions in university admissions [127, 4]. This work complements a large body of work on inter-generational mobility in economics [73, 68, 64]. Extending our predictive estimands and the corresponding measures of income risk to an inter-generational context would be an interesting direction for future work.

Limits of Prediction in Social Science

Another important literature emphasizes the limits of predictability of future life outcomes. Narayanan [215] called predicting social outcomes "fundamentally dubious". A large-scale prediction competition, the Fragile Families Challenge, found that predictive accuracy across a variety of social outcomes and algorithms was low across the board [263]. Flexible machine learning models hardly performed better than linear regression on a handful of features. More broadly, Liao et al [189] and Raji et al [240] outline a large taxonomy of basic ML functionality failures in real-world deployments.

We instead emphasize a potentially under-utilized role for prediction in social science: approximating a conditional expectation. This follows the exhortation in Lundberg et al [197] regarding prediction in sociology: to clearly state the statistical estimand of interest. We would like to characterize the distribution of prediction errors around the conditional expectation of future income and how it evolves from one period to the next. If the absolute size of these errors for the best possible predictor given the feature set is large, then this is not a functionality failure, but an accurate statement of income risk in the population. Likewise, if the model has substantially larger prediction errors for one sub-group compared to another, then the relative distribution of these residuals tell us about the inequality in income risk across groups.

Predicting Future Income

In this work, we solve prediction problems for income h periods into the future conditional on current and past income and other covariates. Surprisingly, we have found very few published papers that solve this kind of income forecasting problem. The only such example

to our knowledge is Gerardi et al [106], who use unpenalized linear regression to predict future income conditional on current income, housing wealth, and demographic variables. See Section 2.4 for a discussion of the performance of linear regression in our setting.

A very large literature in machine learning considers income prediction problems framed as classification tasks. Most of this work is centered on the Adult dataset [176], first used to assess the performance of tree-based ensembles [22]. See Ding et al [80] for a review of recent research using this dataset, especially on algorithmic fairness, and several associated limitations. The standard task on Adult is classifying whether or not income falls below or above $50,000. In contrast, we consider income prediction as a regression problem, and introduce a dynamic dimension by forecasting future income conditional on current and past income. Furthermore, the emphasis of our work is different but complementary to the fairness literature using Adult — if our prediction algorithms have larger prediction errors for certain sub-populations we interpret this as a substantive result about the relative income risk faced by those sub-populations.

2.3 Defining Income Shocks

Let y_{it} denote log income of individual i at time t and let x_{it} denote covariates such as age, education, calendar year, and wealth. We assume that we observe N i.i.d. samples of the trajectories $\tau_i \coloneqq \{(y_{it}, x_{it})\}_{t=1}^{T_i}$ from the same joint distribution — we make no assumptions within a trajectory on the relationship between y_{it} and x_{it} or their evolution over time. The τ_i can either be interpreted as draws from an underlying joint distribution or as samples from some finite population of individuals. In what follows, we omit the i subscripts when clear from context. In practice, draws across individuals are unlikely to be entirely independent — common violations might include individuals within the same household or firm, and we plan to address these limitations in future work.

We define an *income shock* at time t to be the difference between observed income y_t and expected income given all information available before time t. Define the information set $\mathcal{I}_{t-1} = \{y_{t-1}, x_{t-1}, y_{t-2}, x_{t-2}, ...\}$. Then the income shock at time t is:

$$\Delta_t \coloneqq y_t - \mathbb{E}[y_t|\mathcal{I}_{t-1}]. \tag{2.1}$$

We define the persistence of the time t income shock as the change in expected future income upon adding the new information (y_t, x_t) into the information set. We write the horizon-h persistence of the shock Δ_t for all $h \geq 1$ as:

$$\phi_{t,h} \coloneqq \mathbb{E}[y_{t+h}|\mathcal{I}_t] - \mathbb{E}[y_{t+h-1}|\mathcal{I}_{t-1}]. \tag{2.2}$$

As a concrete example, let $\mathbb{E}[y_t|\mathcal{I}_{t-1}] = 1.0$ and realized income $y_t = 2.0$. Then the total shock at time t is $\Delta_t = 1.0$. Now we use the realized y_t (and x_t) to update the expectations for the future to measure how long the shock lasts. If the updated conditional expectation $\mathbb{E}[y_{t+1}|\mathcal{I}_t] = 1.5$, then the portion of the total shock Δ_t that is expected to remain after

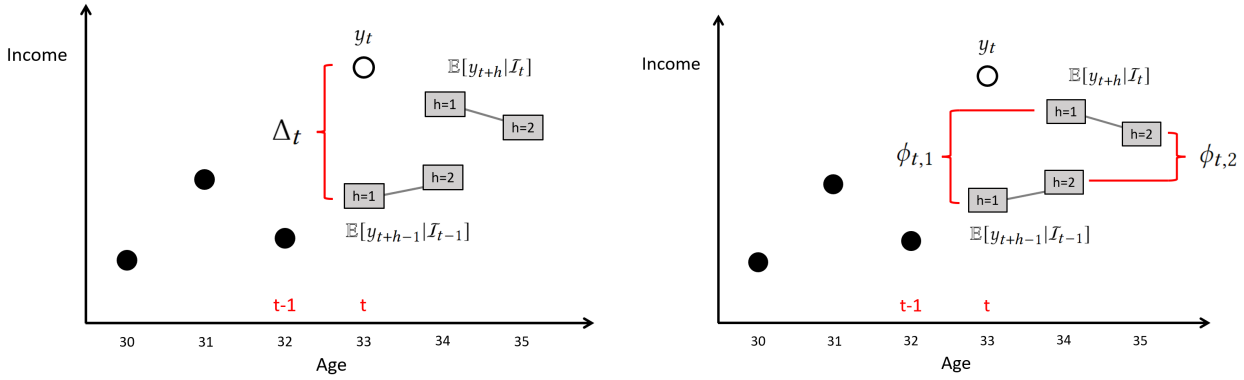


Figure 2.1: An illustration of our estimands. The black dots represent the conditioning set $\mathcal{I}_{t-1}$. The grey boxes represent the predictions one and two periods ahead. At time t, we observe the new observation (shown as an open circle) y_t. The difference between the new observation and the previous prediction is Δ_t as shown in the lefthand diagram. When we add y_t to the conditioning set and update the predictions, we get the persistence at each horizon as shown in the righthand diagram.

one period is $\phi_{t,1} = \mathbb{E}[y_{t+1}|\mathcal{I}_t] - \mathbb{E}[y_t|\mathcal{I}_{t-1}] = 0.5$. If the original and updated 2-step-ahead conditional expectations are $\mathbb{E}[y_{t+2-1}|\mathcal{I}_{t-1}] = 1.0$ and $\mathbb{E}[y_{t+2}|\mathcal{I}_t] = 1.1$, then the amount of the total shock expected to persist two periods into the future is $\phi_{t,2} = 0.1$. So in this example, while the time t unexpected change in income was large, only half of the shock is expected to persist one period into the future, and only 10% of the shock is expected to persist two periods into the future. See Figure 2.1 for an illustration.

The quantities Δ_t and $\phi_{t,h}$ are our *non-parametric estimands*. They are not observed directly, and we would like to estimate them from data. However, first we briefly justify our choice of these quantities.

Theoretical justification

Simple theoretical models for household responses to income shocks usually imply that consumption depends on the expected present value of future income, sometimes called *permanent income*. Given a discount rate γ, permanent income at time t is:[1]

$$y_t^{\text{perm}} := \mathbb{E}\left[\sum_{k=t}^{\infty} \gamma^{k-t} y_k \middle| \mathcal{I}_t\right]. \tag{2.3}$$

[1]Typically, permanent income would be discounted by $1/r$, where r is the rate of return on assets and therefore represents the relative value of money now versus money in the future. In the simplest models, $\gamma = 1/r$ in equilibrium.

Then the unexpected change to permanent income at time t is

$$y_t^{\text{perm}} - \mathbb{E}[y_t^{\text{perm}}|\mathcal{I}_{t-1}] = \Delta_t + \sum_{h=1}^{\infty} \gamma^h \phi_{t,h}. \tag{2.4}$$

So in this sense, the objects Δ_t and $\phi_{t,h}$ are precisely the relevant theoretical objects for studying a household's response to income shocks. Equation (2.4) suggests one way to summarize these shocks in a single measurement. Indeed, the framework of updating future expectations as new information arrives is exactly the motivation for the definition of permanent income in Flavin 1981 [95].

More sophisticated economic models suggest that current consumption choices may depend on the whole conditional distribution of $P(y_{t+h}|\mathcal{I}_{t-1})$ rather than just the conditional mean. This suggests a straightforward extension of our procedure using conformalized quantile regression [250] that we hope to pursue in future work.

Comparison to parametric estimands

It is helpful to compare our non-parametric estimands to the commonly used transient-persistent model [37, 10]. This model imposes the following structural assumptions on the income process:

$$\begin{aligned} y_t &= p_t + \epsilon_t, \\ p_t &= f(p_{t-1}) + \eta_t, \end{aligned}$$

for some measurable function f and where $\mathbb{E}[\epsilon|p_t] = 0$ and $\mathbb{E}[\eta_t|p_{t-1}] = 0$. First, notice that the expected system transitions in the implied autoregressive model, obtained via the standard trick of re-writing the partially-observed non-linear system as an infinite-order autoregressive model, is exactly the conditional expectation $\mathbb{E}[y_t|\mathcal{I}_{t-1}]$. See for example [183] for discussion in the controls setting.

The classical parametric model used in [37] makes the additional functional form assumption:

$$p_t = p_{t-1} + \eta_t.$$

where the variance of η and ϵ are independent of p_t. Note that in this case, our non-parametric definition for shock persistence exactly corresponds to the persistent shock in the model; $\phi_{t,h} = p_t - p_{t-1}$ because $\mathbb{E}[y_{t+h}|p_t] = p_t$. However, the classical model imposes several additional testable implications: that households across the income distribution face shocks of the same size; that there are no interactions between age, demographics and shock size or persistence; that persistent shocks are perfectly-persistent into the future; and that there is no asymmetry in the persistence of positive and negative shocks. Using our non-parametric estimands that avoid making such strong assumptions, we will demonstrate substantial deviations from this simple model for labor income in Iceland.

Limitations of any particular estimator

The exact interpretation of our non-parametric estimand depends entirely on our dataset, our definition of the relevant random variables, and the contents of the conditioning set, $\mathcal{I}_t$. For example, as we discuss in Section 2.4, we have to choose the definition of the periods t; are these yearly or monthly shocks? Monthly shocks have to account for seasonal variation, whereas for yearly shocks the business cycle becomes a central object of concern. Likewise, we have to choose a definition of income; does y_t represent labor income or total family income? Any of these choices may not be inherently right or wrong, but will have different implications for the relevant downstream economic analysis.

More generally, as commonly done in the economics literature [3, 37, 10, 74], we measure shocks in terms of a *statistical* expectation. The relevant theoretical objects of interest in Section 2.3, are *household* expectations because a household's behavioral responses to income risk depend on their own beliefs about the future. This can introduce substantial measurement error along across at least two dimensions. First, we do not have access to important private information — for example an individual's plan to leave their job next year to go back to school. Second, our predictions are formed using hundreds of thousands of observations from across the entire population of Iceland, information that any given individual might not have. This discrepancy must be kept in mind when performing any later economic analysis using our estimated shocks. For example, we may try to impose some structural assumptions on the nature of this measurement error, and *partially identify* behavioral estimands to account for the additional uncertainty. Or otherwise, we have to more narrowly interpret our estimated income shocks as a measure of *aggregate* labor income risk across Iceland, capturing heterogeneity across observables in the tax data, rather than the personal uncertainty about future income faced by any particular individual.

2.4 The Income Prediction Problem

Our non-parametric estimands for income shocks, Δ_t, $\phi_{t,h}$, can all be computed given the conditional expectations $\mathbb{E}[y_{t+h}|\mathcal{I}_t]$ for all t and all $h \geq 1$. The conditional expectation is equivalent to the best mean-squared error predictor of y_{t+h} over all possible functions of the features in the set $\mathcal{I}_t$. Therefore, we have reduced the problem of estimating income shocks and their persistence to a series of prediction problems for which we can use off-the-shelf supervised learning tools. In this section, we discuss how we solve these prediction problems in practice with a large administrative tax record dataset from Iceland.

Data and sample selection

We use income measurements from Icelandic income tax data, made available to us through collaboration with Statistics Iceland. In our Icelandic tax data, the period t is measured in years, and for every individual in every year from 1981-2018 we observe (log) labor income y_t and a collection of other demographic and financial variables. We transform all income

observations to 2018 US dollars, adjusting for Icelandic CPI [135] and the exchange rate between the dollar and the Icelandic Króna [225]. While we observe nearly the entire population of Iceland during this timeframe, we restrict our sample to reflect *in-employment* labor income risk. This involves three sample selection steps: (1) we only include individual-year observations with labor income strictly greater than zero; (2) we only include those individual-year observations for which we observe non-zero income for at least six consecutive periods before and at least twelve consecutive periods after; and (3) we only include observations for individuals aged thirty and older (to avoid income changes due to switching in and out of higher education). This leaves 508,235 individual-year observations across 62,387 individuals.

Note that the choice to study in-employment labor risk and unemployment risk separately is common [76, 287, 37, 212]. Our choice to focus on in-employment risk is mostly for purposes of presentation and for comparison with [37]. Unemployment risk is also of central interest, and re-estimating income shocks with unemployment will be the object of future work. Furthermore, the particular choice of zero for the minimum threshold might include individuals who are unemployed for part of the year; for discussions on alternative choices of the minimum threshold see Nakajima and Smirnyagin [212]. Likewise, the choice to study *labor* income as opposed to total income after taxes and transfers has consequences for the interpretation of our estimand. In general, these choices for sample selection and the definition of income do not change the non-parametric estimands outlined in Section 2.3 but are enormously important for substantive economic analysis of the results.

For covariates x_t, we include age, education, gender, total assets (net of debt), and housing wealth.[2] Education is binned into five categories: incomplete compulsory education, compulsory education only, upper secondary only, undergraduate only, and beyond undergraduate. With no essential loss of generality, instead of fitting a separate model for each t, we will fit a single predictor, but additionally condition on calendar year. Thus our complete feature set includes dummies for calender year t, current income and covariates (y_t, x_t), and six lags of income and covariates, $\{(y_{t-\ell}, x_{t-\ell})\}_{\ell=1}^{6}$. Note that this is only an approximation of the information set $\mathcal{I}_t$, which should include as many lags as are available. However, we can justify this theoretically with relatively mild assumptions on the mixing of the stochastic process for income. In practice, we also found that including more lags does not improve mean-squared error in cross-validation.

Training

As we would like to use highly-flexible regularized function classes for prediction, we leverage both sample-splitting and cross-validation to prevent over-fitting. Note that while each training sample corresponds to an individual-year observation, the observations within an individual trajectory are highly-correlated. Therefore, we perform all sample-splitting and cross-validation at the *individual level*. First, we randomly divide the full population of

[2]Notably missing from the tax data is information on race or ethnicity, presumably due to extremely low rates of immigration. During the timeframe of our dataset, more than 92% of the population were ethnically Icelandic.

individuals into two halves. Within each half, we train models predicting y_{t+h} for $h = \{1, ..., 12\}$, using the feature set described above. Prior to training, all features were shifted and scaled to have mean zero and standard deviation one. We considered a variety of regularized linear models, random forests, and gradient-boosted tree regressors, over a range of hyperparameter values. We chose the best performing model using 5-fold cross-validation. Gradient-boosted trees consistently performed the best in cross-validation across all horizons.

The output of this process is our best approximation of the conditional expectations $\mathbb{E}[y_{t+h}|\mathcal{I}_t]$ for all h from each of the two halves. We then compute the estimated income shocks Δ_t and persistence profiles ϕ_{th} by applying the models trained in one half to the individual-year observations in the opposite half. This way, the income shock for each observation is estimated using a model that was never previously trained using that observation.

Remark: We claim that this process gives the best approximation of the conditional expectation in the population. This does not mean that our trained models are the best predictors of future income on never-before-seen observations from a different population. Applying the predictors outside our dataset could face significant distribution shift, perhaps most notably the massive impact of COVID-19 in 2020 and onward. Instead, we rely on the fact that we *randomly* split individuals into two halves from a known population. This means the strong uniform convergence guarantees that come from the i.i.d. assumption in supervised learning apply exactly. As a result, however, any insights about income risk from our procedure are only guaranteed to describe the population of Iceland during the timeframe of our sample — extrapolating outside this population would require additional statistical assumptions.

Model assessment

Before presenting our results on income shocks, we first assess our models' predictive performance. First, we emphasize the advantage of using a highly flexible model class by comparing our final gradient-boosted tree models to two simpler benchmarks: a simple random-walk baseline that always outputs most-recent income and ordinary-least-squares linear regression. We are inspired to include the random-walk baseline by a famous macroeconomics result that a random walk beat existing models for predicting exchange rate out-of-sample [200], and the linear regression model due to its strong performance in the Fragile Families Challenge [263].

Figure 2.2 compares the mean-squared error of the best performing gradient-boosting model and the two baselines. In particular, we plot the MSE of predictions on data points that were not used in training; for each data point, we make predictions for all horizons h, using the models trained in the opposite split. The gradient-boosted trees model perform much better than the random walk, and modestly better than the linear model, achieving between 7 to 19% reduction in MSE. Note that the magnitude of the average prediction errors across the whole population is quite large. For one year ahead, the MSE suggests that the average magnitude of prediction errors is around 0.2 in logs. In levels, this corresponds to an error of about 22% of income. For twelve years ahead, even the best performing model has average prediction errors of around 40% of income. Recall that prediction error one

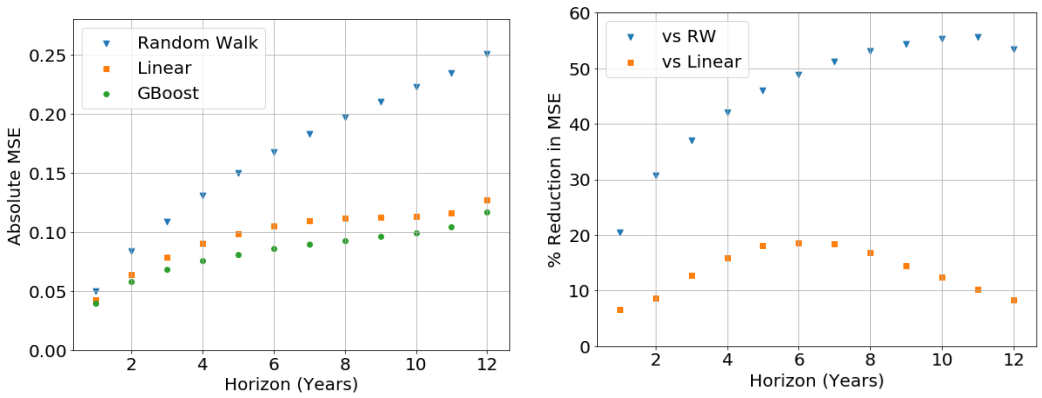


Figure 2.2: Mean-squared error of hold-out predictions. The top diagram plots the mean squared error of the gradient-boosted trees model, the linear model, and the random walk baseline for horizons 1 to 12. The bottom diagramplots the percent reduction in mean-squared error achieved by the flexible gradient-boosted tree model relative to the linear model and random walk baseline.

period ahead is exactly the definition of the income shock Δ_t and so, assuming that we have a good approximation of the conditional expectation, the large absolute mean-squared error indicates a fairly substantial amount of income risk. However, the *average* squared-error can be misleading, and we will show later that the largest prediction errors are concentrated at the bottom of the income distribution.

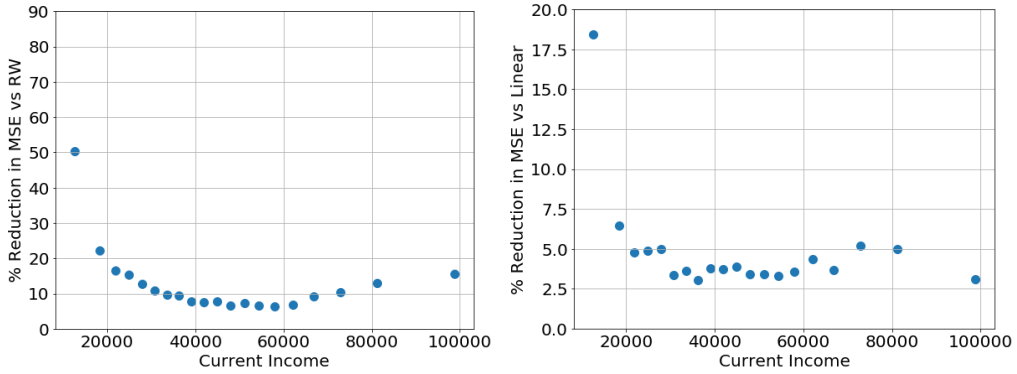


Figure 2.3: Percentage reduction in mean-squared error for predictions $h = 1$ year ahead across bins of current income. The top and bottom plots compare the gradient-boosted model to the random walk baseline and linear model respectively.

The importance of flexible models becomes more clear in Figures 2.3 and 2.4. These figures plot the relative improvement of our gradient-boosted model against the linear and random walk baselines across the distribution of current income. We proceed by binning:

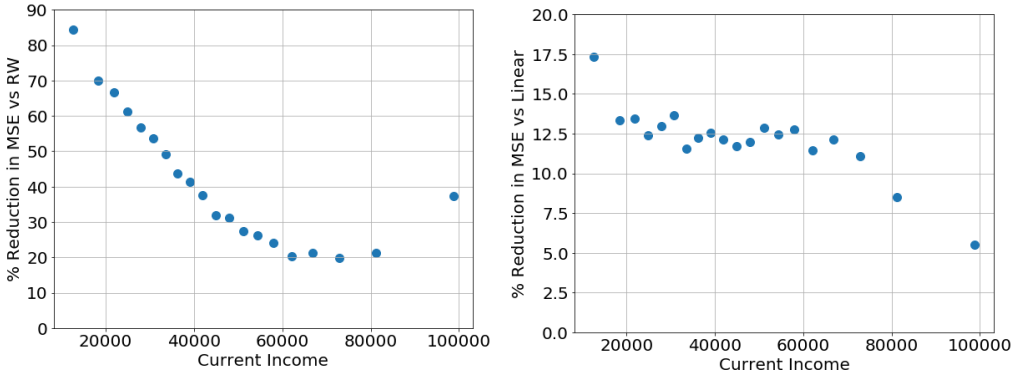


Figure 2.4: Percentage reduction in mean-squared error for predictions $h = 10$ year ahead across bins of current income. The top and bottom plots compare the gradient-boosted model to the random walk baseline and linear model respectively.

we split the observations into 20 equally-sized bins based on quantiles (from 5% to 95%) of log current income. Each dot in these figures corresponds to one of these bins. The x-axis is average current income within bins, with values in levels — recall that this represents inflation-adjusted income in 2018 US dollars. Figure 2.3 plots the reduction in MSE achieved by the gradient-boosted model compared to the two baselines for predictions $h = 1$ year ahead. Figure 2.4 plots the reduction in MSE achieved by the gradient-boosted model compared to the two baselines for predictions $h = 10$ years ahead. Note that the flexible model is especially important when predicting future income for houses at the bottom of the income distribution. The flexible model achieves nearly 20% improvement versus the linear model for individuals who make less than $20,000 (2018 US dollars) a year.

By definition, conditional on any value of the features, our prediction errors should be mean-zero if we have achieved a good approximation of the conditional expectation.[3] We explore this in Figure 2.5, where we compare the distribution of prediction errors across current income for both the linear and gradient-boosted trees models. We use the same buckets of current income, but the y-axis now plots the deciles and mean of prediction error within each bucket. Note that the distribution of prediction errors for the linear model in the upper plot is asymmetric with non-zero mean, and with the most substantial deviation for households with current income less than $20,000. The 90-10 interquantile range is only slightly smaller in the lower plot, but most importantly the distribution has approximately mean-zero everywhere, further validating our approximation of the conditional expectation.

[3]To see this, note that $\mathbb{E}[y_t - \mathbb{E}[y_t|\mathcal{I}_{t-1}]|\mathcal{I}_{t-1}] = \mathbb{E}[y_t|\mathcal{I}_{t-1}] - \mathbb{E}[y_t|\mathcal{I}_{t-1}] = 0$. Or more intuitively: if the prediction errors were not mean-zero conditional on a particular input, we could always improve the MSE by shifting all predictions for that input.

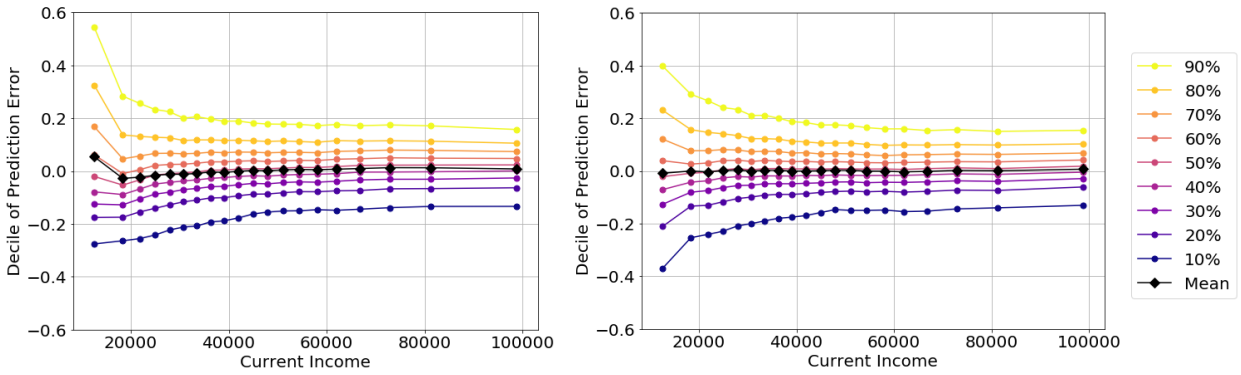


Figure 2.5: Deciles and mean of the distribution of prediction errors within bins of current income. The upper and lower diagrams plot the distribution of errors for the linear model and gradient-boosted model respectively.

2.5 Shocks

With our approximations of the conditional expectations $\mathbb{E}[y_{t+h}|\mathcal{I}_t]$ in hand for all h and t, we can estimate the shocks $\Delta_t \coloneqq y_t - \mathbb{E}[y_t|\mathcal{I}_{t-1}]$ and their persistence $\phi_{t,h} \coloneqq \mathbb{E}[y_{t+h}|\mathcal{I}_t] - \mathbb{E}[y_{t+h-1}|\mathcal{I}_{t-1}]$ for every individual in every year of our sample. This produces a concrete artifact as output: income shock estimates attached to every observation that can be used in downstream economic research tasks. In this section, we use the shock data to provide an initial characterization of labor income risk in Iceland.

The distribution of total shocks Δ_t, is exactly equal to the distribution of prediction errors, but now we analyze them *substantively* instead of as a diagnostic tool for model fitting. From the bottom diagram in Figure 2.5, we can see that low-income households face a much wider distribution of shocks. The 10-30% quantile shocks and the 70-90% quantile shocks are all at least twice as large for individuals at the bottom of the income distribution compared to the middle and top. This is a substantial deviation from the classical autoregressive model discussed in Section 2.3 that predicts an equal amount of income risk across the income distribution. Furthermore, notice that if we had used linear regression for prediction, then from the upper diagram in Figure 2.5 we would have incorrectly concluded that low income individuals face much larger positive shocks than negative shocks.

Using our methodology, we can also assess how these shocks persist over time. Figure 2.6 plots the persistent shocks $\phi_{t,h}$ as a function of the total income shock Δ_t. The upper diagram shows the results for $h = 1$, and the lower for $h = 10$. We begin by summarizing some observations for the $h = 1$ case. First, notice the asymmetry between positive and negative income shocks, a result that mirrors recent findings of asymmetry in consumption responses [66]. For positive total income shocks, there is a clear and roughly linear relationship between the total shock size and the persistence one period ahead. A substantial and fairly consistent proportion of total income shocks are persistent. Negative income shocks, on the other hand,

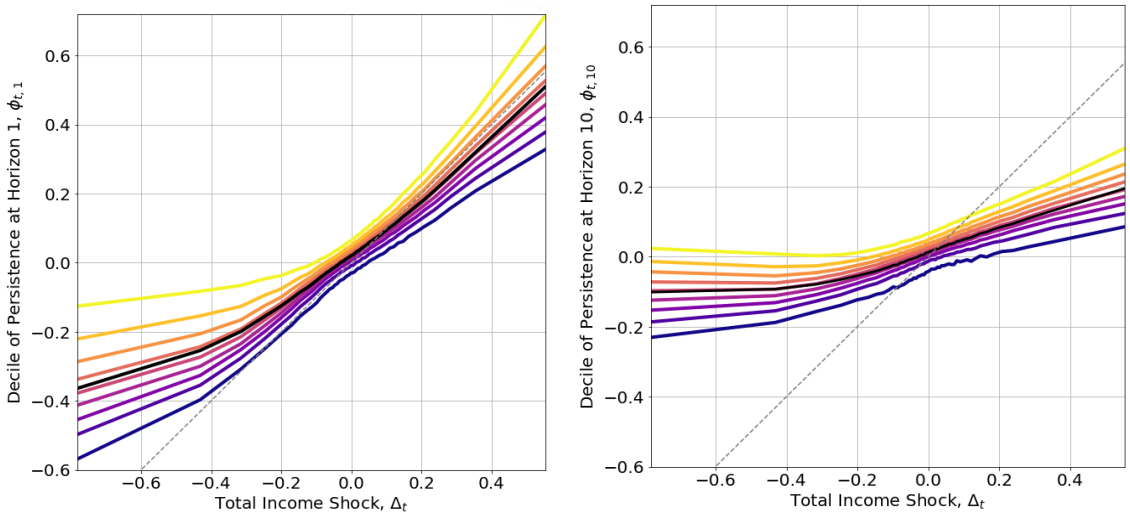


Figure 2.6: The upper and lower diagrams compare Δ_t and $\phi_{t,h}$ for the estimated income shocks computed using the predictions from our model on held-out samples for $h = 1$ and $h = 10$ respectively. We divide the obvservations into 50 bins according to Δ_t; the x-axis plots the mean value within each of these bins. The y-axis gives the 10% through 90% deciles of $\phi_{t,h}$ within these bins, each as a different line. We plot $y = x$ as a dashed line for reference to indicate perfect persistence.

are typically less persistent on average and the heterogeneity in persistence for negative shocks (e.g. as represented by 90-10 interquantile range) also appears to be much larger. That is, the degree of persistence of negative income shocks varies more — especially for the lowest income individuals. Furthermore, as the total income shock becomes more negative, the relationship between the shock size and persistence appears less linear.

The degree of persistence drops off rapidly at longer horizons. The lower diagram of Figure 2.6 plots the deciles for $h = 10$ case; the y-axis now corresponds to the change in expected income 10 years into the future upon receiving the shock, $\phi_{t,10} = \mathbb{E}[y_{t+10}|I_t] - \mathbb{E}[y_{t+10-1}|I_{t-1}]$. Notice that the relationship between the total shock size and persistence 10 years into the future is much flatter, although still noticeably asymmetric.

These results contrast sharply with the AR(1) income process specification from Section 2.3, in which permanent income is perfectly persistent and so we should not see any drop over time. The classical model also does not predict a gap in persistence between positive and negative income shocks, nor does it predict any heterogeneity in the degree of persistence across the distribution of shocks. Each of these features of our shock series is a substantively interesting fact about labor income risk in Iceland that cannot be explained by income processes predominantly adopted in macroeconomic structural models.

2.6 Discussion

Roles for Prediction in Social Science

Our work emphasizes an under-utilized role for prediction in the social sciences. While machine learning models cannot predict future life outcomes with high accuracy [215, 263], they can instead be used to approximate conditional expectations, and the distribution of prediction errors can be of scientific interest in its own right. In this sense, we join Lundberg et al [197] in stressing the importance of clearly defining a statistical estimand.

To estimate a conditional expectation, we need to select the best predictor relatively from among all functions of the input features, a task for which supervised learning algorithms together with sample-splitting and cross-validation are well-suited. We can at least partially validate our model by checking for conditionally mean-zero prediction errors in held out data. Here flexible regressors like gradient-boosted trees play an important role, as simpler prediction models like linear regressions are observably mis-specified in our setting, as we illustrated in Figure 2.5. Our narrowly-scoped usage of these predictors contrasts with typical applied settings, where a predictor is trained from historical data, and then deployed in real-time on newly collected data that will not generally be drawn from the same distribution as the training data. We hope to have demonstrated the utility of our methodology by illustrating the substantial inequality in the size and persistence of labor income shocks in Iceland, especially for low income individuals.

What can we do with this shock series?

One benefit of our procedure is that we produce a concrete research artifact: a series of income shocks and their persistence for every individual in every year of our sample. These shocks are interesting in their own right for studying labor income risk, see our discussion above about the distribution of shocks over quantiles of current income, and the asymmetry and heterogeneity of shock persistence. However, the principle goal is to use these shocks in downstream scientific tasks. In this section, we briefly highlight directions for future work.

First, there is a large literature on scarring during business cycles. In the United States, individuals who first entered the labor market during or immediately before the Great Recession faced worse outcomes that persisted even after the economy recovered [257]. Because we condition on both age and year, we can directly assess the size and persistence profile of shocks that occur in the Great Recession in Iceland, which could provide valuable additional evidence on scarring.

Second, we can estimate the response of household consumption to these shocks. One approach to studying the consumption response would be to estimate the average derivative of consumption with respect to these shocks and their persistence over time. Furthermore, since our shocks are computed using the full heterogeneity across observables, we would be able to break down how these consumption responses differ across income, age, education, assets, etc.

Finally, structural macroeconomic models typically use a simple autoregressive model or first-order discrete Markov chain for the income process when modelling household behavior. Typically, the parameters of this income process are estimated separately, and then the macroeconomic model is calibrated using simulated draws. Our estimates of expected future income give us a way to potentially test these macroeconomic models directly with data, subject to the limitations described above on the difference between our predictions formed with tax data, and the private future expectations of individuals.

Part II

Causal Inference with No Unobserved Confounders

Chapter 3

Duality for Balancing Weights

In Part 2, I transition from discussing pure prediction using machine learning to considering causal inference. We begin by assuming that there are no unobserved confounders, making it possible to use machine learning on observables to perform causal inference. The core issue is that the distributions of observables under the treated and control populations will generally be different, invalidating the critical assupmtion for prediction: that the train and test distributions be the same.

In Chapter 3, I will begin by estimating the difference between the train and test distributions in terms of the density ratio. In particular, we consider so-called "balancing weights" estimators, connect them to estimation of the density ratio through a duality argument. This has implications for our underlying statistical assumptions: assumptions on the density ratio versus assumptions on the outcome function end up producing identical estimates. In Chapter 4, we apply the estimates for the density ratio to causal inference using machine learning, and show that correcting for covariate shift is equivalent to "undersmoothing" — that is we would like to overfit in the training data to reduce bias, but in a principled way based on how the treated and control populations differ.

3.1 Introduction

Using covariates to transfer outcome information from one setting to another is a central task in domain adaptation, observational causal inference, and missing data imputation. These tasks share a common structure: we observe covariates and outcomes for a source data set and want to predict outcomes given covariates in a target data set, which might have a different covariate distribution than the source. One standard approach is to reweight the source distribution to have a similar covariate distribution to the target. When the source and target distributions have common support, using the density ratio for weights leads to unbiased estimation, known as *importance weighting* for domain adaptation under covariate shift [289] and *inverse propensity score weighting* (IPW) for observational causal inference [255].

Importance weighting has several drawbacks. First, using the density ratio for weights can lead to extremely large variance and unstable estimation [162, 69]. Second, the density ratio is notoriously difficult to estimate, and simple plug-in estimates do not guarantee covariate balance between the reweighted source distribution and target distributions [see 29].

Due to these drawbacks, in practice we would like to use weights with smaller variance than the density ratio that directly target a specified level of covariate balance [e.g., 117, 136]. In general, such a bias-variance trade-off only exists if we assume restrictions on the outcome model; without restrictions, only the density ratio can guarantee finite bias. This motivates the so-called *minimax balancing weights* estimators, which we study in this paper. These estimators find the minimum dispersion weights that constrain the worst-case bias between groups over an outcome function class [see 329, 130, 323, 152].

Summary of Contributions

We begin by reviewing existing balancing weights estimators, which achieve a smaller mean squared error than importance weighting by introducing an assumption on the outcome model. We argue that the outcome assumption implies two new results.

First, we use convex duality to show that the minimax optimization problem for balancing weights can be replaced with a simple convex loss over the outcome function class. Our dual formulation in Section 3.4 shows that the minimax weights are always a (rescaled and recentered) function from that class. For example, if the outcomes are bounded, the corresponding weights will be bounded. If the outcome function belongs to an RKHS with some kernel, the corresponding weights will belong to an RKHS with the same kernel. The outcome assumption pins down the shape of the balancing weights.

Second, we show that after making an outcome assumption, we do not need the density ratio to exist, i.e., we do not need to make the additional "overlap" assumption that is common in causal inference. Instead, there is an explicit quantity, the minimum achievable bias, which depends on the outcome function class, and which acts as a quantitative measure of the degree of overlap violations. We show that this measure can be more appropriate

than an overlap assumption in finite samples for quantifying the underlying difficulty of the reweighting problem.

Finally, given the central role of restrictions on the outcome model in both of the previous results, we briefly consider the setting in which this assumption is incorrect. In particular, we provide simple moment conditions under which we can retain a finite bound on the error when the true outcome model is not in the assumed class.

Related work

Estimators that target balance. Many reweighting estimators in causal inference explicitly target the discrepancy between source and target distribution, also known as *balance* [120, 329, 17, 130, 323, 294, 126, 9]. See [29] for a summary. The literature on domain adaptation also uses worst-case discrepancy between distributions [199, 117, 320, 70]. Some approaches learn representations that minimize these discrepancies [103, 272, 15]. Closely related are estimators that target the density ratio between two groups through a surrogate loss [291, 219, 290].

Overlap in causal inference and adversarial training. Many existing theoretical treatments of balancing weights require the density ratio to exist (called *overlap* in causal inference), which is typically used for proving asymptotic consistency [see 130, 152]. This assumption, however, can be highly restrictive especially in high-dimensions, as illustrated by [72]. See also [169] for a discussion of the implications of overlap violations for causal inference. The same topic arises in adversarial training. See, for example, [87, 36], who generalize ϕ-divergences to distributions that do not have common support. This idea is applied to GANs in [283, 108].

Domain adaptation and causal inference. We emphasize that domain adaptation and causal inference are both special cases of the same problem setup. Related work combines ideas from these two literatures. For example, [269, 147] use integral probability metrics to estimate causal effects without the need for an overlap assumption. The same idea is used in [152] for matching estimators in causal inference. Other work has made the connection between causal inference and adversarial training [319, 228].

3.2 Problem Setup

Let $X \in \mathcal{X}$ denote covariates, and $Y \in \mathbb{R}$ denote outcomes. We study the general class of problems with source and target populations, P and Q, with different joint distributions over X and Y. We observe X in both populations, but only observe the outcomes, Y, for the source population, P. The goal is to estimate the missing mean in the target population, $\mathbb{E}_Q[Y]$. Many important problems share this structure, including causal inference and domain adaptation.

Problem Setting 1 (Causal Inference)**.** Consider the causal inference setting with a binary treatment status variable, T, and potential outcomes $Y(0)$ and $Y(1)$. For the control group,

we observe covariates X and the potential outcome $Y(0)$. In the treated group, we still observe X but do not observe $Y(0)$. Therefore, finding the average treatment effect on the treated is equivalent to solving the problem setup described above, where P corresponds to the $T = 0$ population, Q corresponds to the $T = 1$ population, and Y corresponds to $Y(0)$.

Problem Setting 2 (Domain Adaptation)**.** Consider a classification task with features X, labels Z, and loss function ℓ. In a source environment where we observe both X and Z, we train a model h for predicting Z given X. We would like to estimate the average risk of our classifier in a new environment where we observe X but not Z. This problem is equivalent to the setup above, where P corresponds to the source environment, Q corresponds to the target environment, and Y corresponds to the loss, $\ell(h(X), Z)$.

Ignorability and Overlap

To estimate the mean of Y in Q using the outcomes from P, we require some kind of regularity between the source and target populations. A common assumption is the *ignorability* assumption (also called the *covariate shift* assumption, or *selection on observables*), which requires the relationship between covariates and outcomes to be the same across the two groups:

Assumption 1 (Ignorability)**.** *For all $x \in \mathcal{X}$,*

$$P(Y|X = x) = Q(Y|X = x).$$

Notice that in the causal inference setting described above, Assumption 1 is equivalent to the standard conditional independence assumption, $Y(0) \perp\!\!\!\perp T|X$.

Typically, Assumption 1 is paired with a requirement that the density ratio dQ/dP exists, also known as *overlap* or *continuity* in different literatures:

Definition (Overlap)**.** *We say that* overlap *holds if Q is absolutely continuous with respect to P.*

Importance Weighting

In the special case where Assumption 1 and overlap hold, we can estimate the mean of the missing outcomes by reweighting the observed outcomes with the density ratio. This estimator is called *importance weighting* or *inverse probability weighting* (IPW) and is unbiased:

$$\begin{aligned}\mathbb{E}_P\left[\frac{dQ}{dP}(X)\, Y\right] &= \mathbb{E}_P\left[\frac{dQ}{dP}(X)\, \mathbb{E}_P[Y|X]\right] \\ &= \mathbb{E}_Q[\mathbb{E}_P[Y|X]] = \mathbb{E}_Q[Y],\end{aligned}$$

where we use ignorability for the last equality.

Importance weighting has two main drawbacks. First, the overlap assumption is very strong, especially in high dimensions [72]. But even if overlap holds in the super-population, in finite samples there are usually so-called practical overlap violations: regions of the covariate space that are well-represented in the target population, but very rare in the source population, leading to large importance weights.

Mean Squared Error

Large weights lead to large mean squared error. Consider arbitrary weights $w(X)$. We will expand the mean squared error (MSE) of $\mathbb{E}_P[w(X)Y]$ for estimating $\mathbb{E}_Q[Y]$ using the standard bias-variance decomposition. Define the *outcome function*, $f_0(x) \coloneqq \mathbb{E}_P[Y|X=x] = \mathbb{E}_Q[Y|X=x]$ and likewise, let $\sigma_0^2(x)$ be the conditional variance of Y. Then,

$$\begin{aligned}
\mathrm{MSE}(w) &= \mathbb{E}_P[(w(X)Y - \mathbb{E}_Q[Y])^2] \\
&= (\mathbb{E}_P[w(X)Y] - \mathbb{E}_Q[Y])^2 + \mathrm{Var}_P[w(X)Y] \\
&= (\mathbb{E}_P[w(X)f_0(X)] - \mathbb{E}_Q[f_0(X)])^2 \qquad (3.1) \\
&\quad + \mathbb{E}_P[w(X)^2\sigma_0^2(X)]. \qquad (3.2)
\end{aligned}$$

The MSE depends on two quantities: (1) the imbalance of the mean of the outcome function f_0 between the re-weighted source distribution and the target distribution; and (2) the variability of the weights under the source distribution, which amplifies the noise in the outcomes. With practical overlap violations in high dimensional problems, $w = dQ/dP$ can be enormous, and (3.2) will result in a large MSE [see, for example, 162]. Balancing weights, introduced in Section 3.3, explicitly target the trade-off between bias and variance, as discussed extensively in [152].

Notation

We now introduce formal notation used for the remainder of the paper. Let $(\mathcal{X}, \mathcal{S})$ be a measurable space.[1] Let P and Q be given probability measures on $(\mathcal{X}, \mathcal{S})$. Let f_0 be a real-valued measurable function on $\mathcal{X}$. Denote $\mathcal{M}(\mathcal{X})$ the space of signed finite measures on $(\mathcal{X}, \mathcal{S})$ and $\mathcal{M}(P)$ those absolutely continuous with respect to P. Denote $\mathcal{P}(\mathcal{X})$ the space of probability measures on $(\mathcal{X}, \mathcal{S})$ and $\mathcal{P}(P)$ those absolutely continuous with respect to P.

With a slight abuse of notation, for measurable $f : \mathcal{X} \to \mathbb{R}$ and *both* $M \in \mathcal{M}(\mathcal{X})$ and $M \in \mathcal{P}(\mathcal{X})$, we will write $\mathbb{E}_M[f] \coloneqq \int_{\mathcal{X}} f(x)dM(x)$. We assume $\mathbb{E}_P[|f_0|] < \infty$ and $\mathbb{E}_Q[|f_0|] < \infty$.

While our setting is quite general, it may be helpful for the reader to keep in mind the case where $\mathcal{X}$ is finite and discrete with cardinality n. In this case, P and Q are probability vectors of length n and measurable functions are simply vectors in $\mathbb{R}^n$. Likewise, $\mathcal{M}(\mathcal{X})$ is just $\mathbb{R}^n$.

[1] To side-step topological issues, we assume that $\mathcal{X}$ is a separable Banach space.

3.3 Balancing Weights

In this section, we briefly review the existing work on balancing weights estimators and then introduce our main contributions. Balancing weights estimators find weights $w(X)$ with minimum dispersion, subject to a balance constraint between the target covariate distribution and the reweighted source distribution. In general, we will consider weights such that $\mathbb{E}_P[w(X)] = 1$, i.e., we always end up with the same "size" population that we started with. The problem of choosing weights can be reformulated as finding a measure $R \in \mathcal{M}(P)$ such that $\int_{\mathcal{X}} dR(x) = 1$, with $w \coloneqq dR/dP$. This corresponds to the intuition behind reweighting as creating a "pseudo-population" based on P intended to match Q. We will therefore often use w and R interchangably.

A simple balancing weights estimator might constrain the mean of the covariates to match within tolerance δ, similarly to [329]. For example, let $\mathcal{X} = \mathbb{R}^d$. We could find the minimum variance w such that

$$\|\mathbb{E}_R[X] - \mathbb{E}_Q[X]\|_2 \leq \delta, \tag{3.3}$$

where, as a reminder, $w = dR/dP$. If the outcome function $f_0(X)$ is linear with bounded coefficients, then this constraint will bound the bias term (3.2) and the tuning parameter δ lets us trade-off bias and variance to achieve a smaller MSE than the importance weights.

Assumptions on the Outcome Function

More generally, we may not want to assume that f_0 is linear. But this presents a difficulty: without any further restrictions, for any $w \neq dQ/dP$, there always exists an adversarial f_0 that can make the bias term (3.1) arbitrarily large. Only the density ratio guarantees bounded bias for *any* f_0, but often at the cost of high variance.

Therefore, in practical settings, we instead restrict the outcome function in order to control the error. To make progress, we assume that f_0 belongs to some function class $\mathcal{F}$:

Assumption 2. *The outcome function f_0 belongs to $\mathcal{F}$ where $\mathcal{F}$ is a closed and convex set of measurable real-valued functions such that for all $f \in \mathcal{F}$, $\mathbb{E}_P[|f|] < \infty, \mathbb{E}_Q[|f|] < \infty$, and $-f \in \mathcal{F}$.*

For the causal inference problem setting, Assumption 2 requires making an assumption about the relation of the potential outcome $Y(0)$ to the covariates. For the domain adaptation problem setting, the assumption is about the relationship between the accuracy of a predictor, $\ell(h(X), Z)$, to its input features, X.

Many choices of $\mathcal{F}$ in Assumption 2 are quite general and justifiable with domain knowledge. Some examples for $0 < B < \infty$ are:

$$\begin{aligned}
\text{Bounded functions: } & \mathcal{F}_\infty \coloneqq \{f : \|f\|_\infty \leq B\} \\
\text{Lipschitz functions: } & \mathcal{F}_{\text{Lip(c)}} \coloneqq \{f : \|f\|_{\text{Lip(c)}} \leq B\} \\
\text{RKHS functions: } & \mathcal{F}_\mathcal{H} \coloneqq \{f : \|f\|_\mathcal{H} \leq B\},
\end{aligned}$$

where $\|\cdot\|_{\text{Lip(c)}}$ denotes the Lipschitz constant with respect to a metric c and $\|\cdot\|_{\mathcal{H}}$ denotes the norm in some Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}$.

Under Assumption 2, the bias is bounded by the worst-case discrepancy in means over $\mathcal{F}$. This quantity is called an *integral probability metric* (IPM), defined for any set of functions, $\mathcal{G}$, and any $M, N \in \mathcal{M}(\mathcal{X})$ as:[2]

$$\text{IPM}_{\mathcal{G}}(M, N) \coloneqq \sup_{g \in \mathcal{G}} \big\{ |\mathbb{E}_M[g] - \mathbb{E}_N[g]| \big\}. \tag{3.4}$$

The bias term (1) for a re-weighted population R under Assumption 2 is upper-bounded by:

$$|\mathbb{E}_Q[f_0] - \mathbb{E}_R[f_0]| \leq \text{IPM}_{\mathcal{F}}(Q, R). \tag{3.5}$$

This value is always finite by our assumptions on $\mathcal{F}$ and we can trade it off against the variance of the weights.

Before introducing the general form of balancing weights in Section 3.3, we define two quantities that will be useful in our discussion below, the maximum and minimum bias.

Definition (Maximum and minimum bias)**.** *The maximum bias, $\delta_{\max}$, is the bias under uniform weights (when $R = P$). The minimum bias, $\delta_{\min}$, is the smallest bias achieveable by reweighting P.*

$$\delta_{\max} \coloneqq IPM_{\mathcal{F}}(Q, P) \tag{3.6}$$

$$\delta_{\min} \coloneqq \inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \big\{ IPM_{\mathcal{F}}(Q, R) \big\}. \tag{3.7}$$

Since $R = P$ is feasible for (3.7), $\delta_{\min} \leq \delta_{\max}$. In the special case where overlap holds, $R = Q$ is also feasible, which implies $\delta_{\min} = 0$.

Minimax Balancing Weights

Assumption 2 and the resulting IPM bound on the bias (3.5) lead to a generalized balancing weights estimator as discussed in [152] and [29]. Define $\sigma^2 \coloneqq \sup_{x \in \mathcal{X}} \sigma_0^2(x)$, where we assume $0 < \sigma^2 < \infty$. We can plug these bounds into the MSE to arrive at the following optimization problem:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \left\{ \text{IPM}_{\mathcal{F}}(Q, R)^2 + \sigma^2 \mathbb{E}_P \left[\left(\frac{dR}{dP} \right)^2 \right] \right\} \tag{3.8}$$

A solution always exists because the objective is finite for $R = P$, which is feasible. For $\sigma^2 > 0$, the problem is strongly convex in R and has a unique solution. Since the IPM term is itself a supremum, this estimator is sometimes referred to as minimax balancing weights.

[2] If $g \in \mathcal{G} \implies -g \in \mathcal{G}$ then the absolute value can be omitted.

Furthermore, $\exists \delta > 0$ such that (3.8) has the same minimizer as:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \mathbb{E}_P \left[\left(\frac{dR}{dP} \right)^2 \right] \quad \text{such that } \mathrm{IPM}_{\mathcal{F}}(Q, R) \leq \delta. \tag{3.9}$$

We view σ^2 and δ as exchangeable tuning parameters: σ^2 represents the importance of reducing the variance of the weights; δ represents the level of acceptable bias. For $\sigma^2 \in (0, \infty)$, the corresponding δ lies in $(\delta_{\min}, \delta_{\max})$.

Our Contributions

In this paper, we start from the premise that Assumption 2 is necessary to achieve a reasonable MSE in high dimensions leading to estimators (3.8) and (3.9). Our main argument is that Assumption 2 immediately implies two additional results.

First, we derive a general duality result that lets us rewrite problems (3.8) and (3.9) as a single convex optimization problem over $\mathcal{F}$. Therefore, we can solve the minimax balancing weights problem by optimizing a simple convex loss over a function class. Furthermore, this reformulation shows that the optimal weights are always a rescaled and recentered member of $\mathcal{F}$.

Second, we no longer need an overlap assumption. Before restricting f_0, we saw that only $w = dQ/dP$ could guarantee finite bias. Therefore, to bound the MSE we needed the density ratio to exist. But once we assume that $f_0 \in \mathcal{F}$, we no longer need the density ratio to exist, and we can simply minimize our bound on the MSE directly. Moreover, we argue that Assumption 2 provides us with a more appropriate *quantitative* measure of overlap — the minimum bias, $\delta_{\min}$ — that precisely characterizes the difficulty of translating results from one distribution to another.

3.4 Duality Theory for Balancing Weights

In this section, we derive a dual characterization of the solution, R^*, to problems (3.8) and (3.9) and the corresponding minimax weights $w^* = dR^*/dP$.

The Variance of the Weights

Our dual derivation uses the fact that the variance term can be written as a special case of a class of information-theoretic divergences called ϕ-divergences [285]. These have a variational representation that will allow us to simplify the minimax problems (3.8) and (3.9) into a single convex loss.

Definition (ϕ-Divergence)**.** *For any convex function ϕ with $\phi(1) = 0$, the ϕ-divergence between $M \in \mathcal{M}(\mathcal{X})$ and $N \in \mathcal{P}(\mathcal{X})$ is:*

$$D_\phi(M||N) \coloneqq \mathbb{E}_N\left[\phi\left(dM/dN\right)\right],$$

where $D_\phi(M||N) = \infty$ if M is not absolutely continuous with respect to N.

Notice that we can subtract off a constant to re-center our variance term in (3.9) without affecting the minimizer over R. We can then rewrite the objective as the divergence between R and P with $\phi(x) = x^2 - 1$. This is known as the χ^2 divergence, and we denote it $D_2(R||P)$:

$$\mathbb{E}_P\left[\left(\frac{dR}{dP}\right)^2 - 1\right] = D_2(R||P).$$

Variational representations. It is possible to express ϕ-divergences in a dual form, called a *variational representation*, as a supremum over measurable functions. Let $M \in \mathcal{M}(\mathcal{X})$ and let $N \in \mathcal{P}(\mathcal{X})$. Let ϕ^* denote the convex conjugate of ϕ. Then [168] and [220] show that:

$$D_\phi(M||N) = \sup_f \left\{\mathbb{E}_M[f] - \mathbb{E}_N[\phi^*(f)]\right\}, \tag{3.10}$$

where the supremum is over all real-valued measurable functions on $\mathcal{X}$. If we additionally assume, as we do for R, that $\mathbb{E}_M[1] = 1$, then we have the tighter representation,

$$D_\phi(M||N) = \sup_f \left\{\mathbb{E}_M[f] - \Lambda_N^\phi[f]\right\} \tag{3.11}$$
$$\text{where } \Lambda_N^\phi[f] \coloneqq \inf_{\lambda\in\mathbb{R}}\{\lambda + \mathbb{E}_N[\phi^*(f-\lambda)]\}.$$

This result, using the infimum over λ in the spirit of [260], appears to have been independently proposed by [7] and [35]. Under minimal conditions on ϕ, the suprema in (3.10) and (3.11) are achieved by $\phi'(dM/dN)$.

Dual Formulation

We now present our main duality result under Assumption 2 where $f_0 \in \mathcal{F}$.

Theorem 3.4.1. *Under Assumptions 1 and 2, for $\delta > \delta_{\min}$, the optimization problem (3.9) has a unique solution,*

$$\frac{dR^*}{dP} = 1 + \left(\frac{\mathbb{E}_Q[f^*] - \mathbb{E}_P[f^*] - \delta}{Var_P[f^*]}\right)\left(f^* - \mathbb{E}_P[f^*]\right),$$

where, for a unique $\mu \geq 0$ corresponding to δ, f^* achieves the following supremum:

$$\sup_{f\in\mathcal{F}}\left\{\mathbb{E}_Q[f] - \mathbb{E}_P[f] - \frac{\mu}{4}Var_P[f]\right\}. \tag{3.12}$$

The resulting MSE is:

$$MSE(R^*) \leq \delta^2 + \sigma^2 \frac{(\mathbb{E}_Q[f^*] - \mathbb{E}_P[f^*] - \delta)^2}{Var_p[f^*]}. \tag{3.13}$$

Proof Sketch. The full proof is available in the Appendix. Here we provide a brief outline. In the first step, we show that problem (3.9) is equivalent to:

$$\sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_Q[f] + \inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \{(1/\mu) D_2(R||P) - \mathbb{E}_R[f]\} \right\}$$

for some $\mu > 0$ corresponding to δ. In the second step, we apply (3.11) for the χ^2 divergence to show that the inner subproblem has an explicit solution:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \{(1/\mu) D_2(R||P) - \mathbb{E}_R[f]\} = -\mathbb{E}_P[f] - \frac{\mu}{4} \mathrm{Var}_P[f].$$

The theorem then follows from standard convex duality results.

Remark 3.4.1 (The Shape of the Weights)**.** The weights dR^*/dP are equal to f^* multiplied by some scalar s_1 and then shifted by some scalar s_2:

$$\frac{dR^*}{dP} = s_1 + s_2 f^*,$$

where s_1 and s_2 depend on both δ and f^*. Therefore, if we assume $\mathcal{F}$ is the set of quadratic functions, then the balancing weights will also be quadratic, and if we assume $\mathcal{F}$ is an RKHS with a certain kernel, then the balancing weights will belong to an RKHS with that same kernel.

Remark 3.4.2 (Other ϕ-Divergences)**.** We can replace the χ^2 divergence in the balancing weight problems (3.8) and (3.9) with other ϕ-divergences. A duality result corresponding to Theorem 3.4.1 will hold for any convex function ϕ such that $\phi(1) = 0$ with convex conjugate ϕ^* such that $\{\phi^* < \infty\} = \mathbb{R}$. See the Appendix for details. We can use this general formulation to derive corresponding duality results for entropy balancing [120] or other measures of dispersion [see 29].

Remark 3.4.3 (Tuning Parameters)**.** For every δ there is a unique corresponding μ. Therefore, we can treat μ as a tuning parameter instead of δ and solve (3.12) directly. In terms of μ, the solution to (3.9) is:

$$\frac{dR^*}{dP} = 1 + \frac{\mu}{2} (f^* - \mathbb{E}_P[f^*]) \tag{3.14}$$

and there is a closed form relationship between μ and δ given by:

$$\delta = \mathbb{E}_Q[f^*] - \mathbb{E}_P[f^*] - \frac{\mu}{2} \mathrm{Var}_P[f^*].$$

Going forward, we will often use δ and μ interchangeably.

The Full Information Case

To help illustrate Theorem 3.4.1, consider the simplified setting where we know f_0 exactly. This corresponds to a special case of Assumption 2 where $\mathcal{F}$ is the convex hull of $\{f_0, -f_0\}$. Assume without loss of generality that $\mathbb{E}_Q[f_0] \geq \mathbb{E}_P[f_0]$. Then, applying Theorem 3.4.1, we get $f^* = f_0$, and

$$\frac{dR^*}{dP} = 1 + \left(\frac{\mathbb{E}_Q[f_0] - \mathbb{E}_P[f_0] - \delta}{\mathrm{Var}_P[f_0]}\right)(f_0 - \mathbb{E}_P[f_0]).$$

The optimal weights are always a rescaled and recentered version of f_0. In this special case, the dual optimal f^* does not depend on δ; only the scaling factor does. Therefore, the MSE bound (3.13) becomes a quadratic in δ and we can solve for the optimal bias:

$$\delta^* = \left(\frac{\sigma^2}{\mathrm{Var}_P[f_0] + \sigma^2}\right)|\mathbb{E}_Q[f_0] - \mathbb{E}_P[f_0]|,$$

which gives

$$\mathrm{MSE}(w^*) \leq \left(\frac{\sigma^2}{\mathrm{Var}_P[f_0] + \sigma^2}\right)(\mathbb{E}_Q[f_0] - \mathbb{E}_P[f_0])^2.$$

This is an independently interesting result. With complete information, we can analytically find the optimal bias-variance trade-off. Under homoskedasticity, these weights have the smallest possible MSE over all w such that $\mathbb{E}_P[w] = 1$.

The Linear Case

For a second simple example, we return to the linear problem in (3.3). In this case, duality shows that balancing weights are equivalent to fitting a linear model. In fact, for a certain choice of linear $\mathcal{F}$, problem (3.12) is identical to linear regression.

Let $g : \mathcal{X} \to \mathbb{R}^d$ be some feature map. Assume that our balance constraint is:

$$\|\mathbb{E}_R[g(X)] - \mathbb{E}_Q[g(X)]\|_2 \leq \delta.$$

This is equivalent to problem (3.9) using the following linear function class:

$$f_0 \in \mathcal{F}_{\text{lin}} = \left\{\beta^T g(X) : \|\beta\|_2 \leq 1\right\}.$$

Applying Theorem 3.4.1, we know $f^* = (\beta^*)^T g(X) \in \mathcal{F}_{\text{lin}}$ and therefore the optimal weights will be linear. Solving (3.12) via calculus, we get:

$$\beta^* = c_1(\mathrm{Cov}_P[g(X)] + c_2 I)^{-1}(\mathbb{E}_Q[g(X)] - \mathbb{E}_P[g(X)])$$

for some scalars c_1 and c_1 that depend on μ. Notice the dependence on the inverse of the covariance of the features plus a regularization term. This is another way of deriving a

well-known result: for $\mathcal{F}_{\text{lin}}$, problem (3.9) is identical to estimating $\mathbb{E}_Q[Y]$ by fitting a ridge regression in the P population and then applying it to the Q population. See [152] for a direct proof. If we replace the ℓ_2-norm with the ℓ_1-norm, we obtain a similar equivalence for Lasso. When the regularization term is 0, we obtain linear regression as a special case.

Several other papers have recognized the duality between linear regression and balancing weights estimators for a linear function class; see [324, 323, 310, 294, 29]. Theorem 3.4.1 generalizes these existing duality results for linear function classes to general function classes $\mathcal{F}$.

3.5 Outcome Assumptions and Overlap

In this section, we discuss the implications of the outcome assumption for overlap. First, we show that if overlap holds, then under conditions on $\mathcal{F}$, as $\delta \to 0$, the balancing weights converge to the importance weights dQ/dP.

However, when overlap is violated, the only impact on the balancing weights estimator is that the minimum bias (which depends on our function class $\mathcal{F}$) is greater than zero. Due to Assumption 2, for any $\delta > \delta_{\min}$ there still exists a solution to (3.9) that bounds the MSE. If the variance of the outcomes is large, then we naturally want to choose δ larger than $\delta_{\min}$ and the failure of overlap has no impact on our estimator.

Instead, we argue that we should use the minimum bias, $\delta_{\min}$, directly as a measure of practical overlap violations. We illustrate that in finite samples, $\delta_{\min}$ can be large even when overlap holds in the super population, and likewise that $\delta_{\min}$ can be small even when overlap is violated in the super population. Therefore, under Assumption 2, $\delta_{\min}$ is a more precise summary of the underlying difficulty of the reweighting problem.

Convergence to Importance Weights

We begin with an example. Let $\mathcal{X} = \mathbb{R}$. Let P be Gaussian with mean 1 and variance 1, let Q be Gaussian with mean 2 and variance 1, and let p and q denote their densities. Let $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$ so that the outcome function is bounded between -1 and 1. The solution to the dual problem, f^*, and the corresponding optimal weights are illustrated in Figure 3.1.

The weights have a distinctive form. When $\delta = \delta_{max}$, the optimal weights are uniform. As the allowed bias δ decreases, the optimal weights trace out the density ratio dQ/dP but truncated above and below. This is the form of a well-known estimator in the causal inference literature, IPW with a trimmed propensity score [316]: under Assumption 2 with bounded functions, the balancing weights formulation provides formal justification for using the truncated density ratio for weights. As $\delta \to 0$, the optimal weights converge to dQ/dP.

In general, convergence to the importance weights will *always* occur as $\delta \to 0$ under certain conditions on $\mathcal{F}$.

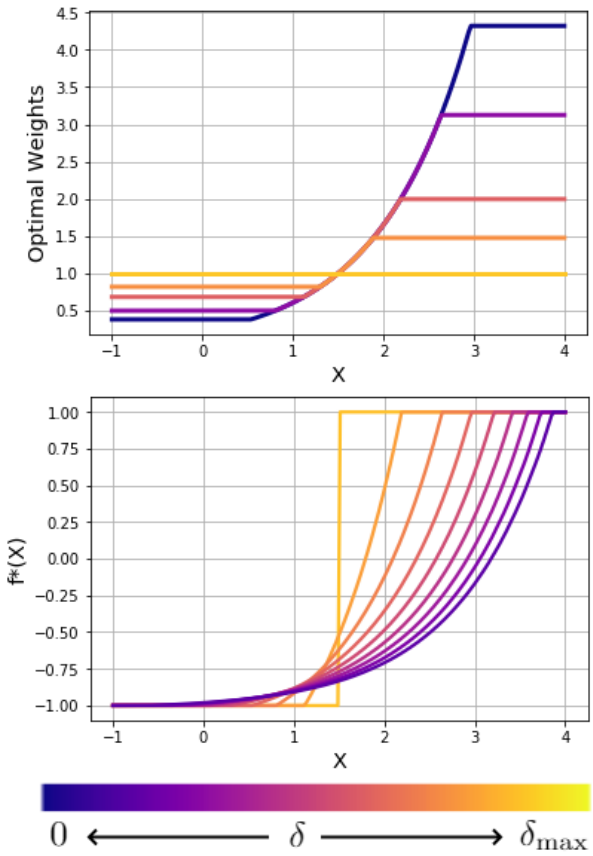


Figure 3.1: The optimal weights and corresponding dual optimal function for the Gaussian example, with δ starting at $\delta_{\max}$ and shrinking towards zero.

Definition (Distribution-defining)**.** $\mathcal{F}$ *is* distribution-defining *if* $\forall M, N \in \mathcal{P}(\mathcal{X})$,

$$IPM_{\mathcal{F}}(M, N) = 0 \iff M = N.$$

For example, $\mathcal{F}_\infty$ and $\mathcal{F}_{\text{Lip(c)}}$ are distribution-defining, as is $\mathcal{F}_{\mathcal{H}}$ for a universal kernel. When $\mathcal{F}$ is distribution-defining then only dQ/dP can achieve worst-case bias zero. Therefore, when overlap holds and $\mathcal{F}$ is distribution-defining, the optimal weights, $w^* \to dQ/dP$ as $\delta \to 0$.

This connection between balancing weights and the density ratio is not new: among others, [323] makes a similar point.

Balancing Weights Without Overlap

What if overlap does not hold? Then if $\mathcal{F}$ is distribution-defining, by definition, $\delta_{\min} > 0$. In this case, problem (3.9) still has a solution that bounds the MSE for any $\delta \geq \delta_{\min}$, but the

failure of overlap precludes us from using $\delta = 0$. However, the motivation behind balancing weights is to avoid using an unbiased estimator: if the variance of the outcomes is sufficiently high, we might still prefer to use $\delta > \delta_{\min}$.

Consider a simple example in which we reweight $P = \text{Uniform}(1, 2)$ to target $Q = \text{Uniform}(1.01, 2.01)$. While Q is not absolutely continuous with respect to P, intuitively, we should be able to find w that achieves small error because the distributions are close to each other. The function class $\mathcal{F}$ provides a formal definition of "close to each other" for the purposes of reweighting.

For these uniform P and Q, there is an irreducible bias, $\delta_{\min}$, for any possible weights:

$$\delta_{\min} = \sup_{f \in \mathcal{F}} \int_1^{1.01} f(x)dx + \sup_{f \in \mathcal{F}} \int_2^{2.01} f(x)dx$$

If $\mathcal{F}$ is unrestricted, then f could take on arbitrarily large values on the intervals $[1, 1.01]$ and $[2, 2.01]$. Therefore, the bias is unbounded without Assumption 2, which typically justifies imposing an overlap assumption. However, if we assume $\mathcal{F} \in \{f : \|f\|_\infty \leq B\}$, for example, then we have $\delta_{\min} = 0.02B$ which may be quite small.

The parameter $\delta > \delta_{\min}$ in problem (3.9) is a tuning parameter that trades off bias and variance. If the variance of the outcomes is very large, then we may prefer to use a value of δ larger than $\delta_{\min}$. In this case, the overlap violation would not have any impact on our estimator at all. On the other hand, if the variance of the outcomes is small relative to $\delta_{\min}$, we may prefer to use a value of δ close to 0. The best we could do would be to set $\delta = \delta_{\min}$; without overlap, the best achievable lower bound for the MSE is $\delta_{\min}^2$.

Quantitative Overlap

In finite samples, we argue that $\delta_{\min}$ will often be a more useful measure of overlap than the existence of the density ratio in a super-population. For example, let $P_{\text{super}} = \mathcal{N}(100, 1)$ and $Q_{\text{super}} = \mathcal{N}(-100, 1)$. Technically, overlap holds and the density ratio exists over all of $\mathbb{R}$. For concreteness, let $\mathcal{F} = \mathcal{F}_{\mathcal{H}}$ be an RKHS with a Gaussian kernel. Then for P_{super} and Q_{super}, $\delta_{\min} = 0$, because $w = dQ/dP$ will perfectly balance the RKHS. However, any finite dataset will have severe practical overlap violations. Let P be a sample of n data points from P_{super} and likewise for Q. With high probability, the points in P and the points in Q will be far apart, and as a result, $\delta_{\min}$ over the RKHS will be large.

On the other hand, if we let $P_{\text{super}} = \text{Uniform}(1, 2)$ and $Q_{\text{super}} = \text{Uniform}(1.01, 2.01)$, overlap does not hold and $\delta_{\min}$ will be non-zero for the super-population. But, for corresponding finite samples P and Q, $\delta_{\min}$ is still likely to be very small. In these examples, super-population overlap is misleading, whereas $\delta_{\min}$ is a precise quantitative summary of the difficulty of the reweighting problem for function class $\mathcal{F}$.

3.6 IHDP Example

In this section, we walk through an example on a real dataset to make the previous two sections more concrete. We apply balancing weights to the Infant Health and Development Program (IHDP) using an RKHS function class.

The IHDP Dataset and Setup

The Infant Health and Development Program (IHDP) data set is a standard observational causal inference benchmark from [129], based on data from a randomized control trial of an intensive home visiting and childcare intervention for low birth weight infants born in 1985. We consider a non-experimental subset of the original data with $n_0 = 608$ children assigned to control, $n_1 = 139$ children assigned to treatment, and $n = 747$ total children. For all children, we have a range of baseline covariates, including both categorical covariates, like the mother's educational attainment, and continuous covariates, like the child's birth weight. Our goal is to estimate the average outcome (a standardized test score) in the absence of the intensive intervention. We observe this outcome for the 608 control children, and want to re-weight these observations to estimate the missing mean for the 139 treated children.

To do so, we use an RKHS as a flexible but tractable functional form for f_0. In particular, we assume that $\mathcal{F} = \mathcal{F}_{\mathcal{H}}^{B} \coloneqq \{f : \|f\|_{\mathcal{H}} \leq B\}$ for $B < \infty$, where $\mathcal{H}$ is the RKHS induced by the Gaussian kernel,

$$\mathcal{K}(x_1, x_2) = \exp\left(-\frac{1}{2}\|x_1 - x_2\|_2^2\right).$$

Define $K \in \mathbb{R}^{n \times n}$ with $K_{ij} = \mathcal{K}(X_i, X_j)$. Then for any $f \in \mathcal{F}$, there exists an $\alpha \in \mathbb{R}^n$ such that $\alpha^T K \alpha \leq B$ and $f(X_j) = \sum_{i=1}^{n} \alpha_i K_{ij}, \forall j$.

Solving the Dual Problem

We compute the minimax balancing weights by solving the dual problem (3.12) directly for many values of the tuning parameter $\mu > 0$. The dual problem can be written as a quadratic optimization problem over the vectors α that characterize the $f \in \mathcal{F}$. See the Appendix for details. We obtain the corresponding optimal weights by plugging the resulting f^* into (3.14).

The balancing weights interpolate between two extremes. See Figure 3.2 for an illustration. At one extreme are the weights with maximum bias and minimum variance. This is achieved at $\mu = 0$, which results in uniform weights and corresponding bias $\delta = \delta_{\max}$.

At the other extreme are the weights with maximum variance and minimum bias. Since some of the covariates are continuous, the data points for the control and treated groups have disjoint support. Therefore, there are no weights that achieve zero bias. Instead, we find weights that achieve the smallest possible bias over $\mathcal{F}_{\mathcal{H}}^{B}$, $\delta_{\min}$, which will correspond to some $\mu = \mu_{\max} < \infty$. We find $\mu_{\max}$ by increasing μ until the bias stops decreasing. The corresponding weights are shown in black in Figure 3.2.

The function class $\mathcal{F}_{\mathcal{H}}^{B}$ has the nice property that the worst-case bias scales with the norm bound B, $\text{IPM}_{\mathcal{F}_{\mathcal{H}}^{B}}(R, Q) = B \cdot \text{IPM}_{\mathcal{F}_{\mathcal{H}}^{1}}(R, Q)$. Furthermore, regardless of B, the optimal weights remain identical. Therefore, we can report the bias as a fraction of the size of functions in $\mathcal{F}$. For the IHDP data, $\delta_{\max} = 0.102B$ and $\delta_{\min} = 0.089B$ with corresponding variances σ^2 (with uniform weights) and $1.7\sigma^2$. For this particular problem, we achieve most of the bias reduction with smaller weights: the intermediate weights in Figure 3.2 have $\delta = 0.090B$ with variance $1.35\sigma^2$, highlighting the relevance of the bias-variance trade-off.

In any real data set with continuous covariates, two finite samples will typically have disjoint support like we have here. The standard approach in causal inference is to assume that overlap holds in the super-populations from which the samples were drawn. In this case, we could approximate the density ratio asymptotically. However, as we emphasized in Section 3.5, the implications of overlap for balancing weights are entirely summarized by $\delta_{\min}$ so we do not need to make such an assumption.

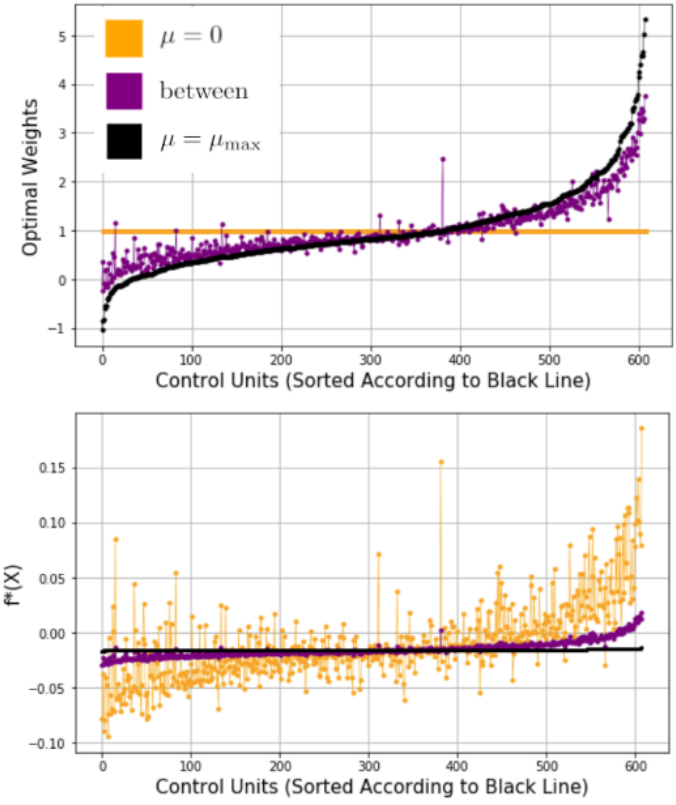


Figure 3.2: The optimal weights and corresponding dual optimal function for the IHDP example for the extreme values of μ and one intermediate value.

Remark 3.6.1 (Computational Advantages of the Dual)**.** For an RKHS, there is a closed form of the IPM, which makes the primal and dual problems equally easy to solve. But

in some situations, it is computationally easier to solve the dual problem (3.12) directly instead of the primal problem (3.9). Consider a class of neural networks parameterized by bounded network weights θ. Then handling the IPM constraint in the primal problem requires adversarial training, as in [151], which can be quite computationally challenging. On the other hand, (3.12) requires training a neural network once with a convex loss function which can be accomplished with off-the-shelf SGD.

3.7 Robustness

Balancing weights rely heavily on the function class in Assumption 2. In this section, we show that with minimal moment conditions we can still retain a bound on the bias even if we have misspecified the function class $\mathcal{F}$. We consider two functions classes. First, a *misspecified* $\mathcal{F}$ for which we solve (3.9) to find R^* such that $\text{IPM}_{\mathcal{F}}(Q, R^*) \leq \delta$. Second, the *true* function class, $\mathcal{G}$ such that $f_0 \in \mathcal{G}$ and $f_0 \notin \mathcal{F}$. To bound the bias, we need to show that

$$\text{IPM}_{\mathcal{F}}(Q, R^*) \leq \delta \implies \text{IPM}_{\mathcal{G}}(Q, R^*) \leq \rho(\delta) \tag{3.15}$$

for some $\rho < \infty$ which has good scaling with δ. Without further assumptions, (3.15) will *not* hold for any $\mathcal{G}$.

IPMs correspond to common perturbations in the robust statistics literature. For example, $\text{IPM}_{\mathcal{F}_\infty}$ and $\text{IPM}_{\mathcal{F}_{\text{Lip(c)}}}$ are equivalent to the total variation (TV) distance and Wasserstein distance respectively. For $\mathcal{F} = \mathcal{F}_\infty$, we can apply Lemma E.2 from [328] to achieve (3.15) for any $\mathcal{G}$. We require an Orlicz norm bound under Q and R^* on $g(X)$ for all $g \in \mathcal{G}$. For a simple example, let $\mathcal{G}$ be linear. Then we get the following result:

Proposition 3.7.1. *Let $TV(Q, R^*) \leq \delta$ and let $f_0 \in \{\beta^T x : \|\beta\| \leq 1\}$. If R^* and Q have bounded covariance, then we have the following upper bound on the bias:*

$$|\mathbb{E}_Q[f_0] - \mathbb{E}_{R^*}[f_0]| \leq \rho_1(\delta),$$

where $\rho_1(\delta) = O(\sqrt{\delta})$.

If instead R^* and Q are sub-Gaussian, then we have the following upper bound on the bias:

$$|\mathbb{E}_Q[f_0] - \mathbb{E}_{R^*}[f_0]| \leq \rho_2(\delta),$$

where $\rho_2(\delta) = O(\delta\sqrt{\log(1/\delta)})$.

For general $\mathcal{G}$, the rate of ρ in terms of δ is similar, but the moment conditions on X become stronger. In practice, these robust statistics results mean that we can make a best guess about $\mathcal{F}$ and as long as Q is sufficiently "nice", the true bias will not be much larger than δ.

Chapter 4

Augmented Balancing Weights as Undersmoothing

4.1 Introduction

Next, we demonstrate the implications of Chapter 3, when applied to machine learning for causal inference. Combining outcome modeling and weighting, as in augmented inverse propensity score weighting (AIPW) and other doubly robust (DR) or double machine learning (DML) estimators, is a core strategy for estimating causal effects using observational data. A growing body of literature finds weights by solving a "balancing weights" optimization problem to estimate weights directly, rather than by first estimating the propensity score and then inverting. DR versions of these estimators are referred to by a number of terms, including *augmented balancing weights* [17, 131], *automatic debiased machine learning* (AutoDML) [55], and *generalized regression estimators* (GREG) [79]; see [30] for a review. Moreover, this strategy has been applied to a wide range of linear estimands via the Riesz representation theorem [131, 56]. In this paper, we consider augmented balancing weights in which the estimators for both the outcome model and the balancing weights are based on penalized linear regressions in some possibly infinite basis; in addition to all high-dimensional linear models, this broad class includes popular nonparametric models such as kernel regression and certain forms of random forests and neural networks.

We first show that, somewhat surprisingly, augmenting any regularized linear outcome regression (the "base learner") with linear balancing weights is numerically equivalent to a single linear outcome regression applied to the target covariate profile. The resulting coefficients are an affine (and often convex) combination of the base learner model coefficients and unregularized OLS coefficients; the hyperparameter for the balancing weights estimator directly controls the regularization path defining the affine combination. In the extreme case where the weighting hyperparameter is set to zero — which we show can easily occur in practice — the entire procedure is equivalent to estimating a single, unregularized OLS regression.

We specialize these results to ridge and lasso regularization (ℓ_2 and ℓ_∞ balancing, respectively) and show that augmenting an outcome regression estimator with balancing weights generally corresponds to a form of *undersmoothing*. Most notably, we show that an augmented balancing weight estimator that use (kernel) ridge regression for both outcome and weighting models — which we refer to as "double ridge" — collapses to a single, undersmoothed (kernel) ridge regression estimator.

We leverage these results to prove novel *statistical* results for double ridge estimators and to make progress towards practical hyperparameter tuning, which remains an open problem in this area. We first make explicit the connection between asymptotic results for double kernel ridge estimators [281] and prior results on optimal undersmoothing for a single kernel ridge outcome model [209], showing that the latter is also semiparametrically efficient. This generalizes the argument in [245] that "OLS is doubly robust" to a much broader class of penalized parametric and non-parametric regression estimators. As a complementary analysis, we next adapt existing finite sample error analysis results for single ridge regression [83] to derive the finite-sample-exact bias and variance of double ridge estimators. Using these expressions, we can compute oracle hyperparameters for any given data-generating process.

Finally, we illustrate our results with several numerical examples. We first explore hyperparameter tuning for double ridge regression in an extensive simulation study on 36 data-generating processes, and compare three practical methods to the optimal hyperparameter computed using our finite sample analysis. Surprisingly, asymptotic theory and our simulation results suggest equating the hyperparameters for the outcome and weighting models. We caution against the naive application of hyperparameter tuning based solely on cross-validating the weighting model, forms of which have been suggested previously. This approach can lead to setting the weighting hyperparameter to exactly zero — and therefore recovering standard OLS — even in scenarios where OLS is far from optimal. We emphasize this point by applying our results to the canonical [184] study, highlighting that researchers can inadvertently recover OLS in practice.

Broadly, our results provide important insights into the nexus of causal inference and machine learning. First, these results open the black box on the growing number of methods based on augmented balancing weights and AutoDML — methods that can sometimes be difficult to taxonomize or understand. We show that, under linearity, these estimators all share an underlying and very simple structure. Our results further highlight that estimation choices for augmented balancing weights can lead to potentially unexpected behavior. At a high level, as causal inference moves towards incorporating machine learning and automation, our work highlights how the traditional lines between weighting and regression-based approaches are becoming increasingly blurred.

Second, our results connect two approaches to "automate" semiparametric causal inference. AutoDML and related methods exploit the fact that we can estimate a Riesz representer without a closed form expression for a wide class of functionals. The estimated Riesz representer then augments a base learner by bias correcting a plug-in estimator of the functional. Older approaches, such as undersmoothing [109, 217], twicing kernels [218], and sieve estimation [216, 273], avoid estimation of the Riesz representer, tuning the base learner

regression fit such that an additional bias correction is not required. Achieving this optimal tuning in practice has long been a hurdle for the implementation of these methods. Subject to certain conditions, both approaches can yield estimators that are asymptotically efficient. We show that if all required tuning parameters are defined in terms of an ℓ_2-norm constraint, then these approaches can be numerically identical even in finite samples. We use these equivalences to make progress toward practical hyperparameter selection and find promising directions for new theoretical analysis.

In Section 4.2 we introduce the problem setup, identification assumptions, and common estimation methods; we also review balancing weights and previous results linking balancing weights to outcome regression models. In Section 4.3 we present our new numerical results, and in Sections 4.4 and 4.5 we cache out the implications for ℓ_2 and ℓ_∞ balancing weights specifically. Building on our numerical results, Section 4.6 explores both asymptotic and finite sample statistical results for kernel ridge regression. Section 4.7 illustrates our results with a simulation study and application to canonical data sets. Section 4.8 offers some other directions for future research. The appendix includes extensive additional technical discussion and extensions.

Related work

Balancing weights and AutoDML. With deep roots in survey calibration methods and the *generalized regression estimator* [GREG; see 79, 196, 105], a large and growing causal inference literature uses balancing weights estimation in place of traditional inverse propensity score weighting (IPW). [30] provide a recent review; we discuss specific examples at length in Section 4.2 below. This approach typically balances features of the covariate distributions in the different treatment groups, with the aim of minimising the maximal design-conditional mean squared error of the treatment effect estimator. Of particular interest here are augmented balancing weights estimators that combine balancing weights with outcome regression; see, for example, [17, 131, 28].

A parallel literature in econometrics instead focuses on so-called *automatic* estimation of the Riesz representer, of which IPW are a special case, where "automatic" refers to the fact that we can estimate the Riesz representer without obtaining a closed form expression. Estimating the Riesz representer directly, under the assumption that it is linear in some basis, dates back at least to [246]; see also [245]. The corresponding augmented estimation framework has more recently come to be known as Automatic Debiased Machine Learning, or AutoDML; see, among others, [61], [63], [55], and [56]. This approach has also been applied in a range of settings, including to corrupted data [6], to dynamic treatment regimes [58], and to address noncompliance [279].

Numerical equivalences for balancing weights. Many seminal papers highlight connections between weighting approaches, such as balancing weights and IPW, and outcome modeling; see [45] for discussion. Most relevant are a series of papers that show numerical equivalences between linear regression and (exact) balancing weights, especially [245, 173,

52], and between kernel ridge regression and forms of kernel weighting [152, 130]. We discuss these equivalences at length in Appendix B.1.

4.2 Problem setup and background

Setup and motivation

The core results in our paper are numeric equivalences for existing estimation procedures, and as such these results hold absent any causal assumptions or statistical model. Nonetheless, a primary motivation for this work is the task of estimating unobserved counterfactual means in causal inference, as well as estimating the broad class of linear functionals described in [57]. We briefly review the corresponding setup, emphasizing that this is purely for interpretation.

Example: Estimating counterfactual means

Let X, Y, Z be random variables defined on $\mathcal{X}, \mathbb{R}, \mathcal{Z}$ with joint probability distribution p. To begin, consider the example of a binary treatment, $\mathcal{Z} = \{0, 1\}$ and covariates X. Define potential or counterfactual outcomes $Y(1)$ and $Y(0)$ under assignment to treatment and control, respectively. Under SUTVA [259], we observe outcomes $Y = ZY(1) + (1 - Z)Y(0)$. To estimate the average treatment effect, $\mathbb{E}[Y(1) - Y(0)]$, we first estimate the means of the partially observed potential outcomes. We initially focus on estimating $\mathbb{E}[Y(1)]$; a symmetric argument holds for $\mathbb{E}[Y(0)]$.

Let $m(x, z) \coloneqq \mathbb{E}[Y \mid X = x, Z = z]$ be the *outcome model*, $e(x) \coloneqq \mathbb{P}[Z = 1 \mid X = x]$ be the *propensity score*, and $\alpha(x, z) = z/e(x)$ be the *inverse propensity score weights* (IPW). Under the additional assumptions of *conditional ignorability*, $Y(1) \perp\!\!\!\perp Z \mid X$, and *overlap*, $\mathbb{E}[\alpha(X, Z)^2] < \infty$, we have that $\mathbb{E}[Y(1)]$ is identified by $\mathbb{E}[m(X, 1)]$, a linear functional of the observed data distribution.

There are three broad strategies for estimating $\mathbb{E}[Y(1)]$. First, the identifying functional above suggests estimating the outcome model, $m(x, 1)$ among those units with $Z = 1$, and plugging this into the *regression functional*, $\mathbb{E}[m(X, 1)]$. Second, the equality $\mathbb{E}[m(X, 1)] = \mathbb{E}[Z/e(X)Y] = \mathbb{E}[\alpha(X, Z)Y]$ suggests estimating the inverse propensity score weights, $\alpha(x, z) = z/e(x)$, and plugging these into the *weighting functional*. Finally, we can combine these two via the *doubly robust functional* [248]:

$$\mathbb{E}[m(X, 1) + \alpha(X, Z)(Y - m(X, 1)].$$

This functional has the attractive property of being equal to $\mathbb{E}[m(X, 1)]$ even if either one of α or m is replaced with an arbitrary function of X and Z, hence the term "doubly robust." Doubly robust estimators have been studied extensively in semiparametric theory; note that $m(X, 1) + \alpha(X, Z)(Y - m(X, Z)) - \psi(m)$ coincides with the efficient influence function for $\psi(m)$ under a nonparametric model (see Kennedy 2022 [167] for a review of the relevant theory). See [60, 167] for recent overviews of the active literature in causal inference and machine learning focused on estimating versions of this functional.

General class of functionals via the Riesz representer

Our results apply well beyond the example above. In particular, they apply to any functional of the form

$$\psi(m) = \mathbb{E}[h(X_i, Z_i, m)], \tag{4.1}$$

where $\mathcal{Z}$ is an arbitrary set; Z a random variable with support $\mathcal{Z}$; and h is a real-valued, mean-squared continuous linear functional of m [57, 131, 55]. Following [55, 56], we can generalize the weighting functional to this general class of estimands via the *Riesz representer*, which is a function $\alpha(X, Z) \in L_2(p)$ such that, for all square-integrable functions $f \in L_2(p)$:

$$\mathbb{E}[h(X, Z, f)] = \mathbb{E}[\alpha(X, Z) f(X, Z)]. \tag{4.2}$$

As in the counterfactual mean example, we can identify the more general target functional in (4.2) via the outcome regression functional in (4.1), via the Riesz representer functional in (4.2) with $f = m$, or via the doubly robust functional

$$\mathbb{E}[h(X, Z, m) + \alpha(X, Z)(Y - m(X, Z))]. \tag{4.3}$$

Estimators of this DR functional are *augmented* in the sense that they augment the "plug-in," "outcome regression," or "base learner" estimator of $\mathbb{E}[h(X, Z, m)]$ with appropriately weighted residuals; or, equivalently, that augment the weighting estimator with an appropriate outcome regression. This is the class of estimators to which our results apply. In future work we will explore whether we can extend our results to a different class of functionals that admit DR functional forms, first introduced by [246], and to the superset of such functionals characterized by [258].

Balancing weights: Background and general form

The core idea behind balancing weights is to estimate the Riesz representer directly — rather than via an analytic functional form (e.g., by estimating the propensity score and inverting it). As a result, balancing weights do not require a known analytic form for the Riesz representer [56], are often much more stable [329], and offer improved control of finite sample covariate imbalance [323]. We briefly describe two primary motivations for this approach.

First, a central property of the Riesz representer is that the corresponding weights, $w(X, Z) = \alpha(X, Z)$, are the unique weights that satisfy the *population balance property* property in Equation (4.2) for all square-integrable functions $f \in L_2(p)$. For our target estimand $\psi(m)$ we only need to satisfy the condition in Equation (4.2) for the special case of $f = m$. If we are willing to assume that m lies in a model class $\mathcal{F} \subset L_2(p)$, then it suffices to balance functions in that class. This is achieved by minimizing the imbalance over $\mathcal{F}$:

$$\text{Imbalance}_{\mathcal{F}}(w) \coloneqq \sup_{f \in \mathcal{F}} \Big\{ \mathbb{E}[w(X, Z) f(X, Z)] - \mathbb{E}[h(X, Z, f)] \Big\}. \tag{4.4}$$

As we discuss next, balancing weights minimize a (penalized) sample analog of Equation (4.4).

Alternatively, [55] consider finding weights f that minimize the mean-squared error for $\alpha(X, Z)$:

$$\min_{f \in \mathcal{F}} \left\{ \mathbb{E}\left[(f(X,Z) - \alpha(X,Z))^2 \right] \right\}. \tag{4.5}$$

Automatic estimation of the Riesz representer, also known as *Riesz regression* [59], minimizes a sample analog of Equation (4.5). When $\mathcal{F}$ is convex, then up to choice of hyperparameters (see (4.6) below), the solutions to Equations (4.4) and (4.5) are equivalent.

Linear balancing weights

In this paper, we consider the special case in which the outcome models are linear in some basis expansion of X and Z. This is an extremely broad class that encompasses linear and polynomial models of arbitrary functions of X and Z and with dimension possibly larger than the sample size, as well as non-parametric models such as reproducing kernel Hilbert spaces [RKHSs; 116], the Highly-Adaptive Lasso [31], the neural tangent kernel space of infinite-width neural networks [140], and "honest" random forests [5]. However, this class excludes models for m that are fundamentally non-linear in their parameters, like general neural networks or generalized linear models with a non-linear link function.

Under linearity, the imbalance over all $f \in \mathcal{F}$ has a simple closed form. Because our results concern numeric equivalences, we will focus on the finite sample version of the linear balancing weights problem. Let $\mathcal{F} = \{f(x,z) = \theta^\top \phi(x,z) : \|\theta\| \leq 1\}$ where $\|\cdot\|$ can be any norm on $\mathbb{R}^d$. The general setup constrains $\|\theta\| \leq r$; we set $r = 1$ without loss of generality, which simplifies exposition below. Let $\|\cdot\|_*$ be the *dual norm* of $\|\cdot\|$; that is, $\|v\|_* \coloneqq \sup_{\|u\| \leq 1} u^\top v$. Many common vector norms have familiar, closed-form, dual norms, e.g., the dual norm of the ℓ_2-norm is the ℓ_2-norm; and the dual norm of the ℓ_1-norm is the ℓ_∞-norm. Let X_p, Y_p, Z_p be n i.i.d. samples from the distribution p of the observed data. Define the feature map $\phi : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ and let $\phi_j : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ denote the mapping for the jth feature. Define $\Phi_p \coloneqq \phi(X_p, Z_p)$ and let $\Phi_q \coloneqq h(X_p, Z_p, \phi)$ denote the *target features*. We will write $\hat{\mathbb{E}}$ for sample averages; define $\overline{\Phi}_p \coloneqq \hat{\mathbb{E}}[\Phi_p]$ and $\overline{\Phi}_q \coloneqq \hat{\mathbb{E}}[\Phi_q]$. For exposition, we assume that $d < n$ and that Φ_p has rank d. We emphasize that this is not necessary for our results — one can replace $\mathbb{R}^d$ with an infinite-dimensional Hilbert space $\mathcal{H}$ and relax the rank restriction. See Appendix B.2 for a formal presentation of the high-dimensional ($d > n$) setting.

In what follows we write w for the $1 \times n$ vector $w(\Phi_p)$, to highlight the fact that we will estimate w directly rather than as an explicit function of X or Φ_p. Using the derivation above, we can directly calculate the finite sample imbalance as:

$$\widehat{\text{Imbalance}}_{\mathcal{F}}(w) = \|\tfrac{1}{n} w \Phi_p - \bar{\Phi}_q\|_*.$$

Now we can write the penalized sample analog of balancing weights optimization problem in (4.4) equivalently as either:

$$\text{Penalized form:} \qquad \min_{w\in\mathbb{R}^n} \left\{ \|\tfrac{1}{n} w\Phi_p - \bar{\Phi}_q\|_*^2 + \delta_1 \|w\|_2^2 \right\}$$

$$\text{Constrained form:} \qquad \min_{w\in\mathbb{R}^n} \|w\|_2^2 \quad \text{such that } \|\tfrac{1}{n} w\Phi_p - \bar{\Phi}_q\|_* \le \delta_2.$$

Furthermore, we can write the equivalent problem in (4.5) as:

$$\text{Riesz regression form:} \qquad \min_{\theta\in\mathbb{R}^d} \left\{ \tfrac{1}{n}\theta^\top(\Phi_p^\top\Phi_p)\theta - \tfrac{1}{n} 2\theta^\top \bar{\Phi}_q + \delta_3\|\theta\| \right\}, \tag{4.6}$$

where we use the terminology "Riesz regression" from [59]. For any parameter $\delta_2 > 0$ and corresponding constrained problem solution $\hat{w}$, there exists a parameter $\delta_3 > 0$ such that $\hat{w} = \delta_3 \Phi_p \hat{\theta}$, where $\hat{\theta}$ is the solution to the Riesz regression form. As a result, for any norm $\|\cdot\|$, the penalized and constrained forms will always produce weights that are linear in Φ_p [see 30, Section 9]. Therefore, since the problems are equivalent, we typically use a generic δ to denote the regularization parameter, and will specify the particular form only if necessary. In Appendix B.1 we illustrate several concrete examples for this problem.

Remark 4.2.1 (Intercept)**.** An important constraint in practice is to normalize the weights, $\frac{1}{n}\sum_{i=1}^n w_i = 1$. This corresponds to replacing Φ_p and Φ_q with their centered forms, $\Phi_p - \bar{\Phi}_p$ and $\Phi_q - \bar{\Phi}_p$, in the dual form of the balancing weights problem. This is also equivalent to adding a column of 1s to Φ_p. Appropriately accounting for this normalization, however, unnecessarily complicates the notation. Therefore, without loss of generality, we will assume that the features are centered throughout, that is, $\bar{\Phi}_p = 0$.

Remark 4.2.2 (Equivalence with kernel ridge regression)**.** For the special case of ℓ_2 balancing (as in Appendix B.1) the balancing weights problem is numerically equivalent to directly estimating the conditional expectation $\mathbb{E}[Y_p|\Phi_p]$ via (kernel) ridge regression and applying the estimated coefficients to $\overline{\Phi}_q$. Moreover, the solution to the balancing weights problem has a closed form that is always linear in $\overline{\Phi}_q$; we provide further details in Appendix B.1. For exact balance with $\delta = 0$, the balancing weights problem is equivalent to fitting unregularized OLS; see, for example, [245], [173], and [51].

4.3 Novel equivalence results for (augmented) balancing weights and outcome regression models

Our first main result demonstrates that *any* linear balancing weights estimator is equivalent to applying OLS to the re-weighted features. Our second result provides a novel analysis of

augmented balancing weights, demonstrating that augmenting any linear balancing weights estimator with a linear outcome regression estimator is equivalent to a plug-in estimator of a new linear model with coefficients that are a weighted combination of estimated OLS coefficients and the coefficients of the original linear outcome model.

Weighting alone

Our first result is that estimating $\psi(m)$ with any linear balancing weights is equivalent to fitting OLS for the regression of Y_p on Φ_p and then applying those coefficients to the re-weighted target feature profile. The key idea for this result begins with the simple unregularized regression prediction for $\psi(m)$, $\overline{\Phi}_q\hat{\beta}_{\text{ols}}$.

Proposition 4.3.1. *Let $\hat{w}^\delta := \hat{\theta}^\delta\Phi_p^\top$, $\hat{\theta}^\delta \in \mathbb{R}^d$, be any linear balancing weights, with corresponding weighted features $\hat{\Phi}_q^\delta := \frac{1}{n}\hat{w}^\delta\Phi_p$. Let $\hat{\beta}_{ols} = (\Phi_p^\top\Phi_p)^\dagger\Phi_p^\top Y_p$ be the OLS coefficients of the regression of Y_p on Φ_p. Then:*

$$\begin{aligned}\hat{\mathbb{E}}\left[\hat{w}^\delta \circ Y_p\right] &= \hat{\Phi}_q^\delta\hat{\beta}_{ols} \\ &= \left(\bar{\Phi}_p + \widehat{\Delta}^\delta\right)\hat{\beta}_{ols},\end{aligned}$$

where $\widehat{\Delta}^\delta = \hat{\Phi}_q^\delta - \bar{\Phi}_p$ is the mean feature shift implied by the balancing weights and where superscript δ indicates possible dependence on a hyperparameter. We have assumed without loss of generality that $\bar{\Phi}_p = 0$, but we sometimes use $\hat{\Delta}$ notation to demonstrate the role of mean feature shift in various expressions. We use the symbol $\circ$ to denote element-wise multiplication.

Note that here we have written the OLS coefficients using the pseudo-inverse †. For clarity in the main text, we focus on the full rank setting, where $(\Phi_p^\top\Phi_p)^\dagger = (\Phi_p^\top\Phi_p)^{-1}$; we provide a proof for the general setting in Appendix B.2.

We can interpret this result via a contrast with standard regularization. Regularized regression models navigate a bias-variance trade-off by regularizing estimated coefficients $\hat{\beta}_{\text{reg}}$ relative to $\hat{\beta}_{\text{ols}}$, leading to $\overline{\Phi}_q\hat{\beta}_{\text{reg}}$. The balancing weights approach instead keeps $\hat{\beta}_{\text{ols}}$ fixed and regularizes the target feature distribution by penalizing the implied feature shift, $\widehat{\Delta}^\delta = \hat{\Phi}_q^\delta - \overline{\Phi}_p$.

We emphasize that this is a new and quite general result. As we discuss in Appendix B.1, it has been shown previously that for exact balancing weights, $\hat{\mathbb{E}}[\hat{w}_{\text{exact}}Y_p] = \overline{\Phi}_q\hat{\beta}_{\text{ols}}$. However, Proposition 4.3.1 holds for any weights of the form $w = \theta\Phi_p^\top$ with arbitrary $\theta \in \mathbb{R}^d$. In Sections 4.4 and 4.5, we consider the particular form of $\hat{\Phi}_q^\delta$ for ℓ_2 and ℓ_∞ balancing, respectively.

Augmented balancing weights

We can immediately extend this to augmented balancing weights, which regularize *both* the coefficients and the feature shift. Let $\hat{\beta}_{\text{reg}}^\lambda$ be the coefficients of any regularized linear model

for the relationship between Y_p and Φ_p, where the superscript λ indicates dependence on a hyperparameter (e.g., estimated by regularized least squares). We consider augmenting $\hat{\mathbb{E}}\left[\hat{w}^\delta \circ Y_p\right]$ with $\hat{\beta}^\lambda_{\text{reg}}$ using the doubly robust functional representation in Equation (4.3). The augmented estimator is:

$$\hat{\mathbb{E}}[\Phi_q \hat{\beta}^\lambda_{\text{reg}}] + \hat{\mathbb{E}}[\hat{w}^\delta \circ (Y_p - \Phi_p \hat{\beta}^\lambda_{\text{reg}})] = \hat{\mathbb{E}}[\hat{w}^\delta \circ Y_p] + \hat{\mathbb{E}}\left[\left(\Phi_q - \hat{\Phi}^\delta_q\right)\hat{\beta}^\lambda_{\text{reg}}\right]. \tag{4.7}$$

Many recently proposed estimators have this form; see e.g., [17, 30]. If the weighting model and outcome model have different bases, our result applies to a shared basis by either combining the dictionaries as in [55] or by applying an appropriate projection as in [131].

We apply Proposition 4.3.1 to the first term of the right-hand side of (4.7) to yield the following result. As this result is purely numerical, it applies to arbitrary vectors $\hat{\beta}^\lambda_{\text{reg}} \in \mathbb{R}^d$, but substantively we think of $\hat{\beta}^\lambda_{\text{reg}}$ as the estimated coefficients from an outcome model.

Proposition 4.3.2. *For any $\hat{\beta}^\lambda_{reg} \in \mathbb{R}^d$, and any linear balancing weights estimator with estimated coefficients $\hat{\theta}^\delta \in \mathbb{R}^d$, and with $\hat{w}^\delta \coloneqq \hat{\theta}^\delta \Phi_p^\top$ and $\hat{\Phi}^\delta_q \coloneqq \frac{1}{n}\hat{w}^\delta \Phi_p$, the resulting augmented estimator*

$$\begin{aligned}
&\hat{\mathbb{E}}[\hat{w}^\delta \circ Y_p] + \hat{\mathbb{E}}\left[\left(\Phi_q - \hat{\Phi}^\delta_q\right)\hat{\beta}^\lambda_{reg}\right] \\
&= \hat{\mathbb{E}}\left[\hat{\Phi}^\delta_q \hat{\beta}_{ols} + \left(\Phi_q - \hat{\Phi}^\delta_q\right)\hat{\beta}^\lambda_{reg}\right] \\
&= \hat{\mathbb{E}}[\Phi_q \hat{\beta}_{aug}],
\end{aligned}$$

where the jth element of $\hat{\beta}_{aug}$ is:

$$\begin{aligned}
\hat{\beta}_{aug,j} &\coloneqq \left(1 - a^\delta_j\right)\hat{\beta}^\lambda_{reg,j} + a^\delta_j \hat{\beta}_{ols,j} \\
a^\delta_j &\coloneqq \frac{\widehat{\Delta}^\delta_j}{\Delta_j},
\end{aligned}$$

where $\Delta_j = \overline{\Phi}_{q,j} - \overline{\Phi}_{p,j}$ is the observed mean feature shift for feature j; and $\widehat{\Delta}^\delta_j = \hat{\Phi}^\delta_{q,j} - \overline{\Phi}_{p,j}$ is the feature shift for feature j implied by the balancing weights model. Finally, $a^\delta \in [0,1]^d$ when the covariance matrix is diagonal, $(\Phi_p^\top \Phi_p) = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$, with $\sigma_j^2 > 0$.

This is our central numerical result for augmented balancing weights: when both the outcome and weighting models are linear, the augmented estimator is equivalent to a linear model applied to the target features Φ_q, with coefficients that are element-wise affine combinations of the base learner coefficients, $\hat{\beta}^\lambda_{\text{reg}}$, and the coefficients $\hat{\beta}_{\text{ols}}$ from an OLS regression of Y_p on Φ_p. (The coefficients are additionally *convex* combinations of $\hat{\beta}^\lambda_{\text{reg}}$ and $\hat{\beta}_{\text{ols}}$ when the covariance matrix is diagonal.) In Sections 4.4 and 4.5 below, we analyze some of the properties of the augmented estimator for ℓ_2 and ℓ_∞ balancing weights problems respectively.

The regularization parameter for the balancing weights problem, δ, parameterizes the path between $\hat{\beta}^{\lambda}_{\text{reg}}$ and $\hat{\beta}_{\text{ols}}$. To see this, consider the cases where $\delta \to 0$ and $\delta \to \infty$. As $\delta \to 0$ the balancing weights problem prioritizes minimizing balance over controlling variance, and $\widehat{\Delta}^{\delta}_j \to \Delta_j$ for all j. (Recall that we assume $\overline{\Phi}_{p,j} = 0$ for all j. Thus, $\Delta_j = \overline{\Phi}_{q,j}$ and $\widehat{\Delta}^{\delta}_j = \hat{\Phi}^{\delta}_{q,j}$. So $\widehat{\Delta}^{\delta}_j \to \Delta_j$ is equivalent to $\hat{\Phi}^{\delta}_q \to \overline{\Phi}_{q,j}$.) In this case, $a^{\delta}_j = \widehat{\Delta}^{\delta}_j / \Delta_j \to 1$, and the weights fully "de-bias" the original outcome model by recovering unregularized regression, $\hat{\beta}_{\text{aug}} \to \hat{\beta}_{\text{ols}}$. In Section 4.7, we will see that when chosen by cross-validation, δ sometimes equals exactly 0 in applied problems; thus even when $\hat{\beta}^{\lambda}_{\text{reg}}$ is a sophisticated regularized estimator, the final augmented point estimate can nonetheless be numerically equivalent to the simple OLS plug-in estimate. Conversely, as $\delta \to \infty$, the balancing weights problem prioritizes controlling variance, leading to uniform weights and $\widehat{\Delta}_j \to 0$. In this case, $a^{\delta}_j = \widehat{\Delta}^{\delta}_j / \Delta_j \to 0$, the weighting model does very little, and $\hat{\beta}_{\text{aug}} \to \hat{\beta}^{\lambda}_{\text{reg}}$.

It is also instructive to consider two other extremes: unregularized outcome model and unregularized balancing weights. First, consider the special case of fitting an unregularized linear regression outcome model, i.e., $\hat{\beta}^{\lambda}_{\text{reg}} = \hat{\beta}_{\text{ols}}$. Then Proposition 4.3.2 reproduces the result, originally due to [245], that "OLS is doubly robust" [see also 173]. This is because $\hat{\beta}_{\text{aug}} = \hat{\beta}_{\text{ols}}$ for arbitrary linear weights $\hat{\theta}^{\delta} \in \mathbb{R}^d$. Thus, OLS augmented by *any* choice of linear balancing weights collapses to OLS alone. Equivalently, we can view OLS alone as an augmented estimator that combines an OLS base learner with linear balancing weights.

A similar result holds for unregularized balancing weights, i.e., exact balancing weights. Let $\hat{w}_{\text{exact}}$ be the solution to a balancing weights problem in Section 4.2 with hyperparameter $\delta = 0$, and let $\hat{\beta}^{\lambda}_{\text{reg}} \in \mathbb{R}^d$ be arbitrary coefficients. Then from the balance condition, $\hat{\Phi}_q = \overline{\Phi}_q$, $a^{\delta}_j = 1$ for all j, and we have that $\hat{\beta}_{\text{aug}} = \hat{\beta}_{\text{ols}}$. Thus, the augmented exact balancing weights estimator also collapses to the OLS regression estimator. Equivalently, the augmented exact balancing weights estimator collapses to the *unaugmented* exact balancing weights estimator. [324] use a very similar result to argue that entropy balancing, a form of exact balancing weights, is doubly robust.

4.4 Augmented ℓ_2 Balancing Weights

In this section, we study ℓ_2 balancing weights estimators, which are commonly used in the context of kernel balancing [116, 130, 152, 27] and for panel data methods [1, 28]. We first show that the regularization path a^{δ}_j from Proposition 4.3.2 follows typical ridge regression shrinkage, with a smooth decay. Moreover, augmenting with ℓ_2 balancing weights is equivalent to boosting with ridge regression, and always overfits relative to the unaugmented outcome model alone. We then show that when the outcome model used to augment ℓ_2 balancing weights is also a ridge regression (which we refer to as "double ridge"), the augmented estimator is itself equivalent to a single, generalized ridge regression, albeit undersmoothed relative to the base learner. These results extend immediately to the RKHS setting of "double kernel ridge" estimation, combining kernel balancing weights and kernel ridge regression. In

Section 4.6, we show the implications of these numeric results for undersmoothing in the statistical sense.

While the following results hold for arbitrary covariance matrices, in the main text we simplify the presentation by assuming that $\Phi_p^\top \Phi_p$ is diagonal; that is, $(\Phi_p^\top \Phi_p) = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$, with $\sigma_j^2 > 0$. This is without loss of generality for ℓ_2 balancing since the ℓ_2-norm is rotation invariant.

General linear outcome model

Following Remark 4.2.2 above, ℓ_2 balancing weights, including kernel balancing weights, have a closed form that is always linear in $\overline{\Phi}_q$. Our next result applies this closed form to Proposition 4.3.2 to derive the regularization path that results from augmenting an arbitrary linear outcome model with ℓ_2 balancing weights. Although this is an immediate consequence of Proposition 4.3.2, the resulting form of the augmented estimator has unique structure that warrants a new result.

Proposition 4.4.1. *Let $\hat{w}_{\ell_2}^\delta$ be (penalized) linear balancing weights with regularization parameter δ and $\mathcal{F} = \{f(x) = \theta^\top \phi(x) : \|\theta\|_2 \leq 1\}$. Then $\frac{1}{n}\hat{w}_{\ell_2}^\delta = \overline{\Phi}_q(\Phi_p^\top \Phi_p + \delta I)^{-1}\Phi_p^\top$. Therefore, the augmented ℓ_2 balancing weights estimator with outcome model $\hat{\beta}_{reg}^\lambda \in \mathbb{R}^d$ has the form*

$$\hat{\mathbb{E}}[\Phi_q \hat{\beta}_{reg}^\lambda] + \hat{\mathbb{E}}[\hat{w}_{\ell_2}^\delta (Y_p - \Phi_p \hat{\beta}_{reg}^\lambda)] = \hat{\mathbb{E}}[\Phi_q \hat{\beta}_{\ell_2}],$$

where the jth coefficient of $\hat{\beta}_{\ell_2}$ is given by

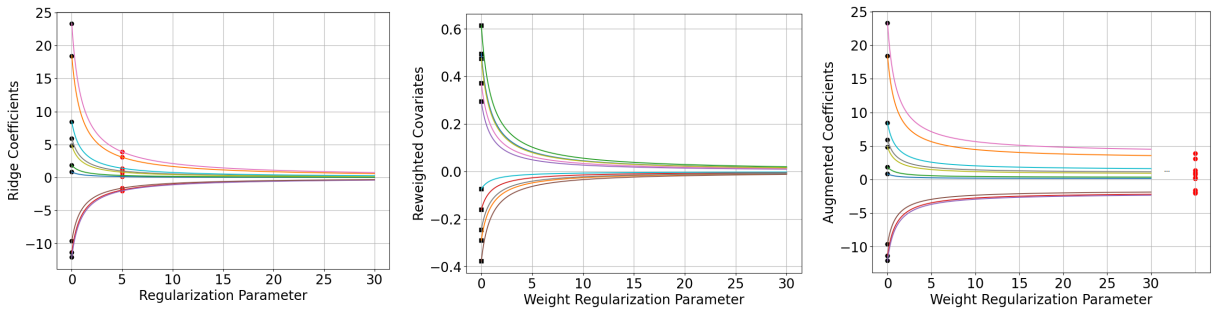
$$\begin{aligned} \hat{\beta}_{\ell_2,j} &\coloneqq \left(1 - a_j^\delta\right) \hat{\beta}_{reg,j}^\lambda + a_j^\delta \hat{\beta}_{ols,j} \\ a_j^\delta &\coloneqq \frac{\sigma_j^2}{\sigma_j^2 + \delta}. \end{aligned} \tag{4.8}$$

In this case, the a_j^δ are exactly equal to the standard regularization path of ridge regression. To see this, recall that ridge regression with penalty δ shrinks the $\hat{\beta}_{\text{ols}}$ coefficients as follows:

$$\hat{\beta}_{\text{ridge},j}^\delta = \left(\frac{\sigma_j^2}{\sigma_j^2 + \delta}\right) \hat{\beta}_{\text{ols},j} = a_j^\delta \hat{\beta}_{\text{ols},j}. \tag{4.9}$$

This is identical to the expression in (4.8) but with $\hat{\beta}_{\text{reg}}^\lambda$ set to 0: Ridge regression shrinks $\hat{\beta}_{\text{ols}}$ towards 0 with regularization path a_j^δ, while ℓ_2 augmenting shrinks $\hat{\beta}_{\text{ols}}$ towards $\hat{\beta}_{\text{reg}}^\lambda$ with the same regularization path.

As an illustration, the right panel of Figure 4.1 shows $\hat{\beta}_{\ell_2}$ (on the y-axis) for ten covariates, with δ increasing from 0 (on the x-axis). The dots on the left pick out $\hat{\beta}_{\text{ols}}$; when $\delta = 0$, then $a_j^0 = 1$ and $\hat{\beta}_{\ell_2} = \hat{\beta}_{\text{ols}}$. The limit on the right shows $\hat{\beta}_{\text{reg}}^\lambda$. The smooth regularization path is characteristic of ridge regression shrinkage.



(a) Outcome model (b) Weighting model (c) Augmented model

Figure 4.1: Regularization paths for "double ridge" augmented ℓ_2 balancing weights. Panel (a) shows the coefficients $\hat{\beta}^{\lambda}_{\text{reg}}$ of a ridge regression of Y_p on Φ_p with hyperparameter λ. The black dots on the left are the OLS coefficients, with $\lambda = 0$. The red dots at $\lambda = 5$ illustrate the coefficients at a plausible hyperparameter value, $\hat{\beta}^{5}_{\text{reg}}$. Panel (b) shows re-weighted covariates, $\hat{\Phi}^{\delta}_{q}$, for the ℓ_2 balancing weights problem with hyperparameter δ; the black dots show exact balance, which corresponds to OLS. As δ increases, the weights converge to uniform weights and $\hat{\Phi}^{\delta}_{q}$ converges to $\overline{\Phi}_p$, which we have centered at zero. Panel (c) shows the augmented coefficients, $\hat{\beta}_{\ell_2}$ as a function of the weight regularization parameter δ. The black dots on the left are the OLS coefficients. As $\delta \to \infty$, the coefficients converge to $\hat{\beta}^{5}_{\text{reg}}$. All three regularization paths have essentially identical qualitative behavior.

We can also view $\hat{\beta}_{\ell_2}$ as the output of a single iteration of a ridge boosting procedure, fit using Y_p and Φ_p alone. See [46] and [232] for detailed discussion; [218] makes a similar connection in the context of twicing kernels.

Proposition 4.4.2. *Let $\check{Y}_p = Y_p - \Phi_p \hat{\beta}^{\lambda}_{reg}$ be the residuals from the base learner. Let $\hat{\beta}^{\delta}_{boost}$ be the coefficients from the ridge regression of $\check{Y}_p$ on Φ_p with hyperparameter δ. Then, $\hat{\beta}_{\ell_2} = \hat{\beta}^{\lambda}_{reg} + \hat{\beta}^{\delta}_{boost}$, and $\|Y_p - \Phi_p \hat{\beta}_{\ell_2}\|_2^2 \leq \|Y_p - \Phi_p \hat{\beta}^{\lambda}_{reg}\|_2^2$.*

So for a fixed δ, the augmented ℓ_2 balancing estimator is equivalent to estimating a new outcome model coefficient estimator $\hat{\beta}_{\ell_2}$ that *overfits* relative to $\hat{\beta}^{\lambda}_{\text{reg}}$ (in the sense of having smaller in-sample training error), and then applying that model to Φ_q.

Surprisingly — and in contrast to the general result in Proposition 4.3.2 — the augmented coefficients $\hat{\beta}_{\ell_2}$ are the same for *every* target covariate profile Φ_q. To see this, note that Proposition 4.4.1 shows that ℓ_2 balancing weights are always linear in $\overline{\Phi}_q$. Therefore, the corresponding regularization path a^{δ}_{j} does not depend on the target profile Φ_q; it depends only on δ and the source distribution variances σ^2_j. This property is closely related to *universal adaptability* in the computer science literature on multi-group fairness [172]. The particular Φ_q may nonetheless impact the choice of δ in hyperparameter selection, e.g., via cross-validating imbalance, which in turn influences the degree of overfitting; we do find this to be the case theoretically in Section 4.6.

Ridge regression outcome model

Proposition 4.4.1 holds for arbitrary linear outcome model coefficient estimators $\hat{\beta}^{\lambda}_{\text{reg}} \in \mathbb{R}^d$; we now state the corresponding result for a "double ridge" estimator, where the base learner outcome model is itself fit via ridge regression. The key takeaway is that the implied augmented coefficients are *undersmoothed* relative to the base learner ridge coefficients.

For this section, we will consider the following generalized ridge regression, sometimes known as "adaptive" ridge regression [115]. Let $\Lambda \in \mathbb{R}^{d \times d}$ be a diagonal matrix with jth diagonal entry $\lambda_j \geq 0$. Then the generalized ridge coefficients are:

$$\begin{aligned}\hat{\beta}^{\Lambda}_{\text{ridge}} &\coloneqq \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \|\Phi_p \beta - Y_p\|_2^2 + \beta^\top \Lambda \beta \\ &= (\Phi_p^\top \Phi_p + \Lambda)^{-1} \Phi_p^\top Y_p.\end{aligned}$$

Standard ridge regression is the special case where the λ_j all take the same value and so $\Lambda = \lambda I$. As above, the generalized ridge coefficients can be rewritten as shrinking the OLS coefficients:

$$\hat{\beta}^{\Lambda}_{\text{ridge},j} = \left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda_j}\right) \hat{\beta}_{\text{ols},j}. \tag{4.10}$$

We now demonstrate that the augmented ℓ_2 balancing weights estimator with base learner $\hat{\beta}^{\Lambda}_{\text{ridge}}$ is equivalent to a plug-in estimator using generalized ridge with *smaller* hyperparameters, $\hat{\beta}^{\Gamma}_{\text{ridge}}$, where Γ is a diagonal matrix with jth diagonal entry $\gamma_j \in [0, \lambda_j]$.

Proposition 4.4.3. *Let $\hat{\beta}^{\Lambda}_{ridge}$ denote the coefficients of a generalized ridge regression of Y_p on Φ_p with hyperparameters Λ, and let $\hat{w}^{\delta}_{\ell_2}$ denote ℓ_2 balancing weights with hyperparameter δ defined in Section 4.2. Define the diagonal matrix Γ with jth diagonal entry:*

$$\gamma_j \coloneqq \frac{\delta \lambda_j}{\sigma_j^2 + \lambda_j + \delta} \leq \lambda_j.$$

Then:

$$\hat{\mathbb{E}}[\Phi_q \hat{\beta}^{\Lambda}_{ridge}] + \hat{\mathbb{E}}[\hat{w}^{\delta}_{\ell_2}(Y_p - \Phi_p \hat{\beta}^{\Lambda}_{ridge})] = \hat{\mathbb{E}}[\Phi_q \hat{\beta}^{\Gamma}_{ridge}].$$

Furthermore, $\hat{\beta}^{\Gamma}_{ridge}$ are standard ridge regression coefficients (i.e., γ_j is a constant for all j) when $\lambda_j = \lambda$ and $\sigma_j = \sigma$ for all j.

The same result holds for kernel ridge regression; see Appendix B.2.

In this setting, augmenting with balancing weights is equivalent to undersmoothing the original outcome model fit. In particular, we can use the expansion in Equation (4.10) to see the undersmoothing in $\hat{\beta}^{\Gamma}_{\text{ridge}}$ explicitly:

$$\frac{\sigma_j^2}{\sigma_j^2 + \gamma_j} = \underbrace{\left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda_j}\right)}_{\text{outcome model}} \underbrace{\left(\frac{\sigma_j^2 + \lambda_j + \delta}{\sigma_j^2 + \delta}\right)}_{\text{augmentation}},$$

where the first term is the shrinkage from the original generalized ridge model alone, and the second term is due to augmenting with ℓ_2 balancing weights. Importantly, the second term is in $[1, \frac{\sigma_j^2+\lambda_j}{\sigma_j^2}]$ and therefore partially reverses the shrinkage of the original estimate. In Section 4.6, we connect this to undersmoothing in the statistical sense.

4.5 Augmented ℓ_∞ balancing weights

In this section, we study ℓ_∞ balancing weights estimators, which are widely used in the balancing weights literature [329, 17] and in the AutoDML literature [55]. In the main text, we consider the special case where the covariance matrix $\Phi_p^\top \Phi_p$ is diagonal; that is, $(\Phi_p^\top \Phi_p) = \text{diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$, with $\sigma_j^2 > 0$. Unlike with ℓ_2 balancing, this is no longer without loss of generality.

For diagonal covariance, we first show that ℓ_∞ balancing has a closed form: it is equivalent to applying a soft-thresholding operator to the feature shift from $\overline{\Phi}_p$ to $\overline{\Phi}_q$. We then write the resulting augmented estimator as applying coefficients $\hat{\beta}_{\ell_\infty}$ to Φ_q and show that $\hat{\beta}_{\ell_\infty}$ is a sparse, element-wise convex combination of the base learner coefficients and OLS coefficients. When the outcome model is also fit via the lasso, we use the resulting representation to demonstrate a familiar "double selection" phenomenon [26], where $\hat{\beta}_{\ell_\infty}$ inherits the non-zero coefficients of both the base learner and the weighting model. This is a form of undersmoothing in the ℓ_0 "norm," in the sense that $\hat{\beta}_{\ell_\infty}$ always has at least as many non-zero coefficients as the base learner, $\hat{\beta}_{\text{reg}}$.

Weighting alone

We first define the soft-thresholding operator and show that the ℓ_∞ balancing problem has a closed form solution.

Definition (Soft-thresholding operator)**.** *For $t > 0$, define the soft-thresholding operator,*

$$\mathcal{T}_t(z) := \begin{cases} 0 & \text{if } |z| < t \\ z - t & \text{if } z > t \\ z + t & \text{if } z < -t \end{cases}.$$

Proposition 4.5.1 (ℓ_∞ Balancing)**.** *If $\Phi_p^\top \Phi_p$ is diagonal, the solution $w_{\ell_\infty}^\delta$ to the ℓ_∞ optimization problem (B.3) is:*

$$\begin{aligned} \tfrac{1}{n} w_{\ell_\infty}^\delta &= \Phi_p (\Phi_p^\top \Phi_p)^{-1} \left[\overline{\Phi}_p + \mathcal{T}_\delta(\overline{\Phi}_q - \overline{\Phi}_p) \right] \\ &= \Phi_p (\Phi_p^\top \Phi_p)^{-1} \left[\overline{\Phi}_p + \mathcal{T}_\delta(\Delta) \right] \end{aligned}$$

where $\Delta = \overline{\Phi}_q - \overline{\Phi}_p$, where we include $\overline{\Phi}_p$ (equal to 0 by assumption) to emphasize the dependence on feature shift, and with corresponding reweighted features, $\hat{\Phi}_q^\delta = \overline{\Phi}_p + \mathcal{T}_\delta(\overline{\Phi}_q - \overline{\Phi}_p)$.

For intuition, compare the (un-augmented) ℓ_∞ balancing weights estimator to the lasso-based coefficient estimates [125]:

$$\hat{\mathbb{E}}[w^\delta_{\ell_\infty} \circ Y_p] = \mathcal{T}_\delta(\overline{\Phi}_q)^\top \hat{\beta}_{\text{ols}}$$
$$\hat{\mathbb{E}}[\Phi_q \hat{\beta}^\lambda_{\text{lasso}}] = \overline{\Phi}_q^\top \mathcal{T}_\lambda(\hat{\beta}_{\text{ols}}),$$

where we simplify $\hat{\Phi}^\delta_q$ here to emphasize the connections between the methods. Whereas lasso performs soft-thresholding on the OLS coefficients (regularizing the outcome regression), ℓ_∞ balancing performs soft-thresholding on the implied feature shift to the target features.

General linear outcome model

We can then plug the closed-form solution for the weights into Proposition 4.3.2.

Proposition 4.5.2. *Let $\hat{w}^\delta_{\ell_\infty}$ be defined as above. Then the augmented ℓ_∞ balancing weights estimator with outcome model fit $\hat{\beta}^\lambda_{reg} \in \mathbb{R}^d$ has the form,*

$$\hat{\mathbb{E}}[\Phi_q \hat{\beta}^\lambda_{reg}] + \hat{\mathbb{E}}[\hat{w}^\delta_{\ell_\infty}(Y_p - \Phi_p \hat{\beta}^\lambda_{reg})] = \hat{\mathbb{E}}[\Phi_q \hat{\beta}_{\ell_\infty}],$$

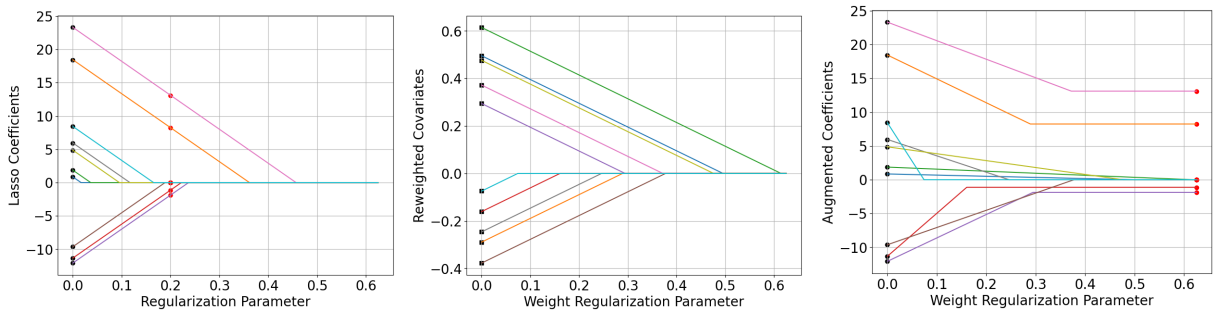
where the jth coefficient of $\hat{\beta}_{\ell_\infty}$ equals:

$$\hat{\beta}_{\ell_\infty,j} = \begin{cases} \hat{\beta}^\lambda_{reg,j} & \text{if } |\Delta_j| < \delta \\ \left|\frac{\delta}{\Delta_j}\right| \hat{\beta}^\lambda_{reg,j} + \left(1 - \left|\frac{\delta}{\Delta_j}\right|\right) \hat{\beta}_{ols,j} & \text{otherwise} \end{cases},$$

where $\Delta_j = \overline{\Phi}_{q,j} - \overline{\Phi}_{p,j}$.

The augmented coefficients $\hat{\beta}_{\ell_\infty}$ are an element-wise convex combination of $\hat{\beta}^\lambda_{\text{reg}}$ and $\hat{\beta}_{\text{ols}}$. For features where the mean feature shift Δ_j is small (relative to δ), $\hat{\beta}_{\ell_\infty}$ is equivalent to the base learner coefficient $\hat{\beta}^\lambda_{\text{reg}}$. The remaining coefficients are interpolated linearly toward the $\hat{\beta}_{\text{ols}}$ coefficients.

Figure 4.2 summarizes these results and their implications for the augmented estimator. As with Figure 4.1, we generate simple simulated data with $d = 10$. In the left panel, we plot the coefficients from lasso regression of Y_p on Φ_p as a function of the lasso regularization parameter. The regularization path begins with the black dots, which represent the OLS coefficients. Each lasso coefficient (represented by a colored line) then shrinks linearly to exactly zero, due to the soft-thresholding operator. The middle panel plots the reweighted covariates using ℓ_∞ balancing weights between Φ_p and Φ_q solved in the constrained form. The black dots represent $\overline{\Phi}_q$, corresponding to exact balance. Then as the weight regularization parameter increases, the reweighted covariates shrink linearly to exactly zero, just as in lasso. The right panel plots coefficients for the augmented estimator that combines a baseline outcome model fit $\hat{\beta}^\lambda_{\text{reg}}$ with ℓ_∞ balancing weights. The lines correspond to $\hat{\beta}_{\ell_\infty}$ as defined



(a) Outcome model (b) Weighting model (c) Augmented model

Figure 4.2: Regularization paths for "double lasso" augmented ℓ_∞ balancing weights. Panel (a) shows the coefficients $\hat{\beta}^{\lambda}_{\text{reg}}$ of a lasso regression of Y_p on Φ_p with hyperparameter λ. The black dots on the left are the OLS coefficients, with $\lambda = 0$. The red dots at $\lambda = 0.2$ illustrate the coefficients at a plausible hyperparameter value, $\hat{\beta}^{0.2}_{\text{reg}}$. Panel (b) shows re-weighted covariates, $\hat{\Phi}^{\delta}_{q}$, for the ℓ_∞ balancing weights problem with hyperparameter δ; the black dots show exact balance, which corresponds to OLS. As δ increases, the weights converge to uniform weights and $\hat{\Phi}^{\delta}_{q}$ converges to $\overline{\Phi}_p$, which we have centered at zero. Panel (c) shows the augmented coefficients, $\hat{\beta}_{\ell_\infty}$ as a function of the weight regularization parameter δ. The black dots on the left are the OLS coefficients. As $\delta \to \infty$, the coefficients converge to $\hat{\beta}^{0.2}_{\text{reg}}$. All three regularization paths show the typical lasso "soft thresholding" behavior. The regularization path for the augmented estimator also shows "double selection" behavior.

in Proposition 4.5.2. The regularization path begins at the black dots, where $\hat{\beta}_{\ell_\infty} = \hat{\beta}_{\text{ols}}$, and eventually converges to $\hat{\beta}^{\lambda}_{\text{reg}}$, showing the usual soft-thresholding behavior. The order at which the coefficients go to zero reflects the size of $\overline{\Phi}_q$, because the regularization path depends on the weight coefficients from the middle panel. Thus, the augmented estimator shrinks $\hat{\beta}_{\text{ols}}$ toward $\hat{\beta}^{\lambda}_{\text{reg}}$ but via a soft-thresholding operator applied to the feature shift, Δ_j.

Lasso outcome model

In the case where $\hat{\beta}^{\lambda}_{\text{reg}}$ is itself fit via lasso, as studied in [55], then we recover a familiar double selection phenomenon [26].

Proposition 4.5.3 (Double Selection)**.** *Let $\hat{\beta}^{\lambda}_{lasso}$ denote the coefficients of lasso regression of Y_p on Φ_p with regularization parameter λ. Denote the indices of the non-zero coefficients as I_λ. Let $\hat{w}^{\delta}_{\ell_\infty}$ be ℓ_∞ balancing weights with parameter δ as in Proposition 4.5.1. Let I_δ denote the non-zero entries of the reweighted covariates $\hat{\Phi}_q$. Assume that $\hat{\beta}_{ols}$ is dense. Then the indices of the non-zero entries of the augmented coefficients $\hat{\beta}_{\ell_\infty}$ are $I_{aug} = I_\lambda \cup I_\delta$.*

The lasso coefficients have a sparsity pattern generated by soft-thresholding the OLS coefficients. The augmented estimator then shrinks from OLS toward $\hat{\beta}^{\lambda}_{\text{reg}}$ by soft-thresholding

the implied feature shift to the target features. As a result, wherever the lasso coefficients are non-zero *or* the weight coefficients are non-zero, the final augmented coefficients are also non-zero. The "included coefficients" for the final estimator are then the union of the coefficients included in either individual model. Therefore, augmenting a lasso outcome model with ℓ_∞ balancing also exhibits a form of undersmoothing in the ℓ_0 "norm", $\|\hat{\beta}_{\ell_\infty}\|_0$, in the sense that there are always at least as many non-zero coefficients as for the unaugmented lasso outcome model. However, this will not correspond to undersmoothing the base learner in the traditional sense, because in general there will not exist a lasso hyperparameter λ that will produce sparsity pattern I_{aug}.

As noted by, for example, [295], the double selection estimator may suffer from imprecision due to adjustment for covariates that are associated with treatment but not outcome. One could in principle remove covariates that are only predictive of the treatment, but this can jeopardize statistical inference. We refer to [207] for a discussion on navigating this trade-off.

4.6 Kernel Ridge Regression: Asymptotic and Finite Sample Analysis

The results above are *numerical*: they hold without any statistical or causal assumptions. However, the connection between augmented estimators and outcome models also presents *statistical* insights that we discuss here. In particular, we leverage the numerical result that double (kernel) ridge regression — which uses ridge regression for fitting both the outcome and weighting models — is equivalent to a single, undersmoothed outcome ridge regression plug-in estimator.

First, we consider an asymptotic analysis in Section 4.6: we use this equivalence to make explicit the connection between asymptotic results for augmented balancing weights with kernel ridge regression and prior results on optimal undersmoothing of a kernel ridge plug-in estimator. As a result, optimally undersmoothed kernel ridge regression inherits whatever guarantees can be proven for augmented ridge regression. An implication is that we can generalize the insight from [245] that OLS is doubly robust to a wider class of non-parametric estimators. This equivalence also suggests an appropriate hyperparameter scheme when the outcome regression is an element of an RKHS.

Second, we consider a finite sample analysis in Section 4.6: we use this equivalence to derive the finite-sample design-conditional mean squared error of augmented kernel ridge regression. We then use this expression to characterize finite-sample-optimal hyperparameter tuning. We turn to hyperparameter tuning in practice in the next section.

Asymptotic Results

We now use our results in Proposition 4.4.3 to make explicit the connection between two otherwise distinct sets of asymptotic results. First, [312] and [278] argue that double kernel ridge regression can deliver $\sqrt{n}$-consistent estimation of functionals in certain scenarios.

[312] also proposes an optimally undersmoothed ℓ_2 balancing weights estimator. Separately, [130] and [209] propose optimally undersmoothed (single) kernel ridge outcome regression. Since, as we have shown in Proposition 4.4.3 (see also Remark 2), these three procedures are equivalent, we can connect these results and show that plug-in estimators based on optimally undersmoothed kernel ridge regression or ℓ_2 balancing weights can be $\sqrt{n}$-consistent. Moveover, results on RKHSs suggest a simple heuristic for hyperparameter choice. We give the high-level argument here and defer additional technical details to Appendix B.5.

To move from numerical results to statistical results, we must place some constraints on the data generating process. Assume that we observe n iid samples of (x_i, y_i, z_i) from p. Define $K \in \mathbb{R}^{n\times n}$ to be the kernel matrix with i,j-th entry $K_{ij} = k((x_i, z_i), (x_j, z_j))$. Let σ_j^2 denote the eigenvalues of K. We assume that $\sigma_j^2 = \sigma^2 > 0$ is constant for all j; we can relax this at the cost of additional complexity. The "single" kernel ridge regression outcome regression estimator with parameter λ has coefficient estimates:

$$\hat{\beta}^{\lambda}_{\text{ridge}} = (K + \lambda I)^{-1} y.$$

Applying Proposition 4.4.3, the augmented "double kernel ridge" estimator with hyperparameter δ is equivalent to a plug-in estimate for a new kernel ridge model:

$$\hat{\beta}_{\text{aug}} = (K + \gamma I)^{-1} y, \qquad \text{with } \gamma = \frac{\lambda\delta}{\sigma^2 + \lambda + \delta}.$$

For statistical guarantees, we must typically allow the hyperparameters to change with n; let γ_n, λ_n and δ_n then denote sequences of hyperparameters. In the doubly robust framework, one can choose λ_n and δ_n in a way that is MSE-optimal for prediction purposes whilst ensuring that the bias of the augmented estimator is small. For two functions of n, f_n and g_n, let $f_n \asymp g_n$ denote that $f_n = O(g_n)$ and $g_n = O(f_n)$. Then due to special properties of RKHS geometry, it follows that δ_n can be of the same order as λ_n, that is $\delta_n \asymp \lambda_n$ [281, Theorem 5.2]. In the next section, we consider setting $\delta = \lambda$ for hyperparameter tuning in practice; our Proposition 4.4.3 then implies that $\gamma_n \asymp \lambda_n^2$. We note more generally that Proposition 4.4.3 implies that $\gamma_n \asymp \lambda_n\delta_n$.

There are two important cases to consider. When the RKHS is finite dimensional, the choice $\lambda_n = \delta_n = n^{-1/2}$ is optimal for controlling the prediction error for both the outcome and weighting models [48, 281]. The augmented estimator is then equivalent to a single ridge regression with hyperparameter $\gamma_n \asymp n^{-1}$, which matches the rate of [130, 209]. Hence, this approach will always undersmooth relative to the MSE-optimal hyperparameter for a single ridge regression.

When the RKHS is infinite-dimensional, we find that the undersmoothed hyperparameter implied by the augmented procedure can take on a range of asymptotic rates, both faster and slower than n^{-1}, depending on effective dimension and smoothness; we give concrete examples in the Appendix. This somewhat contrasts with the results in [130, 209]. In this sense, Proposition 4.4.3 generalizes the standard undersmoothing arguments, which typically change the regularization schedule from $n^{-1/2}$ to n^{-1}.

Remark 4.6.1 (Single-model double robustness)**.** Another interesting implication of the equivalence of these two procedures is that the single kernel ridge procedure is doubly robust, much the same way OLS is. Because estimating the coefficients from an OLS regression of Y onto features of (Z, X) is equivalent to a balancing weights or an IPW estimator based on a model for the inverse weights that is linear in the same features, this procedure is consistent whenever *either* the weights or the outcome model is truly linear–that is, whenever either of these two linear models is correctly specified [245]. Similarly, the single kernel ridge procedure is doubly robust in that it is consistent if either the true outcome regression or the inverse propensity score is consistently estimated. However, valid inference in the case where the inverse weight model but *not* the outcome model is truly linear will typically require different tuning parameter selection.

Finite Sample Mean-Squared Error

We now use our numerical equivalences to write out the exact finite-sample mean squared error of the augmented kernel ridge estimator: by re-writing the augmented balancing weights estimator as a single outcome model, we can immediately leverage existing results from [83].

Following their setup, we define the diagonal matrix $\hat{\Sigma} \coloneqq \frac{1}{n}\Phi_p^T\Phi_p$; if $\hat{\Sigma}$ is not diagonal, we can apply a rotation (without loss of generality). We consider ridge regression with rescaled hyperparameter λ and solution $(\hat{\Sigma} + \lambda I)^{-1}\Phi_p Y_p/n$; this is equivalent to standard ridge regression above with hyperparameter $n\lambda$, and also accommodates kernel ridge regression with appropriate choice of Φ_p. Assume that $Y_p = \Phi_p\beta_0 + \epsilon$ with $\beta_0 \in \mathbb{R}^d$, and where $\epsilon \in \mathbb{R}^n$ are iid with mean zero and variance σ^2. Then the exact, design-conditional, squared bias and variance of the ridge regression prediction applied to a new iid sample $(\Phi_{\text{new}}, Y_{\text{new}}) \sim p$ are:

$$B_{\text{p}}^2(\lambda) = \lambda^2\beta_0^T(\hat{\Sigma} + \lambda I)^{-1}\mathbb{E}[\Phi_p^T\Phi_p](\hat{\Sigma} + \lambda I)^{-1}\beta_0$$
$$V_{\text{p}}(\lambda) = \frac{\sigma^2}{n}\text{tr}\left[\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}\mathbb{E}[\Phi_p^T\Phi_p](\hat{\Sigma} + \lambda I)^{-1}\right].$$

Applying Proposition 4.4.3, we can simlarly derive the squared bias and variance of an augmented ridge estimator for our linear functional estimand; we denote these quantities B_q^2 and V_q respectively. We express the bias and variance in terms of the two hyperparameters, λ and δ:

Proposition 4.6.1. *Let σ_j^2 denote the eigenvalues of $\hat{\Sigma}$ and define $\Gamma_{\lambda,\delta}$ to be the diagonal matrix with non-zero entries $\gamma_j \coloneqq \frac{\delta\lambda}{\sigma_j^2+\delta+\lambda}$. Then,*

$$B_q^2(\lambda, \delta) = \beta_0^T(\hat{\Sigma} + \Gamma_{\lambda,\delta})^{-1}\Gamma_{\lambda,\delta}\mathbb{E}[\Phi_q]^T\mathbb{E}[\Phi_q]\Gamma_{\lambda,\delta}(\hat{\Sigma} + \Gamma_{\lambda,\delta})^{-1}\beta_0$$
$$V_q(\lambda, \delta) = \frac{\sigma^2}{n}tr\left[\hat{\Sigma}(\hat{\Sigma} + \Gamma_{\lambda,\delta})^{-1}\mathbb{E}[\Phi_q]^T\mathbb{E}[\Phi_q](\hat{\Sigma} + \Gamma_{\lambda,\delta})^{-1}\right].$$

In the next section, we compare — numerically and via simulation — existing hyperparameter selection schemes to the optimal trade-off between B_q^2 and V_q. However, first

we note that the analysis above opens up exciting new avenues for both theoretical and methodological work. One could theoretically analyze the mean squared error to understand how the optimal δ scales with the problem parameters; for example, by using proportionate asymptotics from random matrix theory as in the high-dimensional ridge regression literature [124]. Second, our analysis here suggests a novel, more complex hyperparameter selection scheme based directly on the finite sample analysis. We leave this to future work.

4.7 Numerical illustrations and hyperparameter tuning

This section illustrates our results in practice. We first explore hyperparameter tuning for double ridge regression, comparing practical methods to the optimal hyperparameter computed using our results from Proposition 4.6.1. Following our asymptotic results in 4.6, we recommend equating the weighting and outcome model hyperparameters in practice. We then apply both double ridge and lasso-augmented ℓ_∞-balancing to two versions of the canonical [184] application. An important theme throughout is that some approaches for hyperparameter selection frequently lead to $\delta = 0$, which collapses the augmented estimate to OLS alone — even in settings where this is far from optimal. Overall, we take this as a warning that existing hyperparameter tuning schemes can be potentially misleading when applied naively.

Hyperparameter tuning for ridge-augmented ℓ_2 balancing

We begin with practical hyperparameter tuning for the special case of double ridge, building on the MSE expression in Section 4.6. There is an active literature on selecting hyperparameters for augmented balancing weights estimators and double machine learning estimators more broadly [152, 310, 30, 19]. We contribute to this literature by comparing practical hyperparameter tuning schemes with an oracle hyperparameter tuning scheme based on Proposition 4.6.1.

Reflecting empirical practice, we focus here on choosing hyperparameters sequentially: we first select the outcome model hyperparameter λ (e.g. by cross-validation) and then select the weighting model hyperparameter δ. Ultimately, we find strong performance for both *CV imbalance* and *CV outcome* hyperparameters, as defined below. We especially recommend the latter as a reasonable starting point in practice. In additional to theoretical support from our asymptotic analysis, the outcome model hyperparameter scheme does not require any additional algorithm or code after having fit the initial outcome model.

Oracle and practical hyperparameter tuning

Oracle hyperparameter. To compute oracle hyperparameters, we first compute the prediction-MSE-optimal λ using the standard ridge regression MSE expression, and then we use Proposition 4.6.1 to compute the corresponding optimal δ for the linear functional

estimand:

$$\lambda^* \coloneqq \text{argmax}_\lambda\{B_p^2(\lambda) + V_p(\lambda)\}$$
$$\delta^* \coloneqq \text{argmax}_\delta\{B_q^2(\lambda^*, \delta) + V_q(\lambda^*, \delta)\}.$$

While there is not a closed form for δ^*, we can nonetheless directly compute this optimal hyperparameter and characterize its behavior under a range of scenarios. We draw several conclusions about optimal δ^* for a wide range of DGPs of the form $Y_p = \Phi_p\beta_0 + \epsilon$. First, δ^* is generally increasing in the noise, σ^2: larger σ^2 typically implies larger δ^*. Second, δ^* generally depends on the target mean, $\mathbb{E}[\Phi_q]$; that is, two DGPs that are identical except for $\mathbb{E}[\Phi_q]$ can have different values of δ^*. The optimal hyperparameter, however, does *not* depend on the magnitude of the shift in the target mean: replacing $\mathbb{E}[\Phi_q]$ with $c\mathbb{E}[\Phi_q]$ for $c \neq 0$, scales both the bias and variance by c^2, leaving δ^* unchanged.

Practical hyperparameter. We compare the oracle hyperparameter with three implementable practical proposals. In all cases, we first pick λ by cross-validating the mean squared error of a ridge outcome model.

- *CV imbalance.* Choose δ by cross-validating the estimated imbalance, $\|\frac{1}{n}\hat{w}\Phi_p - \bar{\Phi}_q\|_2^2$, adapting a proposal from [310].
- *CV Riesz loss.* Choose δ by cross-validating the Riesz loss in Equation (4.6), adapting a proposal from [55]; this is the dual form of cross-validating the estimated imbalance.
- *CV outcome.* Choose δ to be equal to the cross-validated ridge outcome λ, as inspired by the asymptotic theory in [281].

Before presenting simulation results, we provide a preliminary analytic discussion, comparing these practical schemes to the behavior of the oracle δ^*. For the first two proposals: just like the oracle, both depend on the target mean $\mathbb{E}[\Phi_q]$ and are invariant to re-scaling. However, these two approaches are mechanically independent of the outcomes Y_p, unlike the oracle δ^* which, in general, depends on the variance of the outcomes. On the other hand, the last proposal depends on the outcomes Y_p but is mechanically independent of $\mathbb{E}[\Phi_q]$.

This suggests that any one of these tuning parameter approaches cannot perform well across all DGPs. In future work, if we pursue a theoretical analysis of the oracle hyperparameter, e.g. in a proportionate asymptotics framework, we may be able predict when either the outcomes or the covariate shift is more important. In this work we begin by demonstrating that no one tuning scheme does uniformly best in simulations.

Simulation study

To assess the behavior of these hyperparameter tuning schemes, we conduct an extensive simulation study using 36 distinct data-generating processes, 30 synthetic and 6 semi-synthetic;

Method	# of DGPs		Relative MSE			Prop.($\delta = 0$)
	Best	Worst	Median	Best	Worst	
CV Outcome	10	3	0.58	0.097	2×10^5	0
CV Imbalance	25	2	0.39	0.001	2×10^5	0
CV Riesz Loss	1	31	3,454	0.23	3×10^7	0.56

Table 4.1: Mean-squared error (relative to the oracle) for four hyperparameter selection methods for *double ridge regression* from a numerical investigation of 36 data generating processes (30 synthetic and 6 semi-synthetic). The final column is the proportion of draws where the hyperparameter $\delta = 0$.

see Appendix B.3 for a detailed discussion. For each DGP, we directly compute the oracle hyperparameter using the results in Section 4.6. We then compute values from the three practical hyperparameter tuning methods discussed above. The mean squared error that we consider is design-conditional, and so we draw samples of the covariates for each DGP only once.

Table 4.1 presents a summary of the MSE for the three methods across the 36 DGPs. Overall, we find that the *CV outcome* approach of choosing $\delta = \lambda$ and the *CV imbalance* approach both perform well in practice: these two achieve the lowest MSE in 35 of the 36 DGPs, with CV imbalance performing slightly better on average. By contrast, selecting δ via CV for the Riesz loss has numerical stability problems that compromises performance. The performance for the *outcome* and *balance* approaches, on the other hand, seem to degrade gracefully and rarely perform catastrophically. Taken together, these preliminary findings suggest researchers should begin with these two tuning methods as defaults.

Recovering the OLS point estimate. As we discuss above (see, e.g., Figure 4.1), when $\delta = 0$ the point estimate for the augmented balancing weights estimator is numerically identical to the OLS point estimate. Thus, when a hyperparameter tuning procedure chooses $\delta = 0$ in practice, researchers are simply estimating the equivalent of OLS — even if they are unaware they are doing so. This is especially problematic in settings where OLS is far from optimal [though see 174, 124, for counterexamples]. In our synthetic and semi-synthetic DGPs, $\delta = 0$ is never optimal, and is usually associated with a very large error driven by extreme variance. Thus the fact that hyperparameter tuning procedures can return $\delta = 0$ in these DGPs represents a pathological case.

In our simulation study, we find that, when cross validating the Riesz loss, over half of all draws returned $\delta = 0$. By contrast, none of the other methods returned $\delta = 0$ in the synthetic DGPs, though, as we discuss below, we do observe exact zeros for δ occasionally when cross-validating imbalance in the standard LaLonde dataset. This further highlights the numeric instability of hyperparameter tuning via CV for the Riesz loss, at least in the settings we consider here. We further suggest that in these cases, practitioners assess the

sensitivity of the $\delta = 0$ results to the particular tuning procedure used or to the random choice of cross-validation splits.

Application to LaLonde 1986

We now illustrate our equivalence and hyperparameter tuning results on real-world datasets. Following [55], we focus on the canonical [184] data set evaluating a job training program in the National Supported Work (NSW) Demonstration. The primary outcome of interest is annual earnings in 1978 dollars.

For these illustrations, we estimate the Average Treatment Effect on the Treated (ATT), $\mathbb{E}[Y(1) - Y(0) \mid Z = 1]$. We recover the missing conditional mean $\mathbb{E}[Y(0) \mid Z = 1]$ using the setup from Example 3 in Appendix B.1, where the source and target populations are the control and treated units respectively. Thus Φ_p and Φ_q correspond to the feature expansion $\phi(X)$ applied to the covariates in the control group and treated group respectively. We consider two different features expansions of the original covariates: (1) a "short" set of 11 covariates used in [77];[1] and (2) an expanded, "long" set of 171 interacted features used in [91].

Our goal is to explicate how augmented estimators under different hyperparameter tuning schemes undersmooth in practice in both low and high-dimensional settings. In some cases, the augmented estimator collapses to exactly OLS as we document above.

High-dimensional setting

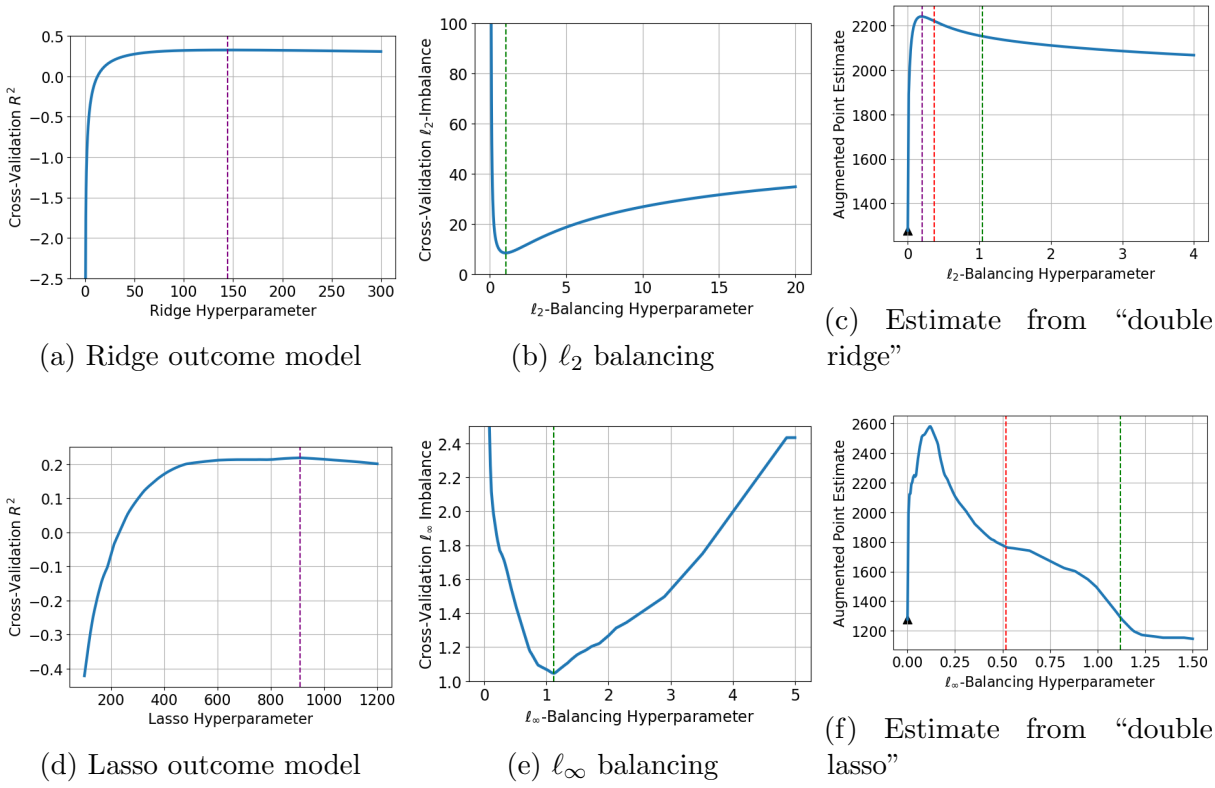
Following [55], we first consider the expanded set of 171 features for [184] used in [91]. Figure 4.3 shows estimates for ridge-augmented ℓ_2 balancing (top row) and lasso-augmented ℓ_∞ balancing (bottom row). The left two panels of each row show the cross-validation curves for the outcome regression and balancing weights, respectively. The right panels show the point estimate as a function of the weighting hyperparamter δ, holding the outcome model hyperparameter λ fixed; the black triangle represents the OLS plug-in point estimate. For context, the corresponding experimental estimate is \$1,794 [see 77]. The green and red dotted lines correspond to hyperparameters chosen by cross-validating balance and the Riesz loss, respectively. For the double ridge estimate, the purple line corresponds to $\delta = \hat{\lambda}$, the outcome hyperparameter selected via cross validation.

Figure 4.3 highlights that both the imbalance and the point estimate are highly nonlinear close to zero. Thus, even small departures from OLS (at $\delta = 0$) lead to large changes in the point estimate. We can also assess the sensitivity of the point estimate to the hyperparameter selection scheme. In this case, choosing δ via CV balance leads to meaningfully larger choices than via other methods.

Finally, the selected δ is always strictly greater than zero for this high-dimensional dataset. However, we find this is sensitive to small perturbations in the problem parameters. For

[1] These are: age, years of education, Black indicator, Hispanic indicator, married indicator, 1974 earnings, 1975 earnings, age squared, years of education squared, 1974 earnings squared, and 1975 earnings squared.

example, when we perturb $\mathbb{E}[\Phi_q]$ by adding a small value to all the even elements, then the cross-validated ℓ_2 Riesz loss chooses $\delta = 0$ in 38% of draws of the cross-validation splits. As suggested by our simulation results, this is likely to result in extremely large mean squared error.



(a) Ridge outcome model

(b) ℓ_2 balancing

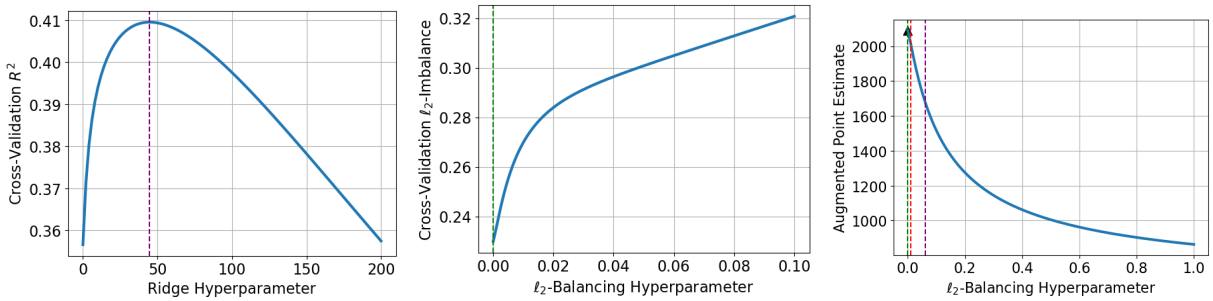
(c) Estimate from "double ridge"

(d) Lasso outcome model

(e) ℓ_∞ balancing

(f) Estimate from "double lasso"

Figure 4.3: Augmented balancing weights estimates for the [184] data set with the expanded set of 171 features used in [91]; the top row shows ridge-augmented ℓ_2 balancing, and the bottom row shows lasso-augmented ℓ_∞ balancing. Panels (a) and (d) show the 3-fold cross-validated R^2 for the ridge- and lasso-penalized regression of Y_p on Φ_p among control units across the hyperparameter λ; the purple dotted lines show the CV-optimal value for each. Panel (b) and (e) show the 3-fold cross-validated imbalance for ℓ_2 and ℓ_∞ balancing weights across the hyperparameter δ; the green dotted lines show the CV-optimal value for each. Panels (c) and (f) show the point estimates for the augmented estimators across the weighting hyperparameter δ; the black triangles correspond to the OLS point estimate; the green and red dotted lines correspond to the cross-validated balance and Riesz loss respectively; the purple line corresponds to the cross-validated ridge hyperparameter (for $\delta = \hat{\lambda}$). The variance-based hyperparameter for ridge is $\hat{\sigma}^2/n^2 = 104.8$ and for lasso is 137.5. The corresponding point estimates are 1923.6 and 725.8 respectively, essentially equal to the plug-in outcome model estimates.



(a) Ridge outcome model (b) ℓ_2 balancing (c) Augmented estimate

Figure 4.4: Ridge-augmented ℓ_2 balancing weights ("double ridge") for [184] with the original 11 covariates. Panel (a) shows the 3-fold cross-validated R^2 for the Ridge-penalized regression of Y_p on Φ_p among control units across the hyperparameter λ; the purple dotted line shows the CV-optimal value, $\hat{\lambda}$. Panel (b) shows the 3-fold cross-validated imbalance for ℓ_2 balancing weights across the hyperparameter δ; the green dotted line shows the CV-optimal value, which is $\delta = 0$ or exact balance. Panel (c) shows the point estimate for the augmented estimator across the weighting hyperparameter δ; the black triangle corresponds to the OLS point estimate, the green dotted line corresponds to cross-validated balance, the red dotted line corresponds to cross-validated Riesz loss, and the purple dotted line corresponds to the ridge outcome hyperparameter.

Low-dimensional setting: Recovering OLS

Finally, we apply double ridge to the "short" version of the [184] data set with 11 features. Figure 4.4 shows the cross-validation curves for the outcome and weighting models, as well as the point estimate as a function of the balance hyperparameter, with the OLS estimate given by the black triangle. As above, the green, red, and purple dotted lines correspond to hyperparameters chosen by cross-validating balance, cross-validating the Riesz loss, and choosing $\delta = \lambda$ respectively.

Unlike for the "long" dataset in Figure 4.3, Figure 4.4 does not display quite as stark nonlinearity around zero. Importantly, however, setting δ by cross-validating imbalance or the Riesz loss yields $\delta = 0$ (up to numerical imprecision), which reduces the augmented estimator to exactly the estimate from a simple OLS regression — even though the base learner ridge outcome model is heavily regularized. By contrast, our preferred hyperparameter tuning scheme of choosing $\delta = \lambda$ results in an estimate that is roughly \$400 dollars smaller than the OLS estimate.

4.8 Discussion

We have shown that augmenting a plug-in regression estimator with linear balancing weights results in a new plug-in estimator with coefficients that are shrunk towards — in some cases all the way to — the estimates from OLS fit on the same observations. We generalize this equivalence for different choices of outcome and weighting regressions. In the asymptotic setting, we draw the explicit connection between augmented estimators and undersmoothing for the special case of kernel ridge regression. Then we derive the design-conditional finite sample MSE for the double ridge estimator, and use it to solve numerically for oracle hyperparameters. We compare the oracle hyperparameters with three practical tuning schemes and then illustrate our results on the canonical LaLonde data set.

There are many promising avenues for future research. The fundamental connection between doubly robust estimation and undersmoothing opens up several theory directions. While we focus on the special case of kernel ridge regression in Section 4.6, we anticipate that these connections will hold more broadly. Similarly, while our focus in this paper has been on interpreting balancing weights as a form of linear regression, the converse is also valid: we could instead focus on how many outcome regression-based plug-in estimators are, in fact, a form of balancing weights; see [191] for connections between outcome modeling and density ratio estimation.

We also anticipate that the MSE we derive in Section 4.6 is a starting place for future theoretical analysis that can inform practice. We demonstrate in our simulation study that existing hyperparameter selection methods cannot perform uniformly well over all DGPs. We expect that analyzing the optimal hyperparameters, for example in a proportionate asymptotics regime, can either help devise new tuning schemes or inform which tuning method will work best on the dataset at hand.

We conjecture that these results may provide new insights into the estimation of causal effects in the proximal causal inference framework [297]. This framework uses proxy variables to identify causal effects in the presence of unmeasured confounding. Estimation has been complicated by the fact that, in the absence of strong parametic assumptions, estimators of proximal causal effects are solutions to ill-posed Fredholm integral equations. [107] and [153] recently proposed tractable nonparametric estimators in this setting. They use an "adversarial" version of double kernel ridge regression — allowing the weighting and outcome models to have different bases — to estimate the solution to the required Fredholm integral equations. Our results apply immediately to standard augmented estimators with different bases for the outcome and weighting models, either via a union basis [55] or by applying an appropriate projection as in [131], and extending these results to proximal causal effect estimators might help in constructing new proximal balancing weights, matching, or regression estimators with attractive asymptotic properties.

Finally, many common panel data estimators are forms of augmented balancing weight estimation [1, 28, 12]. We plan to use the numeric results here to better understand connections between methods and to inform inference.

Part III

Causal Inference with Unobserved Confounders

Chapter 5

Dynamic Sensitivity Analysis: The Tabular Case

In this part, we discuss the setting where we cannot observe all relevant confounding variables. The existence of these unobserved confounders make identifying the exact causal effect impossible and instead we will perform *sensitivity analysis*. We parameterize the strength of unobserved confounding, and this solve an optimization problem to get upper and lower bounds on the causal effect.

Notably in this section we turn to *dynamic* causal effects. We are interested in intervening in a system that evolves over time, and get upper and lower bounds on the causal effect on outcomes of interest at many time steps into the future. To do this, we change our formalism and turn to the language of reinforcement learning. While we will not explicitly refer to potential outcomes, our counterfactuals are now give in terms of different policies running in a Markov decision process. However, all of the key objects — like the density ratio — are the same as they were above. In Chapter 5, we consider the tabular case (with discrete actions and covariates), and in Chapter 6 we extend our results to the full continuous case with machine learning function approximation.

5.1 Introduction

Due to cost, feasibility, or safety concerns, practitioners often need to evaluate a sequential decision-making strategy using only previously-collected observational data. In reinforcement learning (RL), this problem is called off-policy policy evaluation (OPE). When the policy used to collect the data is unknown, there might exist unobserved variables correlated with both the policy and the outcomes. In this case, the causal effect of future interventions is unidentified and naive estimates for a new policy will be biased.

What kind of so-called unobserved confounders arise in Markov decision processes (MDPs)? Unobserved variables of interest in a medical setting are almost always highly persistent. For example, consider electronic medical records that do not document socio-economic status. A patient's socio-economic status is unlikely to change between visits to the hospital. In macroeconomics on the other hand, unobserved shocks are often assumed to be drawn iid every period. Consider the Federal Reserve Board adjusting monetary policy in response to oil price shocks. Events like earthquakes in oil fields might reasonably be assumed to occur independently across quarters.

Recent work develops OPE methods that are robust to unobserved confounding [214, 158]. Given an observational data set and a hypothetical confounder, these methods adapt importance sampling approaches to calculate worst-case estimates for the value of a new policy. A practitioner can assess the sensitivity of their results to unobserved variables by increasing the strength of confounding and computing how quickly the worst-case bounds degrade.

However, the existing literature arrives at radically different conclusions. [158] - henceforth KZ - finds that it is possible to efficiently construct non-conservative bounds in the infinite horizon setting. On the other hand, [214] - henceforth NKYB - only finds non-trivial bounds when confounding is restricted to a single time step. Furthermore, both approaches find the finite horizon case with confounding at each step to be computationally intractable.

The natural questions are: 1) what is responsible for the substantial gap between the conservativeness of the existing bounds? and 2) how can we compute tractable lower bounds for the finite horizon case?

Summary of our Results:

We identify a key assumption under which it is possible to obtain sharp lower bounds on the expected value in a confounded MDP, even as the horizon grows. When the unobserved confounding variables are drawn iid each period, the marginal dynamics over the observed state themselves form an MDP. In this case, OPE methods can be applied to the marginal MDP after appropriate adjustments for confounding. Such an assumption is made in KZ.

But if the unobserved state might be persistent over time, the problem is a genuine partially-observed MDP (POMDP). Marginal transition probabilities for the observed state will not be Markovian in general. Medical applications, which frequently feature persistent unobserved variables, fall under this category. As a result, existing bounds that target this setting, such as NKYB, are more conservative. In this paper, we focus on the case where the

marginal problem is an MDP and demonstrate enormous performance differences compared to setting with persistent unobservables.

We derive an expression for the bias of common estimands under confounding in the marginal MDP setting. We show how to express OPE "direct methods" in this form. Then we demonstrate how to adapt direct methods to give worst-case bounds in the finite horizon case. Our method is sufficiently generic that any approach which regresses a function against states and actions can be plugged into our framework to get bounds.

Finally, we show that model-based OPE methods provide sharper lower bounds on the value function. We can compute these bounds in a computationally efficient way by combining techniques from the robust MDP literature with sensitivity models from causal inference. A model-based approach provides a natural way for domain experts to provide guidance on reasonable limits for the strength of confounding on outcomes. We evaluate our methods with existing OPE benchmarks.

5.2 Related Work

Off-policy evaluation There are several classes of popular OPE algorithms. [306] provides a summary and empirically compares their performance. These classes include: importance sampling (IS) [236, 122], model-free direct methods like Fitted Q-Evaluation [186], model-based methods [229, 110], and hybrid methods [300, 145, 156]. [306] shows that, typically, either simple methods like FQE or hybrid methods have the best performance in practice.

Recently, a variety of marginalized importance sampling (MIS) methods [192, 302, 211] have been developed, which have the potential to solve the poor empirical performance of standard IS. This approach is adopted by KZ.

Causal inference and sensitivity analysis

Estimating the causal effect of a treatment on some outcome is the object of study in causal inference [128, 138, 234]. The line of work on dynamic treatment regimes [210, 181] is the most relevant to RL. Work in this area frequently assumes an unconfoundedness condition, which guarantees that the causal effect of a treatment is identified. For example, unconfoundedness will hold if the data come from a randomized control trial.

If unconfoundedness might be violated, then a researcher can assess the robustness of their causal estimates via sensitivity analysis [253, 97]. In recent work, [315, 155] give bounds for treatment effects subject to a sensitivity model. Other work develops bounds for the effectiveness of a single-step policy in the presence of unobserved confounders [159, 150].

Off-policy evaluation with unobserved confounders

Besides NKYB and KZ, most work in RL with unobserved confounders assumes that the causal effects are identified, i.e. assumptions are made about latent structure such that the true effect of interest can be recovered [32, 224]. For POMDPs, [299] analyze the bias for importance sampling in the presence of confounders, and give some conditions under which this bias can be corrected.

5.3 Problem Setting and Notation

Markov Decision Processes

Let $(\mathcal{X}, \mathcal{A}, P, R, \chi, \gamma)$ be an Markov decision process (MDP) where $\mathcal{X}$ is the set of states and $\mathcal{A}$ is the set of actions, which we assume are finite. Let $\mathcal{P}(S)$ denote all probability distributions on a set S. $P : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathcal{X})$ is the transition function, $R : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ is the reward function, $\chi \in \mathcal{P}(\mathcal{X})$ is the initial state distribution, and $\gamma \in [0, 1)$ is the discount factor. A (stationary) policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ assigns probabilities to each action given a state. We are interested in the expected value of policy π:

$$V_T^\pi = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right],$$

where $x_0 \sim \chi$, $a_t \sim \pi(\cdot|x_t)$, $x_{t+1} \sim P(\cdot|x_t, a_t)$, $r_t = R(x_t, a_t, x_{t+1})$ and $T \leq \infty$.

Confounded Off-policy Evaluation

In this paper, we consider MDPs with unobserved confounding variables. Specifically, we assume the state space is partitioned into observed state $\mathcal{X}$ and unobserved state $\mathcal{U}$. The full-information MDP is $(\mathcal{X} \times \mathcal{U}, \mathcal{A}, P, R, \chi, \gamma)$.

In the confounded off-policy evaluation problem, we have access to a dataset $\mathcal{D}_{\pi_b} = \{\tau_i\}_{n=1}^N$, collected according to a stationary *behavior* policy, $\pi_b : \mathcal{X} \times \mathcal{U} \to \mathcal{P}(\mathcal{A})$. Each $\tau_i = \{(x_t^i, a_t^i, x_{t+1}^i, r_t^i)\}_{t=0}^{T-1}$ denotes an observed trajectory where $(x_0, u_0) \sim \chi$, $a_t \sim \pi_b(\cdot|x_t, u_t)$, $(x_{t+1}, u_{t+1}) \sim P(\cdot|x_t, u_t, a_t)$, and $r_t = R(x_t, a_t, x_{t+1})$. Note that while R is only a function of the observed state, it can still rely on u_t via x_{t+1}.

Our goal is to estimate the expected return $V_T^{\pi_e}$ for a stationary *evaluation* policy, $\pi_e : \mathcal{X} \to \mathcal{P}(\mathcal{A})$, which does not depend on the unobserved state.

5.4 Two Types of Unobserved State

We begin by making a distinction between unobserved states that are dependent over time, and unobserved states that are drawn iid each time step.

Assumption 1 (IID Confounders). The unobserved state u_t is drawn iid for all $t \geq 0$ and therefore the transition dynamics can be factored as:

$$P(x', u'|x, u, a) = P(x'|x, u, a)p(u')$$

This corresponds to the "memoryless" unobserved confounding assumption in KZ. Under Assumption 1, the marginal observed state transition probabilities are Markovian:

$$P(x'|x, a) = \sum_{u \in \mathcal{U}} p(u)P(x'|x, u, a)$$

and the value of evaluation policy π_e in the true MDP is equal to the value of π_e in the marginal MDP, $(\mathcal{X}, \mathcal{A}, P, R, \chi, \gamma)$, where we abuse notation slightly to let P and χ denote the corresponding marginal quantities over the observed state.

While we have reduced the problem to finding the value of π_e in the marginal MDP, this value is *not identified* given the dataset $\mathcal{D}_{\pi_b}$ because the unobserved state u affects both the choice of action $\pi_b(a|x,u)$ and the transitions $P(x'|x,u,a)$. For example, any dataset collected under policy π_b will be consistent with a set of many possible marginal transition probabilities $P(x'|x,a)$. However, standard OPE algorithms for MDPs can be adapted to this setting via some strategy to control for confounding.

If the unobserved state is persistent, then the problem is no longer a marginal MDP plus causal uncertainty. Consider the simplest such scenario where u_0 is drawn from some initial distribution and $u_t = u_0, \forall t$. In this setting, $P(x_{t+1}|x_t, a_t)$ is non-stationary in general and $P(x_{t+1}|x_t, a_t, ..., x_0, a_0)$ is not Markovian due to the dependence via u induced by conditioning on x. Therefore, the problem is a partially-observed MDP (POMDP).

For the POMDP case, even when $\pi_b(a|x,u) = \pi_b(a|x,u'), \forall a, x, u, u'$ (as in a randomized trial), many OPE algorithms are biased because the observed state and actions do not themselves constitute an MDP. A notable exception is IS methods. When $\pi_b(a|x,u) = \pi_b(a|x,u')$, the problem satisfies Assumption 1 in [299] for POMDPs.

When the behavior policy varies over u, the value is not identified and one must further adapt IS methods as in NKYB. However, as we will demonstrate, without Assumption 1 these bounds are too conservative for practical use - even when confounding is limited to a single time step. Therefore, in this paper, we develop lower bounds on the value of a policy given Assumption 1, and show that the bounds are far less sensitive to confounding. It is crucial to remember that Assumption 1 is not reasonable in some settings, especially medical ones, and given the substantial gap in performance, we suspect that new algorithms or sensitivity models need to be developed to make the persistent confounder case work in practice.

5.5 Estimation with Unobserved Confounders

Bias due to Spurious Correlation

Under Assumption 1, we can explicitly quantify the bias due to unobserved confounding. For comparison, we begin with a quantity that *is* identified under confounding: the behavior policy conditional on the observed state. Consider the naive empirical estimate, $\hat{\pi}_b(a|x)$, for $\pi_b(a|x)$ given $\mathcal{D}_{\pi_b}$.

Lemma 5.5.1. *Under Assumption 1, $\hat{\pi}(a|x)$ is an unbiased estimator of $\pi_b(a|x)$.*

Proof.

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{\pi_b}}[\hat{\pi}(a|x)] &= \sum_{u\in\mathcal{U}} p(u|x)\pi_b(a|x,u) \\ &= \sum_{u\in\mathcal{U}} p(u)\pi_b(a|x,u) = \pi_b(a|x).\end{aligned}$$ □

On the other hand, consider estimating the expectation of a function of x, a, and x', conditional on x, a, i.e. $m_f(x,a) \coloneqq \mathbb{E}[f(x,a,x')|x,a]$. Define the corresponding naive estimator, $\hat{m}_f(x,a)$ as above.

Proposition 5.5.2. *Under Assumption 1 and given a function $f : \mathcal{X}\times\mathcal{A}\times\mathcal{X}\to\mathbb{R}$,*

$$m_f(x,a) = \mathbb{E}_{\mathcal{D}_{\pi_b}}\left[\frac{\pi_b(a|x)}{\pi_b(a|x,u)} f(x,a,x')\Big| x,a\right].$$

Proof sketch. Conditional on x and a, the distribution of u in $\mathcal{D}_{\pi_b}$ is $p(u|x,a)$ and

$$p(u|x,a) = \frac{\pi_b(a|x,u)}{\pi_b(a|x)} p(u)$$

by Bayes rule. Then reweight accordingly. □

As an immediate corollary of Proposition 1, $\hat{m}_f(x,a)$ is not, in general, an unbiased estimator of $m_f(x,a)$. For a relevant example, let $f(x,a,x') = \mathbf{1}(x'=i)$ for some $i\in\mathcal{X}$. Then $m_f(x,a) = P(i|x,a)$, the marginal probability of transitioning to state i. Unless $\pi_b(a|x,u) = \pi_b(a|x,u')$ or $P(x'|x,a,u) = P(x'|x,a,u'), \forall u,u'$, the naive estimator of the transition probabilities is biased. Furthermore, since $\pi_b(a|x,u)$ is unobserved, the observed data is consistent with many possible $P(x'|x,a)$.

Sensitivity Model

While estimands like $P(x'|x,a)$ are not point-identified under Assumption 1, it is possible to give upper and lower bounds that are consistent with the observed data. However, without further assumptions these bounds are typically vacuous. Therefore, we follow the sensitivity analysis approach and specify limits on the impact of the unobserved state. The idea is that we will construct a worst-case estimate given a fixed level of confounding and study how the estimate changes as the degree of confounding is increased.

We control the dependence of the behavior policy on the unobserved state via a parameter Γ. This is a popular technique in the causal inference literature, described in [253]. In particular, we follow [293] and have Γ bound the odds ratio between the unobserved behavior policy and the observed marginal behavior policy:

Assumption 2 (Policy Confounding Bound). Given $\Gamma\geq 1$, for all $x\in\mathcal{X}$, $u\in\mathcal{U}$, and $a\in\mathcal{A}$:

$$\frac{1}{\Gamma} \leq \left(\frac{\pi_b(a|x,u)}{1-\pi_b(a|x,u)}\right) \Big/ \left(\frac{\pi_b(a|x)}{1-\pi_b(a|x)}\right) \leq \Gamma$$

Note that Assumption 2 implies the bounds:

$$\alpha(x,a) \leq \frac{\pi_b(a|x)}{\pi_b(a|x,u)} \leq \beta(x,a)$$

where

$$\alpha(x,a) \coloneqq \pi_b(a|x) + \frac{1}{\Gamma}(1-\pi_b(a|x))$$
$$\beta(x,a) \coloneqq \Gamma + \pi_b(a|x)(1-\Gamma)$$

5.6 Policy Evaluation with Confounders

In this section, we will show how to compute worst-case value estimates. As long as Assumption 1 holds, by Proposition 1 we have an unbiased expression for regressing *any* observed quantity f against x and a. This expression depends on the unknown probabilities $\pi_b(a|x,u)$ which we can bound using Assumption 2. By choosing different functions f, we can adapt most OPE direct methods as described in [306]. We illustrate this procedure for Fitted Q Evaluation (FQE).

We begin with some notational details. We denote the state and state-action value functions for a policy π and horizon T as:

$$V_T^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r_t \Big| x_0 = x\right]$$
$$Q_T^\pi(x,a) = \mathbb{E}\left[R(x,a,x') + V_{T-1}^\pi(x',u') \big| x, a\right]$$

respectively. Throughout the rest of the paper, we will use the short-hand $g(x,\pi) \coloneqq \sum_{a\in\mathcal{A}} \pi(a|x)g(x,a)$. Denote the Bellman evaluation operator for a policy π as $\mathcal{T}^\pi$, defined as:

$$(\mathcal{T}^\pi g)(x,a) = \mathbb{E}\Big[r(x,a,x') + \gamma g(x',\pi) \Big| x, a\Big]$$

where g is any function on $\mathcal{X}\times\mathcal{A}$. The state-action value function Q_T^π can be computed by applying $\mathcal{T}^\pi$ to $Q_0 = 0$, T-times [237]. Furthermore, $V_T^\pi(x) = Q_T^\pi(x,\pi)$, and the expected value is simply the average of the value function over the initial state distribution. Therefore, we can easily compute estimates of the expected value using Q_T^π.

Confounded FQE

FQE iteratively applies an empirical approximation of $\mathcal{T}^\pi$ to compute Q_T^π. Let $Q_0 = 0$ and let $\mathcal{H}$ be some function class. Given a dataset $\mathcal{D}_{\pi_b}$ and an evaluation policy π_e, FQE computes

$$Q_k = \operatorname*{argmin}_{h\in\mathcal{H}} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=0}^{T-1}(h(x_t^i,a_t^i) - y_t^i)^2$$
$$\text{where } y_t^i = r(x_t^i,a_t^i,x_{t+1}^i) + \gamma Q_{k-1}(x_{t+1}^i,\pi_e).$$

Essentially, regression with the class $\mathcal{H}$ approximates the conditional expectation of the function

$$f(x, a, x') = r(x, a, x') + \gamma Q_{k-1}(x', \pi_e)$$

and $\mathcal{T}^{\pi_e} Q_{k-1}(x, a) = \mathbb{E}[f(x, a, x')|x, a]$. With unobserved confounding, regression using the data $\mathcal{D}_{\pi_b}$ no longer gives an unbiased estimate of $\mathcal{T}^{\pi_e} Q_{k-1}(x, a)$. Instead, we can apply Proposition 1 with the function f defined above to get:

$$\mathcal{T}^{\pi_e} Q_{k-1}(x, a) = \mathbb{E}_{\mathcal{D}_{\pi_b}} \left[\frac{\pi_b(a|x)}{\pi_b(a|x, u)} f(x, a, x') \Big| x, a \right].$$

We can then use Assumption 2 to bound the unobserved $\pi_b(a|x, u)$. For example, we immediately get the following naive bound.

Proposition 5.6.1. *Let $y \coloneqq f(x, a, x')$. Under Assumptions 1 and 2, For all $x \in \mathcal{X}$ and $a \in \mathcal{A}$,*

$$(\mathcal{T}^{\pi_e} Q_{k-1})(x, a) \geq \\ \mathbb{E}_{\mathcal{D}_{\pi_b}} \left[\left(\beta(x, a) \mathbf{1}(y < 0) + \alpha(x, a) \mathbf{1}(y \geq 0) \right) y \Big| x, a \right].$$

This naive bound is too conservative to use in practice, especially as the horizon grows. To get a better bound, we can solve an optimization problem over all possible values of $\pi_b(a|x, u)$ which are consistent with the observed data. Fix x and a. Let $\pi_b(a|x)$ and $\hat{P}(x'|x, a)$ be the nominal behavior policy and nominal transition probabilities respectively. The basic unknown quantities are $p(u)$, $\pi_b(a|x, u)$, and $P(\cdot|x, u, a) \in \mathcal{P}(\mathcal{X}), \forall u$. We have the following observable implications:

Lemma 5.6.2. *Under Assumption 1, $\forall x \in \mathcal{X}, a \in \mathcal{A}, x' \in \mathcal{X}$,*

$$\sum_{u \in \mathcal{U}} p(u) \pi_b(a|x, u) = \pi(a|x), \text{ and}$$

$$\sum_{u \in \mathcal{U}} p(u) \pi_b(a|x, u) P(x'|x, u, a) = \pi(a|x) \hat{P}(x'|x, a).$$

For a fixed x and a, let $\mathcal{B}_{xa}$ be the set of possible $\pi_b(a|x, \cdot)$ such that Lemma 2 and Assumption 2 hold. Then:

$$\mathcal{T}^{\pi_e} Q_{k-1}(x, a) \geq \\ \min_{\pi_b(a|x,\cdot) \in \mathcal{B}_{xa}} \mathbb{E}_{\mathcal{D}_{\pi_b}} \left[\frac{\pi_b(a|x)}{\pi_b(a|x, u)} f(x, a, x') \Big| x, a \right]$$

Unfortunately, when computing a regression in practice, this requires introducing a new optimization variable for the unknown values of u for every data point. Instead we use a clever reparameterization to remove the dependence on u that KZ introduced for MIS.

Reparameterization

Define

$$g(x, a, x') \coloneqq \sum_{u \in \mathcal{U}} \left(\frac{p(u|x, a)P(x'|x, u, a)}{\hat{P}(x'|x, a)} \right) \frac{1}{\pi_b(a|x, u)}$$

and the corresponding set

$$\tilde{\mathcal{B}}_{xa} \coloneqq \{g(x, a, \cdot) : \pi_b(a|x, u) \in \mathcal{B}_{xa}\}.$$

The idea is that $g(x, a, x')$ is equal to $1/\pi_b(a|x, u)$ convolved with an unknown density. Since both $\pi_b(a|x, u)$ and $p(u|x, a)P(x'|x, u, a)$ are unknown, optimizing over $\tilde{\mathcal{B}}_{xa}$ is equivalent to optimizing over $\mathcal{B}_{xa}$ where we replace $\pi_b(a|x)/\pi_b(a|x, u)$ with $\pi_b(a|x)g(x, a, x')$. We have the following constraints:

Lemma 5.6.3. *Under Assumptions 1 and 2,*
$\forall x \in \mathcal{X}, a \in \mathcal{A}, x' \in \mathcal{X}$,

$$\alpha(x, a) \leq \pi_b(a|x)g(x, a, x') \leq \beta(x, a),$$

and

$$\sum_{x' \in \mathcal{X}} \pi_b(a|x)g(x, a, x')\hat{P}(x'|x, a) = 1.$$

Now we are ready to state our confounded FQE bound:

Theorem 5.6.4. *Under Assumptions 1 and 2,*
$\forall x \in \mathcal{X}, a \in \mathcal{A}$,

$$\begin{aligned}
&\mathcal{T}^{\pi_e} Q_{k-1}(x, a) \geq \\
&\quad \min_{\pi_b(a|x,\cdot) \in \mathcal{B}_{xa}} \mathbb{E}_{\mathcal{D}_{\pi_b}} \left[\frac{\pi_b(a|x)}{\pi_b(a|x, u)} f(x, a, x') \middle| x, a \right] \\
&= \min_{g(x,a,\cdot) \in \tilde{\mathcal{B}}_{xa}} \mathbb{E}_{\mathcal{D}_{\pi_b}} \left[\pi_b(a|x)g(x, a, x')f(x, a, x') \middle| x, a \right]
\end{aligned}$$

For a given dataset $\mathcal{D}_{\pi_b}$, this bound can be computed with a simple linear program. Fix x and a, and for shorthand, denote the naive estimates of the nominal behavior policy and nominal transition probabilities as $\hat{\pi}_{xa} \in [0, 1]$ and $\hat{P}_{xa} \in [0, 1]^{|\mathcal{X}|}$ respectively. The bound in Theorem 1 can be estimated by the following LP:

$$\begin{aligned}
&\min_{w \in \mathbb{R}^{|\mathcal{X}|}} c^T w \\
&\qquad \text{such that} \\
&\qquad \hat{\pi}_{xa} + \frac{1}{\Gamma}(1 - \hat{\pi}_{xa}) \preceq w \preceq \Gamma + \hat{\pi}_{xa}(1 - \Gamma) \\
&\qquad \text{and } \hat{P}_{xa}^T w = 1,
\end{aligned}$$

where $c(x')$ is the sample average of $r + \gamma Q_{k-1}(x', \pi_e)$ conditional on x and a. Note that $\hat{\pi}_{xa}, \hat{P}_{xa}$, and c are all observables estimated from the data, and Γ is given. Only the vector w is unknown.

Remark 1. Theorem 1 gives a lower bound for a single application of $\mathcal{T}^{\pi_e}$. We get a lower bound on $V_k^{\pi_e}$ by applying $\mathcal{T}^{\pi_e}$ k-times and then averaging over the initial state distribution.

Remark 2.The reparameterized optimization problem in Theorem 1 can in principle be used when regressing a wide variety of functions f against x and a. This provides a blueprint for adapting other OPE methods that solve a regression problem.

5.7 Sharper Bounds with Robust MDPs

Unobserved variables create bias when they are correlated with *both* the behavior policy *and* the state transitions. The sensitivity model in Assumption 2 limits the correlation with the behavior policy. However, in the reparameterization strategy above, we combine our unknowns, $\pi_b(a|x,u)$ and $P(x'|x,u,a)$. Therefore, we cannot leverage any additional information that limits the correlation between u and the transitions. Consider the extreme case, where $P(x'|x,u,a) = P(x'|x,a), \forall u$. In this case, naive OPE estimates will be unbiased even if Γ in Assumption 2 is large. While in observational studies, it is not possible to rule out all correlation between unobservables and the dynamics, we might be able to use domain knowledge on causal mechanisms to restrict the feasible transitions.

One branch of the sensitivity analysis literature, exemplified by [254], suggests using three sensitivity parameters. First, a bound on the correlation between the unobserved confounder and the treatment. Second, a bound on the correlation between the unobserved confounder and the outcome. Third, a parameter representing the distribution of the unobserved confounder. [254] presents the case where u is a binary variable. However, [81] show that for worst-case bounds, this is without loss of generality. Therefore, we assume that $\mathcal{U} = \{0, 1\}$.

Assumption 2 bounds the impact of u on π_b. Following [254], we now introduce two additional parameters:

Assumption 3 (Transition Confounding Bound). Given $\Delta \geq 1$, for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, $x' \in \mathcal{X}$, and $u \in \mathcal{U}$:

$$\frac{1}{\Delta} \leq \left(\frac{P(x'|x,u,a)}{1 - P(x'|x,u,a)} \right) \Big/ \left(\frac{\hat{P}(x'|x,a)}{1 - \hat{P}(x'|x,a)} \right) \leq \Delta$$

Assumption 4. Given a fixed $p \in [0,1], p(u=1) = p$.

For any tuple of sensitivity parameters, (Γ, Δ, p), we will give worst-case bounds on the value function $V_T^{\pi_e}(x)$ using a model-based approach. Each (Γ, Δ, p) has a corresponding set of possible transition probabilities under Assumptions 2, 3, and 4, such that the observable implications in Lemma 2 hold. Finding the worst-case value given an uncertainty set for the dynamics has been extensively explored in the Robust MDP literature [221]. The standard approach is to separate the uncertainty over the state-action pairs, assuming that

the uncertainty sets across x, a pairs are not linked. In our problem, this assumption is violated because of the requirement that $\pi_b(\cdot|x,u)$ is a probability distribution. In the language of robust MDPs, our problem is "s-rectangular" instead of "s,a-rectangular".

Fortunately, s-rectangular MDPs can also be solved efficiently [311]. Let $\mathcal{G}_x$ denote the set of feasible transition probabilities for a fixed x. Let $P_x \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{X}|}$ be the matrix whose rows are $P(\cdot|x,a)$ for each a. Instead of the state-action value function, we iteratively solve for worst-case estimates of the value function:

$$V_k(x) = \min_{P_x \in \mathcal{G}_x} \pi_e(\cdot|x)^T P_x y$$

where $y = (\sum_{a\in\mathcal{A}} \pi_e(a|x)R(x,a,\cdot)) + \gamma V_{k-1}(\cdot)$. When optimizing over the unknown quantities $P(x'|x,u,a)$ and $\pi_b(a|x,u)$ for all x', a, and u, this problem has a linear objective with linear and bilinear equality constraints, so it can be easily solved. We estimate $V_T^{\pi_e}$ by letting $V_0 = 0$, then solving the above minimization problem T-times. As we will show in our evaluation, for all values of the parameters (Γ, Δ, p), the s-rectangular robust MDP formulation provides sharper bounds than the linear program corresponding to Theorem 1.

5.8 Evaluation

We use the benchmarks from OPE-Tools [306] for evaluation. In particular, we adapt their three discrete environments, Graph, Discrete MC, and Gridworld, together with a small toy problem. Note that the data generating processes do not strictly need to be confounded. Our methods bound the worst possible confounded MDP that *could* have generated the data. Therefore, the two relevant, observable reference points are the value of the behavior policy and the nominal value of the evaluation policy. Nonetheless, for completeness we augment the environments with unobserved confounding variables. Our approach takes an existing behavior policy and transition matrix, and adds an additional state variable u which induces a correlation between the policy and transitions based on either the rewards or the optimal value function.

For each environment, we choose a behavior policy π_b and evaluation policy π_e such that the value of π_e without confounding is greater than the value of π_b. This way, it is possible to find which level of confounding makes it impossible to guarantee that π_e is superior to π_b. Furthermore, the impact of confounding can be compared relative to the difference in values between the two policies. See Table 5.1 for a summary of the four test environments and the Appendix for full details.

Lower Bounds with Confounding

For our first experiment, we collect trajectories from each of the four environments using their respective behavior policies. For each environment, we collect 30,000/horizon trajectories, keeping the number of data points the same across environments. Then, we compute our

Environment	Horizon	States	Actions	$V_T^{\pi_b}$	$V_T^{\pi_e}$	Sparse Rewards?
toy	5	3	2	0.3397	0.4990	No
ope-graph	4	8	2	-0.1786	0.7174	No
ope-mc	20	22	2	-18.1890	-15.7381	Yes
ope-gridworld	8	16	4	-0.4994	-0.3569	No

Table 5.1: Characteristics of the four test environments.

confounded FQE and robust MDP lower bounds for values of Γ and Δ ranging between 1.1 (barely confounded) and 10 (highly confounded). For the robust MDP bounds, we fix the parameter $p = 0.5$, i.e. each period the unobserved state is equally likely to be $u = 0$ or $u = 1$. The robust MDP bounds are not very sensitive to this parameter and this choice doesn't impact the qualitative results, although corroborating results are in the Appendix. Our lower bounds for the four environments are plotted in Figure 5.1.

The confounded FQE bounds are the black curve at the bottom of each plot. Without any additional restrictions of the transition dynamics, these bounds degrade the quickest as Γ increases. This curve intersects the value of π_b at $\Gamma = 6$ for ope-graph, and $\Gamma < 3$ for the remaining environments. Qualitatively, this means strong requirements on confounding are required for the FQE bounds to guarantee that the evaluation policy is better than the behavior policy. Compare this, for example, to the other curves in ope-graph and ope-mc which are greater than V^{π_b} for all values of Γ.

The curves above the confounded FQE curve correspond to our robust MDP bounds. In all cases as Δ grows, the corresponding lower bounds get worse. As mentioned previously, for ope-graph and ope-mc, any value of Δ guarantees that $V^{\pi_e} > V^{\pi_b}$. For toy and ope-gridworld, consider the $\Delta = 2$ curve, which is third from the top. For the toy environment, assuming $\Delta = 2$ substantially increases the Γ at which the curve crosses the dotted π_b line compared to the FQE curve. For ope-gridworld, the $\Delta = 2$ curve lies above V^{π_b} for all Γ. These examples highlight the qualitative and quantitative importance of limiting the degree of confounding on the transition probabilities.

Tightness

Our confounded FQE and robust MDP methods provide lower bounds on the expected value subject to their respective sensitivity models. A natural question is: how far are these bounds from the infimum over all full-information MDPs consistent with the observed data, subject to the given sensitivity model? We split our analysis of tightness into two parts, the single-step case and the multi-step case.

A single iteration of our bounds requires solving a minimization problem. The tightest possible bound on V_T^{π} is the minimum over all valid full-information MDPs. But our robust MDP solution produces candidate transition probabilities $P(x'|x, u, a)$ and behavior policy

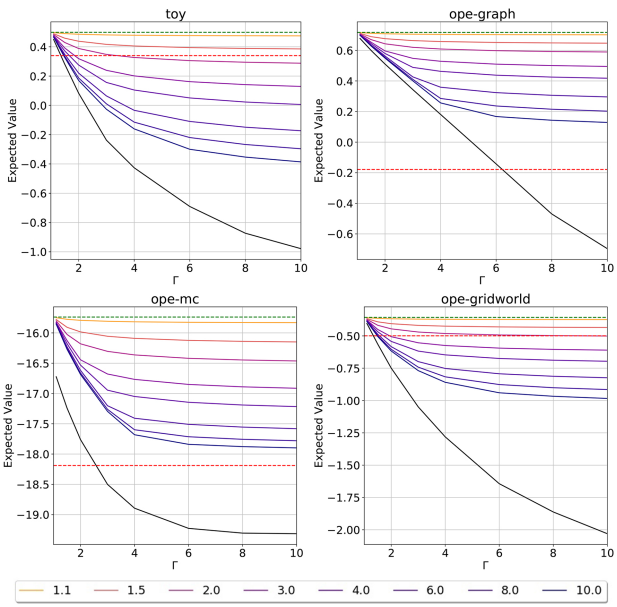


Figure 5.1: Lower bounds on the expected value of π_e. For reference, in each environment, we plot the value of π_e without confounding (the dotted line at the top) and the value of π_b (the dotted line below). The black line at the bottom is the confounded-FQE bound. Each other line corresponds to a robust MDP bound for a single value of the transition confounding parameter Δ, with light to dark lines going from 1.1 to 10.

$\pi_b(a|x,u)$ corresponding to some valid full-information MDP. Therefore, since it is a lower bound, it must achieve the true minimum and so a single iteration of the robust MDP approach is tight.

On the other hand, our confounded FQE bound solves a minimization problem separately for each state-action pair without enforcing that $\pi_b(\cdot|x,u)$ be a density across actions. We quantify the impact on performance by comparing the FQE bound to our robust MDP bound as Δ goes to infinity. We present results for the ope-graph and ope-mc environments in Figure 5.2. The qualitative findings for the other environments are similar.

For the ope-graph environment, the gap between the FQE bound (the black line at the bottom) and the robust MDP bounds for large Δ are negligible until $\Gamma \geq 8$, at which point the gap grows. For the ope-mc environment, the gap begins substantial and grows slightly larger as Γ grows. For this particular environment, the robust MDP lower bounds always guarantee that the evaluation policy is at least as good as the behavior policy. However, the FQE lower bound can only provide this same guarantee for $\Gamma < 3$. Therefore, it appears that enforcing the density constraint across actions can matter in practice, so for cases where we do not wish to make any assumptions on the transitions, we prefer our robust MDP bounds with very large values of Δ.

When confounding occurs in more than one time step, our robust MDP bound is computed

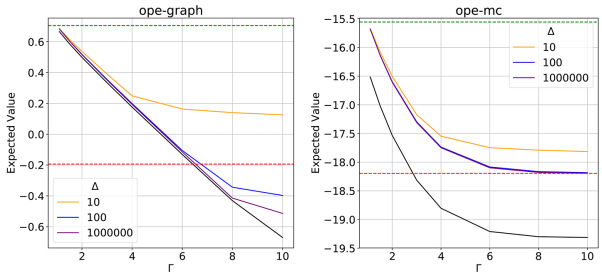


Figure 5.2: Lower bounds on the expected value as Δ grows large. The black line at the bottom is the confounded FQE bound. The upper dashed line is value of π_e with no confounding. The lower dashed line is the value of π_b.

iteratively with different minimization problems solved at each time step. The candidate transitions and behavior policy that correspond to each minima may differ, so the lower bounds are potentially loose. Theoretically, the looseness of our bound is characterized by Theorem 4 of [221]. In particular, as the horizon goes to infinity, our lower bound converges to the best possible lower bound - the rate of convergence can be found in the proof of the theorem.

To test this empirically, we use the full-information transitions and behavior policy from the final iteration of our robust MDP method as a candidate. Because the candidate MDP is consistent with the observed data subject to the sensitivity model, if the value of this MDP matches our lower bound, than our lower bound must be tight. For the toy, ope-graph, and ope-mc environments, we use the same experimental setup as we did for the results in Figure 5.1. The gap between the candidate MDP value and our lower bounds are reported in Table 5.2. For these environments, the value of the candidate MDP differs by less than 10^{-8} from our lower bound. For the ope-gridworld environment, we find our lower bound is not tight at small horizons, so we ran experiments with a short, medium, and long horizon. As predicted by the theory, the bound improves for large T as value iteration approaches its fixed point.

Assumption 1 and Comparison with NKYB

Assumption 1 - that the unobserved state is drawn iid each period - is crucial to the quality of the bounds above. We demonstrate this by comparing our bounds to those in NKYB, which do not assume iid confounders. In order to compare to NKYB, we have to alter the experimental setup above in two ways. First, NKYB only supports confounding that occurs in a single time step. The initial time step is confounded, but for the remainder of the horizon, the behavior policy only uses the observed state. We compute the analogue for our robust MDP algorithm by computing $T-1$ iterations of unconfounded value iteration followed by a single iteration of our lower bound.

Second, NKYB uses a similar but more restrictive sensitivity model. Our sensitivity

env	$\Gamma = 2$, $\Delta = 2$	$\Gamma = 10$, $\Delta = 10$
toy	$<$ 1e-8	$<$ 1e-8
ope-graph	$<$ 1e-8	$<$ 1e-8
ope-mc	$<$ 1e-8	$<$ 1e-8
ope-gridworld T=28	2.03e-3	3.06e-2
ope-gridworld T=208	4.75e-3	2.87e-2
ope-gridworld T=508	2.65e-5	2.97e-4

Table 5.2: The difference between our robust MDP bound and the value of π_e in the candidate MDP defined by the transition probabilities from the last iteration of our bound. The first three environments use the default horizons given in Table 5.1.

parameter restricts the odds ratio between the confounded policy for a given value of u and the policy averaged over all u. Their sensitivity parameter restricts the odds ratio for the confounded policy between any values of u, which grows roughly like the square of ours. For this comparison, we can calculate the true sensitivity parameters for each confounded environment under the different sensitivity models. We provide a performance comparison using the true sensitivity parameters for each environment in Table 5.3. Even with confounding restricted to a single time step, the NKYB bounds, which do not assume iid confounders, are enormously conservative.

This is a key result. Even for a single time-step, policy evaluation is highly sensitive to persistent unobserved variables. The ability of our robust MDP bounds to guarantee improvement over the behavior policy in Figure 5.1, even over longer horizons, depends crucially on our Assumption 1. In turn, this highlights the fact that off-policy evaluation with confounding in settings where Assumption 1 fails is far more difficult and requires a different algorithmic approach. As mentioned in the introduction, while iid confounders are feasible in certain settings - like unobserved oil supply shocks for macroeconomic policy - Assumption 1 is *not* reasonable for many applications, especially in medicine.

The results in Table 5.3 might hinge on the different sensitivity models, so we perform a robustness check which uses identical values of Γ and which should therefore be very favorable for the NKYB bounds. The toy and ope-graph NKYB bounds improve, but the ope-mc and ope-gridworld bounds remain unusable.

Horizon and Comparison with KZ

Many of the details above depend on the horizon. For example, our robust MDP bounds become tight as the horizon increases and NKYB restricts confounding to a single time-step. Therefore, in this section we assess how our lower bounds change as the horizon increases. This also provides a convenient setting to compare with the infinite horizon bounds in KZ.

Comparing with KZ requires modification of our initial experimental setup. We use

env	Nominal	NKYB	Ours
toy	0.5189	0.0436	0.25372
ope-graph	0.7008	0.0280	0.3994
ope-mc	-15.6941	-64.5040	-15.9647
ope-gridworld	-0.3588	-2.3914	-0.4112

Table 5.3: The value of π_e without confounding and the corresponding lower bounds from NKYB and our robust MDP procedure. For each bound and each environment, we use the true parameter value for the respective sensitivity models.

the ope-graph and ope-gridworld environments. In order to generate a non-trivial steady-state distribution, we remove the terminating states and alter the transition probabilities accordingly. Furthermore, to match KZ's approach, we modify the rewards to only depend on the current state. We then calculate our bounds for 1 to 200 time steps. For both environments, $T = 200$ is long enough to spend a majority of the time close to steady-state. We also adopt a discount rate of $\gamma = 0.95$ so that $T = 200$ is well beyond the effective horizon. We produce bounds for $\Gamma = 1.5, 2$, and 10 using our robust MDP method with Δ set to 1,000,000.

Since we use the same marginal sensitivity model, we can use KZ's method to calculate infinite horizon bounds for the same values of Γ. Their method computes bounds on the long-run average value, i.e. the expectation of the rewards with respect to the steady-state distribution, instead of the discounted value. Therefore we use the discounted sum of rewards as the per-state reward for KZ's method. The results are plotted in Figure 5.3. The dotted black curve at the top is the value of π_e without confounding at each horizon. The curves below are the lower bounds for $\Gamma = 1.5, 2$, and 10 respectively. The dots on the far right are the corresponding KZ infinite horizon bounds. In all cases, the gap between our bounds and the unconfounded value grow at the horizon increase. This is not because our bounds are loose - as value iteration reaches its fixed point, our bounds are provably tight as mentioned - but because confounding over many time periods is a more difficult problem. This phenomenon is especially pronounced for $\Gamma = 10$: at long horizons, a smaller value of the sensitivity parameters becomes much more valuable.

The infinite horizon bounds follow roughly the same qualitative behavior as ours but are much looser. This is presumably due to the fact that the long-run average of discounted rewards is a different estimand than the average discounted sum of rewards. With no confounding, the difference is small (compared the uppermost line and uppermost dot). But as the level of confounding increases, the long-run average becomes more sensitive. The magnitude of the difference is surprising and perhaps worth studying in future work.

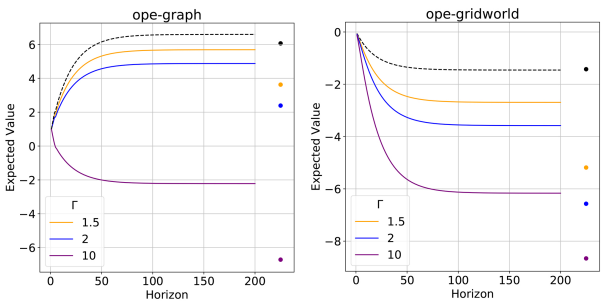


Figure 5.3: Robust MDP lower bounds as the horizon grows. The dotted curve is the nominal value of π_e. The dots on the right are KZ's infinite horizon bounds.

5.9 Conclusion

To summarize: our first key contribution is to develop a method for computing finite horizon lower bounds for policy evaluation with unobserved confounders that are drawn iid each period. We find that our model-based robust MDP approach can give substantially sharper bounds by leveraging assumptions about the transition probabilities. To be clear on this point: the argument is not that a plug-in estimator using a model of the dynamics is inherently more efficient. When using observational data to estimate a dynamic causal effect, understanding the dynamics of the system and the causal mechanisms are critically important. Quantitatively, we illustrate this by showing that sharp partial identification of the value of a policy requires restricting the set of possible transition probabilities. In practice, such an approach relies on domain-expertise. Practitioners must have enough mechanistic understanding of the dynamics that they are able to specify bounds, Δ, on potential confounding in order to get a reasonable estimate of the expected value.

Our second key contribution is to demonstrate that policy evaluation is far more challenging when there are persistent unobserved confounders. This is responsible for the substantial performance gap between our bounds and those in NKYB. These results are especially relevant for medical applications where unobserved variables are likely to be persistent. For example, any patient variable that may not be recorded, but doesn't change between treatment choices like socio-economic status or undocumented chronic illness. Work published after this paper was completed [179] has taken an initial step to tackle this setting without confounding. An important next step will be to achieve similar results in the observational causal setting.

Chapter 6

Dynamic Sensitivity Analysis: The General Case

6.1 Introduction

Sequential decision-making problems in medicine, economics, and e-commerce require the use of historical observational data when online experimentation is costly, dangerous or unethical. Given the rise of big data, there is great potential to improve decisions based on personalizing treatments to those who most benefit. However, it is also more difficult to ex-ante specify the underlying dynamics when personalizing sequential decision-making from rich data, which precludes performance evaluation via traditional methods based on stochastic simulation. The recent literature on offline reinforcement learning addresses these challenges of evaluating sequential decision rules, given only a historical dataset of observed trajectories. In particular, we focus on methods that target estimation of the Q function leveraging black-box regression, such as fitted-Q-evaluation and fitted-Q-iteration for policy evaluation and optimization, respectively.

However, these methods almost unilaterally all assume full observability of all the covariate information that informed historical treatment decisions. Unfortunately, historical decision-making policies typically made decisions based on additional unobserved variables. Such data was usually collected for convenience from a system that was optimizing for outcomes, or other complex human decisions. Data collected under "business as usual" is neither a randomized controlled trial nor a designed observational study emulating a target trial. This introduces *unobserved confounders*, variables that impact both treatment assignment and outcomes. In the presence of unmeasured confounders, the typical approach of estimating transition probabilities and solving standard Markov decision processes is biased due to incomplete adjustment for confounding.

The default realistic case for observational data is that there were some unobserved confounders; but as datasets grow richer in the era of big data, their influence may be limited. For example, if working with a database of electronic health records, it may become

more plausible that information such as recorded patient vitals explain most of medical decision-making while unobserved confounders such as patient affect may be less important. *Sensitivity analysis* techniques in the causal inference literature assess the impact of potential unobserved confounding. Instead of reporting incorrect point estimates, they report the range of estimates consistent with some potential amount of unobserved confounding, via how it affects the probability of selection into treatment [247, 252, 305]. These estimates can be framed as optimization problems over ambiguity sets, which can be sized by domain expertise, for example by comparing to the informativity of observed covariates. Importantly, such restrictions on the unobserved confounding are untestable from observational data, and ambiguity sets on unobserved confounding differ from uncertainty sets motivated on probabilistic grounds alone, i.e. robustness to finite-sample deviations.

We study robust sequential personalized policy learning under an ambiguity set of the unknown probability of taking actions given *both* observed and unobserved confounders, the *propensity score*. Importantly, we go beyond prior work because we seek not only robust bounds on value, but also robust decisions. Our algorithm links sensitivity analysis under unobserved confounders to the framework of robust Markov decision processes, and uses statistical function approximation to estimate bounds on the worst-case conditional bias of the Q function. More specifically, we use the "marginal sensitivity model" (MSM) of [292], a variant of Rosenbaum's sensitivity model [252], which has been widely used for offline single-timestep policy optimization [13, 206, 325, 315, 154, 160]. Contrary to typical uses of the MSM probing importance sampling-based estimators, we partially identify the Bellman equation for the state-action value function using an MSM with state-conditional restrictions. We develop the first principled and practical methodology for robust sequential policy learning under memoryless unobserved confounders. Recent work has only solved robust policy evaluation (not learning) under the sequential MSM under restrictions such as one-stage unobserved confounders [214], or small, discrete state spaces under additional assumptions [158, 44]. Partially identifying the Bellman equation provides a direct connection to practical policy optimization algorithms such as the fitted-Q-iteration we extend.

Learning from observational data is crucial to make progress on data-driven decision-making in consequential domains where online reinforcement learning is infeasible or costly. For example, the release of electronic health records such as the MIMIC-III critical care database enabled rich data-driven research on medical decision-making: researchers developed an illustrative task for offline reinforcement learning based on managing sepsis via administration of vasopressors and fluids, a complex dynamic task without clinical consensus. This is an important problem: sepsis is one of the foremost drivers of both mortality and hospital costs. But in the causal and reinforcement learning setting, typical performance measures in machine learning such as cross-validation, or simulation of sequential policies using a known generative model are *not valid*. Instead, the performance evaluation of learned sequential decision policies via off-policy evaluation from offline reinforcement learning implicitly requires the untrue assumption of unconfoundedness. [111] thoroughly articulates these challenges of offline evaluation, including the likely presence of no unobserved confounders in this dataset. Importantly, such real-world data is complex, motivating scalable approaches based

on statistical learning for generalization to unseen states. In important settings such as sepsis management, robust information can leverage widely available, but imperfect data, to support more resource-intensive investigations. The wide previous usage of these methods speaks to the importance of the question. As observed data grows richer, robust methods can support state-of-the-art methodology to safely obtain valid partial inferential information from observational data and partially inform managerial insights.

In this paper, we develop methodology for robust bounds and decision rules that can inform managerial decisions in a number of ways. Later on, we revisit sepsis data from MIMIC-III: since our method allows direct comparison to typical fitted-Q-evaluation/iteration methods used in the literature, we show how comparing robust vs. nominal value functions can provide insight or inform future investigation. More broadly, the FDA has recognized a growing need for methods that assess the "robustness and resilience of these [clinical decision support] algorithms to withstand changing clinical inputs and conditions" [92]. A recent working group argues that sensitivity analysis can support product development from real-world evidence and points out the need for comparable methodology for the sequential policy learning setting [82]. Finally, even if robust policies are not deployed directly, robust bounds can be used as prior knowledge to improve the data-efficiency of online experimentation, if it becomes available. We introduce an extension of our methods for warm-starting online reinforcement learning, which also highlights key differences of our structural assumptions from other models for Markov decision processes with unobserved confounders: the online counterpart assuming no memoryless unobserved confounders is a tractable MDP instead of an difficult-to-solve partially observable Markov decision process (POMDP).

Contributions: we develop an algorithm for *efficiently computing* MSM bounds with multi-step confounding, high-dimensional continuous state spaces and function approximation. Our approach leverages the recent characterization of sensitivity in single-step settings as a conditional expected shortfall (also called conditional CVaR or superquantile) [270]. Our algorithm is a simple extension of fitted-Q evaluation/iteration [100, 186] that can be implemented with off-the-shelf supervised learning algorithms, making it easily accessible to practitioners. We solve a key *statistical* challenge by incorporating orthogonalized estimation of the robust Bellman operator, and derive a corresponding *theoretical* analysis, giving sample complexity guarantees for orthogonalized robust FQI based on the richness of the approximating function classes. This reduces the dependence of statistical error in estimating the conditional expected shortfall on estimation of the conditional quantile function. Finally, we show how our model enables warm-starting standard optimistic reinforcement learning from valid robust bounds for safe data-efficiency. Our algorithm enables researchers in the managerial, clinical, and social sciences to assess and report sensitivity to unobserved confounding for dynamic policies learned from observational data, and to learn new policies that are more robust when assumptions on confounders fail.

6.2 Related Work

We first discuss offline reinforcement learning in general, and other approaches for unobserved confounders besides ours based on robustness. Then we discuss other topics such as orthogonalized estimation, robust Markov decision processes, and robust offline reinforcement learning; before summarizing how our work is at the intersection of and relates to these areas.

Policy learning with unobserved confounders in single-timestep and sequential settings. The rapidly growing literature on offline reinforcement learning with unobserved confounders can broadly be divided into three categories. We briefly discuss central differences from our approach to these three broad groups and include an expanded discussion in the appendix. First, some work assumes point identification is available via instrumental variables [308]/latent variable models [32]/front-door identification [275]. Although point identification is nice if available, sensitivity analysis can be used when assumptions of point identification (instrumental-variables, front-door adjustment) *are not true*, as may be the case in practice. Second, a growing literature considers proximal causal inference in POMDPs from temporal structure [298, 33, 303, 274] or additional proxies [205]. Proximal causal inference imposes additional (unverifiable) completeness assumptions on the latent variable structure and is a statistically challenging ill-posed inverse problem. Furthermore, we study a more restricted model of memoryless unobserved confounders that precisely delineates unobserved confounding from general POMDP concerns. As a result, we have an online counterpart that is a marginal MDP, justifying warmstarting approaches. Third, a few approaches compute no-information partial identification (PI) bounds based only on the structure of probability distributions and no more. [121] obtains a partial order on decision rules with only the law of total probability. [54] derives PI bounds with time-varying instrumental variables, based on Manski-Pepper bounds. These can generally be much more conservative than sensitivity analysis, which relaxes strong assumptions.

Overall, developing a *variety* of identification approaches further is crucial both for analysts to use appropriate estimators/bounds, and methodologically to support falsifiability analyses. Other works include [101, 188, 261]. In our work, we consider the marginal sensitivity model. Extending to other sensitivity analysis models may also be of interest [247, 264, 317, 38, 39, 265, 62]. Both the state-action conditional uncertainty sets and the assumption of memoryless unobserved confounders are particularly crucial in granting state-action rectangularity (for binary treatments), and avoiding decision-theoretic issues with time-inconsistent preferences in multi-stage robust optimization [78]. On the other hand, the exact functional form (subject to these structural assumptions) could readily be modified.

Recent work of [230] also proposes a robust fitted-Q-iteration algorithm for RMDPs. Although the broad algorithmic design is similar, we consider a different uncertainty set from their ℓ_1 set, and further introduce orthogonalization. In the single-timestep setting, further improvements are possible when targeting a simpler scalar mean, such as in [84, 85]. By constrast, we need to estimate the *entire robust Q-function*.

Off-policy evaluation in offline reinforcement learning An extensive line of work on off-policy evaluation [144, 301, 192, 296] in offline reinforcement learning studies estimating the policy value of a posited evaluation policy when only data from the behavior policy is available. Most of this literature, implicitly or explicitly, assumes sequential ignorability/sequential unconfoundedness. Methods for policy optimization are also different in the offline setting than in the online setting. Options include direct policy search (which is quite sensitive to functional specification of the optimal policy) [326], off-policy policy gradients which are either statistically noisy [137] or statistically debiased but computationally inefficient [157], or fitted-Q-iteration [186, 88]. Of these, fitted-Q-iteration's ease of use and scalability make it a popular choice in practice. It is also theoretically well-studied [86]. A marginal MDP also appears in [161] but in a different context, without unobserved confounders.

Orthogonalized estimation. Double/debiased machine learning seeks so-called Neyman-orthogonalized estimators of statistical functionals so that the Gateaux derivative of the statistical functional with respect to nuisance estimators is 0 [216, 60, 96]. Nuisance estimators are intermediate regression steps (i.e. the conditional quantile) that are not the actual target function of interest (i.e. the robust Q function). Orthogonalized estimation reduces the dependence of the statistical estimator on the estimation rate of the nuisance estimator. See [167] for tutorial discussion and [149] for a computationally-minded tutorial. There is extensive literature on double robustness/semiparametric estimation in the longitudinal setting, often from biostatistics and statistics [180, 247, 227]. Many recent works have studied double/debiased machine learning in the sequential and off-policy setting [34, 156, 280, 187].

Recent work studies orthogonality/efficiency for partial identification and in other sensitivity models than the one here [38, 39, 265, 62]. [266, 226] study orthogonalization of partial identification or conditional expected shortfall, and we build on some of their analysis in this paper. In particular, we directly apply the orthogonalization given in [226]. [315] study orthogonality under the closely related Rosenbaum model and provide very nice theoretical results. They obtain their orthogonalization via a variational characterization of expectiles. Though [214] consider a restricted model of the *worst* single-timestep confounding, out of all timesteps, it seems likely that sequential orthogonalization under the sequential exogenous confounders assumption is also possible. The single-timestep work of [142] orthogonalizes a marginal CVaR, but they assume the quantile function is known. [84] provide very nice and strong theoretical guarantees and surface additional properties of double validity.

Robust Markov decision processes and offline reinforcement learning. Elsewhere, in the robust Markov-decision process framework [221], the challenge of *rectangularity* has been classically recognized as an obstacle to efficient algorithms although special models may admit non-rectangularity and computational tractability [114]. Many recent algorithmic improvements are tailored for special structure of ambiguity sets [24, 132]. On the other hand, work in robust Markov decision processes has prominently featured the role of uncertainty sets and coherent risk measures, for example in distributionally robust Markov decision processes

[327]. Our work relates sensitivity analysis in sequential causal inference to this line of literature and focuses on algorithms for policy evaluation based on a robust fitted-Q-iteration. Other relevant works include [194], which considers a "soft-robust" criterion that averages the nominal expectation and the robust expectation; however, they study *marginal* CVaR while our later discussion of CVaR is conditional. Studying the conditional expected shortfall (equivalently, CVaR) uncertainty set is a crucial difference from previous work on risk-sensitive MDPs [65].

Importantly, robust offline reinforcement learning frameworks by themselves don't necessarily inform the problem of causal *ambiguity*: it can be more plausible that decision-makers can reason about, or external evidence can inform, restrictions on the underlying selection process rather than "assuming the consequent", i.e. positing restrictions on the bias in transition probabilities directly. On the other hand, our identification argument links ambiguity in the unobserved confounders to an equivalent ambiguity set on transition probabilities: we can also go the other way and relate ambiguity sets that appear in robust MDPs to ambiguity sets on unobserved confounding.

Lastly, we emphasize that the quantile level in our setting with ambiguity in our later conditional CVaR reformulation depends on the *analyst-specified ambiguity* rather than a probabilistic confidence level. More generally, the so-called "pessimism" principle in offline reinforcement learning is well-studied as a tool to relax strong concentrability assumptions [146], but such robustness sets are calibrated to probabilistic confidence levels.

Regarding distributionally robust offline reinforcement learning specifically, [198] studies linear function approximation. [318] studies the sample complexity of tabular robust MDPs under a generative model. The focus of our work is on unobserved confounders, although we reformulate the ambiguity set as a distributionally robust optimization problem. Other, less related, works study distributionally robust online learning [309].

Summary of differences of our work. We connect robustness for causal inference under unobserved confounders to distributionally robust MDPs and orthogonalized estimation, to obtain scalable methods with provable guarantees. In contrast to the line of work developing specialized (first-order) algorithms for (robust) Markov decision-processes, we consider approximate (robust) Bellman operator evaluations in the fitted-Q-evaluation/iteration paradigm. We use the closed-form characterization of the state-conditional solution to derive the infinite-data solution and approximate the estimation of the resulting function from data. Also, methodologically, we leverage orthogonalized estimation, which does not appear in previously mentioned works on distributionally robust offline reinforcement learning and can be of interest beyond our setting of unobserved confounders.

6.3 Problem Setup and Characterization

Problem Setup with Unobserved State

We consider a finite-horizon Markov decision process on a full-information state space, summarized as the tuple $\mathcal{M} = (\mathcal{S} \times \mathcal{U}, \mathcal{A}, R, P, \chi, T)$. We let the product state space of observed and unobserved confounders, $\mathcal{S}$,$\mathcal{U}$, be continuous, and assume the action space $\mathcal{A}$ is finite. The Markov decision process dynamics proceed from $t = 0, \ldots, T-1$ for a finite horizon of length T. Although we focus on presenting the finite-horizon case in the main text, the method and results extend readily to the discounted infinite-horizon case, discussed in the appendix. Let $\Delta(X)$ denote probability measures on a set X. The set of time t transition functions P is defined with elements $P_t : \mathcal{S} \times \mathcal{U} \times \mathcal{A} \to \Delta(\mathcal{S} \times \mathcal{U})$; R denotes the set of time t reward maps with $R_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$; the initial state distribution is $\chi \in \Delta(\mathcal{S} \times \mathcal{U})$. A policy, π, is a set of maps $\pi_t : \mathcal{S} \times \mathcal{U} \to \Delta(\mathcal{A})$, where $\pi_t(a \mid s, u)$ describes the probability of taking actions given states and unobserved confounders. Given the initial state distribution, the Markov decision process dynamics under policy π induce the random variables, for all t, $A_t \sim \pi_t(\cdot \mid S_t, U_t)$, $S_{t+1}, U_{t+1} \sim P_t(\cdot \mid S_t, U_t, A_t)$. When another type of norm is not indicated, we let $\|f\| := \mathbb{E}[f^2]^{1/2}$ indicate the 2-norm.

We consider a *confounded offline* setting: data is collected via an arbitrary behavior policy π^b that potentially depends on U_t, but in the resulting data set, the $\mathcal{U}$ part of the state space is *unobserved*. That is, although the underlying dynamics follow a standard Markov decision process generating the history $\{(S_t^{(i)}, U_t^{(i)}, A_t^{(i)}, S_{t+1}^{(i)})_{t=0}^{T-1}\}_{i=1}^n$, the observational dataset omits the unobserved confounder. The observational dataset comprises of N trajectories including observed confounders only, $\mathcal{D}_{obs} := \{(S_t^{(i)}, A_t^{(i)}, S_{t+1}^{(i)})_{t=0}^{T-1}\}_{i=1}^n$ For example, we might have a data set of electronic medical records and treatment decisions made by doctors; the electronic medical records include an observed set of patient measurements S_t, but the doctors may have made their treatment decisions using additional unrecorded information U_t.

As in standard offline RL, we study policy evaluation and optimization for target policies π^e using data collected under π^b. In our confounded setting, we consider π^e that are a function of the observed state S_t alone.We will use P_π and $\mathbb{E}_\pi$ to denote the joint probabilities (and expectations thereof) of the random variables $S_t, U_t, A_t, \forall t$ in the underlying MDP running policy π. For the special case of the behavior policy π^b, we will write P_{obs}, $\mathbb{E}_{\text{obs}}$ to emphasize the distribution of variables in the observational dataset.

Our objects of interest will be the observed state Q function and value function for the target policy π^e:

$$Q_t^{\pi^e}(s,a) := \mathbb{E}_{\pi^e}[\sum_{j=t}^{T-1} R(S_j, A_j, S_{j+1}) | S_t = s, A_t = a] \tag{6.1}$$
$$V_t^{\pi^e}(s) := \mathbb{E}_{\pi^e}[Q_t^{\pi^e}(S_t, A_t) | S_t = s].$$

We would like to find a policy π^e that is a function of the observed state alone, maximizing $V_t^{\pi^e}$. Throughout, we work primarily in the offline reinforcement learning setting where we

do not have access to online exploration due to cost or safety concerns. With unobserved confounders, we cannot directly evaluate the true expectations above due to biased estimation. Therefore in the remainder of Section 6.3, we introduce confounding-*robust* Q and value functions, which we can estimate from the observational data.

Defining an MDP on Observables

We next articulate the challenges of our setting more specifically and introduce our main structural assumption of memoryless unobserved confounders. For offline policy evaluation/optimization with unobserved confounding, there are two separate concerns: biased estimation from confounded observational data, and partial observability in the presence of unobserved confounders. First, the dependence of π^b on U_t introduces unobserved confounding, sothe distribution of the observed data is biased for estimating the true underlying transition probabilities. Without further assumptions, the observational distribution alone cannot completely adjust for the spurious correlation induced by the behavior policy. Second, even if we knew the true underlying transition probabilities, the existence of the unobserved state would change the policy optimization problem from a tractable MDP to an intractable Partially-Observed Markov decision process (POMDP). Standard RL algorithms like Bellman iteration for MDPs would no longer yield an optimal policy — because, for example, the observed next state S_{t+1} need not be Markovian conditional on only S_t and A_t.

In this section, we *isolate* the confounding concern from the POMDP concern by introducing a "memoryless confounding" assumption. Under this assumption, we will show that policy evaluation over π^e in the underlying MDP is equivalent to policy evaluation in a marginal MDP over the observed state alone. Therefore, the underlying difficulty of decision-making under memoryless unobserved confounders is intermediate between the unconfounded and generic POMDP setting.

Assumption 3 (Memoryless unobserved confounders)**.** *The unobserved state U_{t+1} is independent of S_t, U_t, A_t.*

What settings satisfy this assumption of memoryless unobserved confounders? One such example in the medical setting could be due to a memoryless arrival process of some additional information that affects both treatment and transition to the next state. For example, [322] conducts NLP analysis of clinical notes of electronic health records (which may contain information not entered in the structured EHR) and finds keywords such as "attention", "alertness" are confounders in their setting. Such aspects of patients may vary via memoryless arrival rates of periods of unalertness and may be observable by physicians, but unrecorded in the structured data. On the other hand, baseline unobserved confounders that are fixed throughout the time horizon are not directly handled in this framework, except via interpreting our model as a timewise relaxation.

Under this assumption, the full-information transition probabilities factorize as:

$$\begin{aligned} P_t(s_{t+1}, u_{t+1}|s_t, a_t, u_t) &= P_t(s_{t+1}|s_t, a_t, u_t)P_t(u_{t+1}|s_{t+1}, s_t, a_t, u_t) \\ &= \underbrace{P_t(s_{t+1}|s_t, a_t, u_t)}_{\text{new observed state}} \underbrace{P_t(u_{t+1}|s_{t+1})}_{\text{new unobserved state}} . \end{aligned}$$

In a slight abuse of notation, we will change the subscript on the unobserved state distribution to read $P_{t+1}(u_{t+1}|s_{t+1})$ so that the time subscripts are consistent. Note that under Assumption 3, $P_{\text{obs}}(U_t|S_t)$ is always the same regardless of what policy produced the historical data. Without Assumption 3, $P_{\text{obs}}(U_t|S_t)$ would generally vary with the behavior policy π^b because U_t could depend on S_{t-1}, A_{t-1}, and U_{t-1}.

With memoryless unobserved confounders, observed-state policy evaluation and optimization in the full POMDP reduce to an MDP problem. Define the marginal transition probabilities:

$$P_t(s_{t+1}|s_t, a_t) \coloneqq \int_{\mathcal{U}} P_t(u_t|s_t)P_t(s_{t+1}|s_t, a_t, u_t)du_t \tag{6.2}$$

Then we have the following proposition:

Proposition 6.3.1 (Marginal MDP)**.** *Given Assumption 1, for any policy π^e that is a function of S_t alone, the distribution of $S_t, A_t, \forall t$ in the full-information MDP running π^e is equivalent to the distribution of $S_t, A_t, \forall t$ in the* marginal *MDP, $(\mathcal{S}, \mathcal{A}, R, P, \chi, T)$. That is, $S_0 \sim \chi$, $A_t \sim \pi^e(\cdot \mid S_t)$, $S_{t+1} \sim P_t(\cdot|S_t, A_t)$.*

See the Appendix for a formal derivation. The key takeaway from Proposition 6.3.1 is that if we knew the true marginal transition probabilities, $P_t(S_{t+1}|S_t, A_t)$, then we could apply standard RL algorithms for evaluation or optimization. We have observed-state Q and value functions in the marginal MDP, that satisfy the Bellman evaluation equations,

$$\begin{aligned} Q_t^{\pi^e}(s, a) &= \mathbb{E}_{P_t}[R_t + Q_{t+1}^{\pi^e}(S_{t+1}, \pi_{t+1}^e)|S_t = s, A_t = a], \\ V_t^{\pi^e}(s) &= \mathbb{E}_{A\sim\pi_t^e(s)}[Q_t^{\pi^e}(s, A)] \end{aligned}$$

where we use the short-hands $R_t \coloneqq R_t(S_t, A_t, S_{t+1})$ and $g(S', \pi) \coloneqq \mathbb{E}_{A'\sim\pi(S')}[g(S', A')]$ for any $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Furthermore, by classical results [237], an optimal policy exists among policies defined on the observed state alone, yielding the optimal Q function, $Q_t^*(s, a)$, and value function, $V_t^*(s)$, with corresponding Bellman optimality equations.

Before continuing, we want to emphasize that while Assumption 3 is strong, it has testable implications. In particular, under Assumption 3 the observed-state transition probabilities will be *Markovian*, which can be tested from observed states and actions alone.[1]

[1]It is possible to use observed-state Markovian transitions as the core assumption at the cost of substantially more complexity. See the Appendix for discussion.

Offline RL and Unobserved Confounding

Proposition 6.3.1 establishes that the oracle decision problem, given knowledge of the true marginal transition probabilities, remains a Markov decision process under memoryless confounding. However, while Assumption 3 rules out POMDP concerns, it does not rule out bias from unobserved confounding. In general, it is *not* possible to get unbiased estimates of the true marginal observed-state transitions given data collected under π^b when U_t is unobserved. In particular, $P_{\text{obs}}(S_{t+1}|S_t, A_t) \neq P_t(S_{t+1}|S_t, A_t)$. To see this, first define the marginal behavior policy,

$$\pi_t^b(a_t|s_t) \coloneqq \int_{\mathcal{U}} \pi_t^b(a_t|s_t, u_t) P_t(u_t|s_t) du_t = P_{\text{obs}}(a_t|s_t).$$

Then,

$$\begin{aligned} P_{\text{obs}}(s_{t+1}|s_t, a_t) &= \int_{\mathcal{U}} P_t(s_{t+1}|s_t, a_t, u_t) P_{\text{obs}}(u_t|s_t, a_t) du_t \\ &= \int_{\mathcal{U}} P_t(s_{t+1}|s_t, a_t, u_t) \frac{\pi_t^b(a_t|s_t, u_t)}{\pi_t^b(a_t|s_t)} P_t(u_t|s_t) du_t, \end{aligned} \tag{6.3}$$

where the second equality follows by Bayes rule. The final expression for $P_{\text{obs}}(s_{t+1}|s_t, a_t)$ differs from $P_t(s_{t+1}|s_t, a_t)$ in eq. (6.2) by the unobserved factor $\frac{\pi_t^b(a_t|s_t,u_t)}{\pi_t^b(a_t|s_t)}$. Note that the term $P_{\text{obs}}(u_t|s_t, a_t)$ is the bias from confounding: in the observational distribution conditioning on a_t changes the distribution of the unobserved u_t relative to $P_t(u_t|s_t)$ because a_t is drawn according to $\pi^b(a_t|s_t, u_t)$.

If π^b is independent of u_t, the ratio $\frac{\pi_t^b(a_t|s_t,u_t)}{\pi_t^b(a_t|s_t)}$ will be uniformly 1 and we recover $P_t(s_{t+1}|s_t, a_t)$. However, if $\pi_t^b(a_t|s_t, u_t)$ can be arbitrary, then an estimate of $P_t(s_{t+1}|s_t, a_t)$ using $P_{\text{obs}}(s_{t+1}|s_t, a_t)$ can be arbitrarily biased. This result immediately implies that any regression using P_{obs} will be biased for the corresponding estimand in the marginal MDP.

Proposition 6.3.2 (Confounding for Regression)**.** *Let $f : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ be any function. Given Assumption 3, $\forall s, a$,*

$$\mathbb{E}_{P_t}\big[f(S_t, A_t, S_{t+1})|S_t = s, A_t = a\big] = \mathbb{E}_{obs}\left[\frac{\pi_t^b(A_t|S_t)}{\pi_t^b(A_t|S_t, U_t)} f(S_t, A_t, S_{t+1}) \middle| S_t = s, A_t = a\right].$$

where the first equality follows from Proposition 6.3.1 and the second equality follows from Equation (6.3). This proposition shows that regression of f on states and actions using data collected according to π^b is a biased estimator for the corresponding conditional expectation under the true marginal transition probabilities $P_t(s'|s, a)$ where the exact bias is:

$$\begin{aligned} &\mathbb{E}_{\text{obs}}[f(S_t, A_t, S_{t+1})|S_t = s, A_t = a] - \mathbb{E}_{P_t}\big[f(S_t, A_t, S_{t+1})|S_t = s, A_t = a\big] \\ &\qquad = \mathbb{E}_{\text{obs}}\left[\left(1 - \frac{\pi_t^b(A_t|S_t)}{\pi_t^b(A_t|S_t, U_t)}\right) f(S_t, A_t, S_{t+1}) \middle| S_t = s, A_t = a\right]. \end{aligned}$$

Since the unobserved factor $\frac{\pi_t^b(A_t|S_t)}{\pi_t^b(A_t|S_t,U_t)}$ can be arbitrarily large without further assumptions, to make progress we follow the sensitivity analysis literature in causal inference.

Assumption 4 (Marginal Sensitivity Model)**.** *There exists* Λ *such that* $\forall t, s \in \mathcal{S}, u \in \mathcal{U}, a \in \mathcal{A}$*,*

$$\Lambda^{-1} \leq \left(\frac{\pi_t^b(a \mid s,u)}{1-\pi_t^b(a \mid s,u)}\right) / \left(\frac{\pi_t^b(a \mid s)}{1-\pi_t^b(a \mid s)}\right) \leq \Lambda. \tag{6.4}$$

The parameter Λ for this commonly-used sensitivity model in causal inference [292] has to be chosen with domain knowledge. A common approach is to compare Λ to corresponding values for observed variables, e.g. in a clinical setting, if smoking has an effective $\Lambda = 1.5$, a practitioner might say "I do not believe there exists an unobserved variable with twice the explanatory power of smoking" to justify a choice of $\Lambda = 3$ [133].

Now consider any function $f : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ as in Proposition 6.3.2. For shorthand, we will write $Y_t \coloneqq f(S_t, A_t, S_{t+1})$. We use a generic f here to emphasize that this argument would apply to any model-based or model-free RL algorithms using regression, but later when we introduce our fitted-Q iteration algorithm, we will specialize Y_t to get an empirical estimate of the Bellman operator. Combining Assumption 4 and Proposition 6.3.2, we can express the target expectation $\mathbb{E}_{P_t}[Y_t|S_t, A_t]$ as a weighted regression under the behavior policy with bounded weights. Define the random variable

$$W_t^{\pi^b} \coloneqq \frac{\pi_t^b(A_t|S_t)}{\pi_t^b(A_t|S_t,U_t)}, \qquad \text{where } \mathbb{E}_{P_t}[Y_t|S_t,A_t] = \mathbb{E}_{\text{obs}}[W_t^{\pi^b} Y_t|S_t,A_t] \qquad \text{(proposition 6.3.2).} \tag{6.5}$$

While we cannot estimate $W_t^{\pi^b}$, we can bound it. The weights must satisfy that $\pi_b(a \mid s,u)$ is a valid probability distribution,

$$\mathbb{E}_{\text{obs}}[W_t^{\pi^b}|S_t,A_t] = 1, \tag{6.6}$$

and Assumption 4 implies the following bounds almost everywhere:

$$\alpha_t(S,A) \leq W_t^{\pi^b} \leq \beta_t(S,A), \forall s' \tag{6.7}$$
$$\alpha_t(S,A) \coloneqq \pi_t^b(A_t|S_t) + \Lambda^{-1}(1-\pi_t^b(A_t|S_t)), \quad \beta_t(S,A) \coloneqq \pi_t^b(A_t|S_t) + \Lambda(1-\pi_t^b(A_t|S_t)).$$

So while Proposition 6.3.2 demonstrates that we cannot unbiasedly estimate the value function in the confounded setting, we can instead compute worst-case bounds on the conditional bias subject to the constraints in eqs. (6.6) and (6.7). Next, we will make this precise by showing that Assumption 4 defines a Robust Markov decision process.

Robust Estimands and Bellman Operators

In this section, we introduce our key estimands – the robust Q and value functions. Assumption 4 implies the constraints in eqs. (6.6) and (6.7), which define an uncertainty set for the

true observed-state transition probabilities $P_t(s'|s,a)$. [158] and [44] uses a reparameterization to show that for each weight W_t that satisfies these constraints, there is a corresponding transition probability in the set:

$$\bar{P}_t(\cdot \mid s,a) \in \mathcal{P}_t^{s,a} \coloneqq \Bigg\{ \bar{P}_t(\cdot \mid s,a) \colon \alpha_t(s,a) \leq \frac{\bar{P}(s_{t+1} \mid s,a)}{P_{obs}(s_{t+1} \mid s,a)} \leq \beta_t(s,a), \forall s_{t+1}; \int \bar{P}_t(s_{t+1} \mid s,a) ds_{t+1} = 1 \Bigg\}$$

Define the set $\mathcal{P}_t$ of transition probabilities for all s,a to be the product set over the $\mathcal{P}_t^{s,a}$. Then under Assumptions 3 and 4, the true marginal transition probabilities belong to $\mathcal{P}_t$. While point estimation is not possible, we can find the worst-case values of $Q_t^{\pi^e}$ and $V_t^{\pi^e}$ over transition probabilities in the uncertainty set, $\bar{P}_t \in \mathcal{P}_t$ — a Robust Markov decision process (RMDP) problem [139]. Importantly, the set $\mathcal{P}_t$ is s,a-rectangular, and so we can use the results in [139] to define robust Bellman operators and a corresponding robust Bellman equation.

Denote the robust Q and value functions $\bar{Q}_t^{\pi^e}$ and $\bar{V}_t^{\pi^e}$ and define the following operators:

Definition 1 (Robust Bellman Operators)**.** *For any function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,*

$$(\bar{\mathcal{T}}_t^{\pi^e} g)(s,a) \coloneqq \inf_{\bar{P}_t \in \mathcal{P}_t} \mathbb{E}_{\bar{P}_t}[R_t + g(S_{t+1}, \pi^e_{t+1}) | S_t = s, A_t = a], \tag{6.8}$$

$$(\bar{\mathcal{T}}_t^{*} g)(s,a) \coloneqq \inf_{\bar{P}_t \in \mathcal{P}_t} \mathbb{E}_{\bar{P}_t}[R_t + \max_{A'}\{g(S_{t+1}, A')\} | S_t = s, A_t = a]. \tag{6.9}$$

Proposition 6.3.3 (Robust Bellman Equation)**.** *Let $|\mathcal{A}| = 2$ and let Assumptions 1 and 2 hold. Then applying the results in [139], gives*

$$\bar{Q}_t^{\pi^e}(s,a) = \bar{\mathcal{T}}_t^{\pi^e} \bar{Q}_{t+1}^{\pi^e}(s,a), \quad \bar{V}_t^{\pi^e}(s) = \mathbb{E}_{A \sim \pi_t^e(s)}[\bar{Q}_t^{\pi^e}(s,A)],$$
$$\bar{Q}_t^{*}(s,a) = \bar{\mathcal{T}}_t^{*} \bar{Q}_{t+1}^{*}(s,a), \quad \bar{V}_t^{*}(s) = \mathbb{E}_{A \sim \bar{\pi}_t^*(s)}[\bar{Q}_t^{*}(s,A)],$$

where $\bar{Q}_t^*$ and $\bar{V}_t^*$ are the optimal robust Q and value function achieved by the policy $\bar{\pi}^*$.

Finally, we comment on the tightness of the robust operator. For a fixed s and a, the $\mathcal{P}_t^{s,a}$ is exactly the set of transition probabilities consistent with Assumption 4 and the observational data distribution. However the s,a-rectangular product set $\mathcal{P}_t$ does not explicitly enforce the density constraint on π_t^b *across actions*, and is therefore potentially loose. In the special case where there are only two actions, [85] show that the different minima over $\mathcal{P}_t^{s,a}$ across actions are *simultaneously achievable*, and thus the robust bounds are tight and we get equalities in Proposition 6.3.3. For $|\mathcal{A}| > 2$, the infimum in eq. (6.8) is *not* generally simultaneously realizable — it's easy to construct a counter-example. Nonetheless, the robust Bellman operator corresponds to an s,a-rectangular relaxation of the RMDP, Proposition 6.3.3 will hold with lower bounds instead of equalities, and our results are still guaranteed to be robust.

6.4 Method

In the previous section, we defined our estimands of interest — the robust Q and value functions under the marginal sensitivity model. In this section, we introduce robust policy optimization via function approximation. Our estimation strategy is a robust analog of Fitted-Q Iteration (FQI).

Assume that we observe n trajectories of length T, where the observational dataset $\mathcal{D}_{obs} \coloneqq \{(S_t^{(i)}, A_t^{(i)}, S_{t+1}^{(i)})_{t=0}^{T-1}\}_{i=1}^n$ was collected from the underlying MDP under an unknown behavior policy π^b that depends on the unobserved state. We will write $\mathbb{E}_{n,t}$ to denote a sample average of the n data points collected at time t, e.g. $\mathbb{E}_{n,t}[f(S_t, A_t, S_{t+1})] \coloneqq \frac{1}{n}\sum_{i=1}^n f(S_t^{(i)}, A_t^{(i)}, S_{t+1}^{(i)})$. Nominal (non-robust) FQI [88, 186, 86] successively forms approximations $\hat{Q}_t$ at each time step by minimizing the Bellman error:

$$Y_t(Q) \coloneqq R_t + \max_{a'} \left[Q(S_{t+1}, a')\right], \qquad Q_t(s,a) = \mathbb{E}[Y_t(Q_{t+1})|S_t = s, A_t = a], \tag{6.10}$$

$$\hat{Q}_t \in \arg\min_{q_t \in \mathcal{Q}} \mathbb{E}_{n,t}[(Y_t(\hat{Q}_{t+1}) - q_t(S_t, A_t))^2]. \tag{6.11}$$

The Bayes-optimal predictor of Y_t is the true Q_t function, even though Y_t is a stochastic approximation of Q_t that replaces the expectation over the next-state transition with a stochastic sample thereof (realized from data). In this way, fitted-Q-iteration is *pseudo-outcome* regression, regressing onto a random variable whose conditional expectation is the target function, but is not equivalent to it under additive noise, as is the case with typical regression on observed outcomes. Pseudo-outcome regression has recently been used in causal inference [166, 268], and later in our robust procedure we are therefore able to use analogous arguments to obtain orthogonalized estimation. The procedure for fitted-Q-evaluation is exactly analogous, replacing the maximum over next-timestep actions with evaluation under the evaluation policy. In this manuscript, we present on focusing the fitted-Q-iteration case for succinctness.

In our robust version of FQI, we instead approximate the robust Bellman operator from eq. (6.9). In particular, we will apply Proposition 6.3.2, but impose the constraints in eqs. (6.6) and (6.7) to arrive at the following optimization problem in terms of observable quantities:

Proposition 6.4.1. *Let Q be a real-valued function over states and actions, and define $Y_t(Q)$ as in Equation* (6.11)*. Given Assumption 3 and Assumption 4, the robust $Q(s,a)$ function solves the following optimization problem:*

$$(\bar{\mathcal{T}}_t^* Q)(s,a) = \min_{W_t} \big\{ \mathbb{E}_{obs}\left[W_t Y_t(Q) | S_t = s, A_t = a\right] : \\ \mathbb{E}_{obs}\left[W_t | S_t = s, A_t = a\right] = 1, \;\; \alpha_t(S,A) \le W_t \le \beta_t(S,A), \, a.e. \big\}.$$

Next, in Section 6.4, we show that the optimization problem in Proposition 6.4.1 admits a closed form as a conditional expectation of observables. Then in Section 6.4, we incorporate this insight into an orthogonalized confounding-robust FQI algorithm with function approximation.

Closed-Form for the Robust Bellman Operator

Solving the optimization problem in Proposition 6.4.1 for each s, a pair isn't feasible for large state and action spaces. In this section, we use recent results to derive a *closed-form* expression for the minimum in Proposition 6.4.1 in order to derive a feasible algorithm leveraging function approximation. This is an application of the results in [249] and [85].

The closed-form state-action conditional solution to Proposition 6.4.1 is written in terms of a superquantile (also called conditional expected shortfall, or covariate-conditional CVaR). The conditional expected shortfall is the conditional expectation of exceedances of a random variable beyond its conditional quantile. Define $\tau \coloneqq \Lambda/(1+\Lambda)$. For any function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we define the observational $(1-\tau)$-level conditional quantile of the Bellman target:

$$Z_t^{1-\tau}(Y_t(Q) \mid s, a) \coloneqq \inf_z \{z \colon P_{\text{obs}}(Y_t(Q) \geq z \mid S_t = s, A_t = a) \leq 1 - \tau\}.$$

We use the following shorthands when clear from context: $Z_{t,a}^{1-\tau} \coloneqq Z_t^{1-\tau}(Y_t(Q) \mid s, a), \alpha_t \coloneqq \alpha_t(S, A), \beta_t \coloneqq \beta_t(S, A)$. We can learn the conditional quantile functions by minimizing the *pinball loss* over a function class $\mathcal{Z}$:

$$Z_t^{1-\tau}(Y_t(Q) \mid S_t, A_t) \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}[L_\tau(Y_t(Q), z(S_t, A_t))],$$
$$\text{where } L_\tau(y, \hat{y}) \coloneqq \begin{cases} (1-\tau)(\hat{y} - y), & \text{if } y < \hat{y} \\ \tau(y - \hat{y}), & \text{if } y \geq \hat{y} \end{cases}. \tag{6.12}$$

Proposition 6.4.2. *The solution to the minimization problem in Proposition 6.4.1 is:*

$$(\bar{\mathcal{T}}_t^* Q)(s, a) = \mathbb{E}_{obs}\left[\alpha_t Y_t(Q) + \frac{1-\alpha_t}{1-\tau} Y_t(Q) \mathbb{I}\left[Y_t(Q) \leq Z_{t,a}^{1-\tau}\right] \middle| S_t = s, A_t = a\right]. \tag{6.13}$$

Proposition 6.4.2 suggests a simple two-stage procedure. First, estimate $Z_t^{1-\tau}$, and then estimate the conditional expectation in eq. (6.13) via regression using the estimated $Z_t^{1-\tau}$. We do so to develop robust policy evaluation and optimization algorithms in the next section. We first describe the basic method, its improvement via orthogonalization, and lastly sample splitting/cross-fitting.

Improving estimation: the orthogonalized pseudo-outcome

The two-stage procedure depends on the conditional quantile function $Z_t^{1-\tau}$, a *nuisance function* that must be estimated but is not our substantive target of interest. To avoid transferring biased first-stage estimation error of $Z_t^{1-\tau}$ to the Q-function, we introduce orthogonalization. Orthogonalized estimators remove the first-order dependence of estimating the target on the error in nuisance functions. An important literature from biostatistics and econometrics on Neyman-orthogonality (also called double/debiased machine learning, and related to semiparametric statistics) derives bias adjustments [167, 216, 60, 180]. (See

Algorithm 1 Confounding-Robust Fitted-Q-Iteration

1: Estimate the marginal behavior policy $\pi_t^b(a|s)$. Compute $\{\alpha_t(S_t^{(i)}, A_t^{(i)})\}_{i=1}^n$ as in Equation (6.7). Initialize $\hat{\overline{Q}}_T = 0$.
2: **for** $t = T-1, \ldots, 1$ **do**
3: Compute the nominal outcomes $\{Y_t^{(i)}(\hat{\overline{Q}}_{t+1})\}_{i=1}^n$ as in eq. (6.11).
4: For $a \in \mathcal{A}$, where $A_t^{(i)} = a$, fit $\hat{Z}_t^{1-\tau}$ the $(1-\tau)$th conditional quantile of the outcomes $Y_t^{(i)}$.
5: Compute pseudooutcomes $\{\tilde{Y}_t^{(i)}(\hat{Z}_t^{1-\tau}, \hat{\overline{Q}}_{t+1})\}_{i=1}^n$ as in eq. (6.14).
6: For $a \in \mathcal{A}$, where $A_t^{(i)} = a$, fit $\hat{\overline{Q}}_t$ via least-squares regression of $\tilde{Y}_t^{(i)}$ against $(S_t^{(i)}, A_t^{(i)})$.
7: Compute $\pi_t^*(s) \in \arg\max_a \hat{\overline{Q}}_t(s, a)$.
8: **end for**

Appendix C.2 for more). In particular, we apply an orthogonalization of [226] for what they call truncated conditional expectations, $m(\eta, x) = \frac{1}{1-\tau}\mathbb{E}[Y\mathbb{I}[Y \leq Z^{1-\tau}] \mid X = x]$. They show that

$$\frac{1}{1-\tau}\mathbb{E}[Y\mathbb{I}[Y \leq Z^{1-\tau}] - Z^{1-\tau}(\mathbb{I}[Y \leq Z^{1-\tau}] - (1-\tau)) \mid X]$$

is Neyman-orthogonal with respect to error in $Z^{1-\tau}$. Note that this comprises an additive, zero-mean adjustment to the original pseudo-outcome. We apply this orthogonalization to Equation (6.13) to obtain our regression target for robust FQE:

$$\tilde{Y}_t(Z, Q) \coloneqq \alpha_t Y_t(Q) + \tfrac{1-\alpha_t}{1-\tau}\Big(Y_t(Q)\mathbb{I}\left[Y_t(Q) \leq Z_t^{1-\tau}\right] - Z \cdot \{\mathbb{I}[Y_t(Q) \leq Z] - (1-\tau)\}\Big) \quad (6.14)$$

When the quantile functions are consistent, the orthogonalized pseudo-outcome enjoys quadratic, not linear dependence on the first-stage estimation error in the quantile functions. We describe in more detail in the next section on guarantees. The orthogonalized time-t target of estimation is:

$$\hat{\overline{Q}}_t \in \arg\min_{q_t} \mathbb{E}_{n,t}[(\tilde{Y}_t(\hat{Z}_t^{1-\tau}, \hat{\overline{Q}}_{t+1}) - q_t(S_t, A_t))^2]. \quad (6.15)$$

A large literature discusses methods for quantile regression [175, 203, 25], as well as conditional expected shortfall [47, 165] and can guide the choice of function class for quantiles and $\overline{Q}$ appropriately.

We summarize the algorithm in Algorithm 1. In the appendix, we discuss a sample-splitting version in more detail; we describe the approach, which is standard, in the main text for brevity. Lastly, to ensure independent errors in nuisance estimation and the fitted-Q regression, for the theoretical results, we study a cross-time variant of the standard cross-fitting/sample-splitting scheme for orthogonalized estimation and machine learning. Interleaving between timesteps ensures downstream policy evaluation errors are independent of errors in nuisance evaluation at time t. Finally, we note that sample splitting can be

avoided by posing Donsker-type assumptions on the function classes in the standard way. In the experiments (and algorithm description) in the interest of data-efficiency we do not data-split. Recent work of [53] shows rigorously that sample-splitting may not be necessary under stability conditions; extending that analysis to this setting would be interesting future work.

Extension to continuous actions

Although the manuscript focuses on binary or categorical actions, the method can directly be extended to continuous action spaces, at the expense of sharpness results and interpretability of the robust set. [143] proposes a continuous-action sensitivity model which instead directly bounds the density ratio (rather than the odds ratio):

$$\frac{1}{\Lambda} \leq \frac{\pi_t^b(a \mid s)}{\pi_t^b(a \mid s, u)} \leq \Lambda. \tag{6.16}$$

In the continuous setting, densities could be greater than 1, which would violate conditions on the odds ratio. One way to interpret this sensitivity parameter is via implications for the KL-divergence of nominal and complete propensity scores. We can readily apply this to our problem by changing the uncertainty set on W to that implied by the above. Namely, solve the same linear program of Proposition 6.4.1 but enforce that $W_t = \frac{\pi_t^b(a|s)}{\pi_t^b(a|x,u)}$ satisfy the constraints of eq. (6.16) rather than Assumption 4:

$$\begin{aligned}(\bar{\mathcal{T}}_t^* Q)(s,a) = \min_{W_t} \big\{ & \mathbb{E}_{\text{obs}}\left[W_t Y_t(Q) | S_t = s, A_t = a\right] : \\ & \mathbb{E}_{\text{obs}}\left[W_t | S_t = s, A_t = a\right] = 1, \\ & \Lambda^{-1} \leq W_t \leq \Lambda^{-1}, \text{a.e.} \big\}.\end{aligned}$$

That is, the characterization of Proposition 6.4.2 holds, replacing the (α_t, β_t) bounds arising from the MSM with (Λ^{-1}, Λ). The pointwise solution of the (s,a)-conditional optimization problem is structurally the same, i.e. a conditional quantile characterization at a different level. The only difference algorithmically is in the conditional quantile estimation; in the continuous action setting, we would appeal to function approximation and minimize the (orthogonalized) pinball loss of eq. (6.12) with the action as a covariate. In the infinite-data, nonparametric limit, this would be well-specified; in practice, there will be some additional approximation error. Given those conditional quantiles, the rest of the method, (orthogonalization, etc.) proceeds analogously as discussed previously.

6.5 Analysis and Guarantees

We first describe the estimation benefits we receive from orthogonalization before discussing analysis of robust fitted-Q-evaluation and iteration, and insights. (All proofs are in the appendix).

Estimation guarantees

We describe the orthogonalized estimation results, before the results about the full output of the robust fitted-Q-iteration. We also require some regularity conditions for estimation. We assume nonnegative bounded rewards throughout.

Assumption 5 (Boundedness)**.** *Outcomes are nonnegative and bounded:* $0 \leq R_t \leq B_R, \forall t$. *The state space is bounded.*

We assume the transitions are continuously distributed, a common regularity condition for the analysis of quantiles.

Assumption 6 (Bounded conditional density)**.** *Assume that* $P_t(s_{t+1} \mid s_t, a) < M_P, \forall t, s_t, s_{t+1}$ *a.s.*

We let $\hat{\mathbb{E}}_n$ indicate a function obtained by regression, on an appropriate data split independent of the nuisance estimation. Define

$$\begin{aligned}
\hat{\overline{Q}}_t(s,a) &= \hat{\mathbb{E}}_n[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) \mid s, a] && \text{feasible regressed robust Q,} \\
\tilde{\overline{Q}}_t(s,a) &= \hat{\mathbb{E}}_n[\tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid s, a] && \text{oracle-nuisance regressed robust Q} \\
\overline{Q}_t(s,a) &= \mathbb{E}[\tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid s, a] && \text{oracle robust Q.}
\end{aligned}$$

In the above, $\hat{\overline{Q}}_t(s,a) = \hat{\mathbb{E}}_n[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) \mid s, a]$ is the feasible *regressed* robust-Q-estimator with estimated nuisance $\hat{Z}$, while $\tilde{\overline{Q}}(s,a) = \hat{\mathbb{E}}_n[\tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid s, a]$ is the *regressed* robust-Q-estimator with *oracle* nuisance Z, and $\overline{Q}_t(s,a)$ is the true robust Q output at time t (relative to the future Q functions that are the output of the algorithm).

We assume the following regression stability assumption, which appears in [166]. It is a generalization of stochastic equicontinuity and is satisfied, for example, by nonparametric linear smoothers.

Assumption 7 (Regression stability)**.** *Suppose* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *are independent training and test samples, respectively. Let: 1.* $\widehat{f}(x) = \widehat{f}\,(x; \mathcal{D}_1)$ *be an estimate of a function* $f(x)$ *using the training data* $\mathcal{D}_1$*, 2.* $\widehat{b}(x) = \widehat{b}\,(x; \mathcal{D}_1) \equiv \mathbb{E}[\widehat{f}(x) - f(x) \mid \mathcal{D}_1, X = x]$ *the conditional bias of the estimator* $\widehat{f}$*, 3.* $\widehat{\mathbb{E}}_n[Y \mid X = x]$ *denote a generic regression estimator that regresses outcomes on covariates in the test sample* $\mathcal{D}_2$*. Then the regression estimator* $\widehat{\mathbb{E}}_n$ *is defined as stable at* $X = x$ *(with respect to a distance metric* d *) if*

$$\frac{\widehat{\mathbb{E}}_n[\widehat{f}(x)|X=x] - \widehat{\mathbb{E}}_n[f(x)|X=x] - \widehat{\mathbb{E}}_n[\widehat{b}(x)|X=x]}{\sqrt{\mathbb{E}\left(\left[\widehat{\mathbb{E}}_n[f(x)|X=x] - \mathbb{E}[f(x)|X=x]\right]^2\right)}} \xrightarrow{p} 0$$

whenever $d(\widehat{f}, f) \xrightarrow{p} 0$.

Under these regularity conditions, we can show that the bias due to the first-stage estimation of the conditional quantiles is only quadratic in the estimation error of $\hat{Z}_t$.

Proposition 6.5.1 (CVaR estimation error)**.** *Assume Assumptions 5 to 7. For $a \in \mathcal{A}, t \in [T-1]$, if the conditional quantile estimation is $o_p(n^{-\frac{1}{4}})$ consistent, i.e. $\|\hat{Z}_t^{1-\tau} - Z_t^{1-\tau}\|_\infty = o_p(n^{-\frac{1}{4}})$, $\mathbb{E}[\|\hat{Z}_t^{1-\tau} - Z_t^{1-\tau}\|_2] = o_p(n^{-\frac{1}{4}})$, then*

$$\|\hat{\overline{Q}}_t(S,a) - \overline{Q}_t(S,a)\|_2 \leq \|\widetilde{\overline{Q}}_t(S,a) - \overline{Q}_t(S,a)\|_2 + o_p(n^{-\frac{1}{2}}).$$

This implies we can maintain $o_p(n^{-\frac{1}{2}})$ consistent estimation of robust $\overline{Q}$ functions under weaker estimation error requirements on the conditional quantile functions Z.

Next, we describe key assumptions for convergence of fitted-Q-iteration, concentrability which restricts the distribution shift in the sequential offline data vs. optimized policies, and approximate Bellman completeness which assumes the closedness of the regression function class under the Bellman operator. Both these assumptions are standard requirements for fitted-Q-iteration, but certainly not innocuous; they do impose restrictions.

Assumption 8 (concentrability)**.** *Given a policy π, let ρ_t^π denote the marginal distribution at time step t, starting from s_0 and following π, and μ_t denote the true marginal occupancy distribution under π^b. There exists a parameter C such that*

$$\sup_{(s,a,t)\in\mathcal{S}\times\mathcal{A}\times[T-1]} \frac{\mathrm{d}\rho_t^\pi}{\mathrm{d}\mu_t}(s,a) \leq C \quad \text{for any policy } \pi.$$

Assumption 9 (Approximate Bellman completeness)**.** *There exists $\epsilon > 0$ such that, for all $t \in [T-1]$, where ϵ is at most on the order of $O_p(n^{-\frac{1}{2}})$,*

$$\sup_{q_{t+1}\in\mathcal{Q}_{t+1}} \inf_{q_t\in\mathcal{Q}_t} \|q_t - \overline{\mathcal{T}}_t^\star q_{t+1}\|_{\mu_t}^2 \leq \epsilon.$$

concentrability is analogous to sequential overlap or positivity, as it is called in single-timestep causal inference. It assumes a uniformly bounded density ratio between the true marginal occupancy distribution and those induced by arbitrary policies. Approximate Bellman completeness assumes that the function class $\mathcal{Q}$ is approximately closed under the robust Bellman operator. Assuming that ϵ is at most $O_p(n^{-\frac{1}{2}})$ is somewhat restrictive, but is consistent with frameworks for local model misspecification that consider local asymptotics with $O_p(n^{-\frac{1}{2}})$ vanishing bias.

Although we ultimately seek an optimal policy, approaches based on fitted-Q-evaluation and iteration instead optimize the squared loss, which is related to the Bellman error that is a surrogate for value suboptimality.

Definition 2 (Bellman error)**.** *Under data distribution μ_t, define the Bellman error of function $q = (q_0, \ldots, q_{T-1})$ as: $\mathcal{E}(q) = \frac{1}{T}\sum_{t=0}^{T-1} \|q_t - \overline{\mathcal{T}}_t^* q_{t+1}\|_{\mu_t}^2$*

The next lemma, which appears as [86, Lemma 3.2] (finite horizon), [313, Thm. 2] (infinite horizon), justifies this approach by relating the Bellman error to the value suboptimality. Its proof follows immediately by considering the MDP given by the worst-case transition kernel that realizes the optimization in the definition of the robust Bellman operator and is omitted.

Lemma 6.5.2 (Bellman error to value suboptimality). *Under Assumption 8, for any $q \in \mathcal{Q}$, we have that, for π the policy that is greedy with respect to q, $V_1^*(s_1) - V_1^\pi(s_1) \leq 2T\sqrt{C \cdot \mathcal{E}(q^\pi)}$.*

We will describe convergence results based on generic results for loss minimization over a function class of restricted complexity. We use standard covering and bracketing numbers to quantify the functional complexity of infinite function classes.

Definition 3 (Covering numbers, e.g. [304]). *Let $(\mathcal{F}, \|\cdot\|)$ be an arbitrary semimetric space. Then the covering number $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls of radius ϵ needed to cover $\mathcal{F}$.*

Definition 4 (Bracketing numbers). *Given two functions l and u, the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$. An ϵ-bracket is a bracket $[l, u]$ with $\|u - l\| < \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$.*

The covering and bracketing numbers for common function classes such as linear, polynomials, neural networks, etc. are well-established in standard references, e.g. [307, 304]. We assume either that the function class for $\mathcal{Q}, \mathcal{Z}$ is finite (but possibly exponentially large), or has well-behaved *covering* and *bracketing* numbers.

Assumption 10 (Finite function classes.). *The Q-function class $\mathcal{Q}$ and conditional quantile class $\mathcal{Z}$ are finite but can be exponentially large.*

Assumption 11 (Infinite function classes with well-behaved covering number.). *The Q-function class $\mathcal{Q}$, and conditional quantile class $\mathcal{Z}$ have covering numbers $N(\epsilon, \mathcal{Q}, d)$, $N(\epsilon, \mathcal{Z}, d)$ (respectively).*

Theorem 6.5.3 (Fitted Q Iteration guarantee). *Suppose Assumptions 5 to 9 and let B_R be the bound on rewards. Recall that $\mathcal{E}(\hat{Q}) = \frac{1}{T}\sum_{t=0}^{T-1} \left\| \hat{Q}_t - \overline{\mathcal{T}}_t^\star \hat{Q}_{t+1} \right\|_{\mu_t}^2$. Then, with probability greater than $1 - \delta$, under Assumption 10 (finite function class), we have that*

$$\mathcal{E}(\hat{Q}) \leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{56(T^2+1)B_R \log\{T|\mathcal{Q}||\mathcal{Z}|/\delta\}}{3n} + \sqrt{\frac{32(T^2+1)B_R \log\{T|\mathcal{Q}||\mathcal{Z}|/\delta\}}{n}\epsilon_{\mathcal{Q},\mathcal{Z}}} + o_p(n^{-1}),$$

while under Assumption 11 (infinite function class), choosing the covering number approximation error $\epsilon = O(n^{-1})$ such that $\epsilon_{\mathcal{Q},\mathcal{Z}} = O(n^{-1})$, we have that

$$\mathcal{E}(\hat{Q}) \leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{1}{T}\sum_{t=0}^{T-1} \left\{ \frac{56(T-t-1)^2 \log\{T N_{[]}(2\epsilon L_t, \mathcal{L}_{q_t(z'),z}, \|\cdot\|)/\delta\}}{3n} \right\} + o_p(n^{-1}).$$

where $L_t = KB_r(T-t-1)\Lambda$ for an absolute constant K.

Finally, putting the above together with Lemma 6.5.2, our sample complexity bound states that the policy suboptimality is on the order of $O(n^{-\frac{1}{2}})$. Note that this analysis omits estimation error in π^b for simplicity. Note that Lemma C.3.3 of the appendix gives that $N_{[]}\left(2\epsilon L, \mathcal{L}_{q(z'),z}, \|\cdot\|\right) \leq N(\epsilon, \mathcal{Q}\times\mathcal{Z}, \|\cdot\|) \leq N(\epsilon, \mathcal{Q}, \|\cdot\|)N(\epsilon, \mathcal{Z}, \|\cdot\|)$. Therefore ensuring some $\epsilon = cn^{-\frac{1}{2}}$ approximation error (for some arbitrary constant c) can be achieved by fixing $\epsilon' = \frac{\epsilon}{2L}$; i.e. we require finer approximation.

Proof sketch. As appears elsewhere in the analysis of FQI [86], we may obtain the following standard decomposition:

$$\|\hat{\overline{Q}}_{t,\hat{Z}_t} - \overline{\mathcal{T}}^*_{t,\hat{Z}_t}\hat{\overline{Q}}_{t+1}\|^2_{\mu_t} = \mathbb{E}_\mu[\ell(\hat{\overline{Q}}_{t,\hat{Z}_t}, \hat{\overline{Q}}_{t+1}; \hat{Z}_t)] - \mathbb{E}_\mu[\ell(\overline{Q}^\dagger_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] + \|\overline{Q}^\dagger_{t,Z_t} - \overline{\mathcal{T}}^*_t\hat{\overline{Q}}_{t+1}\|^2_{\mu_t}$$

where $\overline{Q}^\dagger_{t,Z_t}$ is the oracle squared loss minimizer, relative to the $\hat{\overline{Q}}_{t+1}$ output from the algorithm. Assumption 9 (completeness) bounds the last term. Our analysis differs onwards with additional decomposition relative to estimated nuisances and applying orthogonality from Proposition 6.5.1.

Finally, we note that our analysis extends immediately to the infinite-horizon case. Crucially, the (s,a)-rectangular uncertainty set admits a stationary worst-case distribution [139].

Bias-variance tradeoff in selection of Λ

We can quantify the dependence of the sample complexity on constants related to problem structure. We consider an equivalent regression target which better illustrates this dependence.

Corollary 6.5.4. *Assume that the same function classes $\mathcal{Q}, \mathcal{Z}$ are used for every timestep, and they are VC-subgraph with dimensions v_q, v_z. Assume that $\epsilon_{\mathcal{Q},\mathcal{Z}} = 0$. Then there exist absolute constants K, k such that*

$$\begin{aligned}\mathcal{E}(\hat{Q}) \leq K\{&\log(v_q + v_z)\\ &+ 2(v_q + v_z)\\ &+ 2((v_q + v_z) - 1)(T-1)\left(\log\left(2KB_r\Lambda(T-1)n/\epsilon\right) - 1\right)\}n^{-1}\\ &+o_p(n^{-1}).\end{aligned}$$

Note that the width of confidence bounds on the robust Q function scale logarithmically in Λ, which illustrates *robustness-variance-sharpness* tradeoffs. Namely, as we increase Λ, we estimate more extremal tail regions, which is more difficult. Sharper tail bounds on conditional expected shortfall estimation would also qualitatively yield similar insights.

Confounding with Infinite Data

While Theorem 6.5.3 analyses the difficulty of estimating the robust value function, here we analyze how the true robust value function differs from the nominal value function at the

population-level for policy evaluation (not optimization). This gives a sense of how potentially conservative the method is, in case unconfoundedness held after all. We consider a simplified linear Gaussian setting.

Proposition 6.5.5. *Let $\mathcal{S} = \mathbb{R}$ and $\mathcal{A} = \{0, 1\}$. Define parameters $\theta_P, \theta_R, \sigma_P \in \mathbb{R}$. Suppose in the observational distribution that $S_{t+1}|S_t, A_t \sim \mathcal{N}(\theta_P S_t, \sigma_P)$, $R(s, a, s') = \theta_R s'$, $\pi_t^e(1|S_t) = 0.5$, and consider some π^b such that $\pi_t^b(A_t|S_t)$ does not vary with S_t. Finally, let $\beta_i := \theta_R \sum_{k=1}^{i} \theta_P^k$ and notice that the nominal, non-robust value functions are $V_{T-i}^{\pi^e}(s) = \beta_i s$ for $i \geq 1$. Then:*

$$|V_0^{\pi^e}(s) - \bar{V}_0^{\pi^e}(s)| \leq (16\theta_P)^{-1}(\textstyle\sum_{i=0}^{T-1} \beta_i)\sigma_P \log(\Lambda).$$

Note that the cost of robustness gets worse as the horizon T increases, depending on the value of θ_P. The parameter θ_P is the autoregressive coefficient for the state transitions — it controls how strongly last period's state impacts this period's state. In the language of linear systems, θ_P will determine whether or not the system is stable. Each of the stability regimes — stable, marginally stable, and unstable — results in different scaling with T for the cost of robustness. For $|\theta_P| < 1$, the term $(\sum_{i=0}^{T-1} \beta_i)/\theta_P$ is asymptotically linear in T; for $|\theta_P| = 1$, the term is quadratic in T; and for $|\theta_P| > 1$, the term scales asymptotically as θ_P^T. In other words, for *stable* systems, unobserved confounding can at worst induce bias that is linear in horizon, but for *unstable* systems, the bias could increase exponentially. In contrast, for the unconfounded problem, unstable systems are typically easier to estimate due to their better signal-to-noise ratio [277]. While this example involves a scalar state for simplicity, we can straightforwardly generalize Proposition 6.5.5 to higher dimensions where the bias will depend on the spectrum of the transition matrix.

On the other hand, the scaling with the *degree* of confounding Λ is *independent of horizon*, and has a modest $\log(\Lambda)$ rate. This is surprising: it suggests that the horizon of the problem presents more of a challenge than the strength of confounding at each time step, and that T and Λ do not interact at the population level — at least in a simple linear-Gaussian setting. Characterizing exactly when the scaling with Λ is horizon-independent is a promising direction for future work.

6.6 Experiments

We first illustrate the benefits of our orthogonalized fitted-Q-iteration in a simulated example, where we know the ground-truth outcomes. Next, we illustrate how the robust fitted-Q-iteration allows robust evaluation of policies learned with methods similar to those used in the literature, and learning robust policies, revisiting the example of sepsis data from MIMIC-III since it has been widely studied in the literature.

Simulation

In this section, we validate the performance of our estimator, including its scaling with the sensitivity parameter Λ and the importance of orthogonalization. Note that our goal is not to evaluate the utility of the marginal sensitivity model itself — we leave that to the existing empirical literature in medicine and social science. Instead, we demonstrate that our robust FQI procedure can successfully solve the MSM, validating our theoretical analysis. We perform simulation experiments in a mis-specified sparse linear setting with heteroskedastic conditional variance. Previous methods for sensitivity analysis in RL, [214, 158, 44], *cannot* solve this continuous state setting with confounding at every time step. We use the following (marginal) data-generating process for the observational data:

$$\mathcal{S} \subset \mathbb{R}^d, \mathcal{A} = \{0, 1\}, S_0 \sim \mathcal{N}(0, 0.01), \qquad \pi^b(1|S_t) = 0.5, \ \forall S_t$$
$$P_{\text{obs}}(S_{t+1}|S_t, A_t) = \mathcal{N}(\theta_\mu S_t + \theta_A a, \max\{\theta_\sigma S_t + \sigma, 0\}), \qquad R(S_t, A_t, S_{t+1}) = \theta_R^T S_{t+1}$$

with parameters $\theta_\mu, \theta_\sigma \in \mathbb{R}^{d\times d}, \theta_R, \theta_A \in \mathbb{R}^d, \sigma \in \mathbb{R}$ chosen such that $AS_t + \sigma > 0$ with probability vanishingly close to 1. The number of features $d = 25$ and θ_μ and θ_σ are chosen to be column-wise sparse, with 5 and 20 non-zero columns respectively. We collect a dataset of size $n = 5000$ from a single trajectory. We then repeat this experiment in a higher-dimensional setting with $d = 100$ and $n = 600$ — the d/n ratio is 300 times worse.

We estimate $\bar{V}_1^*(s)$ for $T = 4$ and several different values of Λ, using both the orthogonalized and non-orthogonalized robust losses. For function approximation of the conditional mean and conditional quantile, we use Lasso regression. Note that while this is correctly specified in the non-robust setting, the CVaR is *non-linear* in the observed state due to the non-linear conditional standard deviation of $\theta_R^T S_{t+1}$, and therefore the Lasso is a misspecified model for the quantile and robust value functions. For details see Appendix C.4 in the Appendix.

We report the mean-squared error (MSE) of the value function estimate over 100 trials, alongside the average ℓ_2-norm parameter error and the percentage of the time a wrong action is taken. The MSE and percentage of mistakes compare the estimated value function/policy to an analytic ground truth and are evaluated on an independently drawn and identically distributed holdout sample of size $n = 200,000$ drawn from the initial state distribution. See the Appendix for details on the ground truth derivation.

The low-dimensional results in Table 6.1 illustrate two important phenomena. First, the MSE increases with Λ. While in practice, we would like to certify robustness for higher levels of Λ, the estimated lower bounds become less reliable. Second, the non-orthogonal algorithm suffers from substantially worse mean-squared error and as a result selects a sub-optimal action more often, especially at high levels of Λ. Orthogonalization has a very large impact not just in theory, but in practice.

The results for the high-dimensional setting are in Table 6.2. In this setting, policy optimization is *substantially* harder — even the nominal policy estimate only picks the true optimal action 72% of the time. However, we still see almost identical behavior as in the low-dimensional setting when comparing the orthogonal and non-orthogonal estimators. Without orthogonalization, performance drops off dramatically as Λ increases, such that for

Λ	Algorithm	MSE($\tilde{V}_0^*$)	ℓ_2 Parameter Error	% wrong action
1	FQI	0.2927	2.506	0%
2	Non-Orthogonal	0.6916	3.458	5e-5%
	Orthogonal	0.4119	2.678	0%
5.25	Non-Orthogonal	10.87	7.263	0.39%
	Orthogonal	0.5552	3.110	0%
8.5	Non-Orthogonal	50.72	17.32	2.5%
	Orthogonal	0.7113	3.410	4e-5%
11.75	Non-Orthogonal	171.1	33.80	5.4%
	Orthogonal	1.336	3.666	6e-4%
15	Non-Orthogonal	432.9	55.86	8.2%
	Orthogonal	2.687	3.931	4e-3%

Table 6.1: Simulation results with $d = 25$ and $n = 5000$, reporting the value function MSE, Q function parameter error, and the portion of the time a sub-optimal action is taken. The results compare non-orthogonal and orthogonal confounding robust FQI over five values of Λ.

Λ	Algorithm	MSE($\tilde{V}_0^*$)	ℓ_2 Parameter Error	% wrong action
1	FQI	0.2300	3.399	28%
2	Non-Orthogonal	0.5496	4.057	31%
	Orthogonal	0.5271	3.522	28%
5.25	Non-Orthogonal	3.160	11.51	43%
	Orthogonal	1.739	3.949	31%
8.5	Non-Orthogonal	7.683	24.04	45%
	Orthogonal	2.723	3.921	31%
11.75	Non-Orthogonal	15.22	48.89	47%
	Orthogonal	3.397	3.725	31%
15	Non-Orthogonal	30.21	88.02	48%
	Orthogonal	3.848	3.462	30%

Table 6.2: Simulation results with $d = 100$ and $n = 600$, reporting the value function MSE, Q function parameter error, and the portion of the time a sub-optimal action is taken. The results compare non-orthogonal and orthogonal confounding robust FQI over five values of Λ.

$\Lambda = 15$, the policy is only slightly better than random choice. Our orthogonalized algorithm has MSE that decays more gracefully with Λ, and picks the correct action at essentially the same rate as the nominal algorithm, even as Λ increases.

Note that these simulation results validate our algorithm for *estimating* the worst-case value function and robust policy. They do not assess how quickly the *ground-truth* population robust value function decays with Λ. See Section 6.5 above for an initial discussion.

Complex real-world healthcare data

In the next computational experiments, we show how our method extends to more complex real-world healthcare data via a case study around the use of MIMIC-III data for off-policy evaluation of learned policies for the management of sepsis in the ICU with fluids and vasopressors [185]. Sepsis is an umbrella term for an extreme response to infection and is a leading cause of mortality, healthcare costs, and readmission. Still, the management of sepsis is complex and there remains substantial uncertainty about clinical guidelines [89]. Practitioners recommend dynamic changes in treatment, i.e. tracking the patient's state over time. For example, giving IV fluids is expected to be beneficial at the very beginning, but there are also expected risks from too much [321].The pioneering efforts in releasing the MIMIC-III database enabled the development of Markov decision process models via model-based approaches or offline reinforcement learning methods [193, 239, 238, 195, 256]. However, a crucial challenge is *off-policy evaluation* for credible, data-driven estimates of the benefits of these learned policies, that are less vulnerable to model assumptions.

Crucial assumptions such as *unconfoundedness* are likely violated in this setting: treatment decisions probably included additional information not recorded in the database. (Indeed, the clinical literature certainly discusses other aspects of patient state and potential actions not included in the data). On the other hand, the comprehensive electronic health record (EHR) contains the most important factors in clinical decision-making such as patient vitals. So, our methods that develop *robust bounds* for off-policy evaluation of complex sequential policies can be applicable here, in highlighting the sensitivity of current learned policies to potential violations of sequential unconfoundedness. Since many research works used fitted-Q-iteration, we compare confounding-robust policies vs. naive policies for prescriptive insights.

We now describe the specific MDP data primitives. Following the data preprocessing of [170] and cohort definition of [177], the data covers an observation period of 72 hours past the onset of sepsis. Observed actions, administration of fluids or vaso-pressors, were categorized by volume and segmented into quantiles per each action type based on observational frequency. This leads to 25 possible discrete actions. Demographic and contextual features include age, gender, weight, ventilation and re-admission status. Other time-varying features include patient information such as blood pressure, heart rate, INR, various blood cell counts, respiratory rate, and different measures of oxygen levels (see [170, Table 2] for exact description). The reward function takes on three values: $R = \{-1, 0, +1\}$ where -1 indicates patient death, $+1$ indicates leaving the hospital; and 0 for all other events.

Fitted-Q Iteration with Gradient Boosting

For this case study, we perform flexible non-parametric regression using gradient-boosted trees in place of the simple linear models in our earlier simulations [98, 125]. Features include the full state vector and indicators for each action.

We begin with nominal (non-robust) estimation using standard fitted-Q iteration with gradient-boosted regression as our approximating function class. Implementing the robust

estimator for MSM parameter Λ requires only a few simple modifications of nominal FQI with off-the-shelf tools. First, we estimate the behavior policy π^b using a gradient-boosted classifier. Then within the FQI loop, we estimate a conditional quantile model using gradient-boosted regression with the quantile loss, which is supported natively in the `scikit-learn` package. Finally, we use the estimated quantiles to compute the orthogonalized pseudooutcomes, and fit a model for the Q function with gradient-boosted regression. We compute the value functions and optimal policies for a time horizon up to $T = 11$.

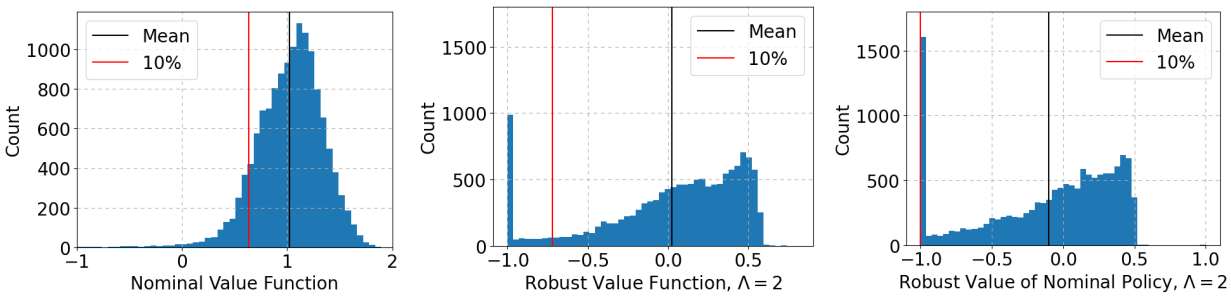
MIMIC Results

This case study is not meant to be a medical analysis, but concretely illustrates why caution is needed for interpreting offline RL applied to healthcare settings. In Figure 6.1a, we plot the distribution of the initial state value function, $V_0(s)$, with horizon $T = 11$ from non-robust FQI over the initial states in our dataset. The expected outcome under the nominal optimal policy is strongly positive for the majority of the population, including the 10% quantile.

By contrast, we plot the value function for the robust optimal value function (with $\Lambda = 2$) in Figure 6.1b. By construction, the robust value estimates are far more pessimistic. The *average* value of the robust optimal policy is still greater than zero, with a fairly substantial mass around $+0.5$. However, there is also a large negative tail with a strongly negative 10% quantile. We have truncated the plot at -1.0, which represents death, and notice that there are nearly 1000 starting states with value function ≤ -1.0. The more pessimistic outlook of the robust optimal value function represents the fact that some of the positive outcomes in the historical data could be due to spurious correlations with unobservables instead of a causal effect of the observed treatment.

We can also perform robust policy evaluation on the nominal optimal policy. We plot the corresponding value function over the initial states in Figure 6.1c. First, note that the expected robust value of the nominal optimal policy is actually negative. In other words, given only a modestly strong unobserved confounder ($\Lambda = 2$), it's possible that the nominal optimal policy *does more harm than good*. Furthermore, the number of initial states whose value is ≤ -1.0 has grown from about 1000 to about 1600, which now subsumes the 10% quantile. So under robust evaluation, not only does the nominal optimal policy have a slightly negative expected value for this distribution of patients, but it also substantially worsens the tail risk of death.

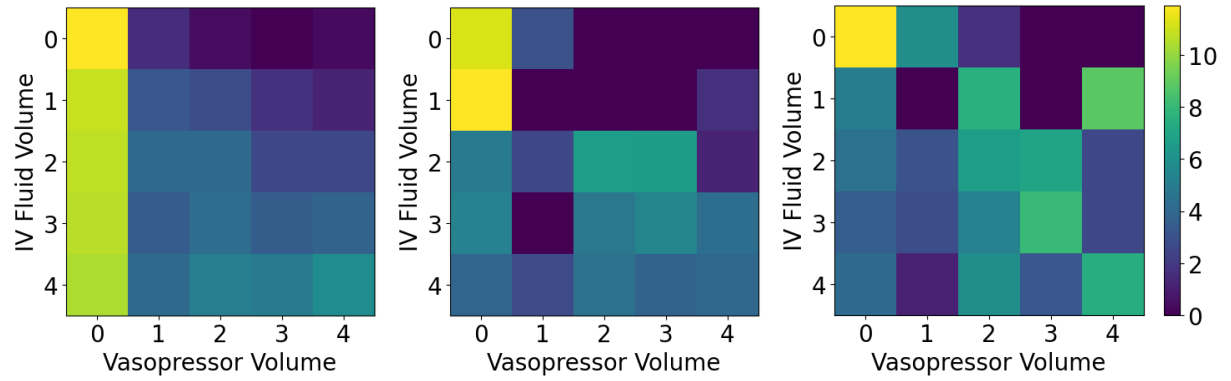
Beyond the value function, we also explore at a high level how robustness changes the actions suggested by the optimal policy. In Figure 6.2, we compare the counts of actions taken in the historical data with the optimal actions from the nominal and robust policy. Figure 6.2a shows log counts of the historical actions, which include a large number of patients with no treatment, many patients being treated with fluid but not vasopressors, and then a smaller number of patients receiving a variety of vasopressor intensities. The nominal optimal policy falls roughly the same pattern but made sharper; most patients are given either no treatment or the lowest level of IV fluid. Of the others, the majority are given a medium or large volume of both fluid and vasopressors. In contrast, the robust optimal policy makes two



(a) (b) (c)

Figure 6.1: Histograms of initial state value functions over the observed initial states in the MIMIC-III dataset. From left to right, the nominal value; the robust value for $\Lambda = 2$; and the robust value of the nominal optimal policy for $\Lambda = 2$. Each histogram includes a solid vertical line for the mean and the 10% quantile.

key changes: there are more patients assigned to no treatment at all, but also more patients assigned to higher levels of vasopressors.



(a) Historical (b) Nominal Policy (c) Robust Policy, $\Lambda = 2$

Figure 6.2: Log of one plus counts of actions in the MIMIC-III dataset. The left panel plots the log counts of the actual actions observed, while the middle and right panels plot the log counts of the nominal and robust policy actions, respectively, given the observed states.

Finally, in Figure 6.3 we plot how the robust optimal actions change as the sensitivity parameter Λ is increased. At the far left, we have $\Lambda = 1$, which corresponds to the nominal policy, where a substantial fraction of patients are assigned to receiving only IV fluid. As Λ increases, the number of untreated increases dramatically, while the number treated with only fluid drops. At the same time, the number treated with both vasopressors and fluids increases by over ten times from $\Lambda = 1$ to $\Lambda = 2.5$. Note that we end the plot at $\Lambda = 2.5$.

We find that at higher values — even $\Lambda = 3$ — the robust value is mostly negative, with a large mass below -1.0. This reflects the fact that off-policy evaluation of the MIMIC-III data is highly sensitive to unobserved variables.

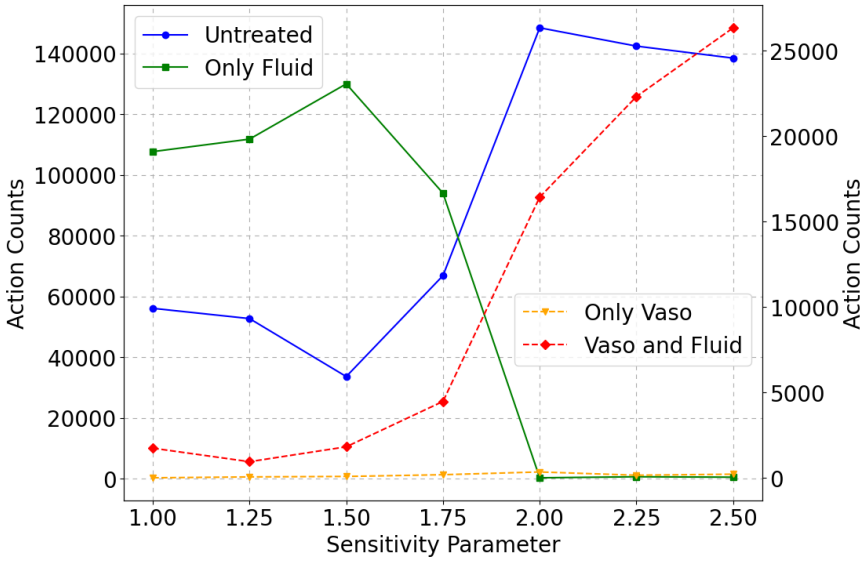


Figure 6.3: Counts of actions taken by the robust optimal policy over the states seen in the observed data as a function of the sensitivity parameter Λ. We combine the actions into four coarse groups: no treatment, only IV fluid, only vasopressors, and both fluid and vasopressors.

6.7 Conclusion

We developed a robust fitted-Q-iteration algorithm under memoryless unobserved confounders, leveraging function approximation, conditional quantiles, and orthogonalization. Importantly, our algorithm can be implemented using only off-the-shelf tools by changing only a few lines of code of standard FQI, making it easily accessible to practitioners. We derived sample complexity guarantees, demonstrated the effectiveness of our algorithm and the benefits of orthogonality in simulation experiments, and then provided a case-study with complex real-world healthcare data. Finally, we showed how to use our robust bounds to warm-start online reinforcement learning, demonstrating substantial performance benefits, whereas naive use of the offline data for warm-starting can actually hurt performance. Interesting directions for future work include falsifiability-based analyses to draw on competing identification proposals, extending to other models, model-selection procedures for the conditional quantile and mean models, and a formal theoretical analysis of warm-starting with our robust bounds.

Chapter 7

Conclusion and Looking Forward

In this dissertation, I considered roles for machine learning in macroeconomics. I began with a case study on a pure prediction problem in Icelandic tax data, and then considered the use of machine learning in observational causal inference — first, with all relevant variables observed, and then with potentially unobserved confounders. While this dissertation is largely technical in nature, much of the work has had direct applications in my applied macro research. To conclude, I will give an overview of some successful applications, and briefly discuss promising future work.

Broadly speaking, there are two broad avenues for applications of my work in macroeconomics. The first is "direct" causal inference applications in empirical macroeconomics. One promising example, is to estimate the marginal propensity to consume (MPC) from income and expenditure data, ala [37, 104]. From Chapter 2, we have some evidence that non-linearity is relevant for the bottom of the income distribution, and that that non-linearity is appropriately modelled by gradient boosted tree methods. The next step would be to estimate a nonparametric scalar estimand for the MPC, such as the average derivative effect. Methodology for this setting follows directly from Chapter 4, by using the Riesz representer corresponding to the average derivative. This is a good starting point, but note that for MPC estimation, one usually pairs an estimate of the average derivative effect with an instrumental variable (IV) for identification. Using machine learning for nonparametric IV regression requires more work than the presentation in Chapter 4, and I have results in this direction forthcoming. Another "direct" causal inference problem with a very similar structure is monetary policy impulse response estimation using "local projections" [148]. In this application, the equivalence results for linear regression in Chapter 4 are particularly relevant, because with them we can rewrite local projections as a weighting estimator. In doing so, we can more clearly connect common macro time series estimators with explicit potential outcomes and selection mechanisms, in the spirit of the project laid out in [241]. Likewise, our CVAR-based sensitivity analysis from Chapter 6 can be applied to these impulse responses.

The second broad way to apply the work in this dissertation is to the development of economic theory. An individual use of observational causal inference as outlined in this

dissertation can only go so far in the study of macro policy. Very often we do not have data for the *specific* intervention we are interested in, but instead have to synthesize evidence from many related interventions. We do this by building a structural/theoretical model of the underlying mechanisms. Perhaps surprisingly, the methodology in this dissertation is quite useful at helping to build these theoretical models. Chapter 2 for example describes a prediction methodology that estimates conditional expectations to construct income shocks. Summary statistics of these conditional expectations and shocks can be used to discipline income processes that are fed into theoretical models. Similarly, via the Riesz representer, other moments used in the calibration of structural models can be estimated using machine learning together with the techniques in Chapter 3 and 4. In forthcoming work, I apply this to the classic "age-time-cohort" problem for income data. In future work, I would like to study how the choice and construction of these data moments might enable predictive guarantees for new interventions and new settings. This is the ultimate goal: to be able to use the interventions we have observed to design new policies, and there is no way to do this without understanding the underlying mechanisms.

Bibliography

[1] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program". In: *Journal of the American statistical Association* 105.490 (2010), pp. 493–505.

[2] Rediet Abebe, Jon Kleinberg, and S Matthew Weinberg. "Subsidy allocations in the presence of income shocks". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 7032–7039.

[3] John M Abowd and David Card. *On the covariance structure of earnings and hours changes*. 1986.

[4] Krishna Acharya et al. "Wealth dynamics over generations: Analysis and interventions". In: *arXiv preprint arXiv:2209.07375* (2022).

[5] Abhineet Agarwal et al. "Hierarchical Shrinkage: Improving the accuracy and interpretability of tree-based models." In: *Proceedings of the 39th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022, pp. 111–135. URL: `https://proceedings.mlr.press/v162/agarwal22b.html`.

[6] Anish Agarwal and Rahul Singh. "Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy". In: *arXiv preprint arXiv:2107.02780* (2021).

[7] Rohit Agrawal and Thibaut Horel. "Optimal Bounds between f-Divergences and Integral Probability Metrics". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 115–124.

[8] Joshua D Angrist and Jörn-Steffen Pischke. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics". In: *Journal of economic perspectives* 24.2 (2010), pp. 3–30.

[9] David Arbour, Drew Dimmery, and Arjun Sondhi. "Permutation weighting". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 331–341.

[10] Manuel Arellano, Richard Blundell, and Stéphane Bonhomme. "Earnings and consumption dynamics: a nonlinear panel data framework". In: *Econometrica* 85.3 (2017), pp. 693–734.

[11] Manuel Arellano et al. *Heterogeneity of Consumption Responses to Income Shocks in the Presence of Nonlinear Persistence*. Tech. rep. Mimeo, University of Chicago, 2021.

[12] Dmitry Arkhangelsky et al. "Synthetic difference-in-differences". In: *American Economic Review* 111.12 (2021), pp. 4088–4118.

[13] Peter M Aronow and Donald KK Lee. "Interval estimation of population means under unknown but bounded probabilities of sample selection". In: *Biometrika* 100.1 (2013), pp. 235–240.

[14] Hero Ashman and Seth Neumuller. "Can income differences explain the racial wealth gap? A quantitative analysis". In: *Review of Economic Dynamics* 35 (2020), pp. 220–239.

[15] Serge Assaad et al. "Counterfactual Representation Learning with Balancing Weights". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 13–15 Apr 2021, pp. 1972–1980. URL: `https://proceedings.mlr.press/v130/assaad21a.html`.

[16] Susan Athey and Guido W Imbens. "Machine learning methods for estimating heterogeneous causal effects". In: *stat* 1050.5 (2015), pp. 1–26.

[17] Susan Athey, Guido W Imbens, and Stefan Wager. "Approximate residual balancing: debiased inference of average treatment effects in high dimensions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.4 (2018), pp. 597–623.

[18] Adrien Auclert and Matthew Rognlie. *Inequality and aggregate demand*. Tech. rep. National Bureau of Economic Research, 2018.

[19] Philipp Bach et al. "Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study". In: *arXiv preprint arXiv:2402.04674* (2024).

[20] Scott R Baker. "Debt and the response to household income shocks: Validation and application of linked financial account data". In: *Journal of Political Economy* 126.4 (2018), pp. 1504–1557.

[21] Peter L Bartlett et al. "Benign overfitting in linear regression". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070.

[22] Eric Bauer and Ron Kohavi. "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants". In: *Machine learning* 36 (1999), pp. 105–139.

[23] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. "On regularization algorithms in learning theory". In: *Journal of complexity* 23.1 (2007), pp. 52–72.

[24] Bahram Behzadian, Marek Petrik, and Chin Pang Ho. "Fast Algorithms for $L\infty$-constrained S-rectangular Robust MDPs". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25982–25992.

[25] Alexandre Belloni and Victor Chernozhukov. "l1-penalized quantile regression in high-dimensional sparse models". In: *The Annals of Statistics* 39.1 (2011), pp. 82–130.

[26] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. "Inference on treatment effects after selection among high-dimensional controls". In: *The Review of Economic Studies* 81.2 (2014), pp. 608–650.

[27] Eli Ben-Michael, Avi Feller, and Erin Hartman. "Multilevel calibration weighting for survey data". In: *arXiv preprint arXiv:2102.09052* (2021).

[28] Eli Ben-Michael, Avi Feller, and Jesse Rothstein. "The augmented synthetic control method". In: *Journal of the American Statistical Association* 116.536 (2021), pp. 1789–1803.

[29] Eli Ben-Michael et al. "The Balancing Act in Causal Inference". In: (2021).

[30] Eli Ben-Michael et al. "The Balancing Act in Causal Inference". In: *arXiv preprint arXiv:2110.14831* (2021).

[31] David Benkeser and Mark Van Der Laan. "The highly adaptive lasso estimator". In: *2016 IEEE international conference on data science and advanced analytics (DSAA)*. IEEE. 2016, pp. 689–696.

[32] Andrew Bennett and Nathan Kallus. "Policy Evaluation with Latent Confounders via Optimal Balance". In: *Advances in neural information processing systems* 32 (2019).

[33] Andrew Bennett et al. "Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1999–2007.

[34] Aurélien F Bibaut et al. "More efficient off-policy evaluation through regularized targeted learning". In: *arXiv preprint arXiv:1912.06292* (2019).

[35] Jeremiah Birrell, Markos A Katsoulakis, and Yannis Pantazis. "Optimizing variational representations of divergences and accelerating their statistical estimation". In: *arXiv preprint arXiv:2006.08781* (2020).

[36] Jeremiah Birrell et al. "(f, Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics". In: *arXiv preprint arXiv:2011.05953* (2020).

[37] Richard Blundell, Luigi Pistaferri, and Ian Preston. "Consumption inequality and partial insurance". In: *American Economic Review* 98.5 (2008), pp. 1887–1921.

[38] Matteo Bonvini and Edward H Kennedy. "Sensitivity analysis via the proportion of unmeasured confounding". In: *Journal of the American Statistical Association* (2021), pp. 1–11.

[39] Matteo Bonvini et al. "Sensitivity Analysis for Marginal Structural Models". In: *arXiv preprint arXiv:2210.04681* (2022).

[40] Michel Broniatowski and Amor Keziou. "Minimization of φ-divergences on sets of signed measures". In: *Studia Scientiarum Mathematicarum Hungarica* 43.4 (2006), pp. 403–442.

[41] David Bruns-Smith, Avi Feller, and Emi Nakamura. "Using Supervised Learning to Estimate Inequality in the Size and Persistence of Income Shocks". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 1747–1756.

[42] David Bruns-Smith and Angela Zhou. "Robust fitted-q-evaluation and iteration under sequentially exogenous unobserved confounders". In: *arXiv preprint arXiv:2302.00662* (2023).

[43] David Bruns-Smith et al. "Augmented balancing weights as linear regression". In: *arXiv preprint arXiv:2304.14545* (2023).

[44] David A Bruns-Smith. "Model-Free and Model-Based Policy Evaluation when Causality is Uncertain". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 1116–1126.

[45] David A Bruns-Smith and Avi Feller. "Outcome Assumptions and Duality Theory for Balancing Weights". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 11037–11055.

[46] Peter Bühlmann and Bin Yu. "Boosting with the L 2 loss: regression and classification". In: *Journal of the American Statistical Association* 98.462 (2003), pp. 324–339.

[47] Zongwu Cai and Xian Wang. "Nonparametric estimation of conditional VaR and expected shortfall". In: *Journal of Econometrics* 147.1 (2008), pp. 120–130.

[48] Andrea Caponnetto and Ernesto De Vito. "Optimal rates for the regularized least-squares algorithm". In: *Foundations of Computational Mathematics* 7 (2007), pp. 331–368.

[49] Christopher D Carroll. "A theory of the consumption function, with and without liquidity constraints". In: *Journal of Economic perspectives* 15.3 (2001), pp. 23–45.

[50] Christopher D Carroll. "How does future income affect current consumption?" In: *The Quarterly Journal of Economics* 109.1 (1994), pp. 111–147.

[51] Ambarish Chattopadhyay, Christopher H Hase, and José R Zubizarreta. "Balancing vs modeling approaches to weighting in practice". In: *Statistics in Medicine* 39.24 (2020), pp. 3227–3254.

[52] Ambarish Chattopadhyay and Jose R Zubizarreta. "On the implied weights of linear regression for causal inference". In: *arXiv preprint arXiv:2104.06581* (2021).

[53] Qizhao Chen, Vasilis Syrgkanis, and Morgane Austern. "Debiased Machine Learning without Sample-Splitting for Stable Estimators". In: *arXiv preprint arXiv:2206.01825* (2022).

[54] Shuxiao Chen and Bo Zhang. "Estimating and improving dynamic treatment regimes with a time-varying instrumental variable". In: *arXiv preprint arXiv:2104.07822* (2021).

[55] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. "Automatic debiased machine learning of causal and structural effects". In: *Econometrica* 90.3 (2022), pp. 967–1027.

[56] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. "Debiased machine learning of global and local parameters using regularized Riesz representers". In: *The Econometrics Journal* (2022).

[57] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. "Learning L2-continuous regression functionals via regularized riesz representers". In: *arXiv preprint arXiv: 1809.05224* 8 (2018).

[58] Victor Chernozhukov et al. "Automatic debiased machine learning for dynamic treatment effects and general nested functionals". In: *arXiv preprint arXiv:2203.13887* (2022).

[59] Victor Chernozhukov et al. *Automatic Debiased Machine Learning via Riesz Regression*. 2024. arXiv: `2104.14737 [math.ST]`.

[60] Victor Chernozhukov et al. "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1 (Jan. 2018), pp. C1–C68.

[61] Victor Chernozhukov et al. "Locally robust semiparametric estimation". In: *Econometrica* 90.4 (2022), pp. 1501–1535.

[62] Victor Chernozhukov et al. *Long story short: Omitted variable bias in causal machine learning*. Tech. rep. National Bureau of Economic Research, 2022.

[63] Victor Chernozhukov et al. "Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 3901–3914.

[64] Raj Chetty et al. "Is the United States still a land of opportunity? Recent trends in intergenerational mobility". In: *American Economic Review* 104.5 (2014), pp. 141–147.

[65] Yinlam Chow et al. "Risk-sensitive and robust decision-making: a cvar optimization approach". In: *Advances in Neural Information Processing Systems*. 2015, pp. 1522–1530.

[66] Dimitris Christelis et al. "Asymmetric consumption effects of transitory income shocks". In: *The Economic Journal* 129.622 (2019), pp. 2322–2341.

[67] Lawrence Christiano, Martin S Eichenbaum, and David Marshall. *The permanent income hypothesis revisited*. 1987.

[68] Miles Corak. "Income inequality, equality of opportunity, and intergenerational mobility". In: *Journal of Economic Perspectives* 27.3 (2013), pp. 79–102.

[69] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. "Learning Bounds for Importance Weighting." In: *Nips*. Vol. 10. Citeseer. 2010, pp. 442–450.

[70] Nicolas Courty, Rémi Flamary, and Devis Tuia. "Domain adaptation with regularized optimal transport". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2014, pp. 274–289.

[71] Alexander D'Amour et al. "Fairness is not static: deeper understanding of long term fairness via simulation studies". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 525–534.

[72] Alexander D'Amour et al. "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics* 221.2 (2021), pp. 644–654.

[73] Mariacristina De Nardi. "Wealth inequality and intergenerational links". In: *The Review of Economic Studies* 71.3 (2004), pp. 743–768.

[74] Mariacristina De Nardi and Giulio Fella. "Saving and wealth inequality". In: *Review of Economic Dynamics* 26 (2017), pp. 280–300.

[75] Mariacristina De Nardi, Giulio Fella, and Gonzalo Paz Pardo. *The implications of richer earnings dynamics for consumption and wealth*. Tech. rep. National Bureau of Economic Research, 2016.

[76] Angus Deaton and Christina Paxson. "Intertemporal choice and inequality". In: *Journal of political economy* 102.3 (1994), pp. 437–467.

[77] Rajeev H Dehejia and Sadek Wahba. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs". In: *Journal of the American statistical Association* 94.448 (1999), pp. 1053–1062.

[78] Erick Delage and Dan A Iancu. "Robust multistage decision making". In: *The operations research revolution*. INFORMS, 2015, pp. 20–46.

[79] Jean-Claude Deville and Carl-Erik Särndal. "Calibration estimators in survey sampling". In: *Journal of the American statistical Association* 87.418 (1992), pp. 376–382.

[80] Frances Ding et al. "Retiring adult: New datasets for fair machine learning". In: *Advances in neural information processing systems* 34 (2021), pp. 6478–6490.

[81] Peng Ding and Tyler J VanderWeele. "Sensitivity analysis without assumptions". In: *Epidemiology (Cambridge, Mass.)* 27.3 (2016), p. 368.

[82] Peng Ding et al. "Sensitivity Analysis for Unmeasured Confounding in Medical Product Development and Evaluation Using Real World Evidence". In: *arXiv preprint arXiv:2307.07442* (2023).

[83] Edgar Dobriban and Stefan Wager. "High-dimensional asymptotics of prediction: Ridge regression and classification". In: *The Annals of Statistics* 46.1 (2018), pp. 247–279.

[84] Jacob Dorn and Kevin Guo. "Sharp sensitivity analysis for inverse propensity weighting via quantile balancing". In: *Journal of the American Statistical Association* just-accepted (2022), pp. 1–28.

[85] Jacob Dorn, Kevin Guo, and Nathan Kallus. "Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding". In: *arXiv preprint arXiv:2112.11449* (2021).

[86] Yaqi Duan, Chi Jin, and Zhiyuan Li. "Risk bounds and rademacher complexity in batch reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2892–2902.

[87] Paul Dupuis and Yixiang Mao. "Formulation and properties of a divergence used to compare probability measures without absolute continuity". In: *arXiv preprint arXiv:1911.07422* (2019).

[88] Damien Ernst et al. "Clinical data based optimal STI strategies for HIV: a reinforcement learning approach". In: *Proceedings of the 45th IEEE Conference on Decision and Control*. IEEE. 2006, pp. 667–672.

[89] Laura Evans et al. "Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021". In: *Intensive care medicine* 47.11 (2021), pp. 1181–1247.

[90] Andreas Fagereng, Martin B Holm, and Gisle J Natvik. "MPC heterogeneity and household balance sheets". In: *American Economic Journal: Macroeconomics* 13.4 (2021), pp. 1–54.

[91] Max H Farrell. "Robust inference on average treatment effects with possibly more covariates than observations". In: *Journal of Econometrics* 189.1 (2015), pp. 1–23.

[92] FDA. *AIML_SaMD_Action_Plan*. `https://www.fda.gov/media/145022/download`. (Accessed on 09/06/2023). Jan. 2021.

[93] Jesús Fernández-Villaverde, Samuel Hurtado, and Galo Nuno. "Financial frictions and the wealth distribution". In: *Econometrica* 91.3 (2023), pp. 869–901.

[94] Simon Fischer and Ingo Steinwart. "Sobolev norm learning rates for regularized least-squares algorithms". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 8464–8501.

[95] Marjorie A Flavin. "The adjustment of consumption to changing expectations about future income". In: *Journal of political economy* 89.5 (1981), pp. 974–1009.

[96] Dylan J Foster and Vasilis Syrgkanis. "Orthogonal statistical learning". In: *arXiv preprint arXiv:1901.09036* (2019).

[97] AlexanderM Franks, Alexander D'Amour, and Avi Feller. "Flexible sensitivity analysis for observational studies without observable implications". In: *Journal of the American Statistical Association* (2019).

[98] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[99] Milton Friedman. "The permanent income hypothesis". In: *A theory of the consumption function*. Princeton University Press, 1957, pp. 20–37.

[100] Justin Fu et al. "Benchmarks for deep off-policy evaluation". In: *arXiv preprint arXiv:2103.16596* (2021).

[101] Zuyue Fu et al. "Offline reinforcement learning with instrumental variables in confounded markov decision processes". In: *arXiv preprint arXiv:2209.08666* (2022).

[102] Wayne A Fuller. "Regression estimation for survey samples". In: *Survey Methodology* 28.1 (2002), pp. 5–24.

[103] Yaroslav Ganin et al. "Domain-adversarial training of neural networks". In: *The journal of machine learning research* 17.1 (2016), pp. 2096–2030.

[104] Peter Ganong et al. *Wealth, race, and consumption smoothing of typical income shocks*. Tech. rep. National Bureau of Economic Research, 2020.

[105] Chenyin Gao, Shu Yang, and Jae Kwang Kim. "Soft calibration for selection bias problems under mixed-effects models". In: *arXiv preprint arXiv:2206.01084* (2022).

[106] Kristopher S Gerardi, Harvey S Rosen, and Paul S Willen. "The impact of deregulation and financial innovation on consumers: The case of the mortgage market". In: *The Journal of Finance* 65.1 (2010), pp. 333–360.

[107] AmirEmad Ghassami et al. "Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 7210–7239.

[108] Pierre Glaser, Michael Arbel, and Arthur Gretton. "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support". In: *arXiv preprint arXiv:2106.08929* (2021).

[109] Larry Goldstein and Karen Messer. "Optimal plug-in estimators for nonparametric functional estimation". In: *The annals of statistics* (1992), pp. 1306–1328.

[110] Omer Gottesman et al. "Combining parametric and nonparametric models for off-policy evaluation". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2366–2375.

[111] Omer Gottesman et al. "Guidelines for reinforcement learning in healthcare". In: *Nature medicine* 25.1 (2019), pp. 16–18.

[112] Philippe Goulet Coulombe et al. "How is machine learning useful for macroeconomic forecasting?" In: *Journal of Applied Econometrics* 37.5 (2022), pp. 920–964.

[113] Pierre-Olivier Gourinchas and Jonathan A Parker. "Consumption over the life cycle". In: *Econometrica* 70.1 (2002), pp. 47–89.

[114] Vineet Goyal and Julien Grand-Clement. "Robust Markov Decision Processes: Beyond Rectangularity". In: *Mathematics of Operations Research* (2022).

[115] Yves Grandvalet. "Least absolute shrinkage is equivalent to quadratic penalization". In: *International Conference on Artificial Neural Networks*. Springer. 1998, pp. 201–206.

[116] Arthur Gretton et al. "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[117] Arthur Gretton et al. "Covariate shift by kernel mean matching". In: *Dataset shift in machine learning* 3.4 (2009), p. 5.

[118] Fatih Guvenen et al. *What do data on millions of US workers reveal about life-cycle earnings risk?* Tech. rep. National Bureau of Economic Research, 2015.

[119] Chung Ha. "A noncompact minimax theorem". In: *Pacific Journal of Mathematics* 97.1 (1981), pp. 115–117.

[120] Jens Hainmueller. "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies". In: *Political analysis* (2012), pp. 25–46.

[121] Sukjin Han. "Optimal dynamic treatment regimes and partial welfare ordering". In: *Journal of American Statistical Association (just-accepted)* (2022).

[122] Josiah Hanna, Scott Niekum, and Peter Stone. "Importance sampling policy evaluation with an estimated behavior policy". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2605–2613.

[123] Christopher Harshaw et al. "Balancing covariates in randomized experiments with the Gram–Schmidt Walk design". In: *arXiv preprint arXiv:1911.03071* (2019).

[124] Trevor Hastie et al. "Surprises in high-dimensional ridgeless least squares interpolation". In: *Annals of statistics* 50.2 (2022), p. 949.

[125] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[126] Chad Hazlett. "KERNEL BALANCING". In: *Statistica Sinica* 30.3 (2020), pp. 1155–1189.

[127] Hoda Heidari and Jon Kleinberg. "Allocating opportunities in a dynamic model of intergenerational mobility". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 15–25.

[128] Miguel A Hernán and James M Robins. *Causal inference*. 2010.

[129] Jennifer L Hill. "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* 20.1 (2011), pp. 217–240.

[130] David A Hirshberg, Arian Maleki, and Jose R Zubizarreta. "Minimax linear estimation of the retargeted mean". In: *arXiv preprint arXiv:1901.10296* (2019).

[131] David A Hirshberg and Stefan Wager. "Augmented minimax linear estimation". In: *The Annals of Statistics* 49.6 (2021), pp. 3206–3227.

[132] Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. "Partial Policy Iteration for L1-Robust Markov Decision Processes." In: *J. Mach. Learn. Res.* 22 (2021), pp. 275–1.

[133] Jesse Y Hsu and Dylan S Small. "Calibrating sensitivity analyses to observed covariates in observational studies". In: *Biometrics* 69.4 (2013), pp. 803–811.

[134] R Glenn Hubbard, Jonathan Skinner, and Stephen P Zeldes. "Precautionary saving and social insurance". In: *Journal of political Economy* 103.2 (1995), pp. 360–399.

[135] Statistics Iceland. *Consumer price index*. Jan. 2023. URL: `https://www.statice.is/statistics/economy/prices/consumer-price-index/`.

[136] Kosuke Imai and Marc Ratkovic. "Covariate balancing propensity score". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1 (2014), pp. 243–263.

[137] Ehsan Imani, Eric Graves, and Martha White. "An off-policy policy gradient theorem using emphatic weightings". In: *Advances in Neural Information Processing Systems* 31 (2018).

[138] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[139] Garud N Iyengar. "Robust dynamic programming". In: *Mathematics of Operations Research* 30.2 (2005), pp. 257–280.

[140] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31 (2018).

[141] Tullio Jappelli and Luigi Pistaferri. "Fiscal policy and MPC heterogeneity". In: *American Economic Journal: Macroeconomics* 6.4 (2014), pp. 107–136.

[142] Sookyo Jeong and Hongseok Namkoong. "Assessing External Validity Over Worst-case Subpopulations". In: *arXiv preprint arXiv:2007.02411* (2020).

[143] Andrew Jesson et al. "Scalable sensitivity and uncertainty analysis for causal-effect estimates of continuous-valued interventions". In: *arXiv preprint arXiv:2204.10022* (2022).

[144] Nan Jiang and Lihong Li. "Doubly Robust off-policy Value Evaluation for Reinforcement Learning". In: *Proceedings of the 33rd International Conference on Machine Learning* (2016).

[145] Nan Jiang and Lihong Li. "Doubly robust off-policy value evaluation for reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 652–661.

[146] Ying Jin, Zhuoran Yang, and Zhaoran Wang. "Is pessimism provably efficient for offline rl?" In: *International Conference on Machine Learning*. PMLR. 2021, pp. 5084–5096.

[147] Fredrik D Johansson et al. "Generalization bounds and representation learning for estimation of potential outcomes and causal effects". In: *arXiv preprint arXiv:2001.07426* (2020).

[148] Òscar Jordà. "Local projections for applied economics". In: *Annual Review of Economics* 15.1 (2023), pp. 607–631.

[149] Michael I Jordan, Yixin Wang, and Angela Zhou. "Empirical Gateaux Derivatives for Causal Inference". In: *arXiv preprint arXiv:2208.13701* (2022).

[150] Jongbin Jung et al. "Algorithmic decision making in the presence of unmeasured confounding". In: *arXiv preprint arXiv:1805.01868* (2018).

[151] Nathan Kallus. "Deepmatch: Balancing deep covariate representations for causal inference using adversarial training". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5067–5077.

[152] Nathan Kallus. "Generalized optimal matching methods for causal inference." In: *Journal of Machine Learning Research* 21.62 (2020), pp. 1–54.

[153] Nathan Kallus, Xiaojie Mao, and Masatoshi Uehara. "Causal inference under unmeasured confounding with negative controls: A minimax learning approach". In: *arXiv preprint arXiv:2103.14029* (2021).

[154] Nathan Kallus, Xiaojie Mao, and Angela Zhou. "Interval estimation of individual-level causal effects under unobserved confounding". In: *arXiv preprint arXiv:1810.02894* (2018).

[155] Nathan Kallus, Xiaojie Mao, and Angela Zhou. "Interval estimation of individual-level causal effects under unobserved confounding". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 2281–2290.

[156] Nathan Kallus and Masatoshi Uehara. "Double reinforcement learning for efficient off-policy evaluation in markov decision processes". In: *Journal of Machine Learning Research* 21.167 (2020), pp. 1–63.

[157] Nathan Kallus and Masatoshi Uehara. "Statistically efficient off-policy policy gradients". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5089–5100.

[158] Nathan Kallus and Angela Zhou. "Confounding-robust policy evaluation in infinite-horizon reinforcement learning". In: *Advances in neural information processing systems* 33 (2020).

[159] Nathan Kallus and Angela Zhou. "Confounding-robust policy improvement". In: *arXiv preprint arXiv:1805.08593* (2018).

[160] Nathan Kallus and Angela Zhou. "Minimax-Optimal Policy Learning Under Unobserved Confounding". In: *Management Science* (2020).

[161] Nathan Kallus and Angela Zhou. "Stateful Offline Contextual Policy Evaluation and Learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 11169–11194.

[162] Joseph DY Kang, Joseph L Schafer, et al. "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data". In: *Statistical science* 22.4 (2007), pp. 523–539.

[163] Greg Kaplan and Giovanni L Violante. "A model of the consumption response to fiscal stimulus payments". In: *Econometrica* 82.4 (2014), pp. 1199–1239.

[164] Greg Kaplan, Giovanni L Violante, and Justin Weidner. *The wealthy hand-to-mouth*. Tech. rep. National Bureau of Economic Research, 2014.

[165] Kengo Kato. "Weighted nadaraya–watson estimation of conditional expected shortfall". In: *Journal of Financial Econometrics* 10.2 (2012), pp. 265–291.

[166] Edward H Kennedy. "Optimal doubly robust estimation of heterogeneous causal effects". In: *arXiv preprint arXiv:2004.14497* (2020).

[167] Edward H Kennedy. "Semiparametric doubly robust targeted double machine learning: a review". In: *arXiv preprint arXiv:2203.06469* (2022).

[168] Amor Keziou. "Dual representation of φ-divergences and applications". In: *Comptes rendus mathématique* 336.10 (2003), pp. 857–862.

[169] Shakeeb Khan and Elie Tamer. "Irregular identification, support conditions, and inverse weight estimation". In: *Econometrica* 78.6 (2010), pp. 2021–2042.

[170] Taylor W Killian et al. "An Empirical Study of Representation Learning for Reinforcement Learning in Healthcare". In: *arXiv preprint arXiv:2011.11235* (2020).

[171] Kwangho Kim, Bijan A. Niknam, and José R Zubizarreta. "Scalable kernel balancing weights in a nationwide observational study of hospital profit status and heart attack outcomes". 2022.

[172] Michael P Kim et al. "Universal adaptability: Target-independent inference that competes with propensity scoring". In: *Proceedings of the National Academy of Sciences* 119.4 (2022), e2108097119.

[173] Patrick Kline. "Oaxaca-Blinder as a reweighting estimator". In: *American Economic Review* 101.3 (2011), pp. 532–37.

[174] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. "The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 6863–6878.

[175] Roger Koenker and Kevin F Hallock. "Quantile regression". In: *Journal of economic perspectives* 15.4 (2001), pp. 143–156.

[176] Ronny Kohavi and Barry Becker. "Adult data set". In: *UCI machine learning repository* 5 (1996), p. 2093.

[177] Matthieu Komorowski et al. "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care". In: *Nature medicine* 24.11 (2018), pp. 1716–1720.

[178] Lorenz Kueng. "Excess sensitivity of high-income consumers". In: *The Quarterly Journal of Economics* 133.4 (2018), pp. 1693–1751.

[179] Jeongyeol Kwon et al. "RL for Latent MDPs: Regret Guarantees and a Lower Bound". In: *arXiv preprint arXiv:2102.04939* (2021).

[180] Mark J Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

[181] Eric B Laber et al. "Dynamic treatment regimes: Technical challenges and applications". In: *Electronic journal of statistics* 8.1 (2014), p. 1225.

[182] Sahin Lale et al. "Logarithmic regret bound in partially observable linear dynamical systems". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 20876–20888.

[183] Sahin Lale et al. "Model learning predictive control in nonlinear dynamical systems". In: *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE. 2021, pp. 757–762.

[184] Robert J LaLonde. "Evaluating the econometric evaluations of training programs with experimental data". In: *The American economic review* (1986), pp. 604–620.

[185] Howard Larkin. "Vasopressors or High-Volume IV Fluids Both Effective for Sepsis". In: *JAMA* 329.7 (2023), pp. 532–532.

[186] Hoang Le, Cameron Voloshin, and Yisong Yue. "Batch policy learning under constraints". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3703–3712.

[187] Greg Lewis and Vasilis Syrgkanis. "Double/debiased machine learning for dynamic treatment effects via g-estimation". In: *arXiv preprint arXiv:2002.07285* (2020).

[188] Luofeng Liao et al. "Instrumental variable value iteration for causal offline reinforcement learning". In: *arXiv preprint arXiv:2102.09907* (2021).

[189] Thomas Liao et al. "Are we learning yet? a meta review of evaluation failures across machine learning". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

[190] Rong Rong Lin, Hai Zhang Zhang, and Jun Zhang. "On reproducing kernel Banach spaces: Generic definitions and unified framework of constructions". In: *Acta Mathematica Sinica, English Series* 38.8 (2022), pp. 1459–1483.

[191] Zhexiao Lin and Fang Han. "On regression-adjusted imputation estimators of the average treatment effect". In: *arXiv preprint arXiv:2212.05424* (2022).

[192] Qiang Liu et al. "Breaking the curse of horizon: infinite-horizon off-policy estimation". In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, pp. 5361–5371.

[193] Siqi Liu et al. "Reinforcement learning for clinical decision support in critical care: comprehensive review". In: *Journal of medical Internet research* 22.7 (2020), e18477.

[194] Elita A Lobo, Mohammad Ghavamzadeh, and Marek Petrik. "Soft-Robust Algorithms for Batch Reinforcement Learning". In: *arXiv preprint arXiv:2011.14495* (2020).

[195] MingYu Lu et al. "Is deep reinforcement learning ready for practical applications in healthcare? A sensitivity analysis of duel-DDQN for hemodynamic management in sepsis patients". In: *AMIA Annual Symposium Proceedings*. Vol. 2020. American Medical Informatics Association. 2020, p. 773.

[196] Thomas Lumley, Pamela A Shaw, and James Y Dai. "Connections between survey calibration estimators and semiparametric models for incomplete data". In: *International Statistical Review* 79.2 (2011), pp. 200–220.

[197] Ian Lundberg, Rebecca Johnson, and Brandon M Stewart. "What is your estimand? Defining the target quantity connects statistical evidence to theory". In: *American Sociological Review* 86.3 (2021), pp. 532–565.

[198] Xiaoteng Ma et al. "Distributionally robust offline reinforcement learning with linear function approximation". In: *arXiv preprint arXiv:2209.06620* (2022).

[199] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. "Domain adaptation: Learning bounds and algorithms". In: *arXiv preprint arXiv:0902.3430* (2009).

[200] Richard A Meese and Kenneth Rogoff. "Empirical exchange rate models of the seventies: Do they fit out of sample?" In: *Journal of international economics* 14.1-2 (1983), pp. 3–24.

[201] Costas Meghir and Luigi Pistaferri. "Earnings, consumption and life cycle choices". In: *Handbook of labor economics*. Vol. 4. Elsevier, 2011, pp. 773–854.

[202] Costas Meghir and Luigi Pistaferri. "Income variance dynamics and heterogeneity". In: *Econometrica* 72.1 (2004), pp. 1–32.

[203] Nicolai Meinshausen. "Quantile Regression Forests". In: *Journal of Machine Learning Research* 7 (2006), pp. 983–999.

[204] Zakaria Mhammedi et al. "Learning the linear quadratic regulator from nonlinear observations". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 14532–14543.

[205] Rui Miao, Zhengling Qi, and Xiaoke Zhang. "Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models". In: *arXiv preprint arXiv:2209.10064* (2022).

[206] Luke W Miratrix, Stefan Wager, and Jose R Zubizarreta. "Shape-constrained partial identification of a population mean under unknown probabilities of sample selection". In: *Biometrika* 105.1 (2018), pp. 103–114.

[207] Niloofar Moosavi, Jenny Häggström, and Xavier de Luna. "The costs and benefits of uniformly valid causal inference with high-dimensional nuisance parameters". In: *Statistical Science* 38.1 (2023), pp. 1–12.

[208] Jonathan Morduch and Rachel Schneider. *The financial diaries: How American families cope in a world of uncertainty*. Princeton University Press, 2017.

[209] Wenlong Mou et al. *Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency*. 2023. DOI: 10.48550/ARXIV.2301.06240. URL: https://arxiv.org/abs/2301.06240.

[210] Susan A Murphy. "Optimal dynamic treatment regimes". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2 (2003), pp. 331–355.

[211] Ofir Nachum and Bo Dai. "Reinforcement learning via fenchel-rockafellar duality". In: *arXiv preprint arXiv:2001.01866* (2020).

[212] Makoto Nakajima and Vladimir Smirnyagin. "Cyclical labor income risk". In: *Available at SSRN 3432213* (2019).

[213] Emi Nakamura and Jón Steinsson. "Identification in macroeconomics". In: *Journal of Economic Perspectives* 32.3 (2018), pp. 59–86.

[214] Hongseok Namkoong et al. "Off-policy Policy Evaluation For Sequential Decisions Under Unobserved Confounding". In: *Advances in neural information processing systems* 33 (2020).

[215] Arvind Narayanan. "How to recognize AI snake oil". In: *Arthur Miller Lecture on Science and Ethics* (2019).

[216] Whitney K Newey. "The asymptotic variance of semiparametric estimators". In: *Econometrica: Journal of the Econometric Society* (1994), pp. 1349–1382.

[217] Whitney K Newey, Fushing Hsieh, and James Robins. "Undersmoothing and bias corrected functional estimation". In: (1998).

[218] Whitney K Newey, Fushing Hsieh, and James M Robins. "Twicing kernels and a small bias property of semiparametric estimators". In: *Econometrica* 72.3 (2004), pp. 947–962.

[219] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. "Estimating divergence functionals and the likelihood ratio by convex risk minimization". In: *IEEE Transactions on Information Theory* 56.11 (2010), pp. 5847–5861.

[220] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. "On divergences, surrogate loss functions, and decentralized detection". In: *arXiv preprint math.ST/0510521* (2005).

[221] Arnab Nilim and Laurent El Ghaoui. "Robust control of Markov decision processes with uncertain transition matrices". In: *Operations Research* 53.5 (2005), pp. 780–798.

[222] Pegah Nokhiz et al. "Precarity: Modeling the Long Term Effects of Compounded Decisions on Individual Instability". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 199–208.

[223] Matthew Norton, Valentyn Khokhlov, and Stan Uryasev. "Calculating CVaR and bPOE for common probability distributions with application to portfolio optimization and density estimation". In: *Annals of Operations Research* 299.1 (2021), pp. 1281–1315.

[224] Michael Oberst and David Sontag. "Counterfactual off-policy evaluation with gumbel-max structural causal models". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4881–4890.

[225] OECD. *Conversion rates - exchange rates - OECD data*. 2023. URL: `https://data.oecd.org/conversion/exchange-rates.htm`.

[226] Tomasz Olma. "Nonparametric estimation of truncated conditional expectation functions". In: *arXiv preprint arXiv:2109.06150* (2021).

[227] Liliana Orellana, Andrea Rotnitzky, and James M Robins. "Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: main content". In: *The international journal of biostatistics* 6.2 (2010).

[228] Michal Ozery-Flato et al. "Adversarial balancing for causal inference". In: *arXiv preprint arXiv:1810.07406* (2018).

[229] Cosmin Paduraru. "Off-policy evaluation in Markov decision processes". PhD thesis. Ph. D. Dissertation. McGill University, 2012.

[230] Kishan Panaganti et al. "Robust reinforcement learning using offline data". In: *arXiv preprint arXiv:2208.05129* (2022).

[231] Marios Papachristou and Jon Kleinberg. "Allocating stimulus checks in times of crisis". In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 16–26.

[232] BU Park, YK Lee, and S Ha. "L_2 boosting in kernel regression". In: *Bernoulli* 15.3 (2009), pp. 599–613.

[233] Jonathan A Parker et al. "Consumer spending and the economic stimulus payments of 2008". In: *American Economic Review* 103.6 (2013), pp. 2530–2553.

[234] Judea Pearl et al. "Causal inference in statistics: An overview". In: *Statistics surveys* 3 (2009), pp. 96–146.

[235] Fabian T Pfeffer, Sheldon Danziger, and Robert F Schoeni. "Wealth disparities before and after the Great Recession". In: *The Annals of the American Academy of Political and Social Science* 650.1 (2013), pp. 98–123.

[236] Doina Precup. "Eligibility traces for off-policy policy evaluation". In: *Computer Science Department Faculty Publication Series* (2000), p. 80.

[237] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[238] Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. "Model-based reinforcement learning for sepsis treatment". In: *arXiv preprint arXiv:1811.09602* (2018).

[239] Aniruddh Raghu et al. "Deep reinforcement learning for sepsis treatment". In: *arXiv preprint arXiv:1711.09602* (2017).

[240] Inioluwa Deborah Raji et al. "The fallacy of AI functionality". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 959–972.

[241] Ashesh Rambachan and Neil Shephard. "When do common time series estimands have nonparametric causal meaning". In: *Manuscript, Harvard University* (2021).

[242] Valerie A Ramey. "Macroeconomic shocks and their propagation". In: *Handbook of macroeconomics* 2 (2016), pp. 71–162.

[243] Paria Rashidinejad et al. "Bridging offline reinforcement learning and imitation learning: A tale of pessimism". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11702–11716.

[244] Lydia Reader et al. "Models for understanding and quantifying feedback in societal systems". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1765–1775.

[245] James Robins et al. "Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable". In: *Statistical Science* 22.4 (2007), pp. 544–559.

[246] James Robins et al. "Higher order influence functions and minimax estimation of nonlinear functionals". In: *Probability and statistics: essays in honor of David A. Freedman*. Vol. 2. Institute of Mathematical Statistics, 2008, pp. 335–422.

[247] James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models". In: *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000, pp. 1–94.

[248] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. "Estimation of regression coefficients when some regressors are not always observed". In: *Journal of the American statistical Association* 89.427 (1994), pp. 846–866.

[249] R Tyrrell Rockafellar, Stanislav Uryasev, et al. "Optimization of conditional value-at-risk". In: *Journal of risk* 2 (2000), pp. 21–42.

[250] Yaniv Romano, Evan Patterson, and Emmanuel Candes. "Conformalized quantile regression". In: *Advances in neural information processing systems* 32 (2019).

[251] Christina D Romer and David H Romer. "A new measure of monetary shocks: Derivation and implications". In: *American economic review* 94.4 (2004), pp. 1055–1084.

[252] Paul R Rosenbaum. "Design sensitivity in observational studies". In: *Biometrika* 91.1 (2004), pp. 153–164.

[253] Paul R Rosenbaum. "Overt bias in observational studies". In: *Observational studies*. Springer, 2002, pp. 71–104.

[254] Paul R Rosenbaum and Donald B Rubin. "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 45.2 (1983), pp. 212–218.

[255] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

[256] Erik Rosenstrom et al. "Optimizing the first response to sepsis: An electronic health record-based Markov decision process model". In: *Decision Analysis* 19.4 (2022), pp. 265–296.

[257] Jesse Rothstein. "The lost generation? Labor market outcomes for post great recession entrants". In: *Journal of Human Resources* (2021), 0920–11206R1.

[258] Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. "Characterization of parameters with a mixed bias property". In: *Biometrika* 108.1 (2021), pp. 231–238.

[259] Donald B Rubin. "Randomization analysis of experimental data: The Fisher randomization test comment". In: *Journal of the American statistical association* 75.371 (1980), pp. 591–593.

[260] Avraham Ruderman et al. "Tighter variational representations of f-divergences via restriction to probability measures". In: *arXiv preprint arXiv:1206.4664* (2012).

[261] Soroush Saghafian. "Ambiguous Dynamic Treatment Regimes: A Reinforcement Learning Approach". In: *arXiv preprint arXiv:2112.04571* (2021).

[262] Claudia R Sahm, Matthew D Shapiro, and Joel Slemrod. "Household response to the 2008 tax rebate: Survey evidence and aggregate implications". In: *Tax Policy and the Economy* 24.1 (2010), pp. 69–110.

[263] Matthew J Salganik et al. "Measuring the predictability of life outcomes with a scientific mass collaboration". In: *Proceedings of the National Academy of Sciences* 117.15 (2020), pp. 8398–8403.

[264] Daniel Scharfstein et al. "Global sensitivity analysis for repeated measures studies with informative drop-out: A semi-parametric approach". In: *Biometrics* 74.1 (2018), pp. 207–219.

[265] Daniel O Scharfstein et al. "Semiparametric sensitivity analysis: Unmeasured confounding in observational studies". In: *arXiv preprint arXiv:2104.08300* (2021).

[266] Vira Semenova. "Debiased Machine Learning of Set-Identified Linear Models". In: *arXiv preprint arXiv:1712.10024* (2017).

[267] Vira Semenova. "Debiased machine learning of set-identified linear models". In: *Journal of Econometrics* (2023).

[268] Vira Semenova and Victor Chernozhukov. "Debiased machine learning of conditional average treatment effects and other causal functions". In: *The Econometrics Journal* 24.2 (2021), pp. 264–289.

[269] Uri Shalit, Fredrik D Johansson, and David Sontag. "Estimating individual treatment effect: generalization bounds and algorithms". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3076–3085.

[270] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[271] Dennis Shen et al. "A Tale of Two Panel Data Regressions". In: *arXiv preprint arXiv:2207.14481* (2022).

[272] Jian Shen et al. "Wasserstein distance guided representation learning for domain adaptation". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[273] Xiaotong Shen. "On methods of sieves and penalization". In: *The Annals of Statistics* 25.6 (1997), pp. 2555–2591.

[274] Chengchun Shi et al. "A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 20057–20094.

[275] Chengchun Shi et al. "Off-policy confidence interval estimation with confounded Markov decision process". In: *Journal of the American Statistical Association* (2022), pp. 1–12.

[276] Max Simchowitz, Ross Boczar, and Benjamin Recht. "Learning linear dynamical systems with semi-parametric least squares". In: *Conference on Learning Theory*. PMLR. 2019, pp. 2714–2802.

[277] Max Simchowitz et al. "Learning without mixing: Towards a sharp analysis of linear system identification". In: *Conference On Learning Theory*. PMLR. 2018, pp. 439–473.

[278] Rahul Singh. "Debiased kernel methods". In: *arXiv preprint arXiv:2102.11076* (2021).

[279] Rahul Singh, Liyang Sun, et al. *Double Robustness for Complier Parameters and a Semiparametric Test for Complier Characteristics*. Tech. rep. 2022.

[280] Rahul Singh and Vasilis Syrgkanis. "Automatic Debiased Machine Learning for Dynamic Treatment Effects". In: *arXiv preprint arXiv:2203.13887* (2022).

[281] Rahul Singh, Liyuan Xu, and Arthur Gretton. "Kernel methods for causal functions: Dose, heterogeneous, and incremental response curves". In: *arXiv preprint arXiv:2010.04855* (2020).

[282] Maurice Sion. "On general minimax theorems." In: *Pacific Journal of mathematics* 8.1 (1958), pp. 171–176.

[283] Jiaming Song and Stefano Ermon. "Bridging the gap between f-gans and wasserstein gans". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9078–9087.

[284] Paul Speckman. "Minimax estimates of linear functionals in a Hilbert space". In: *Unpublished Manuscript* (1979).

[285] Bharath K Sriperumbudur et al. "On integral probability metrics,\phi-divergences and binary classification". In: *arXiv preprint arXiv:0901.2698* (2009).

[286] James H Stock and Mark W Watson. "Identification and estimation of dynamic causal effects in macroeconomics using external instruments". In: *The Economic Journal* 128.610 (2018), pp. 917–948.

[287] Kjetil Storesletten, Chris I Telmer, and Amir Yaron. "Cyclical dynamics in idiosyncratic labor market risk". In: *Journal of political Economy* 112.3 (2004), pp. 695–717.

[288] Ludwig Straub. "Consumption, savings, and the distribution of permanent income". In: *Unpublished manuscript, Harvard University* (2019).

[289] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. "Covariate shift adaptation by importance weighted cross validation." In: *Journal of Machine Learning Research* 8.5 (2007).

[290] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. "Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation". In: *Annals of the Institute of Statistical Mathematics* 64.5 (2012), pp. 1009–1044.

[291] Masashi Sugiyama et al. "Direct importance estimation with model selection and its application to covariate shift adaptation." In: *NIPS*. Vol. 7. Citeseer. 2007, pp. 1433–1440.

[292] Zhiqiang Tan. "A Distributional Approach for Causal Inference using Propensity Scores". In: *Journal of the American Statistical Associatioon* (2012).

[293] Zhiqiang Tan. "A distributional approach for causal inference using propensity scores". In: *Journal of the American Statistical Association* 101.476 (2006), pp. 1619–1637.

[294] Zhiqiang Tan. "Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data". In: *Biometrika* 107.1 (2020), pp. 137–158.

[295] Dingke Tang et al. "Ultra-high dimensional variable selection for doubly robust causal inference". In: *Biometrics* 79.2 (2023), pp. 903–914.

[296] Ziyang Tang et al. "Doubly robust bias reduction in infinite horizon off-policy estimation". In: *arXiv preprint arXiv:1910.07186* (2019).

[297] Eric J Tchetgen Tchetgen Tchetgen et al. "An introduction to proximal causal learning". In: *arXiv preprint arXiv:2009.10982* (2020).

[298] Guy Tennenholtz, Shie Mannor, and Uri Shalit. "Off-Policy Evaluation in Partially Observable Environments". In: *arXiv preprint arXiv:1909.03739* (2019).

[299] Guy Tennenholtz, Uri Shalit, and Shie Mannor. "Off-policy evaluation in partially observable environments". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 06. 2020, pp. 10276–10283.

[300] Philip Thomas and Emma Brunskill. "Data-efficient off-policy policy evaluation for reinforcement learning". In: *International Conference on Machine Learning*. PMLR. 2016, pp. 2139–2148.

[301] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. "High confidence policy improvement". In: *International Conference on Machine Learning*. 2015, pp. 2380–2388.

[302] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. "Minimax weight and q-function learning for off-policy evaluation". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9659–9668.

[303] Masatoshi Uehara et al. "Future-dependent value-based off-policy evaluation in pomdps". In: *arXiv preprint arXiv:2207.13081* (2022).

[304] Aad van de Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Mathematics, 1996.

[305] Tyler J VanderWeele and Peng Ding. "Sensitivity analysis in observational research: introducing the E-value". In: *Annals of internal medicine* 167.4 (2017), pp. 268–274.

[306] Cameron Voloshin et al. "Empirical study of off-policy policy evaluation for reinforcement learning". In: *arXiv preprint arXiv:1911.06854* (2019).

[307] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.

[308] Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. "Provably efficient causal reinforcement learning with confounded observational data". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 21164–21175.

[309] Shengbo Wang et al. "A finite sample complexity bound for distributionally robust q-learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 3370–3398.

[310] Yixin Wang and Jose R Zubizarreta. "Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations". In: *Biometrika* 107.1 (2020), pp. 93–105.

[311] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. "Robust Markov decision processes". In: *Mathematics of Operations Research* 38.1 (2013), pp. 153–183.

[312] Raymond KW Wong and Kwun Chuen Gary Chan. "Kernel-based covariate functional balancing for observational studies". In: *Biometrika* 105.1 (2018), pp. 199–213.

[313] Tengyang Xie and Nan Jiang. "Q* approximation schemes for batch reinforcement learning: A theoretical comparison". In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 550–559.

[314] Tengyang Xie et al. "Bellman-consistent pessimism for offline reinforcement learning". In: *Advances in neural information processing systems* 34 (2021), pp. 6683–6694.

[315] Steve Yadlowsky et al. "Bounds on the conditional and average treatment effect with unobserved confounding factors". In: *arXiv preprint arXiv:1808.09521* (2018).

[316] S Yang and P Ding. "Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores". In: *Biometrika* 105.2 (2018), pp. 487–493.

[317] Shu Yang and Judith J Lok. "Sensitivity analysis for unmeasured confounding in coarse structural nested mean models". In: *Statistica Sinica* 28.4 (2018), p. 1703.

[318] Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. "Toward theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics". In: *The Annals of Statistics* 50.6 (2022), pp. 3223–3248.

[319] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. "GANITE: Estimation of individualized treatment effects using generative adversarial nets". In: *International Conference on Learning Representations*. 2018.

[320] Yaoliang Yu and Csaba Szepesvári. "Analysis of kernel mean matching under covariate shift". In: *arXiv preprint arXiv:1206.4650* (2012).

[321] Fernando G Zampieri, Sean M Bagshaw, and Matthew W Semler. "Fluid therapy for critically ill adults with sepsis: a review". In: *Jama* 329.22 (2023), pp. 1967–1980.

[322] Jiaming Zeng et al. "Uncovering interpretable potential confounders in electronic medical records". In: *Nature Communications* 13.1 (2022), p. 1014.

[323] Qingyuan Zhao. "Covariate balancing propensity score by tailored loss functions". In: *The Annals of Statistics* 47.2 (2019), pp. 965–993.

[324] Qingyuan Zhao and Daniel Percival. "Entropy balancing is doubly robust". In: *Journal of Causal Inference* 5.1 (2017).

[325] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. "Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.4 (2019), pp. 735–761.

[326] Ying-Qi Zhao et al. "New statistical learning methods for estimating optimal dynamic treatment regimes". In: *Journal of the American Statistical Association* 110.510 (2015), pp. 583–598.

[327] Zhengqing Zhou et al. "Finite-sample regret bound for distributionally robust offline tabular reinforcement learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 3331–3339.

[328] Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. "Generalized resilience and robust statistics". In: *arXiv preprint arXiv:1909.08755* (2019).

[329] José R Zubizarreta. "Stable weights that balance covariates for estimation with incomplete outcome data". In: *Journal of the American Statistical Association* 110.511 (2015), pp. 910–922.

Appendix A

Additional Materials: Duality for Balancing

A.1 Proof of Theorem 3.4.1

We derive a dual formulation of the optimization problem:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} D_2(R||P)$$
$$\text{such that } \text{IPM}_{\mathcal{F}}(Q, R) \leq \delta,$$

where $\delta > \delta_{\min}$. As a reminder, $D_2(R||P) \coloneqq \mathbb{E}_P[(dR/dP)^2 - 1]$ is the χ^2 divergence and $\text{IPM}_{\mathcal{F}}(Q, R) \coloneqq \sup_{f \in \mathcal{F}}\{\mathbb{E}_Q[f] - \mathbb{E}_R[f]\}$. Note that this problem takes the form of a projection in D_2 of P onto an IPM ball around Q.

By the definition of $\delta_{\min}$, the constraint set is non-empty and convex. D_2 is strictly convex in R and $0 \leq D_2(R||P) < \infty$ so there is a unique solution.

When P already satisfies the IPM constraint, then $R = P$ has objective 0 and we're done (i.e. we don't need to do a projection, P is already on or in the ball). Otherwise, by standard use of the Lagrangian, we claim (details in Section A.1 below) that for some $\mu > 0$ corresponding to δ, this problem is equivalent to:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \big\{(1/\mu) D_2(R||P) + \sup_{f \in \mathcal{F}}\{\mathbb{E}_Q[f] - \mathbb{E}_R[f]\}\big\}.$$

Exchanging hard subproblem for an easy subproblem

The inner supremum is hard to solve for arbitrary $\mathcal{F}$. However, we can make a series of transformations to get an easier subproblem with a closed-form solution:

$$\begin{aligned}
&\inf_{\substack{R\in\mathcal{M}(P)\\ \mathbb{E}_R[1]=1}} \left\{(1/\mu)D_2(R||P) + \sup_{f\in\mathcal{F}}\{\mathbb{E}_Q[f] - \mathbb{E}_R[f]\}\right\} \\
&= \inf_{\substack{R\in\mathcal{M}(P)\\ \mathbb{E}_R[1]=1}} \sup_{f\in\mathcal{F}} \left\{(1/\mu)D_2(R||P) + \mathbb{E}_Q[f] - \mathbb{E}_R[f]\right\} \\
&= \sup_{f\in\mathcal{F}} \inf_{\substack{R\in\mathcal{M}(P)\\ \mathbb{E}_R[1]=1}} \left\{(1/\mu)D_2(R||P) + \mathbb{E}_Q[f] - \mathbb{E}_R[f]\right\} \\
&= \sup_{f\in\mathcal{F}} \left\{\mathbb{E}_Q[f] + \inf_{\substack{R\in\mathcal{M}(P)\\ \mathbb{E}_R[1]=1}} \{(1/\mu)D_2(R||P) - \mathbb{E}_R[f]\}\right\}
\end{aligned}$$

The only non-trivial step is the interchange of the inf and the sup. This follows by Sion's Minimax Theorem [282]. We assumed that $\mathcal{X}$ was a separable Banach space so we have the necessary topological properties. The objective on the second line is continuous and strictly convex in R and is linear in f. The set $\mathcal{F}$ is convex and closed and R is in a linear subspace. Furthermore, we know there is a unique solution R^*, and so we can always find the necessary compact subset of the linear subspace for R to apply the theorem [e.g. ala 119].

Solving the easy subproblem with the variational representation

Next, we apply the variational representation of ϕ-divergences to get a dual formulation of the inner sub-problem over R. Define: $\phi(x) = (1/\mu)(x^2 - 1)$ which has convex conjugate $\phi^*(y) = (\mu/4)y^2 + (1/\mu)$. The Lagrangian for the infimum for a fixed f is:

$$\mathcal{L}_f(R, \lambda) = D_\phi(R||P) - \mathbb{E}_R[f - \lambda] - \lambda$$

We get the first-order condition:

$$\phi'(dR^*/dP) = f - \lambda^* \implies \frac{dR^*}{dP} = \frac{\mu}{2}(f - \lambda^*)$$

where λ^* solves the supremum $\sup_{\lambda\geq 0} g(\lambda)$ over the dual function:

$$\begin{aligned}
g(\lambda) &\coloneqq -\lambda + \inf_{R\in\mathcal{M}(P)} \{D_\phi(R||P) - \mathbb{E}_R[f - \lambda]\} \\
&= -\lambda - \sup_{R\in\mathcal{M}(P)} \{\mathbb{E}_R[f - \lambda] - D_\phi(R||P)\} \\
&= -\lambda - D^*_\phi(f - \lambda),
\end{aligned}$$

where D^*_ϕ is the convex conjugate of the ϕ-divergence as a function of R for a fixed P. We can then use the standard result [Proposition 4.2 40], $D^*_\phi(f) = \mathbb{E}_P[\phi^*(f)]$.

Using this form of the dual function, we can write our subproblem over R as:

$$\inf_{\substack{R\in\mathcal{M}(P)\\ \mathbb{E}_R[1]=1}}\{(1/\mu)D_2(R||P)-\mathbb{E}_R[f]\}=\sup_{\lambda\geq 0}\{-\lambda-\mathbb{E}_P[\phi^*(f-\lambda)]\}$$

Plugging in ϕ^*, we can solve for λ^* by straightforward calculus:

$$\lambda^*=\mathbb{E}_P[f]-\frac{2}{\mu}.$$

Now using the first-order conditions, we can find the optimal R^*:

$$\frac{dR^*}{dP}=\frac{\mu}{2}(f-\mathbb{E}_P[f])+1$$

and after some algebra, a closed form of the subproblem:

$$\inf_{\substack{R\in\mathcal{M}(P)\\ \mathbb{E}_R[1]=1}}\{(1/\mu)D_2(R||P)-\mathbb{E}_R[f]\}=-\mathbb{E}_P[f]-\frac{\mu}{4}\mathrm{Var}_P[f].$$

Writing the original problem as a single optimization problem over $\mathcal{F}$

Finally, we substitute this form of the sub-problem into our original optimization problem:

$$\begin{aligned}&\inf_{\substack{R\in\mathcal{M}(P)\\ \mathbb{E}_R[1]=1}}\Big\{(1/\mu)D_2(R||P)+\sup_{f\in\mathcal{F}}\{\mathbb{E}_Q[f]-\mathbb{E}_R[f]\}\Big\}\\&=\sup_{f\in\mathcal{F}}\Big\{\mathbb{E}_Q[f]+\inf_{\substack{R\in\mathcal{M}(P)\\ \mathbb{E}_R[1]=1}}\{(1/\mu)D_2(R||P)-\mathbb{E}_R[f]\}\Big\}\\&=\sup_{f\in\mathcal{F}}\Big\{\mathbb{E}_Q[f]-\mathbb{E}_P[f]-\frac{\mu}{4}\mathrm{Var}_P[f]\Big\}\end{aligned}$$

and therefore by duality:

$$\frac{dR^*}{dP}=\frac{\mu}{2}(f^*-\mathbb{E}_P[f^*])+1$$

where f^* achieves this supremum.

Recovering δ in terms of μ

Most of the proof of the theorem is complete. We just need to rewrite μ in terms of the original tuning parameter δ. Remember from the projection perspective, that $\mu>0$ corresponds to P outside of the IPM ball. As a result:

$$\delta=\sup_{f\in\mathcal{F}}\{\mathbb{E}_Q[f]-\mathbb{E}_{R^*}[f]\}$$

We just proved that R^* achieves the infimum of the objective which equals the supremum of the dual:

$$\begin{aligned}
&\text{IPM}_{\mathcal{F}}(Q, R^*) + (1/\mu) D_2(R^* || P) \\
&= \mathbb{E}_Q[f^*] - \mathbb{E}_P[f^*] - \frac{\mu}{4}\text{Var}_P[f^*] \\
&= \Big(\mathbb{E}_Q[f^*] - \mathbb{E}_{R^*}[f^*]\Big) + \Big(\mathbb{E}_{R^*}[f^*] - \mathbb{E}_P[f^*] - \frac{\mu}{4}\text{Var}_P[f^*]\Big)
\end{aligned}$$

But then, using the variational representation of the ϕ-divergence and the definition of the IPM we have:

$$\begin{aligned}
&\text{IPM}_{\mathcal{F}}(Q, R^*) + (1/\mu) D_2(R^* || P) \\
&= \sup_{f \in \mathcal{F}} \Big\{\mathbb{E}_Q[f] - \mathbb{E}_{R^*}[f]\Big\} + \sup_f \Big\{\mathbb{E}_{R^*}[f] - \mathbb{E}_P[f] - \frac{\mu}{4}\text{Var}_P[f^*]\Big\} \\
&= \Big(\mathbb{E}_Q[f^*] - \mathbb{E}_{R^*}[f^*]\Big) + \Big(\mathbb{E}_{R^*}[f^*] - \mathbb{E}_P[f^*] - \frac{\mu}{4}\text{Var}_P[f^*]\Big)
\end{aligned}$$

which implies

$$\delta = \sup_{f \in \mathcal{F}} \Big\{\mathbb{E}_Q[f] - \mathbb{E}_{R^*}[f]\Big\} = \mathbb{E}_Q[f^*] - \mathbb{E}_{R^*}[f^*]$$

Finally, substituting the form of R^* in terms of f^* we get

$$\begin{aligned}
\delta &= \mathbb{E}_Q[f^*] - \mathbb{E}_P[f^*] - \frac{\mu}{2}\text{Var}_P[f^*] \\
\implies \mu &= 2\left(\frac{\mathbb{E}_Q[f^*] - \mathbb{E}_P[f^*] - \delta}{\text{Var}_P[f^*]}\right),
\end{aligned}$$

which concludes the proof.

Transformation from δ to μ via Lagrangian

Here we provide the details for our earlier claim that we can rewrite the problem over δ as a problem over μ. The dual function corresponding to the original δ problem is:

$$\begin{aligned}
g(\mu) &= \inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \mathcal{L}(R, \mu) \\
&= \inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \{D_2(R||P) + \mu(\sup_{f \in \mathcal{F}}\{\mathbb{E}_Q[f] - \mathbb{E}_R[f]\} - \delta)\}
\end{aligned}$$

where $\mathcal{L}$ is the Lagrangian. Notice that the original optimization problem has a strictly convex objective. Furthermore, since the function class $\mathcal{F}$ is convex and closed, $\delta > \delta_{\min}$, and the individual constraints for $f \in \mathcal{F}$ are all linear, the feasible set is convex with a non-empty interior. Then by standard convex duality there exists R^* and $\mu^* \geq 0$ such that R^* solves the

original optimization problem, μ^* achieves $\sup_{\mu \geq 0} g(\mu)$, and R^* achieves the infimum inside $g(\mu^*)$.

By complementary slackness, $\mu^* = 0$ only when the worst-bias constraint doesn't bind which only occurs when $R = P$ already satisfies the IPM constraint. Then $R = P$ has minimum variance and we're done. So we only need to consider the case where $\mu^* > 0$.

At $\mu = \mu^*$, the solution to:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \left\{ D_2(R||P) + \mu \sup_{f \in \mathcal{F}} \{\mathbb{E}_Q[f] - \mathbb{E}_R[f]\} \right\}$$

has the same solution as the original problem. Furthermore, since $\mu^* > 0$, we can apply one more transformation without affecting the infimum to get:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \left\{ (1/\mu) D_2(R||P) + \sup_{f \in \mathcal{F}} \{\mathbb{E}_Q[f] - \mathbb{E}_R[f]\} \right\}$$

A.2 General Statement and Proof for Remark 3.1

Theorem (Birrell et al)**.** *Let ϕ be a convex function such that $\phi(1) = 0$ with convex conjugate ϕ^* such that $\{\phi^* < \infty\} = \mathbb{R}$. Let ϕ_μ denote the weighted function $\phi_\mu(x) = (1/\mu)\phi(x)$ and ϕ^*_μ its convex conjugate. Then under Assumption 2, for $\delta > 0$, $\exists \mu \geq 0$ such that the optimization problem,*

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} D_\phi(R||P)$$
$$\text{such that } IPM_{\mathcal{F}}(Q, R) \leq \delta,$$

has a solution,

$$R^* = P \text{ when } \mu = 0,$$
$$\frac{dR^*}{dP} = (\phi^*_\mu)'(f^* - \lambda^*) \text{ otherwise,}$$

where f^* and λ^* achieve the supremum,

$$\sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_Q[f] - \inf_{\lambda \in \mathbb{R}} \{\lambda + \mathbb{E}_P[\phi^*_\mu(f - \lambda)]\} \right\}.$$

The proof begins with an argument identical to 1.5 above which gives the equivalent optimization problem:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \big\{ D_{\phi_\mu}(R||P) + \text{IPM}_{\mathcal{F}}(Q, R) \big\}.$$

From here, conceptually, the proof is similar to above, except we cannot apply our proof directly because general ϕ loses some of the nice properties of the quadratic. Instead, using the theory of infimal convolutions, [36] prove that for Polish $\mathcal{X}$ and any ϕ-divergence such that $\{\phi^* < \infty\} = \mathbb{R}$:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \big\{ D_\phi(R||P) + \text{IPM}_{\mathcal{F}}(Q, R) \big\}$$
$$= \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_Q[f] - \inf_{\lambda \in \mathbb{R}} \{\lambda + \mathbb{E}_P[\phi^*(f - \lambda)]\} \right\}$$

and so in particular, it holds for ϕ_μ above.

This result follows from Theorems 2.15 and 3.3 in [36] with one minor modification: we do not require that $\lim_{y \to -\infty} \phi^*(y) < \infty$ which results in R^* no longer being a probability measure because we lose statement (174) in their proof of Theorem C.6. This is in the spirit of the arguments in [40]. In fact, since the original problem is a projection of P in ϕ-divergence onto an IPM, we can interpret this as a version of [40] Theorem 5.1 which applies for the case of finitely-many linear inequality constraints, generalized to the case of a linear inequality constraint for each f in a convex and closed set $\mathcal{F}$.

A.3 Extension with non-negative weights

We can use the general Theorem to immediately get results for the case where we require non-negative weights. This is identical to taking $\phi(x) = x^2 - 1$, but restricting the domain to $[0, \infty)$. Then we have

$$\phi_\mu^*(y) = \frac{\mu}{4} y^2 \mathbf{1}(y \geq 0) + \frac{1}{\mu}$$

and applying the theorem, we get the dual formulation:

$$\frac{dR^*}{dP} = \frac{\mu}{2}(f^* - \lambda^*)\mathbf{1}(f^* \geq \lambda^*)$$

such that:

$$\mathbb{E}_P\left[\frac{\mu}{2}(f^* - \lambda^*)\mathbf{1}(f^* \geq \lambda^*)\right] = 1$$

and f^* and λ^* solve:

$$\sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_Q[f] - \inf_{\lambda \in \mathbb{R}} \left\{ \lambda + \frac{\mu}{4} \mathbb{E}_P\left[(f - \lambda)^2 \mathbf{1}(f \geq \lambda)\right] \right\} + \frac{1}{\mu} \right\}.$$

The minimax weights are extremely simliar to the minimax weights from Theorem 3.1. In Theorem 3.1 we found a function f^*, such that we de-meaned it, rescaled it, and then shifted the result to get weights with expectation equal to 1. In the case where the weights are non-negative, we can no longer just de-mean and add 1. Instead, we have to optimize over all shifts λ which give expectation 1 *after* truncating at 0.

A.4 Connection to surrogate loss for density ratio estimation

A closely related literature implements regularized estimators of the density ratio via a surrogate loss as first proposed in [219]. Similar connections have been made for balancing weights; see the relation to propensity score estimation in [323, 29]. Here, we compare our dual formulation based on variational representations to density ratio estimation with a surrogate loss.

Consider the variational representation (3.11). Specializing to $\phi(x) = \mu(x^2 - 1)$ for $\mu > 0$, we can write the weighted χ^2 divergence between Q and P as:

$$\frac{1}{\mu} D_2(Q||P) = \sup_f \Big\{ \mathbb{E}_Q[f] - \mathbb{E}_P[f] - \frac{\mu}{4} \mathrm{Var}_P[f] \Big\},$$

where the supremum is over all real-valued measurable functions. If overlap holds then the supremum is achieved by $2\mu(dQ/dP)$, otherwise $D_2(Q||P) = \infty$. The variational representation is identical to (3.12), except that the problem in (3.12) is restricted to functions in $\mathcal{F}$. We immediately have the following corollary:

Corollary A.4.1. *Under the conditions in Theorem 3.4.1,*

$$\frac{2}{\mu}\frac{dQ}{dP} \in \mathcal{F} \implies f^* = \frac{2}{\mu}\frac{dQ}{dP} \text{ and } \frac{dR^*}{dP} = \frac{dQ}{dP}.$$

If the density ratio exists and a scaled version belongs to the outcome function class, then it is minimax optimal to reweight so that there is zero bias. If $\mathcal{F}$ is sufficiently flexible, then this will always hold as $\mu \to \infty$, as in the example from Section 3.5.

There is a clear similarity to density ratio estimation using ϕ-divergences as in [219]. The form of their final estimator looks the same as ours: maximizing a variational representation over a function class. However, our motivations and assumptions are entirely different. [219] assumes that the density ratio exists and that it belongs to $\mathcal{F}$. We do not make any assumption about the density ratio. It could have some arbitrary functional form. We show that the shape of the optimal weights is determined by the shape of f_0, *not* by the shape of dQ/dP.

If the density ratio exists, and our dual problem is solved over all measurable functions, the only unique solution will be dQ/dP. In that sense, once we restrict to optimization over $\mathcal{F}$, it might be helpful to think of the balancing weights as a projection of the density ratio onto our outcome function class. However, in general, we do not need a functional form for the density ratio to specified. In fact, we do not need even require that the density ratio exists.

A.5 RKHS optimization problem

The optimization problem that implements our dual formulation using an RKHS for the IHDP dataset is:

$$\begin{aligned}
&\sup_{f \in \mathcal{F}_{\mathcal{H}}^{B}} \left\{ \mathbb{E}_Q[f] - \mathbb{E}_P[f] - \frac{\mu}{4} \mathrm{Var}_P[f] \right\} \\
&= \sup_{\substack{f = K\alpha \\ \alpha \in \mathbb{R}^n : \alpha^T K \alpha \leq B}} \left\{ f^T e_q - f^T e_p - (f^T I_p f - (f^T e_p)^2) \right\}
\end{aligned}$$

where:

- $e_p \in \mathbb{R}^n$ is a vector equal to $1/n_0$ for those indices corresponding to control group data points and 0 otherwise
- $e_q \in \mathbb{R}^n$ is a vector equal to $1/n_0$ for those indices corresponding to treatment group data points and 0 otherwise
- $I_p \in \mathbb{R}^{n \times n}$ is the matrix with e_p on the diagonal.

We solve this problem using the scipy Python package.

Appendix B

Additional Materials: Augmented Balancing

B.1 Additional background and examples

Examples of general linear functionals via the Riesz representer

Example 1 (Counterfactual mean)**.** *Let* $Z = \{0, 1\}$ *and* $\psi(m) = \mathbb{E}[m(X, 1)]$*. Under SUTVA and conditional ignorability, this estimand is equal to* $E[Y(1)]$*. The Riesz representer is the IPW,* $\alpha(X, Z) = Z/e(X)$*.*

Example 2 (Average derivative)**.** *Let* $Z \in \mathbb{R}$ *and* $\psi(m) = \mathbb{E}\left[\frac{\partial}{\partial z} m(X, Z)\right]$*. Under an appropriate generalization of SUTVA and conditional ignorability, this estimand corresponds to the average derivative effect of a continuous treatment. Under regularity conditions the Riesz representer is given by* $\alpha(X, Z) = -\frac{\frac{d}{dz} p(Z|X)}{p(Z|X)}$ *where* $p(z|x)$ *is the conditional density of* Z *given* X*.*

Example 3 (Distribution shift)**.** *Consider an example without* Z*, following the machine learning literature on covariate shift. Let* p *denote the* source *distribution of the observed data, and let* p^* *over* $\mathcal{X}$ *denote the* target *distribution. The estimand is then* $\psi(m) = \int_{\mathcal{X}} m(x) dp^*(x)$*. In a causal inference setting, this can recover the Average Treatment Effect on the Treated (ATT) under SUTVA and conditional ignorability; i.e., let* p *be the distribution of covariates and outcomes for units assigned to control and* p^* *be the distribution of the covariates for units assigned to treatment. The Riesz representer is the density ratio,* $\alpha(X) = \frac{dp^*}{dp}(X)$*.*

Example 4 (Exact balancing weights)**.** *The most common balancing weights estimation problem finds the minimum weights that exactly balance each element of* Φ*. In the constrained form, exact balancing solves*

$$\begin{aligned} &\min_{w \in \mathbb{R}^n} \|w\|_2^2 \\ &\text{such that } \tfrac{1}{n} w \phi_{pj} = \bar{\phi}_{qj} \quad \text{for all } j \end{aligned} \tag{B.1}$$

Examples of balancing weights

Example 5 (ℓ_2 balancing)**.** *The ℓ_2 balancing weights problem is usually expressed via its penalized form:*

$$\min_{w\in\mathbb{R}^n}\left\{\|\tfrac{1}{n}w\Phi_p-\overline{\Phi}_q\|_2^2+\delta\|w\|_2^2\right\}. \tag{B.2}$$

The automatic form is a ridge-penalized regression for the Riesz representer.

Example 6 (ℓ_∞ balancing)**.** *The constrained form of the ℓ_∞ balancing weights problem is*

$$\begin{aligned}&\min_{w\in\mathbb{R}^n}\|w\|_2^2\\ &\textit{such that } \|\tfrac{1}{n}w\Phi_p-\overline{\Phi}_q\|_\infty\le\delta\end{aligned} \tag{B.3}$$

The automatic form is a lasso-penalized regression for the Riesz representer, sometimes known as the Minimum Distance Lasso *[55].*

Example 7 (Kernel balancing)**.** *As a brief preview of the balancing problem in the infinite-dimensional setting, we provide an example where $\mathcal{F}=\mathcal{H}$ is a reproducing kernel Hilbert space on $\mathcal{X}\times\mathcal{Z}$ with norm $\|\cdot\|_{\mathcal{H}}$ and kernel $\mathcal{K}:(\mathcal{X}\times\mathcal{Z})\times(\mathcal{X}\times\mathcal{Z})\to\mathbb{R}$. Then for any $x_i\in\mathcal{X}, z_i\in\mathcal{Z}$, the representer $\phi(x_i,z_i)\coloneqq\mathcal{K}(x_i,z_i,\cdot,\cdot)\in\mathcal{H}$. Using infinite-dimensional matrix notation, we denote $\Phi_p\in\mathcal{H}^n$ and $\overline{\Phi}_q\in\mathcal{H}$ as above. The penalized balancing weights problem for $\mathcal{F}=\mathcal{H}$ is:*

$$\min_{w\in\mathbb{R}^n}\left\{\|\tfrac{1}{n}w\Phi_p-\overline{\Phi}_q\|_{\mathcal{H}}^2+\delta\|w\|_2^2\right\}. \tag{B.4}$$

See Appendix B.2 for details and references.

Causal inference

We now return to Example 1 above. Here the goal is estimating the unobserved potential outcomes in an observational study. Let Y be the potential outcome under control [with appropriate restrictions, such as SUTVA; 259], let p be the population of individuals in the control condition, and let q be the population of individuals in the treatment condition. Then Y is observed for population p but not for population q, and the missing mean, $\mathbb{E}_q[Y]$, is the average potential outcome under control for the individuals who in fact were treated. Letting Y be the potential outcome under treatment, p the population of individuals in the treatment condition, and q the population of individuals in the control condition, $\mathbb{E}_q[Y]$ is the average potential outcome under treatment for the individuals who in fact received control.

For both examples, the crucial assumption for identification is *conditional ignorability*: the conditional distribution of Y given X is the same in the source and target populations. This is also known as "conditional exchangeability," "selection on observables," or "no unmeasured

confounding." For our purposes, we will require the mean, but not distributional, version of this assumption: $\mathbb{E}_p[Y|X] = \mathbb{E}_q[Y|X]$.

Since we assume the conditional expectations are the same in the two populations, we occasionally denote the common conditional mean functional without subscripts, $\mathbb{E}[Y|X]$. Under this assumption, we can identify $\mathbb{E}_q[Y]$ with the *regression functional*, also known as the *adjustment formula* or *g-formula*:

$$\mathbb{E}_q\left[\mathbb{E}_p[Y|X]\right] = \mathbb{E}_q\left[\mathbb{E}_q[Y|X]\right] = \mathbb{E}_q[Y]. \tag{B.5}$$

A complementary approach instead relies on the density ratio between the marginal covariate distributions in the source and target populations, $\frac{dq}{dp}(X)$, also known as the Radon-Nikodym derivative, importance sampling weights, or inverse propensity score weights (IPW). [1] This is also a special case of a Riesz representer [55]. Under an additional *population overlap assumption* that $q(x)$ is absolutely continuous with respect to $p(x)$, we can identify $\mathbb{E}_q[Y]$ via the *weighting functional*, also known as the *IPW functional*:

$$\mathbb{E}_p\left[\frac{dq}{dp}(X)\, Y\right] = \mathbb{E}_p\left[\frac{dq}{dp}(X)\, \mathbb{E}_p[Y|X]\right] = \mathbb{E}_q[\mathbb{E}_p[Y|X]] = \mathbb{E}_q[Y]. \tag{B.6}$$

Finally, we can combine the regression and weighting functionals to create a third identifying functional, known as the *doubly robust functional* [248]:

$$\mathbb{E}_q\left[\mathbb{E}_p[Y|X]\right] + \mathbb{E}_p\left[\frac{dq}{dp}(X)\left\{Y - \mathbb{E}_p[Y|X]\right\}\right]. \tag{B.7}$$

This functional has the attractive property of being equal to $\mathbb{E}_q[Y]$ even if either one of $\frac{dq}{dp}(X)$ or $\mathbb{E}_p[Y|X]$ is replaced with an arbitrary function of X, hence the term "doubly robust." [2] See [60, 167] for recent overviews of the active literature in causal inference and machine learning focused on estimating versions of Equation (4.3).

The *augmented* estimators that we analyze in this paper are based on estimating this doubly robust functional. These estimators *augment* an estimator of the regression functional based on an outcome regression (or *base learner*) with appropriately weighted residuals. Alternatively, they *augment* an estimator of the weighting functional with an outcome regression-based estimator of the regression functional (subtracting off the implied estimator of $\frac{dq}{dp}(X)\mathbb{E}_p[Y|X]$).

[1] Using Bayes Rule, we can equivalently express $\frac{dq}{dp}(X)$ via the *propensity score* $P(1_p|X)$, where 1_p is the indicator that an observation from the size-proportional mixture distribution of p and q is from population p: $\frac{dq}{dp}(X) = \frac{1-P(1_p|X)}{P(1_p|X)} \frac{P(1_p)}{1-P(1_p)}$

[2] This functional is equal to $\mathbb{E}_q[Y]$ if $\mathbb{E}_p[Y|X]$ is replaced with an arbitrary well-behaved functional of X_p, because the first and last terms cancel and we are left with the weighting functional $\mathbb{E}_p[\frac{dq}{dp}(X)Y]$. It is also equal to $\mathbb{E}_q[Y]$ if $\frac{dq}{dp}(X)$ is replaced with an arbitrary well-behaved functional of X_p, because the $\mathbb{E}_p[h(X)(Y - \mathbb{E}_p[Y|X])]$ is equal to 0 for any h and therefore we are left with the regression functional $\mathbb{E}_q\left[\mathbb{E}_p[Y|X]\right]$.

Recall that under linearity the imbalance over all $f \in \mathcal{F}$ has a simple closed form. For any $f(x,z) = \theta^\top \phi(x,z) \in \mathcal{F}$, $\mathbb{E}[h(X,Z,f)] = \theta^\top \mathbb{E}[h(X,Z,\phi)]$, where $h(X,Z,\phi)$ is short-hand for the vector with jth entry $h(X,Z,\phi_j)$. We can then write the imbalance in terms of a transformed feature space $h(X,Z\phi)$, giving a closed form that we can readily calculate by applying the linear functional ψ from (4.1) to the features ϕ :

$$\begin{aligned}\text{Imbalance}_{\mathcal{F}}(w) &\coloneqq \sup_{f\in\mathcal{F}} \Big\{ \mathbb{E}[w(X,Z)f(X,Z)] - \mathbb{E}[h(X,Z,f)] \Big\} \\ &= \sup_{\|\theta\|\le 1} \Big\{ \theta^\top \mathbb{E}[w(X,Z)\phi(X,Z)] - \theta^\top \mathbb{E}[h(X,Z,\phi)] \Big\} \\ &= \big\| \mathbb{E}[w(X,Z)\phi(X,Z)] - \mathbb{E}[h(X,Z,\phi)] \big\|_*.\end{aligned}$$

Consider the counterfactual mean estimand $\psi(m) = \mathbb{E}[m(X,0)|Z=1]$. We have

$$\text{Imbalance}_{\mathcal{F}}(w) = \big\| \mathbb{E}[w(X,Z)\phi(X,Z)] - \mathbb{E}[\phi(X,0)|Z=1] \big\|_*.$$

For simplicity let $\phi(x,z) = x$. Now we get $\text{Imbalance}_{\mathcal{F}}(w) = |\mathbb{E}[w(X,Z)X] - \mathbb{E}[X|Z=1]|$ and therefore, the balancing optimization problem finds weights w that reweight the total mean $\mathbb{E}[X]$ to approximate the conditional mean $\mathbb{E}[X|Z=1]$.

Equivalences of outcome regression and balancing weighting methods

For the special case of ℓ_2 kernel balancing, the balancing weights problem is numerically equivalent to directly estimating the conditional expectation $\mathbb{E}[Y_p|\Phi_p]$ via kernel ridge regression and applying the estimated coefficients to $\bar{\Phi}_q$. We begin with the special case of unregularized linear regression and then present the more general setting. We initially present the results assuming $d < n$ and that Φ_p has rank d, turning to the high-dimensional case with $d > n$ in Appendix B.2.

Linear regression. Ordinary least squares regression is equivalent to a weighting estimator that exactly balances the feature means. See [102] for discussion in the survey sampling literature; see [245], [1], [173], and [51] for relevant discussions in the causal inference literature.

In particular, let $\hat{w}_{\text{exact}}$ be the solution to the the exact balancing weights problem in Example 4 in the main text. Let $\hat{\beta}_{\text{ols}} = (\Phi_p^\top \Phi_p)^{-1} \Phi_p^\top Y_p$ be the OLS coefficients from the regression of Y_p on Φ_p. We then have the following numerical equivalence:

$$\begin{aligned}\hat{\mathbb{E}}[\Phi_q \hat{\beta}_{\text{ols}}] &= \hat{\mathbb{E}}[\hat{w}_{\text{exact}} \circ Y_p] \\ \bar{\Phi}_q \underbrace{(\Phi_p^\top \Phi_p)^{-1} \Phi_p^\top Y_p}_{\hat{\beta}_{\text{ols}}} &= \underbrace{\bar{\Phi}_q (\Phi_p^\top \Phi_p)^{-1} \Phi_p^\top}_{\frac{1}{n}\hat{w}_{\text{exact}}} Y_p,\end{aligned} \tag{B.8}$$

where the weights have the closed form $\frac{1}{n}\hat{w}_{\text{exact}} = \bar{\Phi}_q (\Phi_p^\top \Phi_p)^{-1} \Phi_p^\top$.

Ridge regression. This equivalence immediately extends to ridge regression [284, 130, 152].[3] Let $\hat{w}^{\delta}_{\ell_2}$ be the minimizer of the ℓ_2 balancing weights problem in Example 5 in the main text, with hyperparameter δ. Let

$$\hat{\beta}^{\delta}_{\text{ridge}} \coloneqq \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \Big\{ \|Y_p - \Phi_p \beta\|_2^2 + \delta \|\beta\|_2^2 \Big\} \tag{B.9}$$

be the ridge regression coefficients from least squares regression of Y_p on Φ_p. We then have the following numerical equivalence:

$$\begin{aligned} \hat{\mathbb{E}}[\Phi_q \hat{\beta}^{\delta}_{\text{ridge}}] &= \hat{\mathbb{E}}[\hat{w}^{\delta}_{\ell_2} \circ Y_p] \\ \overline{\Phi}_q \underbrace{(\Phi_p^\top \Phi_p + \delta I)^{-1} \Phi_p^\top Y_p}_{\hat{\beta}^{\delta}_{\text{ridge}}} &= \underbrace{\overline{\Phi}_q (\Phi_p^\top \Phi_p + \delta I)^{-1} \Phi_p^\top}_{\frac{1}{n}\hat{w}^{\delta}_{\ell_2}} Y_p, \end{aligned} \tag{B.10}$$

where the weights have the closed form $\frac{1}{n}\hat{w}^{\delta}_{\ell_2} = \overline{\Phi}_q (\Phi_p^\top \Phi_p + \delta I)^{-1} \Phi_p^\top$. Thus, the estimate from ridge regression is identical to the estimate using the ℓ_2 balancing weights. We leverage this equivalence in Section 4.3 below.

Kernel ridge regression. In general, the same equivalence holds in the non-parametric setting where ϕ is the feature map induced by an RKHS. In particular, let $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq r\}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) on $\mathcal{X} \times \mathcal{Z}$ with kernel $\mathcal{K}$, $\|\cdot\|_{\mathcal{H}}$ denotes the norm of the RKHS, and $r > 0$. Then the equivalence above holds for $\phi(x, z) \coloneqq \mathcal{K}(x, z, \cdot, \cdot)$. Although ϕ is typically infinite-dimensional, the Riesz Representer Theorem shows that the least squares regression and, equivalently, the balancing optimization problem have closed-form solutions. The least squares regression approach is *kernel ridge regression* and the weighting estimator is *kernel balancing weights* [see 126, 171]. [130] leverage this equivalence to analyze the asymptotic bias of kernel balancing weights. For further discussion of this equivalence see [116, 152].

Finally, we briefly mention some additional papers that discuss relevant equivalences. In the context of panel data, [271] establish connections between different forms of regression, which is especially relevant for our discussion of high-dimensional features in Appendix B.2. In addition, [191] provide an interesting alternative perspective by demonstrating that a large class of outcome regression estimators can be viewed as implicitly estimating the density ratio of the covariate distributions in the two treatment groups. Our results generalize and unify many of these existing numeric equivalences.

[3] See [123] for an interesting connection of this equivalence to experimental design. See [28] and [271] for related applications in the panel data setting.

B.2 Details for when $d > n$

In this section, we extend our results to the high-dimensional setting. In all that follows we will assume that $d > n$ and we will assume $\Phi_p^\top$ has rank n.[4] For $d = \infty$, we replace $\mathbb{R}^d$ with any infinite-dimensional Hilbert space $\mathcal{H}$ and we require the norm defining $\mathcal{F}$ to be the norm of the Hilbert space. In this case, it should be understood that $\Phi_p \in \mathcal{H}^n$.

Balancing weights when $d > n$

In the main text, recall that there are three equivalent versions of the balancing weights problem: the penalized, constrained, and automatic form with hyperparameters $\delta_1, \delta_2, \delta_3 \geq 0$ respectively. When $\Phi_p^\top \Phi_p$ is no longer invertible, a unique solution may fail to exist for certain values of these hyperparameters. We provide the relevant technical caveats here.

We begin by mentioning that for $\delta_1 > 0$, the penalized form of the balancing weights optimization problem is strictly convex, and therefore a unique solution exists, regardless of whether $d > n$. However, when $\delta_1 = 0$, there could potentially be infinite many solutions. In this setting, we choose the one with the minimum norm:

$$\begin{aligned} &\min_{w \in \mathbb{R}^n} \|w\|_2^2 \\ &\text{such that } \|w\Phi_p - \overline{\Phi}_q\|_*^2 = \min_v \|v\Phi_p - \overline{\Phi}_q\|_*^2. \end{aligned} \tag{B.11}$$

If we define $\delta_{\min} := \min_v \|v\Phi_p - \overline{\Phi}_q\|_*^2$, we see that the minimum norm solution in Equation (B.11) corresponds to a solution to the constrained form of balancing weights with $\delta_2 = \delta_{\min}$. Importantly, no solution exists for $\delta_2 < \delta_{\min}$, and we must make the additional restriction that $\delta_2 \geq \delta_{\min}$. In particular, no solution exists for $\delta_2 = 0$ and we cannot achieve exact balance; that is, for all w, $w\Phi_p \neq \bar{\Phi}_q$.

As in the penalized form, the automatic form is strictly convex and a unique solution exists for $\delta_3 > 0$. When $\delta_3 = 0$ we choose the minimum norm solution: by duality this will be equivalent to the minimum norm solution to the penalized problem [see 45].

Note that for $d = \infty$, each "row" of Φ_p is a vector in a Hilbert space $\mathcal{H}$. To solve the balancing weights problem computationally, we need a closed-form solution to the Hilbert space norm $\|\cdot\|_{\mathcal{H}}$. For example, this is a tractable computation when $\mathcal{H}$ is an RKHS.

[278] gives the automatic form of this problem. See also [312], [126], and [152].

Equivalences from Appendix B.1 when $d > n$

We now extend the equivalences from Appendix B.1 to the high-dimensional case. Let $\delta \geq 0$ be the hyperparameter for the penalized form of balancing weights — as we note above, this

[4] Alternatively, we could follow [21] and assume that, almost surely, the projection of Φ_p on the space orthogonal to any eigenvector of $\mathbb{E}[\Phi_p \Phi_p^\top]$ spans a space of dimension n. But as our results are numerical this has no real advantage.

is important to state explicitly, since the constrained form will not have a solution for all values of its hyperparameter. For hyperparameter $\delta > 0$, the solutions to ℓ_2 balancing weights and ridge regression are identical as in Equation (B.9) with no alterations; ridge regression works by default when $d > n$. On the other hand, when $\delta = 0$, there exist infinitely many solutions to the normal equations that define the solution to the OLS optimization problem. Since $(\Phi_p^\top \Phi_p)$ is not invertible, Equation (B.9) does not apply directly. Instead, we introduce the minimum norm solution to OLS:

$$\begin{aligned} &\min_{\beta \in \mathbb{R}^d} \|\beta\|_2^2 \\ &\text{such that } \|\Phi_p \beta - Y_p\|_2^2 = \min_{\beta'} \|\Phi_p \beta' - Y_p\|_2^2. \end{aligned}$$

See [21] for an extensive discussion of this optimization problem and its statistical properties as an OLS estimator. For $d > n$, the minimum norm solution is:

$$\hat{\beta}_{\text{ols}} := (\Phi_p^\top \Phi_p)^\dagger \Phi_p^\top Y_p = \Phi_p^\top (\Phi_p \Phi_p^\top)^{-1} Y_p,$$

where $A^\dagger$ denotes the pseudoinverse of a matrix A. Note that the definition holds in general.[5] In the low-dimensional setting in the main text, $(\Phi_p^\top \Phi_p)$ is invertible, and so $(\Phi_p^\top \Phi_p)^\dagger = (\Phi_p^\top \Phi_p)^{-1}$. The second equality holds only when $d > n$.

A version of Equation (B.8) holds between the minimum norm ℓ_2 balancing weights and minimum norm OLS estimators. Because the minimum norm ℓ_2 balancing weights do not achieve exact balance, we change the notation from $\hat{w}_{\text{exact}}$ to $\hat{w}^0_{\ell_2}$. In this setting, $\|\cdot\|_* = \|\cdot\|_2$ and the minimum-norm balancing weights problem in Equation (B.11) is also a minimum norm linear regression, but of $\bar{\Phi}_q \in \mathbb{R}^d$ on $\Phi_p^\top \in \mathbb{R}^{d \times n}$:

$$\hat{w}^0_{\ell_2} = \Phi_p^\top (\Phi_p^\top \Phi_p)^\dagger \bar{\Phi}_q = (\Phi_p \Phi_p^\top)^{-1} \Phi_p \bar{\Phi}_q.$$

Therefore, Equation (B.8) holds by replacing the inverse with the pseudo-inverse:

$$\begin{aligned} \hat{\mathbb{E}}[\Phi_q \hat{\beta}_{\text{ols}}] &= \hat{\mathbb{E}}[\hat{w}^0_{\ell_2} \circ Y_p] \\ \hat{\mathbb{E}}[\Phi_q \underbrace{(\Phi_p^\top \Phi_p)^\dagger \Phi_p^\top Y_p}_{\hat{\beta}_{\text{ols}}}] &= \hat{\mathbb{E}}[\underbrace{\bar{\Phi}_q (\Phi_p^\top \Phi_p)^\dagger \Phi_p^\top}_{\hat{w}^0_{\ell_2}} \circ Y_p], \\ \hat{\mathbb{E}}[\Phi_q \underbrace{\Phi_p^\top (\Phi_p \Phi_p^\top)^{-1} Y_p}_{\hat{\beta}_{\text{ols}}}] &= \hat{\mathbb{E}}[\underbrace{\bar{\Phi}_q \Phi_p^\top (\Phi_p \Phi_p^\top)^{-1}}_{\hat{w}^0_{\ell_2}} \circ Y_p]. \end{aligned}$$

Propositions 4.3.1 and 4.3.2 when $d > n$

The results in Propositions 4.3.1 and 4.3.2 apply to the setting where $d > n$ without any further alteration using the pseudo-inverse.

[5] For example, when $\Phi_p \in \mathcal{H}^n$ for an infinite-dimensional Hilbert space $\mathcal{H}$, $(\Phi_p^\top \Phi_p)^\dagger$ is guaranteed to exist, since it is bounded and has closed range.

Proof of Proposition 4.3.1.

$$Y_p^\top \Phi_p \hat{\theta}^\delta = Y_p^\top \Phi_p \Phi_p^\dagger \Phi_p \hat{\theta}^\delta = Y_p^\top \Phi_p (\Phi_p^\top \Phi_p)^\dagger \Phi_p^\top \Phi_p \hat{\theta}^\delta = \hat{\beta}_{\text{ols}} \hat{\Phi}_q,$$

where the first two equalities follow from the pseudoinverse identities $A = AA^\dagger A$ and $A^\dagger = (A^\top A)^\dagger A^\top$ for any matrix A. □

Likewise Proposition 4.3.2 holds exactly for $\hat{\beta}_{\text{ols}}$ defined with the pseudoinverse.

The RKHS Setting

The results for $d = \infty$ can be computed efficiently for reproducing kernel Hilbert spaces. For notational simplicity, we will consider the distribution shift setting from Example 3. Let $\mathcal{H}$ be a possibly-infinite-dimensional RKHS on $\mathcal{X}$ with kernel $\mathcal{K}$ and induced feature map via the representer theorem, $\phi : \mathcal{X} \to \mathcal{H}$ with $\phi(x) = \mathcal{K}(x, \cdot)$. Let $\|\cdot\|_{\mathcal{H}}$ denote the norm of $\mathcal{H}$. Let K_p be the matrix with entries $\mathcal{K}(x_i, x_j)$, where $x_i, x_j \in \mathcal{X}$ are the ith and jth entries of X_p. Then $\Phi_p \Phi_p^\top = K_p$ is invertible.

We will write out the versions of the main results for $\mathcal{F} = \mathcal{H}$ to demonstrate how to compute the corresponding results for RKHSs even though $d = \infty$. Denote the solution to the regularized least squares problem in $\mathcal{H}$ with $\lambda \geq 0$:

$$\hat{f}^\delta \coloneqq \text{argmin}_{f \in \mathcal{H}} \|f(X_p) - Y_p\|_2^2 + \lambda \|f\|_{\mathcal{H}}^2.$$

This is equivalent to the following problem by the representer theorem:

$$\begin{aligned} \hat{\beta}_{\mathcal{H}}^\delta &\coloneqq \text{argmin}_{\beta \in \mathbb{R}^n} \|K\beta - Y_p\|_2^2 + \lambda \beta K \beta \\ &= (K_p + \lambda I)^{-1} Y_p. \end{aligned}$$

Let $K_{x,p} \in \mathbb{R}^n$ be the row vector with entries $\mathcal{K}(x, x_i)$ where x is an arbitrary element of $\mathcal{X}$ and x_i is the ith entry of X_p. Then for any element $x \in \mathcal{X}$, $\hat{f}^\delta(x) = K_{x,p} \hat{\beta}_{\mathcal{H}}^\delta$. In particular, let define $K_{q,p}$ as the matrix with (i,j)th entry $\mathcal{K}(x_{qi}, x_{pj})$ where x_{qi} is the ith sample from the target population and x_{qj} is the jth entry of X_p. Furthermore, define $\bar{K}_{q,p} \coloneqq \hat{\mathbb{E}}[K_{p,q}] \in \mathbb{R}^n$. Then, for any solution $\hat{w}_{\mathcal{H}}^\delta$ to the penalized form of balancing weights with function class $\mathcal{F}$ and hyperparameter $\delta \geq 0$:

$$\hat{w}_{\mathcal{H}}^\delta = (K_p + \lambda I)^{-1} \bar{K}_{q,p}. \tag{B.12}$$

The proof follows from the closed-form of $\text{Imbalance}_{\mathcal{H}}(w)$, known as the Maximum Mean Discrepancy (MMD) [116]; see, e.g., [130, 152, 45].

With these preliminaries, we immediately have the following equivalence from [130], which generalizes Appendix B.1 to the RKHS case:

$$\begin{aligned} \hat{\mathbb{E}}[K_{q,p} \hat{\beta}_{\mathcal{H}}^\delta] &= \hat{\mathbb{E}}[\hat{w}_{\mathcal{H}}^\delta \circ Y_p] \\ \hat{\mathbb{E}}[K_{q,p} \underbrace{(K_p + \delta I) Y_p}_{\hat{\beta}_{\mathcal{H}}^\delta}] &= \hat{\mathbb{E}}[\underbrace{\bar{K}_{q,p} (K_p + \delta I)^{-1}}_{\hat{w}_{\mathcal{H}}^\delta} \circ Y_p]. \end{aligned} \tag{B.13}$$

Likewise, we have the following form for Proposition 4.3.1. Define $\hat{K}_{q,p} \coloneqq \hat{w}_{\mathcal{H}}^{\delta T} K_p$. Then, for any $\delta \geq 0$:

$$\hat{\mathbb{E}}[\hat{w}_{\mathcal{H}}^{\delta} \circ Y_p] = \hat{\mathbb{E}}[\hat{K}_{q,p}\hat{\beta}_{\mathcal{H}}^{0}].$$

The resulting expression for Proposition 4.3.2 is:

$$\hat{\mathbb{E}}[\hat{w}_{\mathcal{H}}^{\delta} \circ Y_p] + \hat{\mathbb{E}}\left[\left(K_{q,p} - \hat{K}_{q,p}^{\delta}\right)\hat{\beta}_{\mathcal{H}}^{\lambda}\right] = \hat{\mathbb{E}}[K_{q,p}\hat{\beta}_{\text{aug}}],$$

where the jth element of $\hat{\beta}_{\text{aug}}$ is:

$$\hat{\beta}_{\text{aug},j} \coloneqq \left(1 - a_j^{\delta}\right)\hat{\beta}_{\mathcal{H},j}^{\lambda} + a_j^{\delta}\hat{\beta}_{\mathcal{H},j}^{0}$$

$$a_j^{\delta} \coloneqq \frac{\widehat{\Delta}_j^{\delta}}{\Delta_j},$$

where $\Delta_j = \bar{K}_{q,p,j} - \bar{K}_{p,j}$ and $\widehat{\Delta}_j^{\delta} = \hat{K}_{q,p,j}^{\delta} - \bar{K}_{p,j}$ with $\bar{K}_{p,j} \coloneqq \hat{\mathbb{E}}[K_p]$.

Identical versions for the RKHS setting apply to Section 4.4. These follow directly from the expressions above so we will omit repeating them explicitly. Importantly, equivalent versions for ℓ_∞ balancing in Section 4.5 do *not* follow immediately because an infinite dimensional vector space equipped with the ℓ_1 norm does not form a Hilbert space. We conjecture that such extensions could be constructed using the Reproducing Kernel Banach Space literature [190].

B.3 Simulation Study Details

Setup

We consider 36 different data generating processes (DGPs) for our simulation study. For each of them, we compare an oracle baseline with three feasible hyperparameter tuning schemes.

For the remainder of this section, we we say "numerically optimize", we mean using the scipy.optimize.minimize solver with tolerance 1e-12. In particular, we pre-compute the SVD of the covariance matrix before solving, so that we can compute the pseudoinverse for all involved expressions in closed form without performing traditional matrix inversion during optimization.

For the oracle hyperparameters, we compute λ^* by numerically optimizing the in-distribution mean squared error for ridge regression from Section 4.6. We then fix $\lambda = \lambda^*$ and then numerically optimize the expression in Proposition 4.6.1 to get the MSE-optimal δ^*

The three feasible hyperparameter tuning schemes we consider, by contrast, use a particular draw of Y_p. In all cases, we choose λ by cross-validating a ridge outcome model. Then we considering choosing δ by: (1) cross-validating balance, (2) cross-validating the Riesz loss, and (3) setting δ equal to the cross-validated ridge λ. In all cases, cross-validation is performed 5-fold via numerical optimization as described above instead of using a grid.

Note that the oracle hyperparameters are only optimal in a setting where we have to pick a single λ and δ for each draw of Y_p. Cross-validation picks a new λ for each Y_p and so can theoretically outperform the oracle. This happens very rarely, but a non-zero amount of the time. Overall, the results in Table 4.1 suggest that this is still a very good baseline. Finally, note that we could have picked λ by cross validation for the oracle, and then solved for the optimal δ^* separately for each Y_p, fixing the CV value of λ. However, we cannot guarantee that the resulting δ will have any optimality properties, since the mean squared error expression is derived by averaging over Y_p draws for fixed values of λ and δ. But this could be an interesting follow up experiment.

In all cases, we take 1000 draws of Y_p and then compute the squared error and take their average as a monte carlo estimate of the mean squared error.

Synthetic DGPs

To compute the oracle λ^* and δ^* from Section 4.6, we need: the true coefficients, β_0, the population covariance matrix, $\mathbb{E}[\Phi_p^T\Phi_p]$, a sample covariance matrix $\hat{\Sigma}$, the conditional variance σ^2, and the target covariance mean, $\mathbb{E}[\Phi_q]$. So for each DGP, we need to specify these five objects.

We consider synthetic DGPs with three basic setups. They all use $n = 2000$ and $d = 50$. For each of the three setups, we draw a random β_0, that is the absolute value of a d-dimensional standard normal, that is then normalized to have 2-norm equal to 1. The three setups each generate a popuation covariance matrix in roughly the same way. In each case, we choose a maximum and minimum eigenvalue, $\eta_{\min}$ and $\eta_{\max}$ respectively. We then generate an equally space grid between $\eta_{\min}^{1/c}$ and $\eta_{\min}^{1/c}$ for some curvature constant c. We choose the eigenvalues of the covariance matrix to be the numbers in this grid raised to the cth power. We then draw a random eigenvector matrix from the special orthogonal group, U, and form the covariance matrix from the eigenvectors and eigenvalues in the standard way. Next, we draw an Φ_p with $n = 2000$ samples from a mean-zero normal distribution with this covariance matrix, and compute $\hat{\Sigma} = \Phi_p^T\Phi_p/n$.

The three basic setups differ in teh choice of $\eta_{\min}, \eta_{\max}$, and c. For setting 1 we choose $\eta_{\min} = 1e-4$, $\eta_{\max} = 3$ and $c = 5000$. For setting 2 we choose $\eta_{\min} = 1e-8, \eta_{\max} = 3$, and $c = 5000$. For setting 3, we choose $\eta_{\min} = 1e-10, \eta_{\max} = 5, c = 10$.

Then for each of these basic setups, we create 10 DGPs, by consider all combinations of a list of $\mathbb{E}[\Phi_q]$ and σ^2. We use $\sigma^2 \in \{0.1, 2\}$. For $\mathbb{E}[\Phi_q]$ we use: the vector of all 0.1, the vector of all 2, and then 3 vectors chosen randomly uniformly between -1 and 1 which are then scaled to have norm 1. Thus the total of $2 \times 5 = 10$ DGPs for each setup.

Semi-Synthetic DGPs

We then also use semi-synthetic DGPs based on Lalonde Long and IHDP Long. For each of these datasets, we recenter Φ_p and Y_p to have mean-zero. Then we choose β_0 to be the coefficients from cross-validated ridge regression of Y_p on Φ_p. We let σ^2 be the variance of

the residuals from this regression, and we choose the population covariance matrix to be $\Phi_p^T \Phi_p / n$. Next, we *redraw* a new matrix of samples from a mean-zero normal distribution with this covariance matrix, and compute the sample covariance $\hat{\Sigma}$ from these *new* samples. For targets, we use the actual $\mathbb{E}[\Phi_q]$, but also include two perturbed versions where all even elements of $\mathbb{E}[\Phi_q]$ are either increased or decreased by a proportion of the norm of $\mathbb{E}[\Phi_q]$. For IHDP, the perturbation is 1/10 times the norm, for Lalonde, the perturbation is 1/100 times the norm.

So since for each semi-synthetic setting we have three values of $\mathbb{E}[\Phi_q]$ and one value of σ^2, that corresponds to 6 DGPs each, for a total of 36 together with the synthetic DGPs.

B.4 Additional Proofs

Closed forms for ℓ_2 and exact balancing weights. We derive the closed form for ℓ_2 balancing weights with parameter δ (with exact balance following as a special case). The optimization problem:

$$\min_{w\in\mathbb{R}^n} \|w^\top \Phi_p - \bar{\Phi}_q\|_2^2 + \delta\|w\|_2^2 = w^\top \Phi_p \Phi_p^\top w - 2w^\top \Phi_p \bar{\Phi}_q + \delta w^\top w$$

has first order condition:

$$2(\Phi_p \Phi_p^\top + \delta I_n)w - 2\Phi_p \bar{\Phi}_q = 0,$$

which gives the solution:

$$\begin{aligned} w^* &= (\Phi_p \Phi_p^\top + \delta I_n)^\dagger \Phi_p \bar{\Phi}_q \\ &= \Phi_p (\Phi_p^\top \Phi_p + \delta I_d)^\dagger \bar{\Phi}_q. \end{aligned}$$

□

Proof of Proposition 4.4.1. We apply Proposition 4.3.2. We have $a_j^\delta = \hat{\Phi}_{q,j}^\delta / \bar{\Phi}_{q,j}^\delta$. Then:

$$\hat{\Phi}_q^\delta = \hat{w}_{\ell_2}^\delta \Phi_p = \overline{\Phi}_q (\Phi_p^\top \Phi_p + \delta I)^{-1} \Phi_p^\top \Phi_p.$$

Since we have assumed that $\Phi_p^\top \Phi_p$ is diagonal, with jth diagonal entry, σ_j^2, we have:

$$\hat{\Phi}_{q,j}^\delta = \left(\frac{\sigma_j^2}{\sigma_j^2 + \delta}\right) \bar{\Phi}_{q,j}.$$

Plugging this back into a_j^δ completes the proof. □

Proof of Proposition 4.4.3. Applying Proposition Proposition 4.4.1:

$$\begin{aligned}\hat{\beta}_{\ell_2,j} &= \left(\frac{\sigma_j^2}{\sigma_j^2+\delta}\right)\hat{\beta}_{\text{ols,j}} + \left(\frac{\delta}{\sigma_j^2+\delta}\right)\hat{\beta}^{\lambda}_{\text{ridge,j}} \\ &= \left(\frac{\sigma_j^2}{\sigma_j^2+\delta}\right)\hat{\beta}_{\text{ols,j}} + \left(\frac{\delta}{\sigma_j^2+\delta}\right)\left(\frac{\sigma_j^2}{\sigma_j^2+\lambda}\right)\hat{\beta}_{\text{ols},j} \\ &= \frac{\sigma_j^2(\sigma_j^2+\lambda+\delta)}{(\sigma_j^2+\delta)(\sigma_j^2+\lambda)}\hat{\beta}_{\text{ols,j}}\end{aligned}$$

Then taking:

$$\frac{\sigma_j^2(\sigma_j^2+\lambda+\delta)}{(\sigma_j^2+\delta)(\sigma_j^2+\lambda)} = \frac{\sigma_j^2}{\sigma_j^2+\gamma_j}$$

and solving for γ_j gives:

$$\gamma_j := \frac{\delta\lambda}{\sigma_j^2+\lambda+\delta}$$

which completes the proof. □

Proof of Proposition 4.5.1. We begin with the constrained form of the balancing problem:

$$\begin{aligned}&\min_{w\in\mathbb{R}^n} \|w\|_2^2 \\ &\text{such that } \left\|w\Phi_p - \bar{\Phi}_q\right\|_\infty \leq \delta.\end{aligned}$$

Note that we can rewrite the norm constraint as two vector-valued linear constraints:

$$\begin{aligned}w\Phi_p &\preceq \bar{\Phi}_q + \delta \\ -w\Phi_p &\preceq -\bar{\Phi}_q + \delta,\end{aligned}$$

which results in the Lagrangian,

$$\mathcal{L}(w,\mu,\nu) = \|w\|_2^2 + \mu^\top\left(w\Phi_p - \bar{\Phi}_q - \delta\right) - \nu^\top\left(w\Phi_p - \bar{\Phi}_q + \delta\right).$$

The first-order conditions for the optimal w^*, μ^*, ν^* are:

$$\begin{aligned}w^* &= -\frac{1}{2}\left(\Phi_p\mu^* - \Phi_p\nu^*\right) \\ \mu_j^*\left(w^*\Phi_{p,j} - \bar{\Phi}_{q,j} - \delta\right) &= 0, \forall j \\ \nu_j^*\left(w^*\Phi_{p,j} - \bar{\Phi}_{q,j} + \delta\right) &= 0, \forall j \\ \mu_j^*, \nu_j^* &\geq 0, \forall j\end{aligned}$$

plus the linear constraints on $w^*\Phi_p$. Note that the linear constraints plus the complimentary slackness conditions imply that one of three mutually-exclusive cases holds for each covariate.

Case 1: $w^*\Phi_{p,j} = \bar{\Phi}_{q,j} - \delta$, in which case $\mu_j^* = 0$. Case 2: $w^*\Phi_{p,j} = \bar{\Phi}_{q,j} + \delta$, in which case $\nu_j^* = 0$. Or Case 3: $w^*\Phi_{p,j} \in (\bar{\Phi}_{q,j} - \delta, \bar{\Phi}_{q,j} + \delta)$, in which case $\mu_j^* = \nu_j^* = 0$.

Define:

$$\theta_j^* \coloneqq \begin{cases} 0 & \text{if } w^*\Phi_{p,j} \in (\bar{\Phi}_{q,j} - \delta, \bar{\Phi}_{q,j} + \delta) \\ -\mu_j^*/2 & \text{if } w^*\Phi_{p,j} = \bar{\Phi}_{q,j} + \delta \\ \nu_j^*/2 & \text{if } w^*\Phi_{p,j} = \bar{\Phi}_{q,j} - \delta. \end{cases}$$

Then we have $w^* = \Phi_p\theta^*$ from the first-order condition, and thus $w^*\Phi_p = (\Phi_p^\top\Phi_p)\theta^*$. Using the fact that $(\Phi_p^\top\Phi_p)$ is diagonal, we get $w^*\Phi_{p,j} = \sigma_j^2\theta_j^*$.

Finally, we can plug this into the three cases that define θ^*. First, $\theta_j^* = 0$ when

$$\begin{aligned} & \sigma_j^2\theta_j^* \in (\bar{\Phi}_q - \delta, \bar{\Phi}_q + \delta) \\ \implies & 0 \in (\bar{\Phi}_q - \delta, \bar{\Phi}_q + \delta) \\ \implies & \bar{\Phi}_q \in (-\delta, \delta). \end{aligned}$$

Second, $\theta_j^* = -\mu_j^*/2$ when $\sigma_j^2\theta_j^* = \bar{\Phi}_{q,j} + \delta$, which implies $\mu_j^* = -2(\bar{\Phi}_{q,j} + \delta)/\sigma_j^2$. We then apply the dual variable constraint:

$$\begin{aligned} & \mu_j^* \geq 0 \\ \implies & -2(\bar{\Phi}_{q,j} + \delta)/\sigma_j^2 \geq 0 \\ \implies & \bar{\Phi}_{q,j} \leq -\delta. \end{aligned}$$

Third, $\theta_j^* = \nu_j^*/2$ when $\sigma_j^2\theta_j^* = \bar{\Phi}_{q,j} - \delta$, which implies $\nu_j^* = 2(\bar{\Phi}_{q,j} - \delta)/\sigma_j^2$. We then apply the dual variable constraint:

$$\begin{aligned} & \nu_j^* \geq 0 \\ \implies & 2(\bar{\Phi}_{q,j} - \delta)/\sigma_j^2 \geq 0 \\ \implies & \bar{\Phi}_{q,j} \geq \delta. \end{aligned}$$

Putting the cases together we get:

$$\theta_j^* \coloneqq \begin{cases} 0 & \text{if } \bar{\Phi}_q \in (-\delta, \delta) \\ (\bar{\Phi}_{q,j} + \delta)/\sigma_j^2 & \text{if } \bar{\Phi}_{q,j} \leq -\delta. \\ (\bar{\Phi}_{q,j} - \delta)/\sigma_j^2 & \text{if } \bar{\Phi}_{q,j} \geq \delta. \end{cases}$$

This is exactly the soft-thresholding operator, which completes the proof. □

Proof of Proposition 4.5.2. To obtain this result from the general form of $a_j^\delta = \widehat{\Delta}_j/\Delta_j$ in Proposition 4.3.2, notice that the implied feature shift, $\widehat{\Delta}_j = \hat{\Phi}_{q,j}^\delta - \overline{\Phi}_{p,j} = \mathcal{T}_\delta(\overline{\Phi}_{q,j} - \overline{\Phi}_{p,j})$ is:

$$\widehat{\Delta}_j = \begin{cases} 0 & \text{if } |\Delta_j| < \delta \\ \Delta_j - \delta & \text{if } \Delta_j > \delta \\ \Delta_j + \delta & \text{if } \Delta_j < -\delta \end{cases}.$$

Thus, for instance, $\frac{\widehat{\Delta}_j}{\Delta} = \frac{\Delta_j - \delta}{\Delta_j} = 1 - \frac{\delta}{\Delta_j}$ when $\Delta_j > \delta$. □

Proof of Proposition 4.5.3. The result follows immediately from Proposition 4.5.2. □

Proof of Proposition B.5.1. Rewriting the definition of γ_n with $\lambda_n = \delta_n$, we have

$$\gamma_n = \frac{\lambda_n^2}{\sigma^2 + 2\lambda_n^2} = \frac{1}{\sigma^2 \lambda_n^{-2} + 2\lambda_n^{-1}}.$$

Because $\sigma^2 x^2 + 2x = O(x^2)$ as a function of x, and because λ_n^{-1} is monotonically increasing, $\sigma^2 \lambda_n^{-2} + 2\lambda_n^{-1} = O(\lambda_n^{-2})$. And $\lambda_n^{-2} = O(\sigma^2 \lambda_n^{-2} + 2\lambda_n^{-1})$ because $\sigma^2 \geq 0$ and $\lambda_n^{-1} > 0$. Thus $\sigma^2 \lambda_n^{-2} + 2\lambda_n^{-1} \asymp \lambda_n^{-2}$.

Finally, note that for any two functions of n, f_n and g_n,

$$f_n \asymp g_n \iff f_n^{-1} \asymp g_n^{-1},$$

and therefore,

$$\gamma_n \asymp \lambda_n^2.$$

□

B.5 Additional Details for Asymptotic Results

Our setup for the RKHS follows [278]. First, assume that the space $\mathcal{X} \times \mathcal{Z}$ is Polish. Let $\mathcal{H}$ be an RKHS on $\mathcal{X} \times \mathcal{Z}$ with corresponding kernel k satisfying standard regularity conditions [278, Assumption 5.2] and let η_j, φ_j denote the eigenvalues and eigenfunctions respectively of its kernel integral operator under p. Next, assume that the eigenvalues satisfy the decay condition $\eta_j \leq C j^{-b}$ for some $b > 1$ and a constant C. The parameter b encodes information on the effective dimension of $\mathcal{H}$. For a bounded kernel, $b > 1$ [94]: the case where $b = \infty$ corresponds to a finite-dimensional RKHS; for the case with $1 < b < \infty$, the η_j must decay at a polynomial rate.

We then assume that for some $c \in [1, 2]$, the outcome function $m(x, z)$ belongs to the set:

$$\mathcal{H}^c := \left\{ f = \sum_{j=1}^{\infty} a_j \varphi_j : \sum_{j=1}^{\infty} \frac{a_j^2}{\eta_j^c} < \infty \right\} \subset \mathcal{H}, \tag{B.14}$$

where c encodes additional *smoothness* of the conditional expectation. If $c = 1$, then by the spectral decomposition of the RKHS, Equation (B.14) is equivalent to requiring $m \in \mathcal{H}$; choosing larger values of c corresponds to m being a smoother element of $\mathcal{H}$, with a "saturation effect" kicking in for $c > 2$ [23]. Varying b (the effective dimension of the RKHS) and c (the additional smoothness of the outcome function) changes the optimal rates for regression, with larger values of both corresponding to faster rates of convergence.

Finally, we assume that the Riesz representer, $\alpha(x, z)$, of our linear functional estimand also belongs to $\mathcal{H}^c$. Under these conditions, [278] demonstrates that an augmented estimator

combining kernel balancing weights and a kernel ridge regression base learner is root-n consistent and asymptotically normal.

Following [48], Theorems 5.1 and 5.2 of [278] use hyperparameter schedules for λ and δ, which depend on the effective dimension b and smoothness c:

$$\lambda_n = \delta_n = \begin{cases} n^{-1/2} & \text{if } b = \infty \\ n^{-\frac{b}{bc+1}} & \text{if } b \in (1,\infty), \quad c \in (1,2] \\ (n/\log(n))^{-b/(b+1)} & \text{if } b \in (1,\infty), \quad c = 1 \end{cases},$$

We can compute the implied augmented hyperparameter sequence γ_n using the following proposition.

Proposition B.5.1. *Let $\lambda_n > 0$ be any monotonically decreasing function of n and let $\delta_n = \lambda_n$. Then:*

$$\gamma_n := \frac{\lambda_n \delta_n}{\sigma^2 + \lambda_n + \delta_n} \asymp \lambda_n^2.$$

The standard ridge regression case corresponds to the finite-dimensional setting with $b = \infty$. When $c > 1$, the optimal rate for λ_n is $n^{-\frac{b}{bc+1}}$; the implied hyperparameter is then order $n^{-2b/bc+1} \in (n^{-2}, n^{-2/3})$ for $c \in (1,2]$ and $b \in (1,\infty)$. Whether or not this smooths more than n^{-1} therefore depends on the relationship between the effective dimension b and the smoothness c. In particular, the implied hyperparameter goes to zero at a slower rate than n^{-1} whenever $c \geq 2 - \frac{1}{b}$. It is unclear whether the rates we find here are the only undersmoothed rates that will yield efficiency for fixed b and c; we leave a thorough investigation to future work.

Appendix C

Additional Materials: Sensitivity Analysis

C.1 Proofs for Section 6.3, Marginal MDP

First, we give a reminder for the main notational device used for the following proofs. We will use P_π and $\mathbb{E}_\pi$ to denote the joint probabilities (and expectations thereof) of the random variables $S_t, U_t, A_t, \forall t$ in the underlying MDP running policy π. That is, $P_\pi(S_t, A_t, S_{t+1})$ is a joint distribution obtained from an MDP running policy π (and we can analogous obtain conditionals, other marginals, etc). In particular, this notation differs from notation about transitions controlled under π, i.e. $P_\pi(S_{t+1} \mid S_t)$, where π indicates expectations over π.

For example, in general due to the unobserved confounders, we will have that $P_{\pi^e}(S_{t+1}|S_t = s, A_t = a) \neq P_{\pi^b}(S_{t+1}|S_t = s, A_t = a)$. Since we are not conditioning on U_t, without further assumptions, these are not Markovian, and so it's important to keep in mind that S_{t+1} has a generally different distribution under π^e than it does under by π^b even after conditioning on S_t and A_t.

Under Assumption 3, the setting with a policy π^e that only depends on the observed state is equivalent to a marginal MDP over the observed state alone:

Proof. Proof of Proposition 6.3.1 First, note that for any π^e and all t, s, a:

$$\begin{aligned} P_{\pi^e}(S_{t+1}|S_t = a, A_t = a) &= \int_{\mathcal{U}} P_{\pi^e}(S_{t+1}|S_t = s, A_t = a, U_t = u) P_{\pi^e}(U_t = u|S_t = s, A_t = a) du \\ &= \int_{\mathcal{U}} P_{\pi^e}(S_{t+1}|S_t = s, A_t = a, U_t = u) P_{\pi^e}(U_t = u|S_t = s) du, \end{aligned}$$

where the second equality uses the fact that π^e is independent of U_t. To complete the proof, we need to show that this resulting value is the same for all possible π^e and equals Equation (6.2). This is always true for the first probability, because it is equal to the transition probability $P_{\pi^e}(S_{t+1}|S_t, A_t = U_t) = P_t(S_{t+1}|S_t, A_t, U_t)$ from the definition of the

full-information MDP. Under Assumption 3, the second term can also be written as a transition probability: $P_{\pi^e}(U_t|S_t) = P_{\pi^e}(U_t|S_t, S_{t-1}, A_{t-1}, U_{t-1}) = P_t(U_t|S_t, S_{t-1}, A_{t-1}, U_{t-1})$. □

The above proof establishes what probabilities are independent of the policy and are only a function of the transition probabilities $P_t(S_{t+1}, U_{t+1}|S_t, U_t, A_t)$. We could use this same idea to prove a more general version of Proposition 6.3.1 that *places assumptions only on the observed states and actions*, but at the cost of substantially more complexity.

For any t, let $H_t = \{S_j, A_j : j \leq t\}$ be the history of the *observed* state and actions up to time t. In the rest of this section, we will use shorthands like $P_\pi(s_{t+1}|s_t, a_t, h_{t-1}) \coloneqq P_\pi(S_{t+1}|S_t = s_t, A_t = a_t, H_{t-1} = h_{t-1})$ whenever clear from the text.

We only require the following Markov assumption on observed states and actions:

Assumption 12 (Observable Markov Property)**.** *For all π and for all t, s, a, h,*

$$P_\pi(s_{t+1}|s_t, a_t, h_{t-1}) = P_\pi(s_{t+1}|s_t, a_t)$$
$$P_\pi(a_t|s_t, h_{t-1}) = P_\pi(a_t|s_t).$$

Note that Assumption 3 implies Assumption 12:

$$\begin{aligned} P_\pi(s_{t+1}|s_t, a_t, h_{t-1}) &= \int_{\mathcal{U}} P_\pi(s_{t+1}|s_t, u_t, a_t, h_{t-1}) P_\pi(u_t|s_t, a_t, h_{t-1}) du \\ &= \int_{\mathcal{U}} P_\pi(s_{t+1}|s_t, u_t, a_t) P_\pi(u_t|s_t, a_t) du \\ &= P_\pi(s_{t+1}|s_t, a_t). \end{aligned}$$

We now prove the following general version of Proposition 6.3.1:

Proposition C.1.1 (Marginal MDP, General)**.** *Let χ^{marg} be the marginal distribution of χ over the observed state. Given Assumption 12, there exists $P_t^{marg} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ such that for any policies π^e and $\pi^{e'}$ that do not depend on U_t and for all s, a, t:*

$$P_t^{marg}(s, a) = P_{\pi^e}(S_{t+1}|S_t = s, A_t = a) = P_{\pi^{e'}}(S_{t+1}|S_t = s, A_t = a).$$

Furthermore, we can define a new MDP, $(\mathcal{S}, \mathcal{A}, R, T^{marg}, \chi^{marg}, H)$, with probabilities under policy π^e denoted $P_{\pi^e}^{marg}$ such that

$$P_{\pi^e}^{marg}(S_0, A_0, ...S_H, A_H) = P_{\pi^e}(S_0, A_0, ...S_H, A_H).$$

The proof uses the following two lemmas:

Lemma C.1.2 (Conditional Mean Independence with Respect to Transitions)**.** *Given Assumption 12,*

$$\int_{\mathcal{U}} P_\pi(u_t|s_t, h_{t-1}) P_\pi(s_{t+1}|s_t, a_t, u_t) du = \int_{\mathcal{U}} P_\pi(u_t|s_t) P_\pi(s_{t+1}|s_t, a_t, u_t) du.$$

Proof. Proof of Lemma C.1.2 Note that the full-information state transitions are Markovian by the definition of an MDP:

$$P_\pi(s_{t+1}|s_t, a_t, u_t, h_{t-1}) = P_\pi(s_{t+1}|s_t, a_t, u_t).$$

The lemma then follows by applying the tower property to both sides of Assumption 3. □

Lemma C.1.3. *Given Assumption 12, for any two π^e and $\pi^{e'}$ which do not depend on U, $\forall s, a,$ and t:*

$$P_{\pi^e}(s_{t+1}|s_t, a_t) = P_{\pi^{e'}}(s_{t+1}|s_t, a_t).$$

Proof. Proof of Lemma C.1.3 The proof proceeds by mutual induction on the statement above and the following statement:

$$P_{\pi^e}(u_t|s_t, h_{t-1}) = P_{\pi^{e'}}(u_t|s_t, h_{t-1}).$$

We will consider π^e and demonstrate the equality with $\pi^{e'}$ by showing that the relevant qualities do not depend on π^e. First, consider $t = 0$. From the definition of the initial state distribution,

$$P_{\pi^e}(u_0|s_0) = \chi(u_0|s_0).$$

which holds for all π^e.

From the definition of the MDP, $P_\pi(s_{t+1}|s_t, u_t, a_t) = P_t(s_{t+1}|s_t, u_t, a_t)$ for any π. Then we have:

$$\begin{aligned}
P_{\pi^e}(s_1|s_0, a_0) &= \int_{\mathcal{U}} P_{\pi^e}(u_0|s_0, a_0) P_{\pi^e}(s_1|s_0, a_0, u_0) du \\
&= \int_{\mathcal{U}} P_{\pi^e}(u_0|s_0, a_0) P_0(s_1|s_0, a_0, u_0) du \\
&= \int_{\mathcal{U}} P_{\pi^e}(u_0|s_0) P_0(s_1|s_0, a_0, u_0) du \\
&= \int_{\mathcal{U}} \chi(u_0|s_0) P_0(s_0, a_0, u_0) du,
\end{aligned}$$

where the third equality uses the fact that π^e does not depend on U. This equality also holds for all π^e and so we have proven the base case.

Now we consider a general t:

$$\begin{aligned}
P_{\pi^e}(u_t|s_t, h_{t-1}) &= \int_{\mathcal{U}} P_{\pi^e}(u_{t-1}|s_t, h_{t-1}) P_{\pi^e}(u_t|s_t, h_{t-1}, u_{t-1}) du \\
&= \int_{\mathcal{U}} P_{\pi^e}(u_{t-1}|s_{t-1}, h_{t-2}) \frac{P_{\pi^e}(s_t, u_t|s_{t-1}, u_{t-1}, a_{t-1})}{P_{\pi^e}(s_t|s_{t-1}, a_{t-1})} du,
\end{aligned}$$

where the second equality follows from applying Bayes rule to both probabilities in the second line. By the inductive hypothesis, $P_{\pi^e}(u_{t-1}|s_{t-1}, h_{t-2})$ does not depend on π^e. The transition probabilities $P_{\pi^e}(s_t, u_t|s_{t-1}, u_{t-1}, a_{t-1})$ do not depend on π^e. And by the inductive hypothesis, $P_{\pi^e}(s_t|s_{t-1}, a_{t-1})$ does not depend on π^e. Therefore, $P_{\pi^e}(u_t|s_t, h_{t-1})$ does not depend on π^e.

Finally,

$$\begin{aligned}
P_{\pi^e}(s_{t+1}|s_t, a_t) &= \int_{\mathcal{U}} P_{\pi^e}(u_t|s_t, a_t) P_{\pi^e}(s_{t+1}|s_t, u_t, a_t) du \\
&= \int_{\mathcal{U}} P_{\pi^e}(u_t|s_t, a_t) P_t(s_{t+1}|s_t, u_t, a_t) du \\
&= \int_{\mathcal{U}} P_{\pi^e}(u_t|s_t) P_t(s_{t+1}|s_t, u_t, a_t) du \\
&= \int_{\mathcal{U}} P_{\pi^e}(u_t|s_t, h_{t-1}) P_t(s_{t+1}|s_t, u_t, a_t) du
\end{aligned}$$

where the third equality follows from the fact that π^e does not depend on U and the fourth equality follows from Lemma 1. We have already shown that $P_t(s_{t+1}|_{\pi^e}(u_t|s_t, h_{t-1})$ does not depend on π^e which concludes the proof.

□

Proof. Proof of Proposition C.1.1 Define $P_t^{\text{marg}} = P_{\pi^e}(s_{t+1}|s_t, a_t)$, which by Lemma C.1.3 is the same for any π^e. From the conditional independence structure of the original MDP together with Assumption 12, we have

$$\begin{aligned}
P_{\pi^e}(S_0, A_0, ..., S_{T-1}, A_{T-1}) &= P_{\pi^e}(S_0) P_{\pi^e}(A_0|S_0) \prod_{t=1}^{T-1} P_{\pi^e}(A_t|S_t) P_{\pi^e}(S_t|S_{t-1}, A_{t-1}) \\
&= \chi^M(S_0) \pi^e(A_0|S_0) \prod_{t=1}^{T-1} \pi^e(A_t|S_t) P_t^{\text{marg}}(S_t|S_{t-1}, A_{t-1}) \\
&= P_{\pi^e}^M(S_0, A_0, ...S_{T-1}, A_{T-1}).
\end{aligned}$$

□

Confounding for Regression

Proof. Proof of Proposition 6.3.2

$$
\begin{aligned}
&\mathbb{E}_{P_t}[f(S_t, A_t, S_{t+1})|S_t = s, A_t = a] \\
&\quad = \int_{\mathcal{S}} f(s, a, s') P_t(s'|s, a) ds' \\
&\quad = \int_{\mathcal{S}} f(s, a, s') \left(\int_{\mathcal{U}} P_t(u|s) P_t(s'|s, a, u) du \right) ds' \\
&\quad = \int_{\mathcal{S}} f(s, a, s') \left(\int_{\mathcal{U}} \frac{\pi^b(a|s)}{\pi^b(a|s, u)} P_{\text{obs}}(U_t = u|S_t = s, A_t = a) P_t(s'|s, a, u) du \right) ds' \\
&\quad = \mathbb{E}_{\text{obs}} \left[\frac{\pi^b(A_t|S_t)}{\pi^b(A_t|S_t, U_t)} f(S_t, A_t, S_{t+1}) \middle| S_t = s, A_t = a \right].
\end{aligned}
$$

□

We conjecture that the same result would hold replacing Assumption 3 with Assumption 12, but it would require showing that

$$
\int_{\mathcal{U}} P_{\pi^e}(u|s) P_t(s'|s, a, u) du = \int_{\mathcal{U}} P_{\pi^b}(u|s) P_t(s'|s, a, u) du
$$

Note that: $P_\pi(u|s) = P_\pi(u|s, h)$ when under the integral with the transitions. So we need to use the fact that this is history-independent.

Proof of Proposition 6.4.1

Proof. The result follows by applying Corollary 4 of [84] to Proposition 6.3.2. □

C.2 Additional discussion

Related Work

Connections to pessimism in offline RL. Pessimism is an important algorithmic design principle for offline RL in the *absence* of unobserved confounders [314, 243, 146]. Therefore, robust FQI with lower-confidence-bound-sized Λ gracefully degrades to a pessimistic offline RL method if unobserved confounders were, contrary to our method's use case, not actually present in the data. Conversely, pessimistic offline RL with *state-wise* lower confidence bounds confers some robustness against unobserved confounders. But state-wise LCBs are viewed as overly conservative relative to a profiled lower bound on the average value [314].

Derivation of the Closed-Form for the Robust Bellman Operator

Proof. Proof of Proposition 6.4.2

[85] show that the linear program in Proposition 6.4.1 has a closed-form solution corresponding to adversarial weights:

$$\tilde{Y}^{-}_{f,t}(s,a) = \mathbb{E}_{\pi^b}\left[W_t^* Y_t | S_t = s, A_t = a\right] \text{ where } W_t^* = \alpha_t \mathbb{I}\left[Y_t > Z_t^{1-\tau}\right] + \beta_t \mathbb{I}\left[Y_t \leq Z_t^{1-\tau}\right].$$

We can derive the form in Proposition 6.4.2 with a few additional transformations. Define:

$$\mu_t(s,a) \coloneqq \mathbb{E}_{\pi^b}[Y_t | S_t = s, A_t = a],$$
$$\text{CVaR}_t^{1-\tau}(s,a) \coloneqq \frac{1}{1-\tau}\mathbb{E}_{\pi^b}\left[Y_t \mathbb{I}\left[Y_t < Z_t^{1-\tau}\right] | S_t = s, A_t = a\right].$$

We use the following identity for any random variables Y and X:

$$\mathbb{E}[Y|X] = \mathbb{E}[Y\mathbb{I}\left[Y > Z^{1-\tau}(Y|X)\right]|X] + \mathbb{E}[Y\mathbb{I}\left[Y \leq Z^{1-\tau}(Y|X)\right]|X]$$

to deduce that

$$\tilde{Y}^{-}_{f,t}(s,a) = \alpha_t \mu_t(s,a) + (\beta_t - \alpha_t)(1-\tau)\text{CVaR}_t^{1-\tau}(s,a),$$

which gives the desired convex combination by noticing that $(\beta_t - \alpha_t)(1-\tau) = (1-\alpha_t)$. □

C.3 Proofs for Robust FQE/FQI

Auxiliary lemmas for robust FQE/FQI

Lemma C.3.1 (Higher-order quantile error terms)**.** *Assume Assumption 6 (i.e. bounded conditional density by M_P), and that $Z_t^{1-\tau}$ is differentiable with respect to s and its gradient is Lipschitz continuous. Then, for $f_t = R_t + \widehat{\overline{Q}}_{t+1}$, if $\hat{Z}_t^{1-\tau}$ is $O_p(w_n)$ sup-norm consistent, i.e. $\sup_{s\in\mathcal{S}} |Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}| = O_p(w_n)$, uniformly over $s \in \mathcal{S}$,*

$$|\mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}[f_t \le \hat{Z}_t^{1-\tau}] - \mathbb{I}[f_t \le Z_t^{1-\tau}]) \mid S = s, A = 1]| = O_p(w_n^2), \tag{C.1}$$

and

$$\mathbb{E}[(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})\left(\mathbb{I}[f \le Z_t^{1-\tau}] - (1-\tau)\right) \mid A = 1] \le M_P \mathbb{E}[(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})^2 \mid A = a]. \tag{C.2}$$

Lemma C.3.1 is a technical lemma which summarizes the properties of the orthogonalized target which lead to quadratic bias in the first-stage estimation error of $\hat{Z}_t$. Equation (C.1) is a slight modification of [226]/[165, A.3]; eq. (C.2) is a slight modification of [267, Lemma 4.1].

Lemma C.3.2 (Bernstein concentration for least-squares loss (under approximate realizability))**.** *Suppose Assumption 10 and that:*

1. *Approximate realizability: $\mathcal{Q}$ approximately realizes $\overline{\mathcal{T}}\mathcal{Q}$ in the sense that $\forall f \in \mathcal{Q}, z \in \mathcal{Z}$, let $q_f^\star = \arg\min_{q\in\mathcal{Q}} \|q - \overline{\mathcal{T}} f\|_{2,\mu}$, then $\left\|q_f^\star - \overline{\mathcal{T}} f\right\|_{2,\mu}^2 \le \epsilon_{\mathcal{Q},\mathcal{Z}}$.*
2. *The dataset $\mathcal{D}$ is generated from P_{obs} as follows: $(s,a) \sim \mu, r = R(s,a), s' \sim P(s' \mid s, a)$.*

We have that $\forall f \in \mathcal{Q}$, with probability at least $1-\delta$,

$$\mathbb{E}_\mu[\ell(\widehat{\mathcal{T}}_{\mathcal{Z}} f; f)] - \mathbb{E}_\mu[\ell(g_f^\star; f)] \le \frac{56 V_{\max}^2 \ln \frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{3n} + \sqrt{\frac{32 V_{\max}^2 \ln \frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{n} \epsilon_{\mathcal{Q},\mathcal{Z}}}$$

Lemma C.3.3 (Stability of covering numbers)**.** *We relate the covering numbers of the squared loss function class, denoted as $\mathcal{L}_{q(z'),z}(q_{t+1})$, to the covering numbers of the function classes $\mathcal{Q}, \mathcal{Z}$. Define the squared loss function class as:*

$$\mathcal{L}_{q(z'),z}(q_{t+1}) = \left\{\ell(q(z'), q_{t+1}; z) - \ell(\overline{Q}^\dagger_{t,Z_t}, q_{t+1}; z) \colon q(z') \in \{\mathcal{Q} \otimes \mathcal{Z}\}, z \in \mathcal{Z}\right\}$$

Then

$$N_{[]}\left(2\epsilon L, \mathcal{L}_{q(z'),z}, \|\cdot\|\right) \le N(\epsilon, \mathcal{Q} \times \mathcal{Z}, \|\cdot\|).$$

Lemma C.3.4 (Difference of indicator functions)**.** *Let $\widehat{f}$ and f take any real values. Then $\left|\mathbb{I}[\widehat{f} > 0] - \mathbb{I}[f > 0]\right| \le \mathbb{I}[|f| \le |\widehat{f} - f|]$*

Proofs of theorems

Proof. Proof of Theorem 6.5.3

The squared loss with respect to a given conditional quantile function Z is:

$$\ell(q, q_{t+1}; Z) = \Bigg(\alpha(R + q_{t+1}) + (1-\alpha)\Big(Z_t^{1-\tau} + \frac{1}{1-\tau}\big((R + q_{t+1} - Z_t^{1-\tau})_- - Z_t^{1-\tau} \cdot (\mathbb{I}\left[R + q_{t+1} \leq Z_t^{1-\tau}\right] - (1-\tau))\big)\Big) - q_t\Bigg)^2$$

We let $\hat{Z}_{t,Q_{t+1}}$ and $Z_{t,Q_{t+1}}$ denote estimated and oracle conditional quantile functions, respectively, with respect to a target function that uses the Q_{t+1} estimate. Where the next-timestep Q_{t+1} function is fixed (as it is in the following analysis) we drop the Q_{t+1} from the subscript.

Define

$$\hat{\overline{Q}}_{t,Z_t} \in \arg\min_q \mathbb{E}_n[\ell(q, \hat{\overline{Q}}_{t+1}; Z_t)]$$

and for $z \in \{\hat{Z}_t, Z_t\}$, define the following *oracle* Bellman error projections $\overline{Q}^{\dagger}_{t,z}$ of the iterates of the algorithm:

$$\overline{Q}^{\dagger}_{t,z} = \arg\min_{q_t \in \mathcal{Q}_t} \|q_t - \overline{\mathcal{T}}^*_{t,z} \hat{\overline{Q}}_{t+1}\|_{\mu_t}.$$

Relating the Bellman error to FQE loss. The bias-variance decomposition implies if U, V are conditionally uncorrelated given W, then

$$\mathbb{E}[(U-V)^2 \mid W] = \mathbb{E}[(U - \mathbb{E}[V \mid W])^2 \mid W] + Var[V \mid W].$$

Hence a similar relationship holds for the robust Bellman error as for the Bellman error:

$$\mathbb{E}[\ell(q, \overline{Q}_{t+1}; Z)^2] = \|q - \overline{\mathcal{T}}^* \overline{Q}_{t+1}\|_{\mu} + Var[W_t^{*,\pi}(Z)(R_t + \bar{V}_{\overline{Q}_{t+1}}(S_{t+1})) \mid S_t, A].$$

which is used to decompose the Bellman error as follows:

$$\|\hat{\overline{Q}}_{t,\hat{Z}_t} - \overline{\mathcal{T}}^*_{t,Z_t} \hat{\overline{Q}}_{t+1}\|^2_{\mu_t} = \mathbb{E}_{\mu}[\ell(\hat{\overline{Q}}_{t,\hat{Z}_t}, \hat{\overline{Q}}_{t+1}; Z_t)] - \mathbb{E}_{\mu}[\ell(\overline{Q}^{\dagger}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] + \|\overline{Q}^{\dagger}_{t,Z_t} - \overline{\mathcal{T}}^*_t \hat{\overline{Q}}_{t+1}\|^2_{\mu_t}.$$

Then,

$$\begin{aligned}
&\|\hat{\overline{Q}}_{t,\hat{Z}_t} - \overline{\mathcal{T}}^*_{t,Z_t} \hat{\overline{Q}}_{t+1}\|^2_{\mu_t} \\
&= \mathbb{E}_{\mu}[\ell(\hat{\overline{Q}}_{t,\hat{Z}_t}, \hat{\overline{Q}}_{t+1}; Z_t)] - \mathbb{E}_{\mu}[\ell(\hat{\overline{Q}}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] && \text{(C.3)} \\
&+ \mathbb{E}_{\mu}[\ell(\hat{\overline{Q}}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] - \mathbb{E}_{\mu}[\ell(\overline{Q}^{\dagger}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] && \text{(C.4)} \\
&+ \|\overline{Q}^{\dagger}_{t,Z_t} - \overline{\mathcal{T}}^*_t \hat{\overline{Q}}_{t+1}\|^2_{\mu_t} && \text{(C.5)}
\end{aligned}$$

We bound eq. (C.3) by orthogonality and eq. (C.4) by Bernstein inequality arguments.

We bound the first term. Let f denote the Bellman residual. Let $x = f$, $(a - x) = Q - f$, $b = Q'$. Since, by expanding the square and Cauchy-Schwarz, we obtain the following elementary inequality:

$$\begin{aligned}(a-x)^2 - (b-x)^2 &= (a-b)^2 + 2(a-b)(b-x) \\ &\leq (a-b)^2 + \sqrt{\mathbb{E}[(a-b)^2]\mathbb{E}[(b-x)^2]}\end{aligned}$$

Applying the above, we have that

$$\begin{aligned}\mathbb{E}_\mu[\ell(\hat{\overline{Q}}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] - \mathbb{E}_\mu[\ell(\overline{Q}^\dagger_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] \leq \\ \underbrace{\|(\hat{\overline{Q}}_{t,Z_t} - \overline{Q}^\dagger_{t,Z_t})\|_2^2}_{o_p(n^{-1}) \text{ by } \textit{Proposition } 6.5.1} + \|(\hat{\overline{Q}}_{t,Z_t} - \overline{Q}^\dagger_{t,Z_t})^2\| \underbrace{\|\hat{\overline{Q}}_{t,Z_t} - \tilde{Y}_t(\hat{\overline{Q}}_{t+1}; Z_t)\|}_{=O_p(n^{-1/2}) \text{ by realizability}}\end{aligned}$$

Therefore

$$\mathbb{E}_\mu[\ell(\hat{\overline{Q}}_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] - \mathbb{E}_\mu[\ell(\overline{Q}^\dagger_{t,Z_t}, \hat{\overline{Q}}_{t+1}; Z_t)] = o_p(n^{-1}).$$

We bound eq. (C.4) by Lemma C.3.2 directly.

Supposing Assumption 10, we obtain that

$$\left\|\hat{Q}_t - \overline{\mathcal{T}}^\star_t \hat{Q}_{t+1}\right\|^2_{\mu_t} \leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{56 V^2_{\max} \ln \frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{3n} + \sqrt{\frac{32 V^2_{\max} \ln \frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{n} \epsilon_{\mathcal{Q},\mathcal{Z}}} + o_p(n^{-1}).$$

Instead, supposing Assumption 11, instantiate the covering numbers choosing $\epsilon = O(n^{-1})$. Lemma C.3.3 bounds the bracketing numbers of the (Lipschitz over a bounded domain) loss function class with the covering numbers of the primitive function classes $\mathcal{Q}, \mathcal{Z}$. Supposing that Bellman completeness holds with respect to $\mathcal{Q}, \mathcal{Z}$, approximate Bellman completeness holds over the ϵ-net implied by the covering numbers with $\epsilon_{\mathcal{Q},\mathcal{Z}} = O(n^{-1})$ and we obtain that:

$$\begin{aligned}\left\|\hat{Q}_t - \overline{\mathcal{T}}^\star_t \hat{Q}_{t+1}\right\|^2_{\mu_t} &\leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{56 V^2_{t,\max} \log\{N(\epsilon, \mathcal{Q}, \|\cdot\|) N(\epsilon, \mathcal{Z}, \|\cdot\|)/\delta\}}{3n} \\ &\quad + \sqrt{\frac{32 V^2_{t,\max} \log\{N(\epsilon, \mathcal{Q}, \|\cdot\|) N(\epsilon, \mathcal{Z}, \|\cdot\|)/\delta\}}{n} \epsilon_{\mathcal{Q},\mathcal{Z}}} + o_p(n^{-1}). \\ &\leq \epsilon_{\mathcal{Q},\mathcal{Z}} + \frac{56 V^2_{t,\max} \log\{N(\epsilon, \mathcal{Q}, \|\cdot\|) N(\epsilon, \mathcal{Z}, \|\cdot\|)/\delta\}}{3n}\end{aligned}$$

□

Proofs of intermediate results

Orthogonality

Proof. Proof of Proposition 6.5.1 We first focus on the case of a single action, $a = 1$. First recall that in the population, $\mathbb{E}[Z_t^{1-\tau} + \frac{1}{1-\tau}(f_t - Z_t^{1-\tau}) \mid s, a] = \frac{1}{1-\tau}\mathbb{E}[f_t \mathbb{I}\left[f_t \leq Z_t^{1-\tau}\right] \mid s, a]$.

In the analysis below we study this truncated conditional expectation representation.

$$\|\hat{\overline{Q}}_t(S,1) - \overline{Q}_t(S,1)\| \lesssim \|\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, A=1]\| + \|\hat{\overline{Q}}_t(S,1) - \overline{Q}_t(S,1)\|$$
$$\text{by Prop. 1 of [166] (regression stability)}$$

Prop. 1 of [166] provides bounds on how regression upon pseudooutcomes with estimated nuisance functions relates to the case with known nuisance functions.

It remains to relate $\|\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, A=1]\|$ to the terms comprising the pointwise bias, which are bounded by Lemma C.3.1. We define these terms as:

$$B_1^1(S) = \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{(f_t - Z_t^{1-\tau})\left(\mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau})\right] - \mathbb{I}\left[f_t \leq Z_t^{1-\tau})\right]\right)\right\} \mid S, A=1\right]$$
$$B_2^1(S) = \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})\left(\mathbb{I}\left[f \leq Z_t^{1-\tau}\right] - (1-\tau)\right)\right\} \mid S, A=1\right].$$

Lemma C.3.1 bounds these terms as quadratic in the first-stage estimation error of $\hat{Z}_t$.

We have that

$$\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, 1] = B_1^1(S) + B_2^1(S).$$

To see this, note:

$$\begin{aligned}
&\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, 1] \\
&= \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{\left(f_t\mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau})\right] - f_t\mathbb{I}\left[f_t \leq Z_t^{1-\tau})\right]\right)\right. \\
&\quad - \left(\hat{Z}_t^{1-\tau} \cdot (\mathbb{I}\left[f \leq \hat{Z}_t^{1-\tau}\right] - (1-\tau)) - Z_t^{1-\tau} \cdot (\mathbb{I}\left[f \leq Z_t^{1-\tau}\right] - (1-\tau))\right) \\
&\quad \left.\pm Z_t^{1-\tau} \cdot \mathbb{I}\left[f \leq \hat{Z}_t^{1-\tau}\right]\right\} \mid S, A=1\Big] \\
&= \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{(f_t - Z_t^{1-\tau})\mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau})\right] - (f_t - Z_t^{1-\tau})\mathbb{I}\left[f_t \leq Z_t^{1-\tau})\right]\right.\right. \\
&\quad \left.\left.+(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})\mathbb{I}\left[f \leq Z_t^{1-\tau}\right] - (Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})(1-\tau)\right\} \mid S, A=1\right] \\
&= \mathbb{E}\left[\frac{1-\tilde{\alpha}}{1-\tau}\left\{(f_t - Z_t^{1-\tau})\left(\mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau})\right] - \mathbb{I}\left[f_t \leq Z_t^{1-\tau})\right]\right)\right. \\
&\quad \left.+(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})\left(\mathbb{I}\left[f \leq Z_t^{1-\tau}\right] - (1-\tau)\right)\right\} \mid S, A=1\Big] \\
&= B_1^1(S) + B_2^1(S)
\end{aligned}$$

Finally, we relate the root mean-squared conditional bias,

$$\|\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, A=1]\|,$$

to the above quadratic error as follows. Using the inequalities $(a+b)^2 \leq 2(a^2+b^2)$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ (for nonnegative a, b), we obtain that

$$\begin{aligned}
&\|\mathbb{E}[\tilde{Y}_t(\hat{Z}_t, \hat{\overline{Q}}_{t+1}) - \tilde{Y}_t(Z_t, \hat{\overline{Q}}_{t+1}) \mid S, A=1]\| \\
&\qquad = \sqrt{\mathbb{E}[(B_1^1(S) + B_2^1(S))^2 \mid A=1]} \\
&\qquad \leq \sqrt{\mathbb{E}[2\{(B_1^1(S))^2 + (B_2^1(S))^2\} \mid A=1]} \\
&\qquad \leq \sqrt{2\mathbb{E}[(B_1^1(S))^2 \mid A=1]} + \sqrt{2\mathbb{E}[(B_2^1(S))^2 \mid A=1]}.
\end{aligned}$$

The result follows by the uniform bounds of Lemma C.3.1.

□

Proof. Proof of Lemma C.3.1

Proof of eq. (C.1):

For $l > 0$, define

$$\mathcal{M}_n^a(l) = \left\{ g : \mathcal{S} \to \mathbb{R} \text{ s.t. } \sup_{s \in \mathcal{S}} |g(s) - Z_t^{1-\tau}(s,a)| \leq l w_n \right\}$$

Define

$$U_n(g,s) \coloneqq |\mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}\left[f_t \leq \hat{Z}_t^{1-\tau}\right] - \mathbb{I}\left[f_t \leq Z_t^{1-\tau}\right]) \mid S=s, A=1]|$$

We will show that for every $l > 0, s \in \mathcal{S}$:

$$\sup_{g \in \mathcal{M}_n(l)} U_n(g,s) = O_p\left(w_n^2\right)$$

Breaking up the absolute value,

$$\begin{aligned}
U_n(g,s) \leq &\mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}\left[Z_t^{1-\tau} \leq f_t \leq g\right]) \mid S=s, A=1] \\
&+ \mathbb{E}[(Z_t^{1-\tau} - f_t)(\mathbb{I}\left[g \leq f_t \leq Z_t^{1-\tau}\right]) \mid S=s, A=1].
\end{aligned}$$

We will bound the first term, bounding the second term is analogous. Define

$$U_{1,n}(g,s) \coloneqq \mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}\left[Z_t^{1-\tau} \leq f_t \leq g\right]) \mid S=s, A=1].$$

Observe that

$$\begin{aligned}
\sup_{g \in \mathcal{M}_n(l)} U_{1,n}(g,s) &= \mathbb{E}[(f_t - Z_t^{1-\tau})(\mathbb{I}\left[Z_t^{1-\tau} \leq f_t \leq Z_t^{1-\tau} + l w_n\right]) \mid S=s, A=1] \\
&\leq M_P l^2 w_n^2
\end{aligned}$$

The result follows.

Proof of eq. (C.2):

The argument follows that of [266]. The difference of indicators is nonzero on the events:

$$\mathcal{E}^- \coloneqq \{f_t - \hat{Z}_t^{1-\tau} < 0 < f_t - Z_t^{1-\tau}\}$$
$$\mathcal{E}^+ \coloneqq \{f_t - Z_t^{1-\tau} < 0 < f_t - \hat{Z}_t^{1-\tau}\}$$

On these events, the estimation error upper bounds the exceedance

$$\{\mathcal{E}^- \cup \mathcal{E}^+\} \implies \{|Z - f| < |Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}|\} \tag{C.6}$$

(since $\mathcal{E}^- \implies \{f - \hat{Z}_t^{1-\tau} < 0 < f - Z_t^{1-\tau}\}$ and $\mathcal{E}^+ \implies \{0 < Z_t^{1-\tau} - f < Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}\}$.)

Then

$$\begin{aligned}\mathbb{E}[(f_t - Z_t^{1-\tau})\mathbb{I}\left[\mathcal{E}^- \cup \mathcal{E}^+\right] \mid S = s, A = 1] &= \int_{-|Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}|}^{|Z_t^{1-\tau} - \hat{Z}_t^{1-\tau}|} (f_t(s, a, s') - Z_t^{1-\tau}) P(s' \mid s, a) ds' \\ &\leq M_P \mathbb{E}[(Z_t^{1-\tau} - \hat{Z}_t^{1-\tau})^2 \mid S = s, A = 1]\end{aligned}$$

Assumption 8 ensures the result holds for state distributions that could arise during policy fitting. The above results hold conditionally on some action $A = 1$ but hold for all actions. □

Other lemmas

Proof. Proof of Lemma C.3.2

Recall that

$$\begin{aligned}\ell(q, q_{t+1}; Z) = \Bigg(&\alpha(R + q_{t+1}) \\ &+ (1 - \alpha)\Big(Z_t^{1-\tau} + \frac{1}{1-\tau}\big((R + q_{t+1} - Z_t^{1-\tau})_- \\ &- Z_t^{1-\tau} \cdot (\mathbb{I}\left[R + q_{t+1} \leq Z_t^{1-\tau} - (1 - \tau)\right])\big)\Big)\Bigg] - q_t\Bigg)^2\end{aligned}$$

Define $f_{q',z} =$ Define X to be the difference of the integrands.

Step 1:

$$Var(X(g, f, z, g_f^*)) \leq 4V_{\max}^2 \|\hat{\overline{Q}}_{t,Z_t} - \overline{Q}_{t,Z_t}^\dagger\|_2^2$$

(by similar arguments as in the original paper). By the same arguments (i.e. adding and subtracting $\overline{\mathcal{T}} f$) we obtain that

$$\|\hat{\overline{Q}}_{t,Z_t} - \overline{Q}_{t,Z_t}^\dagger\|_2^2 \leq 2(\mathbb{E}[X(g, f, z, g_f^*)] + 2\epsilon_{\mathcal{Q},\mathcal{Z}})$$

Therefore,

$$Var(X(g, f, z, g_f^*)) \leq 8V_{\max}^2(\mathbb{E}[X(g, f, z, g_f^*)] + 2\epsilon_{\mathcal{Q},\mathcal{Z}}).$$

Applying (one-sided) Bernstein's inequality uniformly over $\mathcal{Q}, \mathcal{Z}$, we obtain:

$$\begin{aligned}&\mathbb{E}\left[X(g,f,z,g_f^*)\right] - \mathbb{E}_n[X(g,f,z,g_f^\star)] \\ &\leq \sqrt{\frac{16V_{\max}^2\left(\mathbb{E}\left[X(g,f,z,g_f^\star)\right] + 2\epsilon_{\mathcal{F},\mathcal{Z}}\right)\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{n}} + \frac{4V_{\max}^2\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{3n}\end{aligned}$$

Note that $\hat{\overline{Q}}_{t,Z_t}$ minimizes both $\mathbb{E}_n[\ell(q,\hat{\hat{Q}}_{t+1};Z_t)]$ and $\mathbb{E}[(q,\hat{\hat{Q}}_{t+1},Z_t,\overline{Q}^*_{\hat{\hat{Q}}_{t+1}})]$ with respect to q. Therefore, by completeness since the Bayes-optimal predictor is realizable,

$$\mathbb{E}_n[\ell(\hat{\overline{Q}}_{t,Z_t},\hat{\hat{Q}}_{t+1};Z_t)] \leq \mathbb{E}_n[\ell(\overline{Q}^\dagger_{t,Z_t},\hat{\hat{Q}}_{t+1};Z_t)] = 0$$

Therefore (solving for the quadratic formula),

$$\mathbb{E}[X(\hat{\overline{Q}}_{t,Z_t},\hat{Q}_{t+1},Z_t,\overline{Q}^\dagger_{t,Z_t})] \leq \frac{56V_{\max}^2\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{3n} + \sqrt{\frac{32V_{\max}^2\ln\frac{|\mathcal{Q}||\mathcal{Z}|}{\delta}}{n}\epsilon_{\mathcal{F},\mathcal{Z}}}$$

□

Proof. Proof of Lemma C.3.3 We show this result by establishing Lipschitz-continuity of the squared loss function class (with respect to the product function class of $\mathcal{Q}\times\mathcal{Z}$).

We use a stability result on the bracketing number under Lipschitz transformation. Classes of functions $x \mapsto f_\theta(x)$ that are Lipschitz in the index parameter $\theta \in \Theta$ have bracketing numbers readily related to the covering numbers of Θ. Suppose that

$$|f_{\theta'}(x) - f_\theta(x)| \leq d(\theta',\theta)F(x),$$

for some metric d on the index set, function F on the sample space, and every x. Then $(\operatorname{diam}\Theta)F$ is an envelope function for the class $\{f_\theta - f_{\theta_0} : \theta \in \Theta\}$ for any fixed θ_0. We invoke Theorem 2.7.11 of [304] which shows that the bracketing numbers of this class are bounded by the covering numbers of Θ.

Theorem C.3.5 ([304], Theorem 2.7.11)**.** *Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a class of functions satisfying the preceding display for every θ', θ and some fixed function F. Then, for any norm $\|\cdot\|$,*

$$N_{[]}\left(2\epsilon\|F\|,\mathcal{F},\|\cdot\|\right) \leq N(\epsilon,\Theta,d).$$

Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be a class of functions satisfying the preceding display for every s and θ and some fixed envelope function F. Then, for any norm $\|\cdot\|$,

$$N_{[]}\left(2\epsilon\|F\|,\mathcal{F},\|\cdot\|\right) \leq N(\epsilon,\Theta,d).$$

This shows that the bracketing numbers of the loss function class can be expressed via the covering numbers of the estimated function classes $\mathcal{Q}, \mathcal{Z}$, which are the primitive function

classes of estimation, for which results are given in various references for typical function classes.

Denote

$$g(q_{t+1}) = \alpha(s,a)(R + q_{t+1})$$
$$h(z) = (1-\alpha)\Big(\frac{1}{1-\tau}\left(z + (R + q_{t+1} - z)_- - z \cdot (\mathbb{I}\left[R + q_{t+1} \leq z\right] - (1-\tau))\right)\Big)$$

and notate

$$\ell(q, q_{t+1}; z) = (q - g(q_{t+1}) + h(q_{t+1}, z))^2.$$

Note that $\frac{1}{1-\tau} = (1+\Lambda)$. Assuming bounded rewards, define $D_{z,t}, D_{q,t}$ as the diameters of $\mathcal{Q}_t, \mathcal{Z}_t$, respectively and note that $D_{z,t} \approx D_{q,t}$. Note that $h(q_{t+1}, z)$ is $(1-\alpha_{\min})(3(1+\Lambda)+1)$-Lipschitz in z (since the sum of Lipschitz continuous functions is Lipschitz) and it is $(1-\alpha_{\min})\left(1 + (1+\Lambda)(\frac{D_{z,t}}{D_{q,t}} + 1)\right)$-Lipschitz in q_{t+1}. Further, $g(q_{t+1})$ is $\alpha_{\max}$-Lipschitz in q_{t+1}. Therefore, $\ell\left(q, q_{t+1}; z\right)$ is $D_{q,t}$ Lipschitz in q, $L^C_{q,t+1}$-Lipschitz in q_{t+1} and $L^C_{z,t}$-Lipschitz in z, with $L^C_{q,t+1}, L^C_{z,t}$ defined as follows:

$$L^C_{q,t+1} = (2D_{q,t+1} + D_{z,t})(1-\alpha_{\min})\left(\left(1 + (1+\Lambda)(\frac{D_{z,t}}{D_{q,t+1}} + 1)\right) + \alpha_{\max}\right)$$
$$L^C_{z,t} = (2D_{q,t+1} + D_{z,t})(1-\alpha_{\min})(3(1+\Lambda)+1).$$

Therefore we have shown that restrictions of $\ell\left(q, q_{t+1}; z\right)$ to the q_{t+1}, z coordinates are individually Lipschitz. We leverage the fact that a function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz if and only if there exists a constant L such that the restriction of f to every line parallel to a coordinate axis is Lipschitz with constant L. Choosing

$$L_t = \sqrt{3}\max\{D_q, L^C_{q,t+1}, L^C_{z,t}\}$$

gives that $\ell\left(q, q_{t+1}; z\right)$ is L_t-Lipschitz.

□

Proof. Proof of Corollary 6.5.4

Lemma C.3.3 gives that $\ell\left(q, q_{t+1}; z\right)$ is L_t-Lipschitz with $L_t = \sqrt{3}\max\{D_q, L^C_{q,t+1}, L^C_{z,t}\}$.

To interpret the scaling of the result, we can appeal to [304, Thm. 2.6.4] which upper bounds the (log) covering numbers by the VC-dimension. Namely, [304, Thm. 2.6.4] states that there exists a universal constant K such that

$$N\left(\epsilon, \mathcal{F}, L_r(Q)\right) \leq KV(\mathcal{F})(4e)^{V(\mathcal{F})}\left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{F})-1)}.$$

Therefore, achieving an $\epsilon = cn^{-1}$ approximation error on the bracketing numbers of robust Q functions results in an $\log(2L_t n)$ dependence.

Lastly we remark on instantiating L_t. Note that under the assumption of bounded rewards, $D_{q,t+1} = B_r(T-t+1)$. Focusing on leading-order dependence in problem-dependent constants, we have that $L_t = O(B_r(T-t)\Lambda)$. Then $\hat{\mathcal{E}}(\hat{Q}) \leq \epsilon + \sum_{t=1}^{T} K\frac{\log(2B_r(T-t)\Lambda n)}{n}$. Upper bounding the left Riemann sum by the integral, we obtain that

$$\begin{aligned}\sum_{t=1}^{T} K\frac{\log(2KB_r(T-t)\Lambda n/\epsilon)}{n} &\leq \int_1^T K\frac{\log(2KB_r(T-x)\Lambda n/\epsilon)}{n}dx \\ &= \frac{(T-1)}{n}(\log(2KB_r\Lambda(T-1)n/\epsilon) - 1).\end{aligned}$$

□

Confounding with infinite data

First, we prove the following useful result for confounded regression with conditional Gaussian tails:

Lemma C.3.6. *Define:*

$$C(\Lambda) \coloneqq \left(\frac{\Lambda^2-1}{\Lambda}\right)\phi\left(\Phi^{-1}\left(\frac{1}{1+\Lambda}\right)\right),$$

where ϕ and Φ are the standard Gaussian density and CDF respectively. Let $Y_t(Q)$ be conditionally Gaussian given $S_t = s$ and $A_t = a$ with mean $\mu_t(s,a)$ and standard deviation $\sigma_t(s,a)$. Then,

$$(\bar{\mathcal{T}}_t^* Q)(s,a) = \mu_t(s,a) - [1-\pi_t^b(a|s)]C(\Lambda)\sigma_t(s,a).$$

Proof. Proof of Lemma C.3.6

The CVaR for Gaussians has a closed-form [223]:

$$\frac{1}{1-\tau}\mathbb{E}_{\pi^b}\left[Y_t(Q)\mathbb{I}\left[Y_t(Q) < Z_t^{1-\tau}\right]|S_t = s, A_t = a\right] = \mu_t(s,a) - \sigma_t(s,a)\frac{\phi(\Phi^{-1}(1-\tau))}{1-\tau}.$$

Applying this to Proposition 6.4.2 gives the desired result. □

Proof. Proof of Proposition 6.5.5 First, note that R_t is conditionally Gaussian given S_t and A_t with mean $\theta_R\theta_P s$ and standard deviation $\theta_R\sigma_T$. Define $\beta_i \coloneqq \theta_R\sum_{k=1}^{i}\theta_P^k$. Using value iteration, we can show that $V_{T-i}^{\pi^e}(s) = \beta_i s$ for $i \geq 1$. E.g. by induction, $V_{T-1}^{\pi^e}(s) = \theta_R\theta_P s = \beta_1$ and if $V_{T-t+1}^{\pi^e}(s) = \beta_{t-1}s$, then

$$V_{T-t}^{\pi^e}(s) = \theta_P(\theta_R + \gamma\beta_{t-1})s = \beta_t s.$$

Next we will derive the form of the robust value function by induction. For the base case, $t = T-1$, we have:

$$Y_{T-1} = \theta_R s'.$$

Therefore, Y_{T-1} is conditionally gaussian with mean $\theta_R\theta_P s$ and standard deviation $\theta_R\sigma_P$. Applying Lemma C.3.6, we have:

$$\bar{V}^{\pi^e}_{T-1}(s) = \theta_R\theta_P s - 0.5C(\Lambda)\theta_R\sigma_P.$$

Now assume that $\bar{V}^{\pi^e}_{t+1}(s) = \theta_V s + \alpha_V$. Then

$$\begin{aligned} Y_t &= \theta_R s' + (\theta_V s' + \alpha_V) \\ &= (\theta_R + \theta_V)s' + \alpha_V. \end{aligned}$$

Therefore, Y_t is conditionally gaussian with mean $(\theta_R + \theta_V)\theta_P s + \alpha_V$ and standard deviation $(\theta_R + \theta_V)\sigma_P$. Applying Lemma C.3.6, we have:

$$\bar{V}^{\pi^e}_{t}(s) = (\theta_R + \theta_V)\theta_P s + \alpha_V - 0.5C(\Lambda)(\theta_R + \theta_V)\sigma_P, \tag{C.7}$$

which is linear in s with new coefficients $\theta'_V \coloneqq (\theta_R+\theta_V)\theta_T$ and $\alpha'_V \coloneqq \alpha_V - 0.5C(\Lambda)(\theta_R+\theta_V)\sigma_P$.

By rolling out the recursion defined in Equation (C.7), consolidating the coefficients into β_i terms, and then simplifying we get:

$$\bar{V}^{\pi^e}_{0}(s) = V^{\pi^e}_{0}(s) - \frac{1}{2\theta_P}\left(\sum_{i=0}^{T-1}\beta_i\right)\sigma_P C(\Lambda).$$

Finally, that $C(\Lambda) \leq \frac{1}{8}\log(\Lambda)$ can be verified numerically.

□

C.4 Details on experiments

Low-Dimensional Parameter Values

$\theta_A = -0.05$, $\sigma = 0.36$, $\gamma = 0.9$, $H = 4$.

The matrices A and B were chosen randomly with a fixed random seed:

```
np.random.seed(1)
B_sparse0 = np.random.binomial(1,0.3,size=d)
B = 2.2*B_sparse0 * np.array( [ [ 1/(j+k+1) for j in range(d) ]
                                            for k in range(d) ] )

np.random.seed(2)
A_sparse0 = np.random.binomial(1,0.6,size=d)
A = 0.48*A_sparse0 * np.array( [ [ 1/(j+k+10) for j in range(d) ]
                                              for k in range(d) ] )
```

Likewise for θ_R:

```
theta_R = 3 * np.random.normal(size=d)
            * np.random.binomial(1,0.3,size=d)
```

High-Dimensional Parameter Values

$\theta_A = -0.05$, $\sigma = 0.1$, $\gamma = 0.9$, $H = 4$.

The matrices A and B were chosen randomly with a fixed random seed:

```
np.random.seed(1)
B_sparse0 = np.random.binomial(1,0.3,size=d)
B = 2.2*B_sparse0 * np.array( [ [ 1/(j+k+1) for j in range(d) ]
                                            for k in range(d) ] )/1.2

np.random.seed(2)
A_sparse0 = np.random.binomial(1,0.6,size=d)
A = 0.48*A_sparse0 * np.array( [ [ 1/(j+k+10) for j in range(d) ]
                                              for k in range(d) ] )/20
```

Likewise for θ_R:

```
theta_R = 2 * np.random.normal(size=d)
            * np.random.binomial(1,0.3,size=d)
```

Function Approximation

Conditional expectations were approximated with the Lasso using `scikit-learn`'s implementation, with regularization hyperparameter $\alpha =$ 1e-4. Conditional quantiles were approximated with `scikit-learn`'s ℓ_1-penalized quantile regression, regularization hyperparameter *alpha* = 1e-2, using the `highs` solver.

Calculating Ground Truth

To provide ground truth for our sparse linear setting, we analytically derive the form of the robust Bellman operator. Consider the candidate Q function, $Q(s,0) = \beta^\top s + a_0, Q(s,1) = \beta^\top s + a_1$. Then,

$$\begin{aligned} Y_t &= \theta_R^\top S_{t+1} + \gamma\beta^\top S_{t+1} + \theta_A\gamma \max\{1_d^\top\theta_R, 0\} \\ &= \theta_R^\top S_{t+1} + \gamma\beta^\top S_{t+1} + \theta_A\gamma 1_d^\top\theta_R \end{aligned}$$

where we chose simulation parameters such that $\theta_A\gamma \max\{1_d^\top\theta_R, 0\} > 0$. Therefore:

$$Y_t|S_t, A_t \sim \mathcal{N}\left((\theta_R+\gamma\beta)^\top(BS_t+\theta_A A_t) + \theta_A\gamma 1_d^\top\theta_R, \sqrt{\sum_{i=1}^d (\theta_R+\gamma\beta)_i^2 (AS_t+\sigma)_i^2}\right)$$

Since Y_t is conditionally Gaussian, we apply Lemma C.3.6:

$$\begin{aligned} (\bar{\mathcal{T}}_t^* Q)(s,a) &= \mathbb{E}[Y_t|S_t=s, A_t=a] - 0.5C(\Lambda)\sqrt{\mathrm{Var}[Y_t|S_t=s, A_t=a]} \\ &= (\theta_R+\gamma\beta)^\top(BS_t+\theta_A A_t) + \theta_A\gamma 1_d^\top\theta_R - 0.5C(\Lambda)\sqrt{\sum_{i=1}^d (\theta_R+\gamma\beta)_i^2 (AS_t+\sigma)_i^2} \end{aligned}$$

First, note that the slope w.r.t. S_t is not a function of A_t validating our choice of an action-independent β. Second, note that only the last term is non-linear in S_t. So the ground truth for FQI with Lasso adds the first two terms to the closest linear approximation of this last term. Since our object of interest is the average optimal value function at the initial state, we perform this linear approximation in terms of mean squared error at the initial state. In practice, we compute this by drawing $200,000$ samples i.i.d. from the initial state distribution and then doing linear regression on this last term. Plugging the slope and intercept back in is extremely close to the best linear approximation of $(\bar{\mathcal{T}}_t^* Q)(s,a)$.