Statistics meets Optimization: Computational guarantees for statistical learning algorithms

Fanny Yang

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2018-126
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2018/EECS-2018-126.html

August 21, 2018

Copyright © 2018, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

Statistics meets Optimization: Computational guarantees for statistical learning algorithms

by

Fan Yang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrial Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin J. Wainwright, Chair
Professor Peter J. Bickel
Professor Benjamin Recht

Fall 2018

Statistics meets Optimization: Computational guarantees for statistical learning algorithms

Copyright 2018
by
Fan Yang

Abstract

Statistics meets Optimization: Computational guarantees for statistical learning algorithms

by

Fan Yang

Doctor of Philosophy in Engineering – Electrial Engineering and Computer Science

University of California, Berkeley

Professor Martin J. Wainwright, Chair

Modern technological advances have prompted massive scale data collection in many modern fields such as artificial intelligence, and traditional sciences alike. This has led to an increasing need for scalable machine learning algorithms and statistical methods to draw conclusions about the world. In all data-driven procedures, the data scientist faces the following fundamental questions: How should I design the learning algorithm and how long should I run it? Which samples should I collect for training and how many are sufficient to generalize conclusions to unseen data? These questions relate to statistical and computational properties of both the data and the algorithm. This thesis explores their role in the areas of non-convex optimization, non-parametric estimation, active learning and multiple testing.

In the first part of this thesis, we provide insights of different flavor concerning the interplay between statistical and computational properties of first-order type methods on common estimation procedures. The expectation-maximization (EM) algorithm estimates parameters of a latent variable model by running a first-order type method on a non-convex landscape. We identify and characterize a general class of Hidden Markov Models for which linear convergence of EM to a statistically optimal point is provable for a large initialization radius. For non-parametric estimation problems, functional gradient descent type (also called boosting) algorithms are used to estimate the best fit in infinite dimensional function spaces. We develop a new proof technique showing that early stopping the algorithm instead may also yield an optimal estimator without explicit regularization. In fact, the same key quantities (localized complexities) are underlying both traditional penalty-based and algorithmic regularization.

In the second part of the thesis, we explore how data collected adaptively with a constantly updated estimation can lead to signifcant reduction in sample complexity for multiple hypothesis testing problems. In particular, we show how adaptive strategies can be used to simultaneously control the false discovery rate over multiple tests and return the best alternative (among many) for each test with optimal sample complexity in an online manner.

To my parents

Contents

List of Figures

Acknowledgments

This thesis really would not have been possible without the support of many people along the journey in the past five years. The following names that are included does not form a complete list of people I am grateful for and I'm very sorry if you can't find your name here: rest assured that I deeply appreciate the support of anyone who is reading these lines.

First and foremost I owe a lot to my advisor Martin Wainwright. Apart from being a role model in terms of mathematical maturity, presentation in speaking and writing, balance of mountaineering, family and research, he was always open to exploring many different research areas I felt attracted to over the years. He guided me to (hopefully) develop taste for good research problems and how to not abstain from caffein to figure out their answers. Martin has also had a big influence on my future plans and how to deal with decision making processes. I have learnt from him to be more efficient, focusing my thoughts on the important things in life and research which I can actually influence and spending less time thinking about the rest. I want to thank him for always being supportive of me and other students alike to pursue our passions beyond research and for making sure that amidst all the academic craziness, our personal happiness was never sacrificed but always had a positive net gain. He also created an environment of trust, so that I always felt comfortable to talk to him directly about any issues that arised, confident that he would understand or give the right advice. Even when when it was about the objectively unnecessary pains of being an Arsenal fan ... Overall, I am just immensely grateful and feel extremely lucky to have had him as a mentor.

There are many more faculty members whom I am greatly indebted to: Anima Anandkumar, my supervisor during my internship at Amazon, was always very understanding, patient and supportive of my efforts to dive into the applied world of big data pipelines and coding bugs. Working with her opened my eyes to many exciting problems in the real world which I am planning to continue exploring. Teaching CS 189 with Anant Sahai and Jennifer Listgarten ultimately convinced me to pursue an academic career. Their student-oriented teaching philosophy and freedom that I had in the course development made it an exceptionally rewarding experience. I would like to thank Peter Bickel, Benjamin Recht and Aditya Guntobuyina for being on my Quals and thesis committee and giving valuable feedback along the way. Finally, I owe a great debt of gratitude to my master thesis advisors in Munich Volker Pohl and Holger Boche, without whom I might not have even come to Berkeley in the first place. They have spend a lot of time teaching me how to be a researcher and ultimately helped me to jump on the next boat ending up 6000 miles away.

I would also like to give huge credit to all my collaborators during grad school: Sivaraman Balakrishnan who also became a great mentor, Aaditya Ramdas and Kevin Jamieson for the fruitful discussions and fun nights of paper writing and deadline crunching in the Rise Lab. At Amazon, Kamyar Azizzadenesheli and Zack Lipton taught me how to tackle, find and write about relevant practical problems. Most prominently though, thanks to my roommate who became family and partner in crime for the boosting paper Yuting Wei, who was willing to discuss anything, including our research project, no matter how late at night. Working

with your roommate (especially when it's me) is quite intense, but also pretty convenient and it ultimately was a great experience!

Besides my direct academic mentors and collaborators, I consider myself incredibly lucky to have been in such a collegial environment that Berkeley EECS department is, and the support which the community gave me. First of all, shoutouts go to all current and former members of the BLISS (previously WiFo) lab including but not restricted to: Po-Ling Loh, Nihar Shah and Rashmi Vinayak for invaluable advice in the first years, Raaz Dwivedi, Orhan Ocal for being the best core and lunch buddies (although I could not yet convince them to join me in a race), Ashwin and Vidya for being the sweetest couple I could do my first secret engagement shoot for and some fun trips, Vasuki Swamy for helping and convincing me to become a (mostly) vegetarian and be more aware about implicit biases of all kinds, Sang Min Han for many foothill dinners and rides and everyone else which I will not list in fear of forgetting some but who contributed to the fun and relaxed atmosphere in lab.

Thanks also to my EECS friends outside the lab with whom I have had many fun philosophical discussions about our research field and from whom I've learned a lot: Thanks to Philipp Moritz for constant friendship and support as well as knowing everything I ever had a question about, Ludwig Schmidt for awesome hikes together in the Sierras and meaningful conversations about life, the universe and everything, Reinhard Heckel for making me believe in myself and encouraging me to become a (albeit mediocre) triathlete, Aymeric Dieuleveut for teaching me how to throw a frisbee and solving some fun coding problems together, Constantin Berzan and Raunak Bhinge for always having an open ear to whatever issues were on my mind, Cathy Wu and Richard Shin for our lovely group dinners that allowed me to try all the food places I wouldn't have explored otherwise, and coming up with the sophisticated plotting framework together, Nishant Totla for sharing both painful and joyful moments together as fellow Arsenal fans, and yet again, everyone else I cannot name.

A special shoutout also to the extraordinarily competent and supportive staff at EECS: Shirley Salanio who always answered any question in the blink of an eye with the perfect solution to every bureaucratic problem, Sheila Humphreys for her support of me just being myself with all kinds of diverse interests, Kim Kail, Pat Hernan and many more. It's a luxury for us that you're making our life so easy!

My personal well-being was also strongly due to my friends outside of EECS with whom I had the pleasure to share hobbies and activities that are very important to me. Thank you Adam Bloniarz, Peter Hintz, Daniel Greenhouse, Mosa Tsay, Jane Kim, Fanny Kassel, Brady Anderson for many many intimate chamber music reading sessions, visited concerts and making me discover amazing repertoire and recordings that were new to me and changed my life. Thanks also to Alex Appleton, Loi Almeron, Alex Rusciano, Ping Ai, Ye Xia, Ran Gao, Meishan Fan, Michael Xu, Francis Yang for all these fun outdoor trips and Avalon nights with fun non-research conversations. My Amazon intern friends Todor Mihaylov, Jeremy Bernstein, Jean Kossaifi, Tan Nguyen, Ashish Khetan, Stefan Henneking, Youssef Achari made my time in the South Bay so exciting beyond research - thanks for the many late nights at the office with Sushi Burritos, sharing your research insights and practical advice, more or less close foosball matches, the bike rides and fun trips. Finally, special thanks to all my

great friends at home, Lea Kraemer, Alex Palt, Margarethe Woeckel and Lam Duong for being my backbone through all these years during difficult times.

Furthermore, one my biggest mistakes during grad school was not to have started taking PE classes much earlier. Thanks Toni Mar and Elmar Stefke for pushing me over a fitness threshold only above which it seemed in my reach to start training for triathlons which converted me from a fair-weather sports enthusiast to someone who cannot live without regular hard (and easy) workouts anymore. Thanks to Fausto Macciariello, Jonathan McKinley, Rosalie Lawrence, Mariko Stenstedt, Clayton Kinsey, Rob Craven, Dean Harper for accommodating my slow self in races, camp and countless amazing workouts which enabled me to move from the extreme to the moderate tail end of olympic distance triathletes.

Last but not least, none of all this would have been possible without my parents and their sacrifices. Thank you for going out of your way to enable me to become a person who could (more or less) deal with the challenges of a PhD and pursue my dreams and passions without too many burdens and worries of any kind while maintaining a hopefully healthy work-life balance. Your love and incredible support in all situations were invaluable to me. I hope you will one day cease resistance to revert the roles and allow me to make sacrifices of my own to help you fulfill the dreams that you have.

Chapter 1

Introduction

In the modern world, scientists and engineers alike rely on data to draw conclusions and develop new technologies. In general, data consists of either labeled or unlabeled observations which contain information about the underlying true model in nature. If we had all possible observations at our disposal, such as the true probability distribution of a variable of interest or true labels for all images one could ever see, the inferred model would trivially be equivalent to the true one we wish to know. The fundamental question in machine learning and data science is how well one can learn about the true model when only finite and limited observations are available. Controlling the gap between the learned estimator and the true model is one of the essential challenges in machin learning research. The size of the gap depends both on the statistical and algorithmic aspects in the pipeline.

As the name already suggests, data science depends heavily on the data that is available. Hereby it is important to note that the amount of data itself is not necessarily indicative of the quality of the inference. In layman's terms, it is far more relevant how much information is contained in the data and how representative it is. Usually, data is considered as a given and the scientist has to use an optimal procedure to find a good estimator. Not all data is created equal however, and it can be natural to consider procedures where the learning procedure potentially interacts with the data collection process. Regular feedback indicating which samples would be most informative to include in the dataset could reduce the required number of samples to achieve a desired estimation accuracy. This is helpful especially when data is costly, which is for example the case for experiments including human beings or biological organisms.

The data collection process is not the only crucial factor which determines the performance of the learning procedure. Even when the samples are fixed, the gap between the estimated and true model can vary depending on the learning procedure. On a high level, the estimation process usually consists of the following four fundamental building blocks: 1) choosing a family of models among which the one that "fits best" is returned 2) a loss function which defines the best fit 3) an iterative optimization algorithm aimed at finding the optimal model and 4) a stopping criterion which determines when the algorithm should terminate.

The combination of all these choices together with the data collection procedure determine

the final quality of the estimator. The work in this thesis studies the effects of the optimization algorithm including its stopping time and adaptive data collection on common as well as newly proposed machine learning procedures.

In the first part, we discuss topics in parametric and non-parametric estimation via empirical risk minimization. In particular, for the former we prove theoretical guarantees for a standard first-order method to maximize the generally non-convex likelihood of a latent variable model with dependent samples. For non-parametric estimation, we investigate how the effects of early stopping a gradient-descent type method is comparable to well-understood penalty regularization methods. The second part considers common multiple testing problems and explores the benefits of allowing interaction between data collection and learning in that context. In brief, we show how adaptive sampling can lead to significant gains in sample complexity when testing many hypotheses with multiple treatments. The next sections highlight the specific contributions of this thesis in more detail.

1.1 Estimation from samples

A typical data science problem can be mathematically framed as follows: Given data, which consists of samples from a distribution $\mathbb{P}$, we want to learn a parameter in $\mathbb{R}^d$ or a function in some function space which determines the distribution. Sometimes the relation between a covariate X and a target variable Y is of interest for prediction or scientific purposes, at other times the parameter (for example the mean or variance) of the distribution is an interpretable quantity one wants to learn about. We assume that both the "true mapping" f^* from x to y or parameter $\theta^\star$ of a distribution may be computed exactly when the distribution $\mathbb{P}$ is known. Oftentimes, $\theta^\star$ or f^* correspond to the optimizer of a so-called *population loss* involving the distribution. It can be found via iterative optimization algorithms, such as gradient-type methods, which are efficient when the population loss is well-behaved. For first-order methods one often assumes the population loss to be convex and smooth.

In practice, the distribution is unknown and one relies on a finite number of samples n from the distribution instead. As a result, the *empirical loss* must serve as a proxy for the population loss instead and could be non-convex, have multiple local minima or a global minimum far away from the true $\theta^\star, f^*$. Minimizing the empirical loss and obtaining the estimates $\widehat{\theta}, \widehat{f}$ is thus not guaranteed to be a statistically sound or computationally efficient strategy even though the algorithm behaves perfectly on the population loss. In the first part of the thesis, we show how the statistical and computational errors behave as a function of the number of iterations for first-order type algorithms. The results give guidance as to how such methods should be used in practice.

Guarantees for non-convex optimization

A line of recent work has shown that even for problems with an inherent non-convex structure, such as low rank matrix factorization or deep neural networks, first-order optimization

algorithms can still find useful solutions fast. While convex optimization problems are well-understood from a theoretical point of view, until recently non-convex optimization has largely remained an untouched field. The simplicity and effectiveness of these algorithms however has sparked a surge of interest to provide mathematical explanation for this phenomenon. A recent line of work has taken a closer look at specific non-convex instances which are known to exhibit good convergence behavior and often utilized the particular model structure in order to give rigorous performance guarantees.

The expectation-maximization (EM) algorithm is one such algorithm, which behaves like a first-order method and is designed for finding the maximum likelihood estimator (MLE) of probabilistic models with latent variables such as mixture models. Despite its wide usage since its invention, the reason for its good performance in typical problems has not been well understood. In my work with Sivaraman Balakrishnan and Martin Wainwright [85], we focused on the application of EM on the case of non-independent data drawn from a Hidden Markov Model (HMM), also called the Baum-Welch algorithm [9]. The use of EM for fitting HMM models is a workhorse for speech, gesture recognition and bioinformatics for example.

Until our work, it was only known that the Baum-Welch algorithm, similar to a vanilla first order method, can quickly find its nearby stationary point [9]. Furthermore, the maximum likelihood estimator was proven to be consistent and asymptotically normal [14]. However, a corollary of both results merely predicted convergence to a consistent estimator for *arbitrarily close* initializations which is not useful to explain its empirical performance. This is because it usually operates on finite-sample non-convex loss functions. In contrast, we wanted to understand why, how and when Baum-Welch converges with high probability to parameter estimates that are close to the population optimum: How large does the necessary initialization radius have to be for convergence to a statistically good estimate? How many iterations does one need to get how close to the population optimum?

In a nutshell, for a broad class of Hidden Markov models we were able to prove linear convergence of Baum-Welch to a statistically optimal point. This holds for starting points within a big basin of attraction around the population optimum and shows that in fact every local minimum in that basin lies within statistical precision of the population maximum likelihood estimator $\theta^\star$. Therefore, as long as the algorithm is suitably initialized (for example via spectral methods) we are guaranteed to obtain an estimate that is optimally close to $\theta^\star$. To our knowledge, these are the first results which give rigorous model-dependent non-asymptotic statistical bounds and convergence characterization for Baum-Welch estimates.

Implicit regularization via early stopping

For iterative algorithms such as boosting and gradient-type algorithms, a good estimate is often reached much before convergence. In fact, besides wasting computational resources, finding the actual finite-sample optimum in non-parametric spaces could potentially result in suboptimal generalization to unseen data. The classical approach to avoid overfitting is to add a penalty term or constraint to the loss function, referred to as a *penalized estimator*. In practice, another way to control the generalization performance of iterative algorithms is

to monitor the error on a separate validation set. The algorithm then terminates whenever the test error stops decreasing or even starts increasing. It has indeed been observed in the boosting literature that estimators obtained by early stopping gradient descent type methods are comparable to optimizers of penalized likelihoods.

While there is a rich literature on penalized estimators (e.g. [74], [73]), algorithmic regularization is far less understood. Most existing literature addresses consistency and non-asymptotics for specific loss functions. These types of results however did not provide fundamental insights as to why stopping rules can result in consistent and statistically *optimal* estimators for many different models. Our work attempts to move one step further into this direction by understanding the geometric reason underlying algorithmic regularization of iterative first-order methods: How does the effective complexity of the function space evolve with the number of iterations and how does this relate to generalization properties of intermediate iterates?

Some previous works [35] have shed light on this issue using the concept of algorithmic stability. For more explicitly structured function classes, we show how we can in fact analyze early stopping by adopting the perspective of localized complexity measures, which are known to yield tight upper bounds for penalized estimation. Using this technique, we are able to show minimax optimality of early stopped boosted estimators for a variety of loss functions including AdaBoost, LogitBoost and L^2-Boost for arbitrary reproducing kernel Hilbert spaces. Ultimately, the goal is to use the new insights to make rigorous optimality statements about a common practice for classical and deep learning procedures, which is to stop the algorithm once the error on the validation set stops decreasing. This work was developed jointly Yuting Wei (equal contribution) and Martin Wainwright [87].

1.2 Multiple testing with adaptive sampling

In part II of the thesis we go beyond the setting when the algorithm takes given data and outputs an estimate of the truth. Often, data collection and experiments are expensive and the main obstacle to good performance of a machine learning system. In such cases, it is sometimes more efficient to acquire data sequentially. By choosing unlabeled samples which are expected to be the most informative based on data seen so far, one may extract the same amount of information using a much smaller budget. For example, when the task is to find the best alternative (also called *arm*) among many, it is much more efficient to sample the most promising alternatives more often than the ones that are clearly worse.

While adaptive sampling has become standard for finding a best arm for example, it is rarely considered for related classical statistical settings when two or more alternatives are tested against a control. This is in part due to the bias that adaptivity introduces in measuring treatment effects. Furthermore, classical hypothesis tests are primarily designed to control the probability of false alarm, when in truth the control is indeed the best arm. When the null hypothesis is rejected, we conclude that at least one arm is better than the

control. However, when many treatments promise better performance, what we really want to know is which one of them is in fact the best, as illustrated in Figure 1.1.

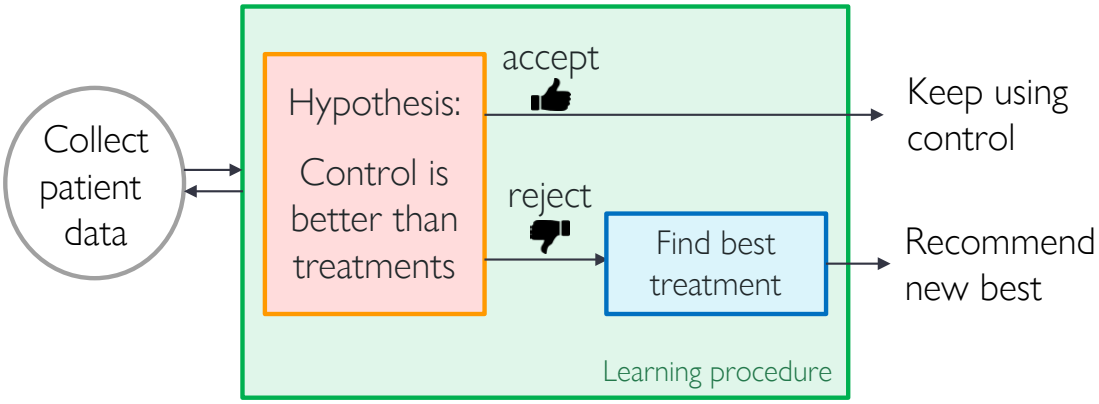


Figure 1.1: Practical goals in a multiple treatment experiment beyond classical hypothesis testing. The aim in our work is to find a procedure which can simultaneously test the null hypothesis and find the best arm efficiently using adaptive sampling.

A standard statistical approach to answer the above question would involve testing hypotheses of all pairs of treatments. Adaptive sampling strategies on the other hand, can essentially pick the hypotheses which compare the most promising arms. In recently published work with Aaditya Ramdas, Kevin Jamieson and Martin Wainwright [86], we show how multi-armed bandit algorithms can in fact be used to simultaneously achieve best-arm detection guarantees and traditional false discovery control for *multiple* experiments. The magic lies in the construction of always valid p-values which allows the algorithm to stop according to a criterion that ensures a desired probability of detecting the right arm.

Furthermore, instead of controlling the false alarm probability for one experiment only, the user generally needs to make immediate decisions for many different such experiments over time. Our doubly-sequential meta-framework controls the false discovery rate of multiple experiments with only as many samples as is needed to also determine the best treatment in each test with desired confidence. It also allows the scientist to employ their favorite best-arm and online FDR strategy independently. In order to cover more practically relevant scenarios, in another paper which is not presented in this thesis [63], we extend existing *online FDR frameworks* by allowing scientists to incorporate prior knowledge and assign decaying weight to discoveries in the past.

Part I

Estimation

Chapter 2

Guarantees for the Baum-Welch Algorithm

2.1 Introduction

Hidden Markov models (HMMs) are one of the most widely applied statistical models of the last 50 years, with major success stories in computational biology [27], signal processing and speech recognition [61], control theory [28], and econometrics [46] among other disciplines. At a high level, a hidden Markov model is a Markov process split into an observable component and an unobserved or latent component. From a statistical standpoint, the use of latent states makes the HMM generic enough to model a variety of complex real-world time series, while the Markovian structure enables relatively simple computational procedures.

In applications of HMMs, an important problem is to estimate the state transition probabilities and the parameterized output densities based on samples of the observable component. From classical theory, it is known that under suitable regularity conditions, the maximum likelihood estimate (MLE) in an HMM has good statistical properties [13]. However, given the potentially nonconvex nature of the likelihood surface, computing the global maximum that defines the MLE is not a straightforward task. In fact, the HMM estimation problem in full generality is known to be computationally intractable under cryptographic assumptions [72]. In practice, however, the Baum-Welch algorithm [9] is frequently applied and leads to good results. It can be understood as the specialization of the EM algorithm [26] to the maximum likelihood estimation problem associated with the HMM. Despite its wide use in many applications, the Baum-Welch algorithm can get trapped in local optima of the likelihood function. Understanding when this undesirable behavior occurs—or does not occur—has remained an open question for several decades.

A more recent line of work [58, 69, 37] has focused on developing tractable estimators for HMMs, via approaches that are distinct from the Baum-Welch algorithm. Nonetheless, it has been observed that the practical performance of such methods can be significantly improved by running the Baum-Welch algorithm using their estimators as the initial point; see, for

instance, the detailed empirical study in Kontorovich et al. ([49]). This curious phenomenon has been observed in other contexts [24], but has not been explained to date. Obtaining a theoretical characterization of when and why the Baum-Welch algorithm behaves well is the main objective of this chapter.

Related work and our contributions

Our work builds upon a framework for analysis of EM, as previously introduced by a subset of the current authors [3]; see also the follow-up work to regularized EM algorithms [90, 83]. All of this past work applies to models based on i.i.d. samples, and as we show in this chapter, there are a number of non-trivial steps required to derive analogous theory for the dependent variables that arise for HMMs. Before doing so, let us put the results of this chapter in context relative to older and more classical work on Baum-Welch and related algorithms.

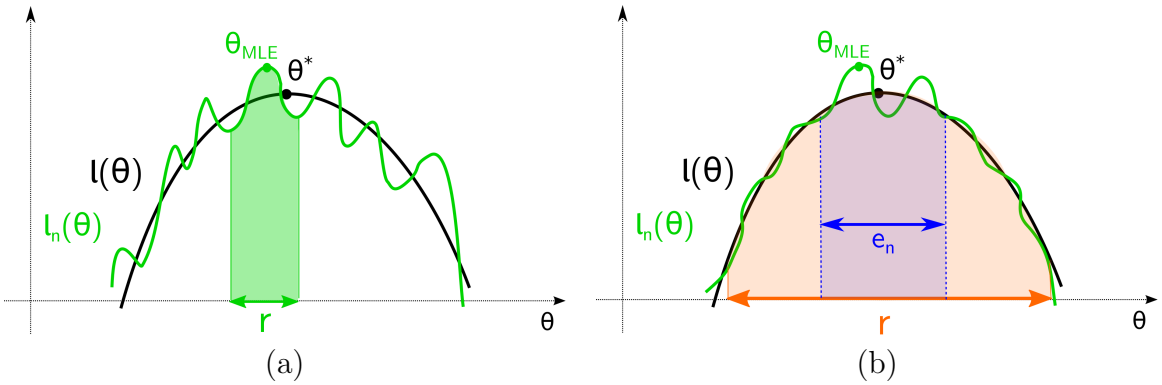


Figure 2.1: (a) A poorly behaved sample likelihood, for which there are many local optima at varying distances from the MLE. It would require an initialization extremely close to the MLE in order to ensure that the Baum-Welch algorithm would not be trapped at a sub-optimal fixed point. (b) A well-behaved sample likelihood, for which all local optima lie within an e_n-ball of the MLE, as well as the true parameter $\theta^\star$. In this case, the Baum-Welch algorithm, when initialized within a ball of large radius r, will converge to the ball of much smaller radius e_n. The goal of this chapter is to give sufficient conditions for when the sample likelihood exhibits this favorable structure.

Under mild regularity conditions, it is well-known that the maximum likelihood estimate (MLE) for an HMM is a consistent and asymptotically normal estimator; for instance, see Bickel et al. [13], as well as the expository works [22, 75]. On the algorithmic level, the original papers of Baum and co-authors [9, 8] showed that the Baum-Welch algorithm converges to a stationary point of the sample likelihood; these results are in the spirit of the classical convergence analysis of the EM algorithm [84, 26]. These classical convergence results only provide a relatively weak guarantee—namely, that if the algorithm is initialized sufficiently close to the MLE, then it will converge to it. However, the classical analysis does not

quantify the size of this neighborhood, and as a critical consequence, it *does not* rule out the pathological type of behavior illustrated in panel (a) of Figure 2.1. Here the sample likelihood has multiple optima, both a global optimum corresponding to the MLE as well as many local optima *far away from the MLE* that are also fixed points of the Baum-Welch algorithm. In such a setting, the Baum-Welch algorithm will only converge to the MLE if it is initialized in an extremely small neighborhood.

In contrast, the goal of this chapter is to give sufficient conditions under which the sample likelihood has the more favorable structure shown in panel (b) of Figure 2.1. Here, even though the MLE does not have a large basin of attraction, the sample likelihood has all of its optima (including the MLE) localized to a small region around the true parameter $\theta^\star$. Our strategy to reveal this structure, as in our past work [3], is to shift perspective: instead of studying convergence of Baum-Welch updates to the MLE, we study their convergence to an ϵ_n-ball of the true parameter $\theta^\star$, and moreover, instead of focusing exclusively on the sample likelihood, we first study the structure of the population likelihood, corresponding to the idealized limit of infinite data. Our first main result (Theorem 1) provides sufficient conditions under which there is a large ball of radius r, over which the population version of the Baum-Welch updates converge at a geometric rate to $\theta^\star$. Our second main result (Theorem 2) uses empirical process theory to analyze the finite-sample version of the Baum-Welch algorithm, corresponding to what is actually implemented in practice. In this finite sample setting, we guarantee that over the ball of radius r, the Baum-Welch updates will converge to an ϵ_n-ball with $\epsilon_n \ll r$, and most importantly, this ϵ_n-ball contains the true parameter $\theta^\star$. Typically this ball also contains the MLE with high-probability, but our theory does *not* guarantee convergence to the MLE, but rather to a point that is close to both the MLE and the true parameter $\theta^\star$ and whose statistical risk is equivalent to that of the MLE upto logarithmic factors.

These latter two results are abstract, applicable to a broad class of HMMs. We then specialize them to the case of a hidden Markov mixture consisting of two isotropic components, with means separated by a constant distance, and obtain concrete guarantees for this model. It is worth comparing these results to past work in the i.i.d. setting, for which the problem of Gaussian mixture estimation under various separation assumptions has been extensively studied (e.g. [25, 76, 10, 57]). The constant distance separation required in our work is much weaker than the separation assumptions imposed in most papers that focus on correctly labeling samples in a mixture model. Our separation condition is related to, but in general incomparable with the non-degeneracy requirements in other work [37, 36, 57].

Finally, let us discuss the various challenges that arise in studying the dependent data setting of hidden Markov models, and highlight some important differences with the i.i.d. setting [3]. In the non-i.i.d. setting, arguments passing from the population-based to sample-based updates are significantly more delicate. First of all, it is not even obvious that the population version of the Q-function—a central object in the Baum-Welch updates— exists. From a technical standpoint, various gradient smoothness conditions are much more difficult to establish, since the gradient of the likelihood no longer decomposes over the samples as in the i.i.d. setting. In particular, each term in the gradient of the likelihood is a function of all

observations. Finally, in order to establish the finite-sample behavior of the Baum-Welch algorithm, we can no longer appeal to standard i.i.d. concentration and empirical process techniques. Nor do we pursue the approach of some past work on HMM estimation (e.g. [37]), in which it is assumed that there are multiple independent samples of the HMM.[1] Instead, we directly analyze the Baum-Welch algorithm that practioners actually use—namely, one that applies to a single sample of an n-length HMM. In order to make the argument rigorous, we need to make use of more sophisticated techniques for proving concentration for dependent data [91, 59].

The remainder of this chapter is organized as follows. In Section 3.2, we introduce basic background on hidden Markov models and the Baum-Welch algorithm. Section 2.3 is devoted to the statement of our main results in the general setting, whereas Section 2.4 contains the more concrete consequences for the Gaussian output HMM. The main parts of our proofs are given in Section 3.5, with the more technical details deferred to the appendices.

2.2 Background and problem set-up

In this section, we introduce some standard background on hidden Markov models and the Baum-Welch algorithm.

Standard HMM notation and assumptions

We begin by defining a discrete-time hidden Markov model with hidden states taking values in a discrete space. Letting $\mathbb{Z}$ denote the integers, suppose that the observed random variables $\{X_i\}_{i\in\mathbb{Z}}$ take values in $\mathbb{R}^d$, and the latent random variables $\{Z_i\}_{i\in\mathbb{Z}}$ take values in the discrete space $[s] := \{1, \ldots, s\}$. The Markov structure is imposed on the sequence of latent variables. In particular, if the variable Z_1 has some initial distribution π_1, then the joint probability of a particular sequence $(z_1, \ldots, z_n)$ is given by

$$p(z_1, \ldots, z_n; \beta) = \pi_1(z_1; \beta) \prod_{i=1}^{n} p(z_i \mid z_{i-1}; \beta), \tag{2.1}$$

where the vector β is a particular parameterization of the initial distribution and Markov chain transition probabilities. We restrict our attention to the homogeneous case, meaning that the transition probabilities for step $(t-1) \to t$ are independent of the index t. Consequently, if we define the transition matrix $A \in \mathbb{R}^{s\times s}$ with entries

$$A(j, k; \beta) := p(z_2 = k \mid z_1 = j; \beta),$$

then the marginal distribution π_i of Z_i can be described by the matrix vector equation

$$\pi_i^T = \pi_1^T A^{i-1},$$

[1] The rough argument here is that it is possible to reduce an i.i.d. sampling model by cutting the original sample into many pieces, but this is not an algorithm that one would implement in practice.

where π_i and π_1 denote vectors belonging to the s-dimensional probability simplex.

We assume throughout that the Markov chain is aperiodic and recurrent, whence it has a unique stationary distribution $\bar{\pi}$, defined by the eigenvector equation $\bar{\pi}^T = \bar{\pi}^T A$. To be clear, both $\bar{\pi}$ and the matrix A depend on β, but we omit this dependence so as to simplify notation. We assume throughout that the Markov chain begins in its stationary state, so that $\pi_1 = \bar{\pi}$, and moreover, that it is reversible, meaning that

$$\bar{\pi}(j)A(j,k) = \bar{\pi}(k)A(k,j) \tag{2.2}$$

for all pairs $j, k \in [s]$.

A key quantity in our analysis is the mixing rate of the Markov chain. In particular, we assume the existence of *mixing constant* $\epsilon_{\text{mix}} \in (0, 1]$ such that

$$\epsilon_{\text{mix}} \leq \frac{p(z_i|z_{i-1};\beta)}{\bar{\pi}(z_i)} \leq \epsilon_{\text{mix}}^{-1} \tag{2.3}$$

for all $(z_i, z_{i-1}) \in [s] \times [s]$. This condition implies that the dependence on the initial distribution decays geometrically. More precisely, some simple algebra shows that

$$\sup_{\pi_1} \left\| \pi_1^T A^t - \pi_1^T \right\|_{\text{TV}} \leq c_0 \rho_{\text{mix}}^t \qquad \text{for all } t = 1, 2, \ldots, \tag{2.4}$$

where $\rho_{\text{mix}} = 1 - \epsilon_{\text{mix}}$ denotes the *mixing rate* of the process, and c_0 is a universal constant. Note that as $\epsilon_{\text{mix}} \to 1^-$, the Markov chain has behavior approaching that of an i.i.d. sequence, whereas as $\epsilon_{\text{mix}} \to 0^+$, its behavior becomes increasingly "sticky".

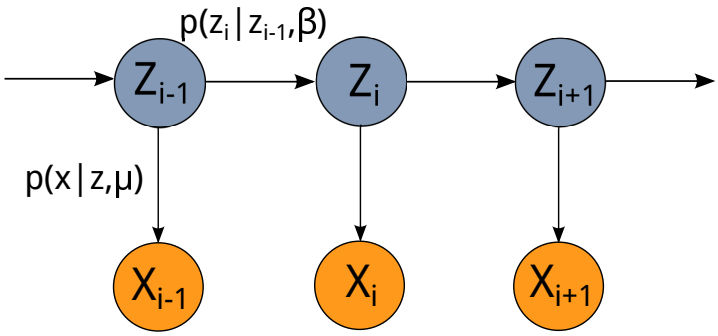


Figure 2.2: The hidden Markov model as a graphical model. The blue circles indicate observed variables Z_i, whereas the orange circles indicate latent variables X_i.

Associated with each latent variable Z_i is an observation $X_i \in \mathbb{R}^d$. We use $p(x_i|z_i;\mu)$ to denote the density of X_i given that $Z_i = z_i$, an object that we assume to be parameterized by a vector μ. Introducing the shorthand $x_1^n = (x_1, \ldots, x_n)$ and $z_1^n = (z_1, \ldots, z_n)$, the joint

probability of the sequence (x_1^n, z_1^n) (also known as the complete likelihood) can be written in the form

$$p(z_1^n, x_1^n; \theta) = \pi_1(z_1) \prod_{i=2}^{n} p(z_i \mid z_{i-1}; \beta) \prod_{i=1}^{n} p(x_i | z_i; \mu), \tag{2.5}$$

where the pair $\theta := (\beta, \mu)$ parameterizes the transition and observation functions. The likelihood then reads

$$p(x_1^n; \theta) = \sum_{z_1^n} p(z_1^n, x_1^n; \theta).$$

For our convenience in subsequent analysis, we also define a form of complete likelihood including an additional hidden variable z_0 which is not associated to any observation x_0

$$p(z_0^n, x_1^n; \theta) = \pi_0(z_0) \prod_{i=1}^{n} p(z_i \mid z_{i-1}; \beta) \prod_{i=1}^{n} p(x_i | z_i; \mu), \tag{2.6}$$

where $\pi_0 = \bar{\pi}$. Note that it preserves the usual relationship $\sum_{z_0^n} p(z_0^n, x_1^n; \theta) = p(x_1^n; \theta)$ between the ordinary and complete likelihoods in EM problems.

A simple example: A special case helps to illustrate these definitions. In particular, suppose that we have a Markov chain with $s = 2$ states. Consider a matrix of transition probabilities $A \in \mathbb{R}^{2 \times 2}$ of the form

$$A = \frac{1}{e^{\beta} + e^{-\beta}} \begin{bmatrix} e^{\beta} & e^{-\beta} \\ e^{-\beta} & e^{\beta} \end{bmatrix} = \begin{bmatrix} \zeta & 1 - \zeta \\ 1 - \zeta & \zeta \end{bmatrix}, \tag{2.7}$$

where $\zeta := \frac{e^{\beta}}{e^{\beta} + e^{-\beta}}$. By construction, this Markov chain is recurrent and aperiodic with the unique stationary distribution $\bar{\pi} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}^T$. Moreover, by calculating the eigenvalues of the transition matrix, we find that the mixing condition (2.4) holds with

$$\rho_{\text{mix}} := |2\zeta - 1| = |\tanh(\beta)|.$$

Suppose moreover that the observed variables in $\mathbb{R}^d$ are conditionally Gaussian, say with

$$p(x_t | z_t; \mu) = \begin{cases} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{ -\frac{1}{2\sigma^2} \|x - \mu\|_2^2 \right\} & \text{if } z_t = 1 \\ \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{ -\frac{1}{2\sigma^2} \|x + \mu\|_2^2 \right\} & \text{if } z_t = 2. \end{cases} \tag{2.8}$$

With this choice, the marginal distribution of each X_t is a two-state Gaussian mixture with mean vectors μ and $-\mu$, and covariance matrices $\sigma^2 I_d$. We provide specific consequences of our general theory for this special case in the sequel.

Baum-Welch updates for HMMs

We now describe the Baum-Welch updates for a general discrete-state hidden Markov model. As a special case of the EM algorithm, the Baum-Welch algorithm is guaranteed to ascend on the likelihood function of the hidden Markov model. It does so indirectly, by first computing a lower bound on the likelihood (E-step) and then maximizing this lower bound (M-step).

For a given integer $n \geq 1$, suppose that we observe a sequence $x_1^n = (x_1, \ldots, x_n)$ drawn from the marginal distribution over X_1^n defined by the model (2.5). The rescaled log likelihood of the sample path x_1^n is given by

$$\ell_n(\theta) = \frac{1}{n} \log \Big(\sum_{z_0^n} p(z_0^n, x_1^n; \theta) \Big)$$

The EM likelihood is based on lower bounding the likelihood via Jensen's inequality. For any choice of parameter θ' and positive integers $i \leq j$ and $a < b$, let $\mathbb{E}_{Z_i^j | x_a^b, \theta'}$ denote the expectation under the conditional distribution $p(Z_i^j \mid x_a^b; \theta')$. With this notation, the concavity of the logarithm and Jensen's inequality imply that for any choice of θ', we have the lower bound

$$\begin{aligned} \ell_n(\theta) &= \frac{1}{n} \log \big[\mathbb{E}_{Z_0^n | x_1^n, \theta'} \frac{p(Z_0^n, x_1^n; \theta)}{p(Z_0^n \mid x_1^n; \theta')} \big] \\ &\geq \underbrace{\frac{1}{n} \mathbb{E}_{Z_0^n | x_1^n, \theta'} \big[\log p(Z_0^n, x_1^n; \theta) \big]}_{Q_n(\theta \,|\, \theta')} + \underbrace{\frac{1}{n} \mathbb{E}_{Z_0^n | x_1^n, \theta'} \big[- \log p(Z_0^n \mid x_1^n; \theta') \big]}_{H_n(\theta')}. \end{aligned}$$

For a given choice of θ', the E-step corresponds to the computation of the function $\theta \mapsto Q_n(\theta \mid \theta')$. The M-step is defined by the EM operator $M_n : \widetilde{\Omega} \mapsto \widetilde{\Omega}$

$$M_n(\theta') = \arg\max_{\theta \in \widetilde{\Omega}} Q_n(\theta \mid \theta'), \tag{2.9}$$

where $\widetilde{\Omega}$ is the set of feasible parameter vectors. Overall, given an initial vector $\theta^0 = (\beta^0, \mu^0)$, the EM algorithm generates a sequence $\{\theta^t\}_{t=0}^{\infty}$ according to the recursion $\theta^{t+1} = M_n(\theta^t)$.

This description can be made more concrete for an HMM, in which case the Q-function takes the form

$$\begin{aligned} Q_n(\theta \mid \theta') = \frac{1}{n} \mathbb{E}_{Z_0 | x_1^n, \theta'} \big[\log \pi_0(Z_0; \beta) \big] + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z_{i-1}, Z_i | x_1^n, \theta'} \big[\log p(Z_i \mid Z_{i-1}; \beta) \big] \\ + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{Z_i | x_1^n, \theta'} \big[\log p(x_i \mid Z_i; \mu) \big], \end{aligned} \tag{2.10}$$

where the dependence of π_0 on β comes from the assumption that $\pi_0 = \bar{\pi}$. Note that the Q-function can be decomposed as the sum of a term which is solely dependent on μ, and another one which only depends on β—that is

$$Q_n(\theta \mid \theta') = Q_{1,n}(\mu \mid \theta') + Q_{2,n}(\beta \mid \theta') \tag{2.11}$$

where $Q_{1,n}(\mu \mid \theta') = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{Z_i|x_1^n,\theta'}\big[\log p(x_i \mid Z_i, \mu)\big]$, and $Q_{2,n}(\beta \mid \theta')$ collects the remaining terms. In order to compute the expectations defining this function (E-step), we need to determine the marginal distributions over the singletons Z_i and pairs (Z_i, Z_{i+1}) under the joint distribution $p(Z_0^n \mid x_1^n; \theta')$. These marginals can be obtained efficiently using a recursive message-passing algorithm, known either as the forward-backward or sum-product algorithm [50, 82].

In the M-step, the decomposition (2.11) suggests that the maximization over the two components (β, μ) can also be decoupled. Accordingly, with a slight abuse of notation, we often write

$$M_n^{\mu}(\theta') = \arg\max_{\mu \in \Omega_\mu} Q_{1,n}(\mu \mid \theta'), \quad \text{and} \quad M_n^{\beta}(\theta') = \arg\max_{\beta \in \Omega_\beta} Q_{2,n}(\beta \mid \theta')$$

for these two decoupled maximization steps, where Ω_β and Ω_μ denote the feasible set of transition and observation parameters respectively and $\widetilde{\Omega} := \Omega_\beta \times \Omega_\mu$. In the following, unless otherwise stated, $\Omega_\mu = \mathbb{R}^d$, so that the maximization over the observation parameters is unconstrained.

2.3 Main results

We now turn to the statement of our main results, along with a discussion of some of their consequences. The first step is to establish the existence of an appropriate population analog of the Q-function. Although the existence of such an object is a straightforward consequence of the law of large numbers in the case of i.i.d. data, it requires some technical effort to establish existence for the case of dependent data; in particular, we do so using a k-truncated version of the full Q-function (see Proposition 2.3.1). This truncated object plays a central role in the remainder of our analysis. In particular, we first analyze a version of the Baum-Welch updates on the expected k-truncated Q-function for an extended sequence of observations x_{1-k}^{n+k}, and provide sufficient conditions for these population-level updates to be contractive (see Theorem 1). We then use non-asymptotic forms of empirical process theory to show that under suitable conditions, the actual sample-based EM updates—i.e., the updates that are actually implemented in practice—are also well-behaved in this region with high probability (see Theorem 2). In subsequent analysis to follow in Section 2.4, we show that this initialization radius is suitably large for an HMM with Gaussian outputs.

Existence of population Q-function

In the analysis of [3], the central object is the notion of a population Q-function—namely, the function that underlies the EM algorithm in the idealized limit of infinite data. In their setting of i.i.d. data, the standard law of large numbers ensures that as the sample size n increases, the sample-based Q-function approaches its expectation, namely the function

$$\bar{Q}(\theta \mid \theta') = \mathbb{E}\big[Q_n(\theta \mid \theta')\big] \;=\; \mathbb{E}\big[\mathbb{E}_{Z_1|X_1,\theta'}\big[\log p(X_1, Z_1; \theta)\big]\big].$$

Here we use the shorthand $\mathbb{E}$ for the expectation over all samples X that are drawn from the joint distribution (in this case $\mathbb{E} := \mathbb{E}_{X_1^n|\theta^*}$).

When the samples are dependent, the quantity $\mathbb{E}\big[Q_n(\theta \mid \theta')\big]$ is no longer independent of n, and so an additional step is required. A reasonable candidate for a general definition of the population Q-function is given by

$$\bar{Q}(\theta \mid \theta') := \lim_{n \to +\infty} [\mathbb{E} Q_n(\theta \mid \theta')]. \tag{2.12}$$

Although it is clear that this definition is sensible in the i.i.d. case, it is necessary for dependent sampling schemes to prove that the limit given in definition (2.12) actually exists.

In this chapter, we do so by considering a suitably truncated version of the sample-based Q-function. Similar arguments have been used in past work (e.g., [22, 75]) to establish consistency of the MLE; here our focus is instead on the behavior of the Baum-Welch algorithm. Let us consider a sequence $\{(X_i, Z_i)\}_{i=1-k}^{n+k}$, assumed to be drawn from the stationary distribution of the overall chain. Recall that $\mathbb{E}_{Z_i^j|x_a^b,\theta}$ denotes expectations taken over the distribution $p(Z_i^j \mid x_a^b, \theta)$. Then, for a positive integer k to be chosen, we define

$$Q_n^k(\theta \mid \theta') = \frac{1}{n}\Big[\mathbb{E}_{Z_0|x_{-k}^k,\theta'} \log p(Z_1; \beta) + \sum_{i=1}^{n} \mathbb{E}_{Z_{i-1}^i|x_{i-k}^{i+k},\theta'} \log p(Z_i \mid Z_{i-1}; \beta) + \sum_{i=1}^{n} \mathbb{E}_{Z_i|x_{i-k}^{i+k},\theta'} \log p(x_i \mid Z_i; \mu)\Big]. \tag{2.13}$$

In an analogous fashion to the decomposition in equation (2.10), we can decompose Q_n^k in the form

$$Q_n^k(\theta \mid \theta') = Q_{1,n}^k(\mu \mid \theta') + Q_{2,n}^k(\beta \mid \theta').$$

We associate with this triplet of Q-functions the corresponding EM operators $M_n^k(\theta')$, $M_n^{\mu,k}(\theta')$ and $M_n^{\beta,k}(\theta')$ as in Equation (2.9). Note that as opposed to the function Q_n from equation (2.10), the definition of Q_n^k involves variables Z_i, Z_{i-1} that are not conditioned on the full observation sequence x_1^n, but instead only on a $2k$ window centered around the index i. By construction, we are guaranteed that the k-truncated population function and its decomposed analogs given by

$$\begin{aligned} \bar{Q}^k(\theta \mid \theta') &:= \lim_{n \to \infty} \mathbb{E} Q_n^k(\theta \mid \theta') = \mathbb{E} Q_{1,n}^k(\mu \mid \theta') + \lim_{n \to \infty} \mathbb{E} Q_{2,n}^k(\beta \mid \theta') \\ &:= \bar{Q}_1^k(\mu \mid \theta') + \bar{Q}_2^k(\beta \mid \theta') \end{aligned} \tag{2.14}$$

are well-defined. In particular, due to stationarity of the random sequences $\{p(z_i \mid X_{i-k}^{i+k})\}_{i=1}^n$ and $\{p(z_{i-1}^i \mid X_{i-k}^{i+k})\}_{i=1}^n$, the expectation over $\{(X_i, Z_i)\}_{i=1-k}^{n+k}$ is independent of the sample size n. Notice that the Baum-Welch algorithm in practice essentially corresponds to using $k = n$.

Our first result uses the existence of this truncated population object in order to show that the standard population Q-function from equation (2.12) is indeed well-defined. In doing so, we make use of the sup-norm

$$\|Q_1 - Q_2\|_\infty := \sup_{\theta,\theta' \in \widetilde{\Omega}} \Big| Q_1(\theta \mid \theta') - Q_2(\theta \mid \theta') \Big|. \tag{2.15}$$

We require in the following that the observation densities satisfy the following boundedness condition

$$\sup_{\theta \in \widetilde{\Omega}} \mathbb{E}\Big[\max_{z_i \in [s]} \big| \log p(X_i \mid z_i, \theta) \big| \Big] < \infty. \tag{2.16}$$

Proposition 2.3.1. *Under the previously stated assumptions, the population function $\bar{Q}$ defined in equation* (2.12) *exists.*

The proof of this claim is given in Section 2.7. It hinges on the following auxiliary claim, which bounds the difference between $\mathbb{E}Q_n$ and the k-truncated Q-function as

$$\|\mathbb{E}Q_n - \bar{Q}^k\|_\infty \leq \frac{c\, s^4}{\epsilon_{\min}^9 \overline{\pi}_{\min}^2} \big(1 - \epsilon_{\min} \overline{\pi}_{\min}\big)^k + \frac{c(\overline{\pi}_{\min}, s, \epsilon_{\min})}{n}, \tag{2.17}$$

where $\overline{\pi}_{\min} := \min_{\beta \in \Omega_\beta, j \in [s]} \overline{\pi}(j \mid \beta)$ is the minimum probability in the stationary distribution, $\epsilon_{\min}$ is the mixing constant from equation (2.3), and $c(\cdot)$ is a constant dependent only on the inherent model parameters. The dependencies on $\epsilon_{\min}$ and $\overline{\pi}_{\min}$ are not optimized here. Since this bound holds for all n, it shows that the population function $\bar{Q}$ can be uniformly approximated by $\bar{Q}^k$, with the approximation error decreasing geometrically as the truncation level k grows. This fact plays an important role in the analysis to follow.

Analysis of updates based on $\bar{Q}^k$

Our ultimate goal is to establish a bound on the difference between the sample-based Baum-Welch estimate and θ^*, in particular showing contraction of the Baum-Welch update towards the true parameter. Our strategy for doing so involves first analyzing the Baum-Welch iterates at the population level, which is the focus of this section.

The quantity $\bar{Q}$ is significant for the EM updates because the parameter θ^* satisfies the self-consistency property $\theta^* = \arg\max_\theta \bar{Q}(\theta \mid \theta^*)$. In the i.i.d. setting, the function $\bar{Q}$ can often be computed in closed form, and hence directly analyzed, as was done in past work [3]. In the HMM case, this function $\bar{Q}$ no longer has a closed form, so an alternative route is needed. Here we analyze the population version via the truncated function $\bar{Q}^k$ (2.14) instead, where k is a given truncation level (to be chosen in the sequel). Although θ^* is no longer a fixed point of $\bar{Q}^k$, the bound (2.17) combined with the assumption of strong concavity of $\bar{Q}^k$ imply an upper bound on the distance of the maximizers of $\bar{Q}^k$ and $\bar{Q}$.

With this setup, we consider an idealized population-level algorithm that, based on some initialization $\widetilde{\theta}^0 \in \Omega = \mathbb{B}_2\big(r; \mu^*\big) \times \Omega_\beta$, generates the sequence of iterates

$$\widetilde{\theta}^{t+1} = \overline{M}^k(\widetilde{\theta}^t) := \arg\max_{\theta \in \widetilde{\Omega}} \bar{Q}^k(\theta \mid \widetilde{\theta}^t). \tag{2.18}$$

where $\widetilde{\Omega} = \Omega_\beta \times \Omega_\mu$ is a larger set than Ω, especially $\Omega_\mu = \mathbb{R}^d$. Since $\bar{Q}^k$ is an approximate version of $\bar{Q}$, the update operator $\overline{M}^k$ should be understood as an approximation to the idealized population EM operator $\overline{M}$ where the maximum is taken with respect to $\bar{Q}$. As part (a) of the following theorem shows, the approximation error is well-controlled under suitable conditions. We analyze the convergence of the sequence $\{\widetilde{\theta}^t\}_{t=0}^\infty$ in terms of the norm $\|\cdot\|_\star : \Omega_\mu \times \Omega_\beta \to \mathbb{R}^+$ given by

$$\|\theta - \theta^*\|_\star = \|(\mu, \beta) - (\mu^*, \beta^*)\|_\star := \|\mu - \mu^*\|_2 + \|\beta - \beta^*\|_2. \tag{2.19}$$

Contraction in this norm implies that both parameters μ, β converge linearly to the true parameter.

Conditions on $\bar{Q}^k$: Let us now introduce the conditions on the truncated function $\bar{Q}^k$ that underlie our analysis. For this purpose, we concentrate on a potentially smaller set

$$\Omega := \mathbb{B}_2(r; \mu^*) \times \Omega_\beta$$

with radius $r > 0$, where Ω_β is the set of allowable HMM transition parameters. The goal is to find the largest $\Omega \subset \widetilde{\Omega}$, in which said conditions are fulfilled. This set Ω is then equivalent to the basin of attraction, i.e. the set in which we can initialize the algorithm and obtain linear convergence to a good optimum.

First, let us say that the function $\bar{Q}^k(\cdot \mid \theta')$ is $(\lambda_\mu, \lambda_\beta)$-*strongly concave* in Ω if for all $\theta' \in \Omega$ we have

$$\bar{Q}_1^k(\mu_1 \mid \theta') - \bar{Q}_1^k(\mu_2 \mid \theta') - \langle \nabla_\mu \bar{Q}_1^k(\mu_2 \mid \theta'), \mu_1 - \mu_2 \rangle \leq -\frac{\lambda_\mu}{2}\|\mu_1 - \mu_2\|_2^2 \tag{2.20a}$$

$$\text{and} \qquad \bar{Q}_2^k(\beta_1 \mid \theta') - \bar{Q}_2^k(\beta_2 \mid \theta') - \langle \nabla_\beta \bar{Q}_2^k(\beta_2 \mid \theta'), \beta_1 - \beta_2 \rangle \leq -\frac{\lambda_\beta}{2}\|\beta_1 - \beta_2\|_2^2 \tag{2.20b}$$

for all $(\mu_1, \beta_1), (\mu_2, \beta_2) \in \Omega$.

Second, we impose *first-order stability* conditions on the gradients of each component of $\bar{Q}^k$:

- For each $\mu \in \Omega_\mu, \theta' \in \Omega$, we have

$$\|\nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta') - \nabla_\mu \bar{Q}_1^k(\mu \mid \mu^*, \beta')\|_2 \leq L_{\mu,1}\|\mu' - \mu^*\|_2 \tag{2.21a}$$

$$\|\nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta') - \nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta^*)\|_2 \leq L_{\mu,2}\|\beta' - \beta^*\|_2, \tag{2.21b}$$

We refer to this condition as L_μ-FOS for short.

- Secondly, for all $\beta \in \Omega_\beta, \theta' \in \Omega$, we require that

$$\|\nabla_\beta \bar{Q}_2^k(\beta \mid \mu', \beta') - \nabla_\beta \bar{Q}_2^k(\beta \mid \mu^*, \beta')\|_2 \leq L_{\beta,1}\|\mu' - \mu^*\|_2 \tag{2.22a}$$

$$\|\nabla_\beta \bar{Q}_2^k(\beta \mid \mu', \beta') - \nabla_\beta \bar{Q}_2^k(\beta \mid \mu', \beta^*)\|_2 \leq L_{\beta,2}\|\beta' - \beta^*\|_2. \tag{2.22b}$$

We refer to this condition as L_β-FOS for short. The experienced reader may find that the (L_μ, L_β)-FOS conditions look intriguingly similar to the Lipschitz gradient conditions often encountered when proving geometric convergence for gradient descent methods. On a high level, smoothness requires function values of one function to be close for any pair of arguments that are close. Although our conditions seem to invoke Lipschitz gradients as well, it is actually of a completely different nature. The important difference arises from the existence of two parameters, as we now clarify.

As opposed to gradient descent, the EM updates optimize over the first parameter θ of a function $\bar{Q}^k(\cdot \mid \theta')$ defined by the second parameter θ' at every time step. If we could access $\bar{Q}^k(\cdot \mid \theta^*)$, EM would converge in one step to the true optimum. Therefore, if we can guarantee that $\bar{Q}^k(\cdot \mid \theta')$ and $\bar{Q}^k(\cdot \mid \theta^*)$ are close in some sense, there should be good reasons to hope that under some more regularity assumptions the maximizers are close as well, i.e. that $\overline{M}^k(\theta')$ is close to θ^*.

The (L_μ, L_β)-FOS conditions are precisely encouraging closeness of these two functions in a first-order sense. In particular, we require the gradients (with respect to the first argument θ) to be Lipschitz in the second argument θ'. Typical smoothness however is a property with respect to a *fixed* function (i.e. a fixed θ' in our case) and thus requires gradients to be Lipschitz in the first argument. Loosely speaking it upper bounds the curvature of said function, and thus is more like a second-order condition by nature. This distinction also explains why (L_μ, L_β)-FOS conditions require to be uniformly satisfied only over the first argument, while one of the second arguments can be fixed at μ^* or β^* respectively. Finally, as we show in Section 2.4, these conditions hold for concrete models.

Convergence guarantee for $\bar{Q}^k$-updates: We are now equipped to state our main convergence guarantee for the updates. It involves the quantities

$$L := \max\{L_{\mu_1}, L_{\mu_2}\} + \max\{L_{\beta_1}, L_{\beta_2}\}, \quad \lambda := \min\{\lambda_\mu, \lambda_\beta\} \quad \text{and} \quad \kappa := \frac{L}{\lambda}, \tag{2.23}$$

with κ generally required to be smaller than one, as well as the additive norm $\|\cdot\|_\star$ from equation (2.19).

Part (a) of the theorem controls the *approximation error* induced by using the k-truncated function $\bar{Q}^k$ as opposed to the exact population function $\bar{Q}$, whereas part (b) guarantees a *geometric rate of convergence* in terms of κ defined above in equation (2.23).

Theorem 1. *(a)* Approximation guarantee: *Under the mixing condition* (2.4)*, density boundedness condition* (2.16)*, and* $(\lambda_\mu, \lambda_\beta)$*-strong concavity condition* (2.20)*, there is a universal*

constant c_0 such that

$$\|\overline{M}^k(\theta) - \overline{M}(\theta)\|_\star^2 \leq \underbrace{\frac{Cs^4}{\lambda\, \epsilon_{\min}^9 \bar{\pi}_{\min}^2}\big(1 - \epsilon_{\min}\bar{\pi}_{\min}\big)^k}_{=:\varphi^2(k)} \qquad \text{for all } \theta \in \Omega, \tag{2.24}$$

where s is the number of states, and $\bar{\pi}_{\min} := \min_{\beta\in\Omega_\beta}\min_{j\in[s]}\bar{\pi}(j;\beta)$.

(b) Convergence guarantee: *Suppose in addition that the (L_μ, L_β)-FOS conditions* (2.21),(2.22) *holds with parameter $\kappa \in (0,1)$ as defined in* (2.23) *for $\theta, \theta' \in \Omega = \mathbb{B}_2\big(r;\mu^*\big)\times\Omega_\beta$, and that the truncation parameter k is sufficiently large to ensure that*

$$\varphi(k) \leq \big(1-\kappa\big)r - \kappa\max_{\beta\in\Omega_\beta}\|\beta - \beta^*\|_2.$$

Then, given an initialization $\widetilde{\theta}^0 \in \Omega$, the iterates $\{\widetilde{\theta}^t\}_{t=0}^\infty$ generated by the $\overline{M}^k$ operator satisfy the bound

$$\|\widetilde{\theta}^t - \theta^\star\|_\star \leq \kappa^t\|\widetilde{\theta}^0 - \theta^\star\|_\star + \frac{1}{1-\kappa}\varphi(k). \tag{2.25}$$

Note that the subtlety here is that $\theta^\star$ is no longer a fixed point of the operator $\overline{M}^k$, due to the error induced by the k^{th}-order truncation. Nonetheless, under the mixing condition, as the bounds (2.24) and (2.25) show, this approximation error is controlled, and decays exponentially in k. The proof of the recursive bound (2.25) is based on showing that

$$\|\overline{M}^k(\theta) - \overline{M}^k(\theta^\star)\|_\star \leq \kappa\|\theta - \theta^*\|_\star \tag{2.26}$$

for any $\theta \in \Omega$. Inequality (2.26) is equivalent to stating that the operator $\overline{M}^k$ is contractive, i.e. that applying $\overline{M}^k$ to the pair θ and $\theta^\star$ always decreases the distance.

Finally, when Theorem 1 is applied to a concrete model, the task is to find a big r and Ω_β such that the conditions in the theorem are satisfied, and we do so for the Gaussian output HMM in Section 2.4.

Sample-based results

We now turn to a result that applies to the sample-based form of the Baum-Welch algorithm—that is, corresponding to the updates that are actually applied in practice. For a tolerance parameter $\delta \in (0,1)$, we let $\varphi_n(\delta, k)$ be the smallest positive scalar such that

$$\mathbb{P}\Big[\sup_{\theta\in\Omega}\|M_n(\theta) - M_n^k(\theta)\|_\star \geq \varphi_n(\delta,k)\Big] \leq \delta. \tag{2.27a}$$

This quantity bounds the approximation error induced by the k-truncation, and is the sample-based analogue of the quantity $\varphi(k)$ appearing in Theorem 1(a). For each $\delta \in (0,1)$, we let $\epsilon_n^{\mu}(\delta, k)$ and $\epsilon_n^{\beta}(\delta, k)$ denote the smallest positive scalars such that

$$\mathbb{P}\big[\sup_{\theta \in \Omega} \|M_n^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta)\|_2 \geq \epsilon_n^{\mu}(\delta, k)\big] \leq \delta, \quad \text{and} \tag{2.27b}$$

$$\mathbb{P}\big[\sup_{\theta \in \Omega} \|M_n^{\beta,k}(\theta) - \overline{M}^{\beta,k}(\theta)\|_2 \geq \epsilon_n^{\beta}(\delta, k)\big] \leq \delta,$$

where $M_n^{\mu,k}(\cdot)$ and $M_n^{\beta,k}(\cdot)$ correspond to the truncated versions of $M_n^{\mu}(\cdot)$ and $M_n^{\beta}(\cdot)$. Furthermore we define $\epsilon_n(\delta, k) := \epsilon_n^{\mu}(\delta, k) + \epsilon_n^{\beta}(\delta, k)$. For a given truncation level k, these values give an upper bound on the difference between the population and sample-based M-operators, as induced by having only a finite number n of samples.

Theorem 2 (Sample Baum-Welch)**.** *Suppose that the truncated population EM operator $\overline{M}^k$ satisfies the local contraction bound* (2.26) *with parameter $\kappa \in (0,1)$ in Ω. For a given sample size n, suppose that (k, n) are sufficiently large to ensure that*

$$\varphi_n(\delta, k) + \varphi(k) + \epsilon_n^{\mu}\big(\delta, k\big) \leq (1-\kappa)\, r - \kappa \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2. \tag{2.28a}$$

Then given any initialization $\widehat{\theta}^0 \in \Omega$, with probability at least $1-2\delta$, the Baum-Welch sequence $\{\widehat{\theta}^t\}_{t=0}^{\infty}$ satisfies the bound

$$\|\widehat{\theta}^t - \theta^\star\|_\star \leq \underbrace{\kappa^t \|\widehat{\theta}^0 - \theta^\star\|_\star}_{\textit{Geometric decay}} + \underbrace{\frac{1}{1-\kappa}\Big\{\varphi_n\big(\delta, k\big) + \varphi(k) + \epsilon_n\big(\delta, k\big)\Big\}}_{\textit{Residual error } e_n}. \tag{2.28b}$$

The bound (2.28b) shows that the distance between $\widehat{\theta}^t$ and $\theta^\star$ is bounded by two terms: the first decays geometrically as t increases, and the second term corresponds to a residual error term that remains independent of t. Thus, by choosing the iteration number T larger than $\frac{\log(2r/\epsilon)}{\log \kappa}$, we can ensure that the first term is at most ϵ. The residual error term can be controlled by requiring that the sample size n is sufficiently large, and then choosing the truncation level k appropriately. We provide a concrete illustration of this procedure in the following section, where we analyze the case of Gaussian output HMMs. In particular, we can see that the residual error is of the same order as for the MLE and that the required initialization radius is optimal up to constants. Let us emphasize here that k as well as the truncated operators are purely theoretical objects which were introduced for the analysis.

2.4 Concrete results for the Gaussian output HMM

We now return to the concrete example of a Gaussian output HMM, as first introduced in Section 2.2, and specialize our general theory to it. Before doing so, let us make some

preliminary comments about our notation and assumptions. Recall that our Gaussian output HMM is based on $s = 2$ hidden states, using the transition matrix from equation (2.7), and the Gaussian output densities from equation (2.8). For convenience of analysis, we let the hidden variables Z_i take values in $\{-1, 1\}$. In addition, we require that the mixing coefficient $\rho_{\text{mix}} = 1 - \epsilon_{\text{mix}}$ is bounded away from 1 in order to ensure that the mixing condition (2.3) is fulfilled. We denote the upper bound for ρ_{mix} as $b < 1$ so that $\rho_{\text{mix}} \leq b$ and $\epsilon_{\text{mix}} \geq 1 - b$. The feasible set of the probability parameter ζ and its log odds analog $\beta = \frac{1}{2} \log \left(\frac{\zeta}{1-\zeta} \right)$ are then given by

$$\Omega_\zeta = \left\{ \zeta \in \mathbb{R} \mid \frac{1-b}{2} \leq \zeta \leq \frac{1+b}{2} \right\}, \quad \text{and} \quad \Omega_\beta = \Big\{ \beta \in \mathbb{R} \ \mid \ |\beta| < \underbrace{\frac{1}{2} \log \big(\frac{1+b}{1-b} \big)}_{\beta_B} \Big\}. \tag{2.29}$$

Explicit form of Baum-Welch updates

We begin by deriving an explicit form of the Baum-Welch updates for this model. Using this notation, the Baum-Welch updates take the form

$$\widehat{\mu}^{t+1} = \frac{1}{n} \sum_{i=1}^{n} (2p(Z_i = 1 \mid x_1^n; \widehat{\theta}^t) - 1) x_i, \tag{2.30a}$$

$$\widehat{\zeta}^{t+1} = \Pi_{\Omega_\zeta} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{Z_i} p(Z_i = Z_{i+1} \mid x_1^n; \widehat{\theta}^t) \right), \text{ and} \tag{2.30b}$$

$$\widehat{\beta}^{t+1} = \frac{1}{2} \log \big(\frac{\widehat{\zeta}^{t+1}}{1 - \widehat{\zeta}^{t+1}} \big), \tag{2.30c}$$

where Π_{Ω_ζ} denotes the Euclidean projection onto the set Ω_ζ. Note that the maximization steps are carried out on the decomposed Q-functions $Q_{1,n}(\cdot \mid \theta^t), Q_{2,n}(\cdot \mid \theta^t)$. In addition, since we are dealing with a one-dimensional quantity β, the projection of the unconstrained maximizer onto the interval Ω_ζ is equivalent to the constrained maximizer over the feasible set Ω_ζ. This step is in general not valid for higher dimensional transition parameters.

Population and sample guarantees

We now use the results from Section 2.3 to show that the population and sample-based version of the Baum-Welch updates are linearly convergent in a ball around θ^* of fixed radius. In establishing the population-level guarantee, the key conditions which need to be fulfilled—and the one that are the most technically challenging to establish— are the (L_μ, L_β)-FOS conditions (2.21), (2.22). In particular, we want to show that these conditions hold with Lipschitz constants L_μ, L_β that decrease exponentially with the separation of the mixtures. As a consequence, we obtain that for large enough separation $\frac{L}{\lambda} < 1$, i.e. the EM operator is contractive towards the true parameter.

In order to ease notation, our explicit tracking of parameter dependence is limited to the standard deviation σ and Euclidean norm $\|\mu^*\|_2$, which together determine the signal-to-noise ratio $\eta^2 := \frac{\|\mu^*\|_2^2}{\sigma^2}$ of the mixture model. Throughout this section, we therefore use c_0, c_1 to denote universal constants and C_0, C_1 for quantities that do not depend on $(\|\mu^*\|_2, \sigma)$, but may depend on other parameters such as $\overline{\pi}_{\min}$, ρ_{mix}, b, and so on.

We begin by stating a result for the sequence $\{\widetilde{\theta}^t\}_{t=0}^{\infty}$ obtained by repeatedly applying the k-truncated population-level Baum-Welch update operator $\overline{M}^k$. Our first corollary establishes that this sequence is linearly convergent, with a convergence rate $\kappa = \kappa(\eta)$ that is given by

$$\kappa(\eta) := \frac{C_1 \eta^2 (\eta^2 + 1)\, \mathrm{e}^{-c_2 \eta^2}}{1 - b^2}. \tag{2.31}$$

Corollary 1 (Population Baum-Welch)**.** *Consider a two-state Gaussian output HMM that is mixing (i.e. satisfies equation* (2.3)*), and with its SNR lower bounded as $\eta^2 \geq C$ for a sufficiently large constant C. Given the radius $r = \frac{\|\mu^*\|_2}{4}$, suppose that the truncation parameter k is sufficiently large to ensure that $\varphi(k) \leq (1-\kappa) r - \kappa \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2$. Then for any initialization $\widetilde{\theta}^0 = (\widetilde{\mu}^0, \widetilde{\beta}^0) \in \mathbb{B}_2\big(r; \mu^*\big) \times \Omega_\beta$, the sequence $\{\widetilde{\theta}^t\}_{t=0}^{\infty}$ generated by $\overline{M}^k$ satisfies the bound*

$$\|\widetilde{\theta}^t - \theta^*\|_\star \leq \kappa^t \|\widetilde{\theta}^0 - \theta^*\|_\star + \frac{1}{1-\kappa} \varphi(k) \tag{2.32}$$

for all iterations $t = 1, 2, \ldots$.

From definition (2.31) it follows that as long as the signal-to-noise ratio η is larger than a universal constant, the convergence rate $\kappa(\eta) < 1$. The bound (2.32) then ensures a type of contraction and the pre-condition $\varphi(k) \leq (1-\kappa) r - \kappa \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2$ can be satisfied by choosing the truncation parameter k large enough. If we use a finite truncation parameter k, then the contraction occurs up to the error floor given by $\varphi(k)$, which reflects the bias introduced by truncating the likelihood to a window of size k. At the population level (in which the effective sample size is infinite), we could take the limit $k \to \infty$ so as to eliminate this bias. However, this is no longer possible in the finite sample setting, in which we must necessarily have $k \ll n$. While large k give a better truncation approximation, it allows for fewer samples which are "sufficiently independent" from each other within the sequence. We can see in the proof of Corollary 2 that $k \gtrsim \log n$ is a good choice to obtain an adequate trade-off.

Corollary 2 (Sample Baum-Welch iterates)**.** *For a given tolerance $\delta \in (0,1)$, suppose that the sample size is lower bounded as $n \geq \frac{C_1}{\|\mu^*\|_2^2 \sigma^2} (\eta^2 + 1)^3 d \log^8(\frac{d}{\delta})$. Then under the conditions of Corollary 1 and $\eta^2 \geq C \log \frac{1}{1-b^2}$, with probability at least $1 - \delta$, we have*

$$\|\widetilde{\theta}^t - \theta^*\|_\star \leq \kappa^t \|\widehat{\theta}^0 - \theta^*\|_\star + \frac{C}{\sigma} \frac{\Big(\frac{\|\mu^*\|_2^2}{\sigma^2} + 1\Big)^{3/2} \sqrt{\frac{d \log^8(n/\delta)}{n}}}{1 - \kappa}. \tag{2.33}$$

Remarks: As a consequence of the bound (2.33), if we are given a sample size $n \gtrsim d \log^8 d$, then taking $T \approx \log n$ iterations is guaranteed to return an estimate $(\widehat{\mu}^T, \widehat{\beta}^T)$ with error of the order $\sqrt{\frac{d \log^8(n)}{n}}$.

In order to interpret this guarantee, note that in the case of symmetric Gaussian output HMMs as in Section 2.4, standard techniques can be used to show that the minimax rate of estimating μ^* in Euclidean norm scales as $\sqrt{\frac{d}{n}}$. If we could compute the MLE in polynomial time, then its error would also exhibit this scaling. The significance of Corollary 2 is that it shows that the Baum-Welch update achieves this minimax risk of estimation up to logarithmic factors.

Moreover, it should be noted that the initialization radius given here is essentially optimal up to constants. Because of the symmetric nature of the population log-likelihood, the all zeroes vector is a stationary point. Consequently, the maximum Euclidean radius of any basin of attraction for one of the observation parameters—that is, either μ^* or $-\mu^*$—can at most be $r = \|\mu^*\|_2$. Note that our initialization radius only differs from this maximal radius by a small constant factor.

Simulations

In this section, we provide the results of simulations that confirm the accuracy of our theoretical predictions for two-state Gaussian output HMMs. In all cases, we update the estimates for the mean vector $\widehat{\mu}^{t+1}$ and transition probability $\hat{\zeta}^{t+1}$ according to equation (2.30); for convenience, we update ζ as opposed to β. The true parameters are denoted by μ^* and ζ^*.

In all simulations, we fix the mixing parameter to $\rho_{\text{mix}} = 0.6$, generate initial vectors $\widehat{\mu}^0$ randomly in a ball of radius $r := \frac{\|\mu^*\|_2}{4}$ around the true parameter μ^*, and set $\widehat{\zeta}^0 = \frac{1}{2}$. Finally, the estimation error of the mean vector μ is computed as $\log_{10} \|\hat{\mu} - \mu^*\|_2$. Since the transition parameter estimation errors behave similarly to the observation parameter in simulations, we omit the corresponding figures here.

Figure 2.3 depicts the convergence behavior of the Baum-Welch updates, as assessed in terms of both the optimization and the statistical error. Here we run the Baum-Welch algorithm for a fixed sample sequence X_1^n drawn from a model with SNR $\eta^2 = 1.5$ and $\zeta = 0.2$, using different random initializations in the ball around μ^* with radius $\frac{\|\mu^*\|_2}{4}$. We denote the final estimate of the $i-$th trial by $\widehat{\mu}_i$. The curves in blue depict the *optimization error*—that is, the differences between the Baum-Welch iterates $\widehat{\mu}_i^t$ using the i-th initialization, and $\widehat{\mu}_1$. On the other hand, the red lines represent the *statistical error*—that is, the distance of the iterates from the true parameter μ^*.

For both family of curves, we observe linear convergence in the first few iterations until an error floor is reached. The convergence of the statistical error aligns with the theoretical prediction in upper bound (2.33) of Corollary 2. The (minimax-optimal) error floor in the curve corresponds to the residual error and the e_n–region in Figure 2.1. In addition, the blue optimization error curves show that for different initializations, the Baum-Welch algorithm

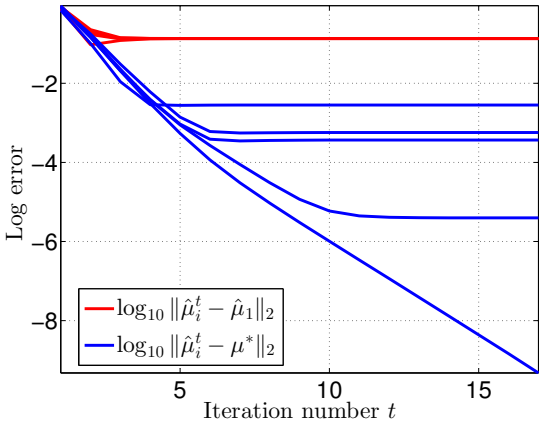


Figure 2.3: Plot of the convergence of the optimization error $\log \|\widehat{\mu}_i^t - \widehat{\mu}_1\|_2$, plotted in blue, and the statistical error $\log \|\widehat{\mu}_i^t - \mu^*\|_2$, plotted in red, for 5 different initializations. The parameter settings were $d = 10$, $n = 1000$, $\rho_{\mathrm{mix}} = 0.6$ and SNR $\frac{\|\mu^*\|_2}{\sigma} = 1.5$. See the main text for further details.

converges to *different stationary points* $\widehat{\mu}_i$; however, all of these points have roughly the same distance from μ^*. This phenomenon highlights the importance of the change of perspective in our analysis—that is, focusing on the true parameter as opposed to the MLE. Given the presence of all these local optima in a small neighborhood of μ^*, the basin of attraction of the MLE must necessarily be much smaller than the initialization radius guaranteed by our theory.

Figure 2.4 shows how the convergence rate of the Baum-Welch algorithm depends on the underlying SNR parameter η^2; this behavior confirms the predictions given in Corollary 2. Lines of the same color represent different random draws of parameters given a fix SNR. Clearly, the convergence is linear for high SNR, and the rate decreases with decreasing SNR.

2.5 Proofs

In this section, we collect the proofs of our main results. In all cases, we provide the main bodies of the proofs here, deferring the more technical details to the appendices.

Proof of Theorem 1

Throughout this proof, we make use of the shorthand $\widetilde{\rho}_{\mathrm{mix}} = 1 - \epsilon_{\mathrm{mix}}\overline{\pi}_{\min}$. Also we denote the separate components of the population EM operators by $\overline{M}(\theta) =: (\overline{M}^{\mu}(\theta), \overline{M}^{\beta}(\theta))^T$ and their truncated equivalents by $\overline{M}^k(\theta) =: (\overline{M}^{\mu,k}(\theta), \overline{M}^{\beta,k}(\theta))^T$. We begin by proving the bound

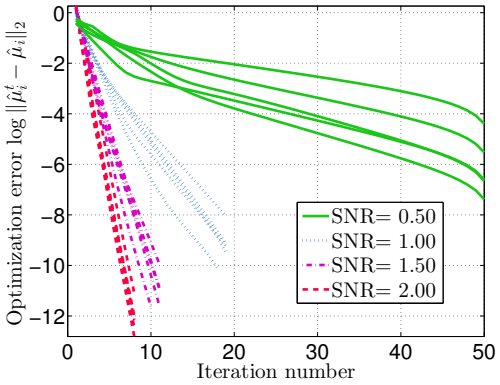


Figure 2.4: Plot of convergence behavior for different SNR, where for each curve, different parameters were chosen. The parameter settings are $d = 10$, $n = 1000$ and $\rho_{\mathrm{mix}} = 0.6$.

given in part (a). Since $\bar{Q} = \lim_{n\to\infty} \mathbb{E}[Q_n]$, we have

$$\|\bar{Q} - \bar{Q}^k\|_\infty = \|\lim_{n\to\infty} \mathbb{E}[Q_n] - \bar{Q}^k\|_\infty \;\leq\; \frac{Cs^4}{\epsilon_{\mathrm{mix}}^9 \bar{\pi}_{\min}^2} \widetilde{\rho}_{\mathrm{mix}}^k,$$

where we have exchanged the supremum and the limit before applying the bound (2.17). The same holds for the separate functions $\bar{Q}_1, \bar{Q}_2$.

Using this bound and the fact that for $\bar{Q}_1$ we have $\bar{Q}_1(\overline{M}^\mu(\theta) \mid \theta) \geq \bar{Q}_1(\overline{M}^{\mu,k}(\theta) \mid \theta)$, we find that

$$\bar{Q}_1(\overline{M}^\mu(\theta) \mid \theta) \geq \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta) - \frac{Cs^4}{\epsilon_{\mathrm{mix}}^9 \bar{\pi}_{\min}^2} \widetilde{\rho}_{\mathrm{mix}}^k.$$

Since $\overline{M}^{\mu,k}(\theta)$ is optimal, the first-order conditions for optimality imply that

$$\langle \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta),\, \theta - \overline{M}^{\mu,k}(\theta) \rangle \leq 0 \qquad \text{for all } \theta \in \Omega.$$

Combining this fact with strong concavity of $\bar{Q}^k(\cdot|\theta)$ for all θ, we obtain

$$\begin{aligned}
\frac{Cs^4}{\epsilon_{\mathrm{mix}}^9 \bar{\pi}_{\min}^2} \widetilde{\rho}_{\mathrm{mix}}^k &\geq \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta) - \bar{Q}_1(\overline{M}^\mu(\theta) \mid \theta) \\
&\geq \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta) - \bar{Q}_1^k(\overline{M}^\mu(\theta) \mid \theta) - \frac{Cs^4}{\epsilon_{\mathrm{mix}}^9 \bar{\pi}_{\min}^2} \widetilde{\rho}_{\mathrm{mix}}^k \\
&\geq \frac{\lambda_\mu}{2} \|\overline{M}^\mu(\theta) - \overline{M}^{\mu,k}(\theta)\|_2^2 - \frac{Cs^4}{\epsilon_{\mathrm{mix}}^9 \bar{\pi}_{\min}^2} \widetilde{\rho}_{\mathrm{mix}}^k
\end{aligned}$$

and therefore $\|\overline{M}^{\mu}(\theta) - \overline{M}^{\mu,k}(\theta)\|_2^2 \leq 4\frac{Cs^4}{\lambda \epsilon_{\min}^9 \bar{\pi}_{\min}^2}\widetilde{\rho}_{\min}^k$. In particular, setting $\theta = \theta^*$ and identifiability, i.e. $\overline{M}^{\mu}(\theta^*) = \theta^*$, yields

$$\|\overline{M}^{\mu,k}(\theta^*) - \theta^*\|_2^2 \leq 4\frac{Cs^4}{\lambda_\mu \epsilon_{\min}^9 \bar{\pi}_{\min}^2}\widetilde{\rho}_{\min}^k,$$

and the equivalent bound can be obtained for $\overline{M}^{\beta,k}(\cdot)$ which yields the claim.

We now turn to the proof of part (b). Let us suppose that the recursive bound (2.26) holds, and use it to complete the proof of this claim. We first show that if $\widetilde{\mu}^t \in \mathbb{B}_2(r; \mu^*)$, then we must have $\widetilde{\mu}^{t+1} \in \mathbb{B}_2(r; \mu^*)$ as well. Indeed, if $\widetilde{\mu}^t \in \mathbb{B}_2(r; \mu^*)$, then we have by triangle inequality and contraction in (2.26)

$$\begin{aligned}
\|\overline{M}^{\mu,k}(\widetilde{\theta}^t) - \mu^*\|_2 &\leq \|\overline{M}^{\mu,k}(\widetilde{\theta}^t) - \overline{M}^{\mu,k}(\theta^*)\|_2 + \|\overline{M}^{\mu,k}(\theta^*) - \mu^*\|_2 \\
&\leq \kappa[\|\widetilde{\mu}^t - \mu^*\|_2 + \|\widetilde{\beta}^t - \beta^*\|_2] + \varphi(k) \\
&\leq \kappa(r + \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2) + \varphi(k) \leq r,
\end{aligned}$$

where the final step uses the assumed bound on φ. For the joint parameter update we in turn have

$$\begin{aligned}
\|\overline{M}^{k}(\widetilde{\theta}^t) - \theta^*\|_\star &\leq \|\overline{M}^{k}(\widetilde{\theta}^t) - \overline{M}^{k}(\theta^*)\|_\star + \|\overline{M}^{k}(\theta^*) - \theta^*\|_\star \\
&\leq \kappa\|\widetilde{\theta}^t - \theta^*\|_\star + \varphi(k). \qquad (2.34)
\end{aligned}$$

By repeatedly applying inequality (2.34) and summing the geometric series, the claimed bound (2.25) follows.

It remains to prove the bound (2.26). Since the vector $\overline{M}^k(\theta^\star)$ maximizes the function $\theta \mapsto \bar{Q}_1^k(\theta \mid \theta^*)$, we have the first-order optimality condition

$$\langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta^*) \mid \theta^\star),\, \overline{M}^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta^*)\rangle \leq 0, \qquad \text{valid for any } \theta.$$

Similarly, we have $\langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta),\, \overline{M}^{\mu,k}(\theta^*) - \overline{M}^{\mu,k}(\theta)\rangle \leq 0$, and adding together these two inequalities yields

$$0 \leq \langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta^*) \mid \theta^*) - \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta),\, \overline{M}^{\mu,k}(\theta^*) - \overline{M}^{\mu,k}(\theta)\rangle$$

On the other hand, by the λ-strong concavity condition, we have

$$\lambda_\mu \|\overline{M}^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta^*)\|_2^2 \leq \langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta^*) - \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta^*) \mid \theta^\star),\, \overline{M}^{\mu,k}(\theta^*) - \overline{M}^{\mu,k}(\theta)\rangle$$

Combining these two inequalities with the (L_μ, L_β)-FOS condition yields

$$\begin{aligned}
\lambda_\mu \|\overline{M}^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta^*)\|_2^2 &\leq \langle \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta^*) - \nabla \bar{Q}_1^k(\overline{M}^{\mu,k}(\theta) \mid \theta),\, \overline{M}^{\mu,k}(\theta^*) - \overline{M}^{\mu,k}(\theta)\rangle \\
&\leq \big[L_{\mu_1}\|\mu - \mu^*\|_2 + L_{\mu_2}\|\beta - \beta^*\|_2\big]\|\overline{M}^{\mu,k}(\theta) - \overline{M}^{\mu,k}(\theta^*)\|_2,
\end{aligned}$$

and similarly we obtain $\lambda_\beta \|\overline{M}^{\beta,k}(\theta) - \overline{M}^{\beta,k}(\theta^*)\|_2 \leq \big[L_{\beta_1}\|\mu - \mu^*\|_2 + L_{\beta_2}\|\beta - \beta^*\|_2\big]$. Adding both inequalities yields the claim (2.26).

Proof of Theorem 2

By the triangle inequality and inequality (2.34), we have with probability at least $1 - 2\delta$ that for any iteration

$$\begin{aligned}\|\widehat{\theta}^{t+1} - \theta^\star\|_\star &\leq \|M_n(\widehat{\theta}^t) - M_n^k(\widehat{\theta}^t)\|_\star + \|M_n^k(\widehat{\theta}^t) - \overline{M}^k(\widehat{\theta}^t)\|_\star + \|\overline{M}^k(\widehat{\theta}^t) - \theta^*\|_\star \\ &\leq \varphi_n(\delta, k) + \epsilon_n(\delta, k) + \kappa\|\widehat{\theta}^t - \theta^\star\|_\star + \varphi(k).\end{aligned}$$

In order to see that the iterates do not leave $\mathbb{B}_2\big(r; \mu^*\big)$, observe that

$$\begin{aligned}\|\widehat{\mu}^{t+1} - \mu^*\|_2 &\leq \|M_n^\mu(\widehat{\theta}^t) - M_n^{\mu,k}(\widehat{\theta}^t)\|_2 + \|M_n^{\mu,k}(\widehat{\theta}^t) - \overline{M}^{\mu,k}(\widehat{\theta}^t)\|_2 + \|\overline{M}^{\mu,k}(\widehat{\theta}^t) - \mu^*\|_2 \\ &\leq \varphi_n(\delta, k) + \epsilon_n^\mu(\delta, k) + \kappa(\|\widehat{\mu}^t - \mu^*\|_2 + \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2) + \varphi(k). \qquad (2.35)\end{aligned}$$

Consequently, as long as $\|\widehat{\mu}^t - \mu^*\|_2 \leq r$, we also have $\|\widehat{\mu}^{t+1} - \mu^*\|_2 \leq r$ whenever

$$\varphi_n(\delta, k) + \varphi(k) + \epsilon_n^\mu(\delta, k) \leq (1 - \kappa)\, r - \kappa \max_{\beta \in \Omega_\beta} \|\beta - \beta^*\|_2.$$

Combining inequality (2.35) with the equivalent bound for β, we obtain

$$\|\widehat{\theta}^t - \theta^*\|_\star \leq \kappa\|\widehat{\theta}^{t-1} - \theta^*\|_\star + \varphi_n(\delta, k) + \epsilon_n(\delta, k) + \varphi(k)$$

Summing the geometric series yields the bound (2.28b).

Proof of Corollary 1

The boundedness condition (Assumption (2.16)) is easy to check since for $X \sim N(\mu^*, \sigma^2)$, the quantity $\sup_{\mu \in \mathbb{B}_2(r; \mu^*)} \mathbb{E}\big[\max\{\|X - \mu\|_2, \|X + \mu\|_2\}\big]$ is finite for any choice of radius $r < \infty$.

By Theorem 1, the k-truncated population EM iterates satisfy the bound

$$\|\widetilde{\theta}^t - \theta^*\|_\star \leq \kappa^t\|\widetilde{\theta}^0 - \theta^*\|_\star + \frac{1}{1 - \kappa}\varphi(k), \qquad (2.36)$$

if the strong concavity (2.20) and FOS conditions (2.21), (2.22) hold with suitable parameters.

In the remainder of proof—and the bulk of the technical work— we show that:

- strong concavity holds with $\lambda_\mu = 1$ and $\lambda_\beta \geq \frac{2}{3}(1 - b^2)$;
- the FOS conditions hold with

$$\begin{aligned}L_{\mu,1} &= c\ (\eta^2 + 1)\varphi_2(\epsilon_{\text{mix}})\eta^2 \mathrm{e}^{-c\eta^2}, \quad \text{and} \quad L_{\mu,2} = c\sqrt{\|\mu^*\|_2^2 + \sigma^2}\varphi_2(\epsilon_{\text{mix}})\eta^2 \mathrm{e}^{-c\eta^2} \\ L_{\beta,1} &= c\frac{1 - b}{1 + b}\varphi_2(\epsilon_{\text{mix}})\eta^2 \mathrm{e}^{-c\eta^2} \quad \text{and} \quad L_{\beta,2} = c\sqrt{\|\mu^*\|_2^2 + \sigma^2}\varphi_2(\epsilon_{\text{mix}})\eta^2 \mathrm{e}^{-c\eta^2},\end{aligned}$$

where $\varphi_2(\epsilon_{\text{mix}}) := \Big(\frac{1}{\log(1/(1-\epsilon_{\text{mix}}))} + \frac{1}{\epsilon_{\text{mix}}}\Big)$. Substuting these choices into the bound (2.36) and performing some algebra yields the claim.

Establishing strong concavity

We first show concavity of $\bar{Q}_1^k(\cdot \mid \theta')$ and $\bar{Q}_2^k(\cdot \mid \theta')$ separately. For strong concavity of $\bar{Q}_1^k(\cdot \mid \theta')$, observe that

$$\bar{Q}_1^k(\mu \mid \theta') = -\frac{1}{2}\mathbb{E}\left[p(z_0 = 1 \mid X_{-k}^k; \theta')\|X_0 - \mu\|_2^2 + (1 - p(z_0 = 1|X_{-k}^k; \theta'))\|X_0 + \mu\|_2^2 + c\right],$$

where c is a quantity independent of μ. By inspection, this function is strongly concave in μ with parameter $\lambda_\mu = 1$.

On the other hand, we have

$$\bar{Q}_2^k(\beta \mid \theta') = \mathbb{E}_{X_{-k}^k|\theta^*} \sum_{z_0,z_1} p(z_0, z_1 \mid X_{-k}^k; \theta') \log\left(\frac{\mathrm{e}^{\beta z_0 z_1}}{\mathrm{e}^{\beta} + \mathrm{e}^{-\beta}}\right).$$

This function has second derivative $\frac{\partial^2}{\partial \beta^2}\bar{Q}_2^k(\beta \mid \theta') = -4\frac{\mathrm{e}^{-2\beta}}{(\mathrm{e}^{-2\beta}+1)^2}$. As a function of $\beta \in \Omega_\beta$, this second derivative is maximized at $\beta = \frac{1}{2}\log\left(\frac{1+b}{1-b}\right)$. Consequently, the function $\bar{Q}_2^k(\cdot \mid \theta')$ is strongly concave with parameter $\lambda_\beta \geq \frac{2}{3}(1 - b^2)$.

Separate FOS conditions

We now turn to proving that the FOS conditions in equations (2.21) and (2.22) hold. A key ingredient in our proof is the fact that the conditional density $p(z_{-k}^k \mid x_{-k}^k; \mu, \beta)$ belongs to the exponential family with parameters $\beta \in \mathbb{R}$, and $\gamma_i := \frac{\langle \mu, x_i \rangle}{\sigma^2} \in \mathbb{R}$ for $i = -k, \ldots, k$ which define the vector $\gamma = (\gamma_{-k}, \ldots, \gamma_k)$ (see [82] for more details on exponential families.) In particular, we have

$$\underbrace{p(z_{-k}^k \mid x_{-k}^k, \mu, \beta)}_{:= p(z_{-k}^k; \gamma, \beta)} = \exp\left\{\sum_{\ell=-k}^{k} \gamma_\ell z_\ell + \beta \sum_{\ell=-k}^{k-1} z_\ell z_{\ell+1} - \Phi(\gamma, \beta)\right\}, \tag{2.37}$$

where the function h absorbs various coupling terms. Note that this exponential family is a specific case of the following exponential family distribution

$$\underbrace{\tilde{p}(z_{-k}^k \mid x_{-k}^k, \mu, \beta)}_{:= \tilde{p}(z_{-k}^k; \gamma, \beta)} = \exp\left\{\sum_{\ell=-k}^{k} \gamma_\ell z_\ell + \sum_{\ell=-k}^{k-1} \beta_\ell z_\ell z_{\ell+1} - \Phi(\gamma, \beta)\right\}. \tag{2.38}$$

The distribution in (2.37) corresponds to (2.38) with $\beta_\ell = \beta$ for all ℓ and the so-called partition function Φ is given by

$$\Phi(\gamma, \beta) = \log \sum_{z} \exp\left\{\sum_{\ell=-k}^{k} \gamma_\ell z_\ell + \sum_{\ell=-k}^{k-1} \beta_\ell z_\ell z_{\ell+1}\right\}.$$

The reason to view our distribution as a special case of the more general one in (2.38) becomes clear when we consider the equivalence of expectations and the derivatives of the cumulant function

$$\left.\frac{\partial \Phi}{\partial \gamma_\ell}\right|_{\theta'} = \mathbb{E}_{Z_{-k}^k | x_{-k}^k, \theta'} Z_\ell \quad \text{and} \quad \left.\frac{\partial \Phi}{\partial \beta_0}\right|_{\theta'} = \mathbb{E}_{Z_{-k}^k | x_{-k}^k, \theta'} Z_0 Z_1, \tag{2.39}$$

where we recall that $\mathbb{E}_{Z_{-k}^k | x_{-k}^k, \theta'}$ is the expectation with respect to the distribution $\tilde{p}(Z_{-k}^k \mid x_{-k}^k; \mu', \beta')$ with $\beta_\ell = \beta'$. Note that in the following any value θ' for $\tilde{p}$ is taken to be on the manifold on which $\beta_\ell = \beta'$ for some β' since this is the manifold the algorithm works on. Also, as before, $\mathbb{E}$ denotes the expectation over the joint distribution of all samples X_ℓ drawn according to $p(\cdot; \theta^*)$, in this case X_{-k}^k.

Similarly to equations (2.39), the covariances of the sufficient statistics correspond to the second derivatives of the cumulant function

$$\left.\frac{\partial^2 \Phi}{\partial \beta_\ell \partial \beta_0}\right|_{\theta} = \operatorname{cov}(Z_0 Z_1, Z_\ell Z_{\ell+1} \mid X_{-k}^k, \theta) \tag{2.40a}$$

$$\left.\frac{\partial^2 \Phi}{\partial \gamma_\ell \partial \gamma_0}\right|_{\theta} = \operatorname{cov}(Z_0, Z_\ell \mid X_{-k}^k, \theta) \tag{2.40b}$$

$$\left.\frac{\partial^2 \Phi}{\partial \beta_\ell \partial \gamma_0}\right|_{\theta} = \operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid X_{-k}^k, \theta). \tag{2.40c}$$

In the following, we adopt the shorthand

$$\begin{aligned}\operatorname{cov}(Z_\ell, Z_{\ell+1} \mid \gamma', \beta') &= \operatorname{cov}(Z_\ell, Z_{\ell+1} \mid X_{-k}^k, \theta') \\ &= \mathbb{E}_{Z_\ell^{\ell+1} | X_{-k}^k, \theta'} (Z_\ell - \mathbb{E}_{Z_\ell^{\ell+1} | X_{-k}^k, \theta'} Z_\ell)(Z_{\ell+1} - \mathbb{E}_{Z_\ell^{\ell+1} | X_{-k}^k, \theta'} Z_{\ell+1})\end{aligned}$$

where the dependence on β is occasionally omitted so as to simplify notation.

Proof of inequality (2.21a)

By an application of the mean value theorem, we have

$$\|\nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta') - \nabla_\mu \bar{Q}_1^k(\mu \mid \mu^*, \beta')\| \leq \underbrace{\left\| \mathbb{E} \sum_{\ell=-k}^{k} \left.\frac{\partial^2 \Phi}{\partial \gamma_\ell \partial \gamma_0}\right|_{\theta=\widetilde{\theta}} (\gamma'_\ell - \gamma^*_\ell) X_0 \right\|}_{T_1}$$

where $\widetilde{\theta} = \theta' + t(\theta^* - \theta')$ for some $t \in (0, 1)$. Since second derivatives yield covariances (see equation (2.40)), we can write

$$T_1 = \left\| \sum_{\ell=-k}^{k} \mathbb{E} X_0 \mathbb{E} \left[\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) \frac{\langle \mu' - \mu^*, X_\ell \rangle}{\sigma^2} \middle| X_0 \right] \right\|_2,$$

so that it suffices to control the expected conditional covariance. By the Cauchy-Schwarz inequality and the fact that $\operatorname{cov}(X,Y) \leq \sqrt{\operatorname{var} X}\sqrt{\operatorname{var} Y}$ and $\operatorname{var}(Z_0 \mid X) \leq 1$, we obtain the following bound on the expected conditional covariance by using Lemma 4 (see Section 2.8)

$$\begin{aligned}\mathbb{E}\left[\left|\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) \mid X_0\right|\right] &\leq \sqrt{\mathbb{E}\left[\operatorname{var}(Z_0 \mid \widetilde{\gamma}) \mid X_0\right]}\sqrt{\mathbb{E}\left[\operatorname{var}(Z_\ell \mid \widetilde{\gamma}) \mid X_0\right]} \\ &\leq \sqrt{\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0)}. \end{aligned} \tag{2.41a}$$

Furthermore, by Lemma 5 and 6 (see Section 2.8), we have

$$|\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma})| \leq 2\rho_{\mathrm{mix}}^{\ell}, \quad \text{and} \quad \left\|\mathbb{E}(\operatorname{var}(Z_0|\widetilde{\gamma}_0))^{1/2} X_0 X_0^T\right\|_{op} \leq C\mathrm{e}^{-c\eta^2}. \tag{2.41b}$$

From the definition of the operator norm, we have

$$\begin{aligned}\left\|\mathbb{E}\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) X_0 X_\ell^T\right\|_{op} &= \sup_{\substack{\|u\|_2=1 \\ \|v\|_2=1}} \mathbb{E}\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma})\langle X_0,\, v\rangle\, \langle X_\ell,\, u\rangle \\ &\leq \sup_{\|v\|_2=1} \mathbb{E}|\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma})|\langle X_0,\, v\rangle^2 \\ &+ \sup_{\|u\|_2=1} \mathbb{E}|\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma})|\langle X_\ell,\, u\rangle^2 \\ &= \|\mathbb{E} X_0 X_0^T \mathbb{E}\big[|\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) \mid X_0\big]\|_{op} \\ &+ \|\mathbb{E} X_\ell X_\ell^T \mathbb{E}\big[\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) \mid X_\ell\big]\|_{op} \\ &\overset{(i)}{\leq} 2\min\{\rho_{\mathrm{mix}}^{|\ell|}\|\mathbb{E} X_0 X_0^T\|_{op}, \|\mathbb{E}\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0)^{1/2} X_0 X_0^T\|_{op}\} \\ &\overset{(ii)}{\leq} 2\min\{(\|\mu^*\|_2^2 + \sigma^2)\rho_{\mathrm{mix}}^{|\ell|}, C'\mathrm{e}^{-c\eta^2}\}, \end{aligned} \tag{2.42}$$

where inequality (i) makes use of inequalities (2.41a) and (2.41b), and step (ii) makes use of the second inequality in line (2.41b).

By inequality (2.42), we find that

$$\begin{aligned} T_1 &\leq \frac{\|\mu' - \mu^*\|_2}{\sigma^2} \sum_{\ell=-k}^{k} \left\|\mathbb{E}\operatorname{cov}(Z_0, Z_\ell \mid \widetilde{\gamma}) X_0 X_\ell^T\right\|_{op} \\ &\leq 2\frac{\|\mu' - \mu^*\|_2}{\sigma^2} \sum_{\ell=-k}^{k} \min\{(\|\mu^*\|_2^2 + \sigma^2)\rho_{\mathrm{mix}}^{|\ell|}, C\mathrm{e}^{-c\eta^2}\} \\ &\leq 4(\eta^2+1)\big(mC\mathrm{e}^{-c\eta^2} + \frac{\rho_{\mathrm{mix}}^m}{1-\rho_{\mathrm{mix}}}\big)\|\mu' - \mu^*\|_2. \end{aligned}$$

where $m = \frac{c\eta^2}{\log(1/\rho_{\mathrm{mix}})}$ is the smallest integer such that $\rho_{\mathrm{mix}}^m \leq C\mathrm{e}^{-c\eta^2}$ The last inequality follows from the proof of Corollary 1 in the paper [3] if $\eta^2 > C$ for some universal constant C. We have thus shown that

$$\|\nabla_\mu \bar{Q}_1^k(\mu \mid \mu', \beta') - \nabla_\mu \bar{Q}_1^k(\mu \mid \mu^*, \beta')\| \leq L_{\mu,1}\|\mu' - \mu^*\|_2,$$

where $L_{\mu,1} = c\ \varphi_1(\eta)\varphi_2(\epsilon_{\mathrm{mix}})\eta^2(\eta^2+1)\mathrm{e}^{-c\eta^2}$ as claimed.

Proof of inequality (2.21b)

The same argument via the mean value theorem guarantees that

$$\|\frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \mu', \beta') - \frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \mu', \beta^*)\| \leq \left\|\mathbb{E}\sum_{\ell=-k}^{k}\frac{\partial^2\Phi}{\partial\beta_\ell\partial\gamma_0}\Big|_{\theta=\widetilde{\theta}}(\beta'-\beta^*)X_0\right\|_2.$$

In order to bound this quantity, we again use the equivalence (2.40) and bound the expected conditional covariance. Furthermore, Lemma 5 and 6 yield

$$\operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}) \overset{(i)}{\leq} 2\rho_{\text{mix}}^\ell \quad \text{and} \quad \left\|\mathbb{E}\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0)X_0X_0^T\right\|_{op} \overset{(ii)}{\leq} c\mathrm{e}^{-c\eta^2}. \tag{2.43}$$

Here inequality (ii) follows by combining inequality (2.54c) from Lemma 5 with the fact that $\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0) \leq 1$.

$$\begin{aligned}
\|\mathbb{E}X_0\operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma})\|_2 &= \sup_{\|u\|_2=1}\mathbb{E}\langle X_0, u\rangle\operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}) \\
&\leq \sup_{\|u\|_2=1}\mathbb{E}|\langle X_0, u\rangle|\mathbb{E}\big[|\operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma})| \mid X_0\big] \\
&\overset{(iii)}{\leq} \sup_{\|u\|_2=1}\mathbb{E}|\langle X_0, u\rangle|\min\{\rho_{\text{mix}}^{|\ell|}, (\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0))^{1/2}\} \\
&\overset{(iv)}{\leq} \min\{\sup_{\|u\|_2=1}\sqrt{\mathbb{E}\langle X_0, u\rangle^2}\rho_{\text{mix}}^{|\ell|}, \sup_{\|u\|_2=1}\sqrt{\mathbb{E}\langle X_0, u\rangle^2\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0)}\} \\
&\overset{(v)}{\leq} \min\{\rho_{\text{mix}}^{|\ell|}\sqrt{\|\mathbb{E}X_0X_0^T\|_{op}}, \sqrt{\|\mathbb{E}\operatorname{var}(Z_0 \mid \widetilde{\gamma}_0)X_0X_0^T\|_{op}}\} \\
&\overset{(vi)}{\leq} \min\{\rho_{\text{mix}}^{|\ell|}\sqrt{\|\mu^*\|_2^2+\sigma^2}, C\mathrm{e}^{-c\eta^2}\}
\end{aligned}$$

where step (iii) uses inequality (2.43); step (iv) follows from the Cauchy-Schwarz inequality; step (v) follows from the definition of the operator norm; and step (vi) uses inequality (2.43) again.

Putting together the pieces, we find that

$$\begin{aligned}
\left\|\mathbb{E}\sum_{\ell=-k}^{k}\frac{\partial^2\Phi}{\partial\beta_\ell\partial\gamma_0}X_0\right\|_2|\beta'-\beta^*| &\leq \sum_{\ell=-k}^{k}\|\mathbb{E}X_0\mathbb{E}[\operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}) \mid X_0]\|_2\,|\beta'-\beta^*| \\
&\leq 4\sqrt{\|\mu^*\|_2^2+\sigma^2}\left(c\,m\,\mathrm{e}^{-c\eta^2} + \frac{\rho_{\text{mix}}^m}{1-\rho_{\text{mix}}}\right)|\beta'-\beta^*|.
\end{aligned}$$

again with $m = \frac{c\eta^2}{\log(1/\rho_{\text{mix}})}$, we find that inequality (2.21b) holds with $L_{\mu,2} = c\varphi_2(\epsilon_{\text{mix}})\sqrt{\|\mu^*\|_2+\sigma^2}\eta^2\mathrm{e}^{-c\eta^2}$, as claimed.

Proof of inequality (2.22a)

By the same argument via the mean value theorem, we find that

$$\big\|\frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \beta', \mu') - \frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \beta', \mu^*)\big\| \leq \Big|\mathbb{E}\sum_{\ell=-k}^{k} \frac{\partial^2\Phi}{\partial\gamma_\ell\partial\beta_0}\Big|_{\theta=\widetilde{\theta}} \frac{\langle\mu' - \mu^*, X_\ell\rangle}{\sigma^2}\Big|.$$

Equation (2.40) guarantees that $\frac{\partial^2\Phi}{\partial\gamma_\ell\partial\beta_0} = \operatorname{cov}(Z_0Z_1, Z_\ell \mid \gamma)$. Therefore, by similar arguments as in the proof of inequalities (2.21), we have

$$\begin{aligned}
T &:= \Big|\sum_{\ell=-k}^{k} \mathbb{E}\langle\mu' - \mu^*, X_\ell\rangle\mathbb{E}[\operatorname{cov}(Z_0Z_1, Z_\ell \mid \widetilde{\gamma}_\ell, \beta')|X_\ell]\Big| \\
&\leq \Big|\sum_{\ell=-k}^{k} \mathbb{E}|\langle\mu' - \mu^*, X_\ell\rangle| \min\{\rho_{\mathrm{mix}}^{|\ell|}, (\operatorname{var}(Z_\ell \mid \widetilde{\gamma}_\ell, \beta'))^{1/2}\}\Big| \\
&\leq \Big|\sum_{\ell=-k}^{k} \min\big\{\rho_{\mathrm{mix}}^{|\ell|}, \sqrt{\mathbb{E}\operatorname{var}(Z_\ell \mid \widetilde{\gamma}_\ell, \beta')}\big\}\sqrt{\mathbb{E}\langle\mu' - \mu^*, X_\ell\rangle^2}\Big| \\
&\leq \sqrt{\|\mu^*\|_2^2 + \sigma^2}\big(mc\,\mathrm{e}^{-c\eta^2} + 2\sum_{\ell=m+1}^{k} \rho_{\mathrm{mix}}^{\ell}\big).
\end{aligned}$$

where we have used inequality (2.54b) from Lemma 6. Finally, again noting that $m = \frac{c\eta^2}{\log(1/\rho_{\mathrm{mix}})}$ yields that the FOS condition holds with $L_{\beta,2} = c\sqrt{\|\mu^*\|_2 + \sigma^2}\varphi_2(\epsilon_{\mathrm{mix}})\eta^2\mathrm{e}^{-c\eta^2}$, as claimed.

Proof of inequality (2.22b)

By the same mean value argument, we find that

$$\|\frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \beta', \mu') - \frac{\partial}{\partial\beta}\bar{Q}_2^k(\beta \mid \beta^*, \mu')\| \leq \Big|\mathbb{E}\sum_{\ell=-k}^{k} \frac{\partial^2\Phi}{\partial\beta_\ell\partial\beta_0}\Big|_{\theta=\widetilde{\theta}} (\beta' - \beta^*)\Big|.$$

By the exponential family view in equality (2.40) it suffices to control the expected conditional covariance. Lemma 5 and 6 guarantee that

$$|\operatorname{cov}(Z_0Z_1, Z_\ell Z_{\ell+1} \mid X_{-k}^k, \widetilde{\gamma})| \leq \rho_{\mathrm{mix}}^{|\ell|}, \quad \text{and} \quad \mathbb{E}\operatorname{var}(Z_0Z_1 \mid \widetilde{\gamma}_0^1, \widetilde{\beta}) \leq c\,\frac{1+b}{1-b}\,\mathrm{e}^{-c\eta^2}. \tag{2.44}$$

Furthermore, the Cauchy-Schwarz inequality combined with the bound (2.53a) from Lemma 4 yields

$$\begin{aligned}
\mathbb{E}\big|\operatorname{cov}(Z_0Z_1, Z_\ell Z_{\ell+1} \mid \widetilde{\gamma})\big| &\leq \sqrt{\mathbb{E}\operatorname{var}(Z_0Z_1 \mid \widetilde{\gamma}, \widetilde{\beta})}\sqrt{\mathbb{E}\operatorname{var}(Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}, \widetilde{\beta})} \\
&\leq \sqrt{\mathbb{E}\operatorname{var}(Z_0Z_1 \mid \widetilde{\gamma}_0^1, \widetilde{\beta})}\sqrt{\mathbb{E}\operatorname{var}(Z_\ell Z_{\ell+1} \mid \widetilde{\gamma}_\ell^{\ell+1}, \widetilde{\beta})} \\
&\leq \mathbb{E}\operatorname{var}(Z_0Z_1 \mid \widetilde{\gamma}_0^1, \widetilde{\beta}).
\end{aligned} \tag{2.45}$$

Combining the bounds (2.44) and (2.45) yields

$$\begin{aligned}\left|\sum_{\ell=-k}^{k}\mathbb{E}\frac{\partial^2\Phi}{\partial\beta_\ell\partial\beta_0}(\beta'-\beta^*)\right| &\leq \sum_{\ell=-k}^{k}\left|\mathbb{E}\operatorname{cov}(Z_0Z_1, Z_\ell Z_{\ell+1}\mid \widetilde{\gamma}_{-k}^{k},\widetilde{\beta})\right||\beta'-\beta^*| \\ &\leq \sum_{\ell=-k}^{k}\min\left\{\rho_{\mathrm{mix}}^{|\ell|}, \mathbb{E}\operatorname{var}(Z_0Z_1\mid\widetilde{\gamma}_0^1,\widetilde{\beta})\right\}|\beta'-\beta^*| \\ &\leq 2\left(c\frac{1+b}{1-b}\,m\mathrm{e}^{-c\eta^2}+\sum_{l=m+1}^{k}\rho_{\mathrm{mix}}^{\ell}\right)|\beta'-\beta^*| \\ &\leq 2c\frac{1+b}{1-b}\varphi_2(\epsilon_{\mathrm{mix}})\eta^2\mathrm{e}^{-c\eta^2}|\beta'-\beta^*|\end{aligned}$$

where the final inequality follows by setting $m=\frac{c\eta^2}{\log(1/\rho_{\mathrm{mix}})}$. Therefore, the FOS condition holds with $L_{\beta,1}=c\frac{1-b}{1+b}\varphi_2(\epsilon_{\mathrm{mix}})\eta^2\mathrm{e}^{-c\eta^2}$, as claimed.

Proof of Corollary 2

In order to prove this corollary, it is again convenient to separate the updates on the mean vectors μ from those applied to the transition parameter β. Recall the definitions of φ, φ_n and ϵ_n from equations (2.24) and (2.27a) respectively, as well as $\widetilde{\rho}_{\mathrm{mix}}=1-\epsilon_{\mathrm{mix}}\overline{\pi}_{\min}$.

Using Theorem 2 we readily have that given any initialization $\widehat{\theta}^0\in\Omega$, with probability at least $1-2\delta$, we are guaranteed that

$$\|\widehat{\theta}^T-\theta^*\|_\star\leq\kappa^T\|\widehat{\theta}^0-\theta^*\|_\star+\frac{\varphi_n(\delta,k)+\epsilon_n(\delta,k)+\varphi(k)}{1-\kappa}. \tag{2.46}$$

In order to leverage the bound (2.46), we need to find appropriate upper bounds on the quantities $\varphi_n(\delta,k)$, $\epsilon_n(\delta,k)$.

Lemma 1. *Suppose that the truncation level satisifes the lower bound*

$$k\geq\log\left(\frac{C_\epsilon n}{\delta}\right)\left(\log\frac{1}{\widetilde{\rho}_{\mathrm{mix}}}\right)^{-1}\qquad \textit{where } C_\epsilon:=\tfrac{C}{\epsilon_{\mathrm{mix}}^3\overline{\pi}_{\min}^3}. \tag{2.47a}$$

Then, when the number of observations n satifies the lower bound in the assumptions of the corollary and the radius is chosen to be $r=\frac{\|\mu^*\|_2}{4}$, we have

$$\epsilon_n^{\mu}\big(\delta,k\big)\leq C_0\frac{1}{\sigma}\big(\frac{\|\mu^*\|_2^2}{\sigma^2}+1\big)^{3/2}\log(k^2/\delta)\sqrt{\frac{k^3d\log n}{n}},\quad \textit{and} \tag{2.47b}$$

$$\epsilon_n^{\beta}\big(\delta,k\big)\leq C_0\,\frac{1}{\sigma}\sqrt{\frac{\|\mu^*\|_2^2}{\sigma^2}+1}\sqrt{\frac{k^3\log(k^2/\delta)}{n}}. \tag{2.47c}$$

Lemma 2. *Suppose that* $\frac{1}{2}\frac{\log\left(C_\epsilon n/\delta\right)}{\log(1/\widetilde{\rho}_{\mathrm{mix}})} \leq k \leq C\frac{\log\left(C_\epsilon n/\delta\right)}{\log(1/\widetilde{\rho}_{\mathrm{mix}})}$ *with* $C > 1$. *Then by choosing* $r = \frac{\|\mu^*\|_2}{4}$ *and* C_1 *large enough, we have*

$$\varphi_n(\delta, k) \leq C_1 \left\{ \sqrt{\frac{d\log^2(C_\epsilon n/\delta)}{\sigma n}} + \sqrt{\frac{\|\mu^*\|_2}{\sigma}\frac{\log^2(C_\epsilon n/\delta)}{n}} + \frac{\|\mu^*\|_2}{\sigma}\sqrt{\frac{\delta}{n}} \right\}. \tag{2.48}$$

See Appendices 2.9 and 2.9, respectively, for the proofs of these two lemmas.

Note that the set for which k simultaneously satisfies the conditions in Lemma 1 and 2 is nonempty. Furtheremore, the choice of k is made purely for analysis purposes – it does not have any consequence on the Using these two lemmas, we can now complete the proof of the corollary. From the definition (2.31) of κ, under the stated lower bound on η^2, we can ensure that $\kappa \leq \frac{1}{2}$. Under this condition, inequality (2.28a) with $r = \|\mu^*\|_2/4$ reduces to showing that

$$\varphi_n(\delta, k) + \epsilon_n^\mu(\delta, k) + \varphi(k) \leq \frac{\|\mu^*\|_2}{8}. \tag{2.49}$$

Now any choice of k satisfying both conditions in Lemmas 1 and 2 guarantees that

$$\varphi_n(\delta, k) + \epsilon_n^\mu(\delta, k) + \epsilon_n^\beta(\delta, k) + \varphi(k) \leq \frac{C}{\sigma}\left(\frac{\|\mu^*\|_2^2}{\sigma^2} + 1\right)^{3/2}\sqrt{\frac{d\log^8(n/\delta)}{n}}. \tag{2.50}$$

Furthermore, as long as $n \geq \frac{C_1}{\|\mu^*\|_2^2\sigma^2}(\eta^2+1)^3 d\log^8(d/\delta)$ for a sufficiently large C_1, we are guaranteed that the bound (2.49) holds. Substituting the bound (2.50) into inequality (2.46) completes the proof of the corollary.

2.6 Discussion

In this chapter, we provided general global convergence guarantees for the Baum-Welch algorithm as well as specific results for a hidden Markov mixture of two isotropic Gaussians. In contrast to the classical perspective of focusing on the MLE, we focused on bounding the distance between the Baum-Welch iterates and the true parameter. Under suitable regularity conditions, our theory guarantees that the iterates converge to an e_n-ball of the true parameter, where e_n represents a form of statistical error. It is important to note that our theory does not guarantee convergence to the MLE itself, but rather to a ball that contains the true parameter, and asymptotically the MLE as well. When applied to the Gaussian mixture HMM, we proved that the Baum-Welch algorithm achieves estimation error that is minimax optimal up to logarithmic factors. To the best of our knowledge, these are the first rigorous guarantees for the Baum-Welch algorithm that allow for a large initialization radius.

2.7 Proof of Proposition 2.3.1

In order to show that the limit $\lim_{n\to\infty} \mathbb{E}Q_n(\theta \mid \theta')$ exists, it suffices to show that the sequence of functions $\{\mathbb{E}Q_1, \mathbb{E}Q_2, \ldots, \mathbb{E}Q_n\}$ is Cauchy in the sup-norm (as defined previously in equation (2.15)). In particular, it suffices to show that for every $\epsilon > 0$ there is a positive integer $N(\epsilon)$ such that for $m, n \geq N(\epsilon)$,

$$\|\mathbb{E}Q_m - \mathbb{E}Q_n\|_\infty \leq \epsilon.$$

In order to do so, we make use of the previously stated bound (2.17) relating $\mathbb{E}Q_n$ to $\bar{Q}^k$. Taking this bound as given for the moment, an application of the triangle inequality yields

$$\|\mathbb{E}Q_m - \mathbb{E}Q_n\|_\infty \leq \|\mathbb{E}Q_m - \bar{Q}^k\|_\infty + \|\mathbb{E}Q_n - \bar{Q}^k\|_\infty \ \leq \ \epsilon,$$

the final inequality follows as long as we choose $N(\epsilon)$ and k large enough (roughly proportional to $\log(1/\epsilon)$).

It remains to prove the claim (2.17). In order to do so, we require an auxiliary lemma:

Lemma 3 (Approximation by truncation)**.** *For a Markov chain satisfying the mixing condition* (2.3)*, we have*

$$\sup_{\theta' \in \Omega} \sup_{x} \sum_{z_i} |p(z_i \mid x_1^n; \theta') - p(z_i \mid x_{i-k}^{i+k}; \theta')| \leq \frac{Cs^2}{\epsilon_{\min}^8 \overline{\pi}_{\min}} \big(1 - \epsilon_{\min} \overline{\pi}_{\min}\big)^{\min\{i, n-i, k\}} \tag{2.51}$$

for all $i \in [0, n]$*, where* $\overline{\pi}_{\min} = \min_{j \in [s], \beta \in \Omega_\beta} \overline{\pi}(j; \beta)$.

See Section 2.10 for the proof of this lemma.

Using Lemma 3, let us now prove the claim (2.17). Introducing the shorthand notation

$$h(X_i, z_i, \theta, \theta') := \log p(X_i \mid z_i; \theta) + \sum_{z_{i-1}} p(z_i \mid z_{i-1}; \theta') \log p(z_i | z_{i-1}, \theta),$$

we can verify by applying Lemma 3 that

$$
\begin{aligned}
&\|\mathbb{E}Q_n - \bar{Q}^k\|_\infty \qquad (2.52)\\
&= \Big| \sup_{\theta,\theta'} \frac{1}{n} \sum_{i=1}^{n} \sum_{z_i} \mathbb{E}(p(z_i \mid X_1^n, \theta') - p(z_i \mid X_{i-k}^{i+k}, \theta')) h(X_i, z_i, \theta, \theta') \Big| \\
&+ \Big| \frac{1}{n} \sup_{\theta,\theta'} \mathbb{E} \sum_{z_0} p(z_0 \mid X_1^n, \theta') \log p(z_0; \theta) \Big| \\
&\leq \sup_{\theta,\theta'} \frac{1}{n} \sum_{i=1}^{n} \sum_{z_i} \sup_x \big| p(z_i \mid x_1^n, \theta') - p(z_i \mid x_{i-k}^{i+k}, \theta') \big| \, \mathbb{E} \, |h(X_i, z_i, \theta, \theta')| + \frac{1}{n} \log \overline{\pi}_{\min}^{-1} \\
&\leq \frac{Cs^3}{\epsilon_{\mathrm{mix}}^8 \overline{\pi}_{\min} n} \big(2 \sum_{i=1}^{k} (1 - \epsilon_{\mathrm{mix}} \overline{\pi}_{\min})^i + (n - 2k)(1 - \epsilon_{\mathrm{mix}} \overline{\pi}_{\min})^k \big) \left[\max_{z_i \in [s]} \mathbb{E} |h(X_i, z_i, \theta, \theta')| \right] + \frac{1}{n} \log \overline{\pi}_{\min}^{-1} \\
&\leq \frac{Cs^3}{\epsilon_{\mathrm{mix}}^8 \overline{\pi}_{\min}} \big(\frac{2}{n \epsilon_{\mathrm{mix}} \overline{\pi}_{\min}} + \frac{n - 2k}{n} (1 - \epsilon_{\mathrm{mix}} \overline{\pi}_{\min})^k \big) \left[\max_{z_i \in [s]} \mathbb{E} |h(X_i, z_i, \theta, \theta')| \right] + \frac{1}{n} \log \overline{\pi}_{\min}^{-1} \\
&\leq \frac{C\, s^4}{\epsilon_{\mathrm{mix}}^9 \overline{\pi}_{\min}^2} \big(1 - \epsilon_{\mathrm{mix}} \overline{\pi}_{\min} \big)^k + \frac{1}{n} \left(\log \overline{\pi}_{\min}^{-1} + \frac{Cs^4}{\epsilon_{\mathrm{mix}}^{10} \overline{\pi}_{\min}^3} \right),
\end{aligned}
$$

using the crude bound

$$
\max_{z_i \in [s]} \mathbb{E} |h(X_i, z_i, \theta, \theta')| \leq \mathbb{E} \max_{z_i \in [s]} \big| \log p(X_i \mid z_i, \theta) \big| + s \log(\overline{\pi}_{\min} \epsilon_{\mathrm{mix}})^{-1} \leq \frac{Cs}{\overline{\pi}_{\min} \epsilon_{\mathrm{mix}}}.
$$

which uses condition (2.16) and where C denotes generic constants which are potentially different each time they appear.

2.8 Technical details for Corollary 1

In this section, we collect some auxiliary bounds on conditional covariances in hidden Markov models. These results are used in the proof of Corollary 1.

Lemma 4. *For any HMM with observed-hidden states (X_i, Z_i), we have*

$$
\mathbb{E} \left[\operatorname{var}(Z_0 Z_1 \mid X_{-k}^k) \right] \leq \mathbb{E} \operatorname{var}(Z_0 Z_1 \mid X_0^1) \qquad (2.53a)
$$

$$
\mathbb{E} \left[\operatorname{var}(Z_0 \mid X_{-k}^k) \mid X_0 \right] \leq \operatorname{var}(Z_0 \mid X_0) \qquad (2.53b)
$$

where we have omitted the dependence on the parameters.

Proof. We use the law of total variance, which guarantees that $\operatorname{var} Z = \mathbb{E}\big[\operatorname{var}(Z \mid X) \big] + \operatorname{var} \mathbb{E}[Z \mid X]$. Using this decomposition, we have

$$
\begin{aligned}
\mathbb{E}[\operatorname{var}(Z_0 \mid X_0^1) \mid X_0] &\leq \operatorname{var}(Z_0 \mid X_0) \\
\mathbb{E}[\operatorname{var}(Z_0 Z_1 \mid X_0^2) \mid X_0^1] &\leq \operatorname{var}(Z_0 Z_1 \mid X_0^1).
\end{aligned}
$$

The result then follows by induction. □

We now show that the expected conditional variance of the hidden state (or pairs thereof) conditioned on the corresponding observation (pairs of observations) decays exponentially with the SNR.

Lemma 5. *For a 2-state Markov chain with true parameter θ^*, we have for $\mu \in \mathbb{B}_2\big(\frac{\|\mu^*\|_2}{4}; \mu^*\big)$ and $\beta \in \Omega_\beta$*

$$\big\|\mathbb{E}X_0X_0^T(\operatorname{var}(Z_0 \mid \gamma_0, \beta))^{1/2}\big\|_{op} \leq c_0\, \mathrm{e}^{-c\eta^2} \tag{2.54a}$$

$$\mathbb{E}\operatorname{var}(Z_\ell \mid \gamma_\ell, \beta) \leq c_0\, \mathrm{e}^{-c\eta^2} \tag{2.54b}$$

$$\mathbb{E}\operatorname{var}(Z_0Z_1 \mid \gamma_0^1, \beta) \leq c_0 \frac{1+b}{1-b}\, \mathrm{e}^{-c\eta^2}. \tag{2.54c}$$

Proof. By definition of the Gaussian HMM example, we have $\operatorname{var}(Z_i \mid \gamma_i) = \frac{4}{(\mathrm{e}^{\gamma_i}+\mathrm{e}^{-\gamma_i})^2}$. Moreover, following the proof of Corollary 1 in the paper [3], we are guaranteed that $\mathbb{E}\operatorname{var}(Z_i \mid \gamma_i) \leq 8\mathrm{e}^{-\frac{\eta^2}{32}}$ and $\|\mathbb{E}X_iX_i^T(\operatorname{var}(Z_i|\gamma_i))^{1/2}\|_{op} \leq c_0\mathrm{e}^{-\frac{\eta^2}{32}}$, from which inequalities (2.54a) and (2.54b) follow.

We now prove inequality (2.54c) for $\beta \in \Omega_\beta$ and $\mu \in \mathbb{B}_2\big(\frac{\|\mu^*\|_2}{4}; \mu^*\big)$. Note that

$$\begin{aligned}
\frac{1}{4}\operatorname{var}(Z_0Z_1 \mid \gamma_0^1, \beta) &= \frac{\mathrm{e}^{2\gamma_1} + \mathrm{e}^{-2\gamma_1} + \mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0}}{\big[\mathrm{e}^{\beta}\big(\mathrm{e}^{\gamma_0+\gamma_1} + \mathrm{e}^{-(\gamma_0+\gamma_1)}\big) + \mathrm{e}^{-\beta}\big(\mathrm{e}^{\gamma_0-\gamma_1} + \mathrm{e}^{-(\gamma_0-\gamma_1)}\big)\big]^2} \\
&\leq \mathrm{e}^{2|\beta|}\frac{\mathrm{e}^{2\gamma_1} + \mathrm{e}^{-2\gamma_1} + \mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0}}{\big(\mathrm{e}^{\gamma_0+\gamma_1} + \mathrm{e}^{-(\gamma_0+\gamma_1)} + \mathrm{e}^{\gamma_0-\gamma_1} + \mathrm{e}^{-(\gamma_0-\gamma_1)}\big)^2} \\
&\leq \left(\frac{1+b}{1-b}\right)\left[\frac{\mathrm{e}^{|\gamma_0|}}{\mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0}} + \frac{\mathrm{e}^{|\gamma_1|}}{\mathrm{e}^{2\gamma_1} + \mathrm{e}^{-2\gamma_1}}\right]
\end{aligned}$$

where γ are now random variables and we used

$$\begin{aligned}
&\big(\mathrm{e}^{\gamma_0+\gamma_1} + \mathrm{e}^{-(\gamma_0+\gamma_1)} + \mathrm{e}^{\gamma_0-\gamma_1} + \mathrm{e}^{-(\gamma_0-\gamma_1)}\big)^2 \\
&\geq \mathrm{e}^{-|\gamma_0|}(\mathrm{e}^{-\gamma_0} + \mathrm{e}^{\gamma_0})(\mathrm{e}^{2\gamma_1} + \mathrm{e}^{-2\gamma_1}) + \mathrm{e}^{-|\gamma_1|}(\mathrm{e}^{-\gamma_1} + \mathrm{e}^{\gamma_1})(\mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0}) \\
&\geq (\mathrm{e}^{-|\gamma_0|} + \mathrm{e}^{-|\gamma_1|})(\mathrm{e}^{2\gamma_0} + \mathrm{e}^{-2\gamma_0})(\mathrm{e}^{2\gamma_1} + \mathrm{e}^{-2\gamma_1}).
\end{aligned}$$

It directly follows that

$$\begin{aligned}
\frac{1}{4}\mathbb{E}\operatorname{var}(Z_0Z_1 \mid \gamma_0^1, \beta) &\leq 2\left(\frac{1+b}{1-b}\right)\mathbb{E}\left[\frac{1}{\mathrm{e}^{\gamma_0} + \mathrm{e}^{-3\gamma_0}}\mathbb{1}_{\gamma_0\geq 0} + \frac{1}{\mathrm{e}^{3\gamma_0} + \mathrm{e}^{-\gamma_0}}\mathbb{1}_{\gamma_0\leq 0}\right] \\
&\leq 2\left(\frac{1+b}{1-b}\right)\left(\mathbb{E}[\mathrm{e}^{-\gamma_0}\mathbb{1}_{\gamma_0\geq 0}] + \mathbb{E}[\mathrm{e}^{\gamma_0}\mathbb{1}_{\gamma_0\leq 0}]\right) \\
&\leq 4\left(\frac{1+b}{1-b}\right)\mathbb{E}[\mathrm{e}^{-\gamma_0}\mathbb{1}_{\gamma_0\geq 0}]
\end{aligned}$$

where the last inequality follows from symmetry of the random variables X_i. One can then bound

$$\mathbb{E}[\mathrm{e}^{-\gamma_0}\mathbb{1}_{\gamma_0\geq 0}] = \mathbb{E}\,\mathrm{e}^{-\frac{\|\mu\|_2 V_1}{\sigma^2}}\mathbb{1}_{V_1\geq 0} \leq 2\mathrm{e}^{-\frac{\eta^2}{32}}$$

by employing a similar procedure as in the proof of Corollary 1 in [3]. Inequality (2.54c) then follows. □

The last lemma provides rigorous confirmation of the intuition that the covariance between any pair of hidden states should decay exponentially in their separation ℓ:

Lemma 6. *For a 2-state Markov chain with mixing coefficient ϵ_{mix} and uniform stationary distribution, we have*

$$\max\Big\{\operatorname{cov}(Z_0, Z_\ell \mid \gamma),\ \operatorname{cov}(Z_0Z_1, Z_\ell Z_{\ell+1} \mid \gamma),\ \operatorname{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \gamma)\Big\} \leq 2\rho_{\mathrm{mix}}^{\ell} \tag{2.55}$$

with $\rho_{\mathrm{mix}} = 1 - \epsilon_{\mathrm{mix}}$ for all $\theta \in \Omega$.

Lemma 6 is a mixing result and its proof is found in Section 2.10.

2.9 Technical details for Corollary 2

In this section we prove Lemmas 1 and 2. In order to do so, we leverage the independent blocks approach used in the analysis of dependent data (see, for instance, the papers [91, 59]). For future reference, we state here an auxiliary lemma that plays an important role in both proofs.

Let $\{X_i\}_{i=-\infty}^{\infty}$ be a sequence sampled from a Markov chain with mixing rate $\rho_{\mathrm{mix}} = 1-\epsilon_{\mathrm{mix}}$, $\overline{\pi}_{\min}$ be the minimum entry of the stationary distribution and $\widetilde{\rho}_{\mathrm{mix}} = 1 - \epsilon_{\mathrm{mix}}\overline{\pi}_{\min}$. Given some functions $f_1 : \mathbb{R}^{2k} \to \mathbb{R}^d$ and $f_2 : \mathbb{R} \to \mathbb{R}^d$ in some function class $\mathcal{F}_1, \mathcal{F}_2$ respectively, our goal is to control the difference between the functions

$$g_1(X) := \frac{1}{n}\sum_{i=1}^{n} f_1(X_{i-k}^{i+k}), \qquad g_2(X) := \frac{1}{n}\sum_{i=1}^{n} f_2(X_i) \tag{2.56a}$$

and their expectation. Defining $m_1 := \lfloor n/4k \rfloor$ and $m_2 := \lfloor n/k \rfloor$, we say that f_1 respectively f_2 is (δ, k)-concentrated if

$$\mathbb{P}\Big[\sup_{f\in\mathcal{F}_1} \|\frac{1}{m_1}\sum_{i=1}^{m_1} f_1(\widetilde{X}_{i;2k}) - \mathbb{E}f_1(\widetilde{X}_{1;2k})\|_2 \geq \epsilon\Big] \leq \frac{\delta}{8k}, \tag{2.56b}$$

$$\mathbb{P}\Big[\sup_{f\in\mathcal{F}_2} \|\frac{1}{m_2}\sum_{i=1}^{m_2} f_2(\widetilde{X}_i) - \mathbb{E}f_2(\widetilde{X}_1)\|_2 \geq \epsilon\Big] \leq \frac{\delta}{2k}$$

where $\{\widetilde{X}_{i;2k}\}_{i\in\mathbb{N}}$ are a collection of i.i.d. sequences of length $2k$ drawn from the same Markov chain and $\{\widetilde{X}_i\}_{i\in\mathbb{N}}$ a collection of i.i.d. variables drawn from the same stationary distribution. In our notation, $\{\widetilde{X}_{i;2k}\}_{i\in\mathbb{N}}$ under $\mathbb{P}$ are identically distributed as $\{X_{i;2k}\}_{i\in\mathbb{N}}$ under $\mathbb{P}_0$.

Lemma 7. *Consider functions f_1, f_2 that are (δ, k)-concentrated* (2.56b) *for a truncation parameter $k \geq \log\left(\frac{Cn}{\pi_{\min}^3 \epsilon_{\mix}^3 \delta}\right)(\log \frac{1}{\widetilde{\rho}_{\mix}})^{-1}$. Then the averaged functions g_1, g_2 from equation* (2.56a) *satisfy the bounds*

$$\mathbb{P}\Big[\sup_{g \in \mathcal{F}_1} \|g_1(X) - \mathbb{E}g_1(X)\|_2 \geq \epsilon\Big] \leq \delta \quad \text{and} \quad \mathbb{P}\Big[\sup_{g \in \mathcal{F}_2} \|g_2(X) - \mathbb{E}g_2(X)\|_2 \geq \epsilon\Big] \leq \delta. \quad (2.57)$$

Proof. We prove the lemma for functions of the type (f_1, g_1); the proof for the case (f_2, g_2) is very similar. In order to simplify notation, we assume throughout the proof that the effective sample size n is a multiple of $4k$, so that the block size $m = \frac{n}{4k}$ is integral. By definition (2.56a), the function g is a function of the sequences $\{X_{1-k}^{1+k}, X_{2-k}^{2+k}, \ldots, X_{n-k}^{n+k}\}$. We begin by dividing these sequences into blocks. Let us define the subsets of indices

$$\begin{aligned} H_i^j &= \{4k(i-1) + k + j \mid 4k(i-1) + 3k + j\}, \quad \text{and} \\ R_i^j &= \{4k(i-1) - k + j \mid 4k(i-1) + k - 1 + j\}. \end{aligned}$$

With this notation, we have the decomposition

$$g(X) = \frac{1}{2}\left(\frac{1}{2k}\sum_{j=1}^{2k} \underbrace{\frac{1}{m}\sum_{i=1}^{m} f(X_{H_i^j})}_{g^{H^j}(X)} + \frac{1}{2k}\sum_{j=1}^{2k} \underbrace{\frac{1}{m}\sum_{i=1}^{m} f(X_{R_i^j})}_{g^{R^j}(X)}\right),$$

from which we find that

$$\begin{aligned} \mathbb{P}\big[\sup_{g \in \mathcal{F}} \|g(X) - \mathbb{E}g(X)\|_2 \leq \epsilon\big] &\geq \mathbb{P}\big(\bigcap_{j=1}^{2k} \{\sup_{g \in \mathcal{F}} \|g^{H^j}(X) - \mathbb{E}g(X)\|_2 \leq \epsilon\} \\ &\qquad \cap \{\sup_{g \in \mathcal{F}} \|g^{R^j}(X) - \mathbb{E}g(X)\|_2 \leq \epsilon\}\big) \\ &\overset{\text{(i)}}{\geq} 1 - 4k\,\mathbb{P}(\sup_{g \in \mathcal{F}} \|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon), \end{aligned}$$

where (i) follows using stationarity of the underlying sequence combined with the union bound.

In order to bound the probability $\mathbb{P}\big[\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon\big]$, it is convenient to relate it to the probability of the same event under the product measure $\mathbb{P}_0$ on the blocks $\{H_1^1, \ldots, H_m^1\}$. In particular, we have $\mathbb{P}(\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon) \leq T_1 + T_2$, where

$$\begin{aligned} T_1 &:= \mathbb{P}_0(\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon), \quad \text{and} \\ T_2 &:= |\mathbb{P}(\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon) - \mathbb{P}_0(\|g^{H^1}(X) - \mathbb{E}g(X)\|_2 \geq \epsilon)|. \end{aligned}$$

By our assumed concentration (2.56b), we have $T_1 \leq \frac{\delta}{8k}$, and so it remains to show that $T_2 \leq \frac{\delta}{8k}$.

Now following standard arguments (e.g., see the papers [59, 91]), we first define

$$\beta(k) = \sup_{A \in \sigma(\mathcal{S}_{-\infty}^0, \mathcal{S}_k^\infty)} |\mathbb{P}(A) - \mathbb{P}_{-\infty}^0 \times \mathbb{P}_1^\infty(A)|, \tag{2.58}$$

where $\mathcal{S}_{-\infty}^0$ and $\mathcal{S}_k^\infty$ are the σ-algebras generated by the random vector $X_{-\infty}^0$ and X_k^∞ respectively, and $\mathbb{P}_{-\infty}^0 \times \mathbb{P}_1^\infty$ is the product measure under which the sequences $X_{-\infty}^0$ and X_1^∞ are independent. Define $\mathcal{S}_i$ to be the σ-algebra generated by $X_{H_i^j}$ for $i = \{1, \ldots, m\}$; it then follows by induction that $\sup_{A \in \sigma(\mathcal{S}_1, \ldots, \mathcal{S}_m)} |\mathbb{P}(A) - \mathbb{P}_0(A)| \leq m\beta(k)$. An identical relation holds over the blocks R_i^j.

For our two-state HMM, Lemma 12 implies that

$$\begin{aligned} \beta(k) = |p(x) - p(x_k^\infty)p(x_{-\infty}^0)| &\leq |p(x_{-\infty}^0 \mid x_k^n) - p(x_{-\infty}^0)| \\ &\leq |p(z_0 \mid x_k^n) - p(z_0)| \\ &\overset{\text{(i)}}{\leq} \frac{3}{\overline{\pi}_{\min}^3 \epsilon_{\text{mix}}^3} \rho_{\text{mix}}^k = \frac{3}{\overline{\pi}_{\min}^3 \epsilon_{\text{mix}}^3} \mathrm{e}^{-k \log(1/\rho_{\text{mix}})}, \end{aligned} \tag{2.59}$$

where step (i) follows from inequality (2.73). From our assumed lower bound on k, we conclude that $m\beta(k) \leq \frac{\delta}{8k}$, which completes the proof. □

In the following sections we apply it in order to prove the bounds on the approximation and sample error of the M-operators.

Proof of Lemma 1

We prove each of the two inequalities in equations (2.47b) and (2.47c) in turn by using suitable choices of the function f in Lemma 7. Throughout, note that our function class is parameterized and $f \in \mathcal{F}$ is equivalent to $\theta \in \Omega = \mathbb{B}_2(r; \mu^*) \times \Omega_\beta$.

Proof of inequality (2.47b)**:** We use the notation from the proof of Lemma 7 and furthermore define the weights $w_\theta(X_{i-k}^{i+k-1}) = p(Z_i = 1 \mid X_{i-k}^{i+k-1}, \theta)$, as well as the function $f_0(X_{i-k}^{i+k-1}, \theta') = (2w_{\theta'}(X_{i-k}^{i+k-1}) - 1)X_i$. It is then possible to write the EM operator explicitly as the average

$$M_n^{\mu,k}(\theta') = \arg\max_{\mu \in \widetilde{\Omega}} \frac{1}{n} \Big[\sum_{i=1}^n \mathbb{E}_{Z_i | X_{i-k}^{i+k}, \theta'} \log p(X_i \mid Z_i, \mu) \Big] = \frac{1}{n} \sum_{i=1}^n f_0(X_{i-k}^{i+k-1}, \theta').$$

We are now ready to apply Lemma 7 with the choices $f_1 = f_0$, $g_1(X) = M_n^{\mu,k}(\theta)$. According to Lemma 7, given that the lower bound on the truncation parameter k holds, we now need

to show that f_0 is (δ, k)-concentrated, that means finding ϵ_n^μ such that

$$\mathbb{P}_0\left[\sup_{\theta\in\Omega}\big\|\frac{1}{m}\sum_{i=1}^m f_0(\widetilde{X}_{i;2k},\theta') - \mathbb{E}f_0(\widetilde{X}_{i;2k},\theta')\big\|_2 \geq \epsilon_n^\mu\right] \leq \frac{\delta}{8k},$$

where $\mathbb{P}_0$ denotes the product measure over the independent blocks and $m := m_1 = \lfloor n/4k \rfloor$.

Let X_i be the middle element of the (i.i.d. drawn) sequence $\widetilde{X}_{i;2k}$ and Z_i, V_i the corresponding latent and noise variable. We can then write $X_i = Z_i + V_i$ where V_i are zero-mean Gaussian random variables with covariance matrix $\sigma^2 I$.

With a minor abuse of notation, let us use $X_{i,\ell}$ to denote ℓ^{th} element in the block $\widetilde{X}_{i;2k} = (X_{i,1}, \ldots, X_{i,2k})^T$, and write $\widetilde{X} = \{\widetilde{X}_{i;2k}\}_{i=1}^n$. In view of Lemma 7, our objective is to find the smallest scalar ϵ_n^μ such that

$$\mathbb{P}\big[\sup_{\theta\in\Omega}\|\frac{1}{m}\sum_{i=1}^m \underbrace{(2w_\theta(\widetilde{X}_{i;2k})-1)X_{i,k} - \mathbb{E}(2w_\theta(\widetilde{X}_{i;2k})-1)X_{i,k}}_{f_\theta(\widetilde{X}_{i;2k})}\|_2 \geq \epsilon_n^\mu\big] \leq \frac{\delta}{8k} \tag{2.60}$$

For each unit norm vector $u \in \mathbb{R}^d$, define the random variable

$$\tilde{V}_m(\widetilde{X}; u) = \sup_{\theta\in\Omega}\frac{1}{m}\sum_{i=1}^m (2w_\theta(\widetilde{X}_{i;2k})-1)\langle X_{i,k}, u\rangle - \mathbb{E}(2w_\theta(\widetilde{X}_{i;2k})-1)\langle X_{i,k}, u\rangle.$$

Let $\{u^{(1)}, \ldots, u^{(T)}\}$ denote a 1/2-cover of the unit sphere in $\mathbb{R}^d$; by standard arguments, we can find such a set with cardinality $\log T \leq d \log 5$. Using this covering, we have

$$\sup_{\theta\in\Omega}\|\frac{1}{m}\sum_{i=1}^m f_\theta(\widetilde{X}_{i;2k})\|_2 = \sup_{\|u\|_2\leq 1}\tilde{V}_m(\widetilde{X}; u) \leq 2\max_{j\in[T]}\tilde{V}_m(\widetilde{X}; u^{(j)}),$$

where the inequality follows by a discretization argument. Consequently, we have

$$\begin{aligned}\mathbb{P}\big[\sup_{\theta\in\Omega}\|\frac{1}{m}\sum_{i=1}^m f_\theta(\widetilde{X}_{i;2k})\|_2 \geq \epsilon_n^\mu\big] &\leq \mathbb{P}\big[\max_{j\in[T]}\tilde{V}_m(\widetilde{X}; u^{(j)}) \geq \frac{\epsilon_n^\mu}{2}\big] \\ &\leq T\max_{j\in[T]}\mathbb{P}\big[\tilde{V}_m(\widetilde{X}; u^{(j)}) \geq \frac{\epsilon_n^\mu}{2}\big].\end{aligned}$$

The remainder of our analysis focuses on bounding the tail probability for a fixed unit vector u, in particular ensuring an exponent small enough to cancel the $T \leq e^{d\log 5}$ pre-factor. By Lemma 2.3.7 of [74], for any $t > 0$, we have

$$\mathbb{P}_X\big[\tilde{V}_m(\widetilde{X}; u) \geq t\big] \leq c\mathbb{P}_{X,\epsilon}\big[V_m(\widetilde{X}; u) \geq \frac{t}{4}\big],$$

where $V_m(\widetilde{X}; u) = \sup_{\theta\in\Omega}\big|\frac{1}{m}\sum_{i=1}^m \epsilon_i(2w_\theta(\widetilde{X}_{i;2k})-1)\langle X_{i,k}, u\rangle\big|$, and $\{\epsilon_i\}_{i=1}^m$ is a sequence of i.i.d. Rademacher variables.

We now require a sequence of technical lemmas; see Section 2.9 for their proofs. Our first lemma shows that the variable $V_m(\widetilde{X}; u)$, viewed as a function of the Rademacher sequence, is concentrated:

Lemma 8. *For any fixed $(\widetilde{X}, u)$, we have*

$$\mathbb{P}_\epsilon\big[V_m(\widetilde{X};u) \geq \mathbb{E}_\epsilon V_m(\widetilde{X};u) + t\big] \leq 2\mathrm{e}^{-\frac{t^2}{16L_m^2(\widetilde{X};u)}}, \tag{2.61}$$

where $L_m(\widetilde{X};u) = \frac{1}{m}\sqrt{\sum_{i=1}^m \langle X_{i,k}, u\rangle^2}$.

Our next lemma bounds the expectation with respect to the Rademacher random vector:

Lemma 9. *There exists a universal constant c such that for each fixed $(\widetilde{X};u)$, we have*

$$\mathbb{E}_\epsilon V_m(\widetilde{X};u) \leq \underbrace{c\frac{\|\mu^*\|_2}{\sigma^2}\sqrt{\log m}\,\Big[\sum_{\ell=1}^{2k}\mathbb{E}_{\tilde\epsilon}\|\frac{1}{m}\sum_{i=1}^m \tilde\epsilon_{i,\ell}X_{i,\ell}\langle X_{i,k},u\rangle\|_2\Big]}_{M_m(\widetilde{X};u)} + \underbrace{\mathbb{E}_g\big|\frac{1}{m}\sum_{i=1}^m g_{i,2k+1}\langle X_{i,k},u\rangle\big|}_{N_m(\widetilde{X};u)} \tag{2.62}$$

where $\epsilon, \tilde\epsilon \in \mathbb{R}^m$ are random vectors with i.i.d. Rademacher components, and g is a random vector with i.i.d. $\mathcal{N}(0,1)$ components.

We now bound the three quantities $L_m(\widetilde{X};u)$, $M_m(\widetilde{X};u)$, and $N_m(\widetilde{X};u)$ appearing in the previous two lemmas. In particular, let us introduce the quantities $L' = cL\|\mu^*\|_2\big(\frac{\|\mu^*\|_2^2}{\sigma^2}+1\big)$, $L'' = L\sqrt{\|\mu^*\|_2^2+\sigma^2}$ and $L = \frac{\sqrt{8}}{1-\rho_{\mathrm{mix}}}$.

Lemma 10. *Define the event*

$$\mathcal{E} = \Bigg\{ L_m(\widetilde{X};u) \leq \tilde c\sqrt{\frac{2(\|\mu^*\|_2^2+\sigma^2)\log\frac{1}{\delta}}{m}}, \quad M_m(\widetilde{X};u) \leq L'k\sqrt{\frac{d\log m\log\frac{k}{\delta}}{m}} \\ \text{and } N_m(\widetilde{X};u) \leq cL''\sqrt{\frac{d\log\frac{1}{\delta}}{m}}\Bigg\}.$$

Then we have $\mathbb{P}\big[\mathcal{E}\big] \geq 1 - \mathrm{e}^{-c'\,d\log\frac{1}{\delta}}$ for $m > d$ and a universal constant $c' > 0$ which increases with the constants c in L', N_m and $\tilde c$ in L_m.

In conjunction, Lemmas 8 and 9 imply that conditionally on the event $\mathcal{E}$, we have

$$\mathbb{E}_\epsilon\big[V_m(\widetilde{X};u)\big] \leq c\sqrt{\|\mu^*\|_2^2+\sigma^2}\big(\frac{\|\mu^*\|_2^2}{\sigma^2}+1\big)k\sqrt{\frac{d\log m\log\frac{k}{\delta}}{m}}.$$

Note that by assumption on n we also have $m \geq d$ so that we can combine this bound with Lemma 10 which yields

$$\begin{aligned} T\,\mathbb{P}_X\big[\tilde V_m(\widetilde{X};u) \geq t\big] &\leq T\,\mathbb{P}_{X,\epsilon}\big[V_m(\widetilde{X};u) \geq \frac{t}{4} \mid \mathcal{E}\big] + T\,\mathbb{P}\big[\mathcal{E}^c\big] \\ &\leq 2\mathrm{e}^{4d-\left(\frac{c}{\tilde c}\right)^2 k^2 d\log m\log\frac{k}{\delta}} + \delta\mathrm{e}^{4d-c'd} \\ &\leq \delta, \end{aligned}$$

where the second inequality follows by setting $t/4 = c\|\mu^*\|_2(\frac{\|\mu^*\|_2^2}{\sigma^2}+1)k\log(\frac{k}{\delta})\sqrt{\frac{d\log m}{m}}$ and the final inequality holds for c', c and $\tilde{c}$ big enough. After rescaling δ by $8k$ and setting $m = \frac{n}{4k}$, the result follows after an application of Lemma 7.

Proof of inequality (2.47c)**:** In order to bound $|M_n^{\beta,k}(\theta) - \overline{M}^{\beta,k}(\theta)|$, we need a few extra steps. First, let us define new weights

$$v_\theta(X_{i-k}^{i+k-1}) = p(Z_0 = Z_1 = 1 \mid X_{i-k}^{i+k-1}, \theta) + p(Z_0 = Z_1 = -1 \mid X_{i-k}^{i+k-1}, \theta),$$

and also write the update in the form

$$\begin{aligned} M_n^{\beta,k}(\theta) &= \arg\max_{\zeta\in\Omega_\zeta} \Big\{ \mathbb{E}_{Z_1|X_{i-k}^{i+k},\theta} \log p(Z_1 \mid \zeta) + \sum_{i=2}^n \mathbb{E}_{Z_{i-1}^i|X_{i-k}^{i+k},\theta} \log p(Z_i \mid Z_{i-1}, \zeta) \Big\} \\ &= \arg\max_{\zeta\in\Omega_\zeta} \Big\{ \frac{1}{2} + \sum_{i=2}^n \mathbb{E}_{Z_{i-1}^i|X_{i-k}^{i+k},\theta} \log p(Z_i \mid Z_{i-1}, \zeta) \Big\} \\ &= \Pi_{\Omega_\zeta}\Big(\frac{1}{n}\sum_{i=2}^n v_\theta(X_{i-k}^{i+k-1}) \Big), \end{aligned}$$

where we have reparameterized the transition probabilities with ζ via the equivalences $\beta = h(\zeta) := \frac{1}{2}\log\left(\frac{\zeta}{1-\zeta}\right)$. Note that the original EM operator is obtained via the transformation $M_n^{\beta,k}(\theta') = h(M_n^{\beta,k}(\theta'))$ and we have $\overline{M}^{\beta,k}(\theta) = \Pi_{\Omega_\zeta}\mathbb{E}v_\theta(X_{i-k}^{i+k-1})$ by definition.

Given this set-up, we can now pursue an argument similar to that of inequality (2.47b). The new weights remain Lipschitz with the same constant—that is, we have the bound $|v_\theta(\widetilde{X}_{i;2k}) - v_{\theta'}(\widetilde{X}_{i;2k})| \leq L\|\tilde{\theta}_i - \tilde{\theta}_i'\|_2$. As a consequence, we can write

$$\mathbb{P}\big[\sup_{\theta\in\Omega} \mid \frac{1}{m}\sum_{i=1}^m v_\theta(\widetilde{X}_{i;2k}) - \mathbb{E}v_\theta(\widetilde{X}_{i;2k})| \geq \epsilon_n^\beta\big] \leq \frac{\delta}{8k},$$

with ϵ_n^β defined as in the lemma statement. In this case, it is not necessary to perform the covering step, nor to introduce extra Rademacher variables after the Gaussian comparison step; consequently, the two constants ϵ_n^β and ϵ_n^μ differ by a factor of $\sqrt{d\log n}$ modulo constants.

Applying Lemma 7 then yields a tail bound for the quantity $\big|\frac{1}{n}\sum_{i=1}^n v_\theta(\widetilde{X}_{i;2k}) - \mathbb{E}v_\theta(\widetilde{X}_{i;2k})\big|$ with probability δ. Since projection onto a convex set only decreases the distance, we find that

$$\mathbb{P}\left[|M_n^{\beta,k}(\theta) - \overline{M}^{\beta,k}(\theta)| \geq C\frac{\sqrt{\|\mu^*\|_2^2+\sigma^2}}{\sigma^2}\sqrt{\frac{k^3\log(k^2/\delta)}{n}} \right] \leq \delta.$$

In order to prove the result, the last step needed is the fact that

$$\frac{1}{2}\Big|\log\frac{x}{1-x} - \log\frac{y}{1-y}\Big| \leq \frac{1}{\tilde{x}(1-\tilde{x})}|x-y| \leq \frac{2}{1-b^2}|x-y| =: L|x-y|$$

for $x, y, \tilde{x} \in \Omega_\zeta$. Since $M_n^{\beta,k}(\theta) \in \Omega_\zeta$ we finally arrive at

$$\mathbb{P}\Big[|M_n^{\beta,k}(\theta) - \overline{M}^{\beta,k}(\theta)| \geq C(1-b^2)\frac{\sqrt{\|\mu^*\|_2^2 + \sigma^2}}{\sigma^2}\sqrt{\frac{k^3 \log\left(\frac{k^2}{\delta}\right)}{n}}\Big] \leq \delta$$

and the proof is complete.

Proof of Lemma 2

We need to show that

$$\mathbb{P}\Big[\sup_{\theta\in\Omega} \|M_n(\theta) - M_n^k(\theta)\|_\star^2 \geq c_1 \varphi_n^2(\delta, k)\Big] \leq \delta$$

with

$$\varphi_n^2(\delta, k) = \frac{Cs^4}{(1-b^2)\epsilon_{\text{mix}}^{10}\bar{\pi}_{\min}^3}\Big[\frac{1}{\sigma}\frac{d\log^2(C_\epsilon n/\delta)}{n} + \frac{\|\mu^*\|_2}{\sigma}\frac{\log^2(C_\epsilon n/\delta)}{n} + \frac{\|\mu^*\|_2^2}{\sigma^2}\frac{\delta}{n}\Big].$$

We first claim that

$$\sup_{\theta\in\Omega} \|M_n(\theta) - M_n^k(\theta)\|_\star^2 \leq \frac{8\|Q_n - Q_n^k\|_\infty}{\lambda}, \quad \text{where} \quad \lambda \geq \frac{2}{3}(1-b^2). \tag{2.63}$$

In Section 2.5. we showed that population operators are strongly concave with parameter at least λ. We make the added observation that using our parameterization, the sample Q functions $Q_n^k(\cdot \mid \theta'), Q_n(\cdot \mid \theta')$ are also strongly concave. This is because the concavity results for the population operators did not use any property of the covariates in the HMM, in particular not the expectation operator, and the single term $\frac{1}{n}\mathbb{E}\sum_{z_0} p(z_0 \mid X_1^n, \beta')\log p(z_0; \beta) = \frac{1}{n}\log\frac{1}{2}$ is constant for all $\beta \in \Omega_\beta$. From this λ-strong concavity, the bound (2.63) follows immediately using the same argumentation as in the proof of Theorem 1.

Given the bound (2.63), the remainder of the proof focuses on bounding the difference $\|Q_n - Q_n^k\|_\infty$. Recalling the shorthand notation

$$h(X_i, z_i, \theta, \theta') = \log p(X_i|z_i, \theta) + \sum_{z_{i-1}} p(z_i|z_{i-1}, \theta')\log p(z_i|z_{i-1}, \theta),$$

we use a similar argumentation as in the Proof of Proposition 2.3.1 equation (2.52) to obtain

$$\begin{aligned}
\|Q_n - Q_n^k\|_\infty &= \Big| \sup_{\theta,\theta' \in \Omega} \frac{1}{n} \sum_{i=1}^{n} \sum_{z_i} (p(z_i|X_1^n, \theta') - p(z_i|X_{i-k}^{i+k}, \theta')) h(X_i, z_i, \theta, \theta') \Big| \qquad (2.64) \\
&+ \Big| \sup_{\theta,\theta' \in \Omega} \frac{1}{n} \sum_{z_0} p(z_0 \mid x_1^n, \theta') \log p(z_0 \mid \theta) \Big| \\
&\leq \frac{2Cs^3}{\epsilon_{\text{mix}}^8 \overline{\pi}_{\min}} \frac{1}{n} \big[\sum_{i=1}^{k} \widetilde{\rho}_{\text{mix}}^i \max_{z_i \in [s]} \big| h(X_i, z_i, \theta, \theta') \big| + \log \overline{\pi}_{\min}^{-1} \big] \\
&+ \frac{Cs^3 \widetilde{\rho}_{\text{mix}}^k}{\epsilon_{\text{mix}}^8 \overline{\pi}_{\min}} \frac{1}{n-2k} \sum_{i=k}^{n-k} \max_{z_i \in [s]} \big| h(X_i, z_i, \theta, \theta') \big| \\
&\leq \underbrace{\frac{2Cs^3}{\epsilon_{\text{mix}}^8 \overline{\pi}_{\min} n} \big[\max_{z_i \in [s], X_1^k} \big| \log p(X_i \mid z_i, \theta) \big| \sum_{i=1}^{k} \widetilde{\rho}_{\text{mix}}^i + \frac{s \log(\overline{\pi}_{\min} \epsilon_{\text{mix}})^{-1}}{\overline{\pi}_{\min} \epsilon_{\text{mix}}} + \log \overline{\pi}_{\min}^{-1} \big]}_{S_1} \\
&+ \frac{Cs^3 \widetilde{\rho}_{\text{mix}}^k}{\epsilon_{\text{mix}}^8 \overline{\pi}_{\min}} \big[\mathbb{E} \max_{z_i \in [s]} \big| \log p(X_i \mid z_i, \theta') \big| + e_{n-2k}(X) + s \log(\overline{\pi}_{\min} \epsilon_{\text{mix}})^{-1} \big]
\end{aligned}$$

where we use $\max_{z_i \in [s]} \big| h(X_i, z_i, \theta, \theta') \big| \leq \max_{z_i \in [s]} \big| \log p(X_i | z_i, \theta) \big| + s \log(\overline{\pi}_{\min} \epsilon_{\text{mix}})^{-1}$, and

$$e_n(X) := \Big| \frac{1}{n} \sum_{i=1}^{n} \max_{z_i \in [s]} \Big| \log p(X_i \mid z_i, \theta) \Big| - \mathbb{E} \max_{z_i \in [s]} \Big| \log p(X_i \mid z_i, \theta) \Big| \Big|.$$

By assumption, we have that $\mathbb{E} \max_{z_i \in [s]} |\log p(X_i \mid z_i, \theta)|$ is bounded by an appropriately large universal constant. We therefore have with probability one that

$$S_1 \leq \frac{Cs^4}{\epsilon_{\text{mix}}^9 \overline{\pi}_{\min}^2} \frac{k}{n} \log(\epsilon_{\text{mix}} \overline{\pi}_{\min})^{-1}.$$

Putting these together, we find that

$$\sup_{\theta \in \Omega} \|M_n(\theta) - M_n^k(\theta)\|_\star^2 \leq \frac{Cs^4}{\lambda \epsilon_{\text{mix}}^9 \overline{\pi}_{\min}^2} \big[\frac{k}{n} \log(\epsilon_{\text{mix}} \overline{\pi}_{\min})^{-1} + \widetilde{\rho}_{\text{mix}}^k e_{n-2k}(X) \big].$$

Suppose that we can show that

$$\mathbb{P}\Big(e_n(X) \geq c_0 \Big(\frac{1}{\sigma} \sqrt{\frac{d \log^2(C_\epsilon n/\delta)}{n}} + \frac{\|\mu^*\|_2}{\sigma} \sqrt{\frac{\log^2(C_\epsilon n/\delta)}{n}} + \frac{\|\mu^*\|_2^2}{\sigma^2} \Big) \Big) \leq \delta, \qquad (2.65)$$

where c_0 is a universal constant and $C_\epsilon = \frac{C}{\epsilon_{\text{mix}}^3 \overline{\pi}_{\min}^3}$. By assumption we have $\frac{1}{2} \frac{\log\left(Cn/\delta\right)}{\log(1/\widetilde{\rho}_{\text{mix}})} \leq k \leq C \frac{\log\left(Cn/\delta\right)}{\log(1/\widetilde{\rho}_{\text{mix}})}$ so that we obtain

$$\sup_{\theta \in \Omega} \|M_n(\theta) - M_n^k(\theta)\|_\star^2 \leq \varphi_n^2(\delta, k)$$

with probability at least $1 - \frac{\delta}{3}$ for an appropriate choice of C.

We now move on to prove the bound (2.65). Observe that we have

$$\begin{aligned} e_n(X) &= \frac{1}{2n\sigma^2} \sum_{i=1}^{n} \left[\max\{\|X_i + \mu\|_2^2, \|X_i - \mu\|_2^2\} - \mathbb{E} \max\{\|X_i + \mu\|_2^2, \|X_i - \mu\|_2^2\} \right] \\ &= \frac{1}{2n\sigma^2} \sum_{i=1}^{n} \left(\|X_i\|_2^2 - \mathbb{E}\|X_i\|_2^2 \right) + \frac{1}{n\sigma^2} \sum_{i=1}^{n} \left(|X_i^T \mu| - \mathbb{E}|X_i^T \mu| \right). \end{aligned}$$

Note that we are again dealing with a dependent sequence so that we cannot use usual Hoeffding type bounds. For some $\tilde{k}$ to be chosen later on, and $m = n/\tilde{k}$ using the proof idea of Lemma 7 with $f_2(X_i) = |X_i^T \mu|$ and $f_2(X_i) = \|X_i\|_2^2$, we can write

$$\begin{aligned} \mathbb{P}(e_n(X) \geq \frac{t}{2\sigma^2}) \leq \tilde{k} \Big(& \underbrace{\mathbb{P}_0\Big(|\frac{1}{m} \sum_{i=1}^{m} \|X_i\|_2^2 - \mathbb{E}\|X_i\|_2^2| \geq \frac{t}{2} \Big)}_{T_1} \\ & + \underbrace{\mathbb{P}_0\Big(|\frac{1}{m} \sum_{i=1}^{m} |X_i^T \mu| - \mathbb{E}|X_i^T \mu|| \geq \frac{t}{4} \Big)}_{T_2} + m\beta(\tilde{k}) \Big), \end{aligned}$$

where $\beta(\tilde{k})$ was previously defined in equation (2.58). We claim that the choices

$$t := c_1 \Big(\sigma \sqrt{\frac{d \log(\tilde{k}/\delta)}{m}} + \sigma \|\mu^*\|_2 \sqrt{\frac{\log(\tilde{k}/\delta)}{m}} + \|\mu^*\|_2^2 \Big), \quad \text{and} \quad \tilde{k} := \frac{C_2 \log(\frac{3n}{\epsilon_{\text{mix}}^3 \overline{\pi}_{\min}^3 \delta})}{\log 1/\widetilde{\rho}_{\text{mix}}},$$

suffice to ensure that $\mathbb{P}(e_n(X) \geq t/(2\sigma^2)) \leq \delta$. Notice that the bound (2.59) implies that

$$m\beta(\tilde{k}) \leq \frac{cm\rho_{\text{mix}}^{\tilde{k}}}{\epsilon_{\text{mix}}^3 \overline{\pi}_{\min}^3} \leq \frac{\delta}{3\tilde{k}}.$$

In the sequel we develop bounds on T_1 and T_2. For T_1, observe that since $X_i \sim Z_i \mu^* + \epsilon_i$ where ϵ_i is a Gaussian vector with covariance $\sigma^2 I$ and Z_i independent under $\mathbb{P}_0$, standard χ^2 tail bounds imply that

$$\mathbb{P}_0 \Big[|\frac{1}{m} \sum_{i=1}^{m} \|X_i\|_2^2 - \mathbb{E}\|X_i\|_2^2| \geq \frac{t}{2} \Big] \leq \frac{\delta}{3\tilde{k}}.$$

Finally, we turn our attention to the term T_2. Observe that,

$$X_i^T \mu \sim \frac{1}{2} \mathcal{N}(\mu^T \mu^*, \sigma^2 \|\mu\|_2^2) + \frac{1}{2} \mathcal{N}(-\mu^T \mu^*, \sigma^2 \|\mu\|_2^2),$$

so that $| X_i^T \mu \sim |\mathcal{N}(\mu^T \mu^*, \sigma^2 \|\mu\|_2^2)|$. Denote $U_i = | X_i^T \mu |$. Letting ϵ denote a Rademacher random variable, observe that

$$\mathbb{E} \exp(tU_i) \stackrel{\text{(i)}}{\leq} \mathbb{E} \exp(2t\epsilon U_i) \stackrel{\text{(ii)}}{\leq} \exp\left(2t^2\sigma^2\|\mu\|_2^2 + 2t\mu^T\mu^*\right),$$

where (i) follows using symmetrization, and (ii) follows since the random variable ϵU_i is a Gaussian mixture. Observe that

$$\mathbb{E}U_i \stackrel{\text{(iii)}}{\leq} |\mu^T\mu^*| + \sigma\|\mu\|_2 \stackrel{\text{(iv)}}{\leq} \underbrace{2(\sigma + \|\mu^*\|_2)\|\mu^*\|_2}_{M},$$

where we have used for (iii) that U_i is a folded normal, and for (iv) that $\|\mu - \mu^*\|_2 \leq \frac{\|\mu^*\|_2}{4}$. Setting $D := 4\sigma\|\mu^*\|_2\sqrt{\frac{\log(6\tilde{k}/\delta)}{m}}$ observe that $\frac{t}{4} \geq 2M + D$ for big enough c_1. Thus, applying the Chernoff bound yields

$$\begin{aligned} T_2 \leq \mathbb{P}_0\Big[|\frac{1}{m}\sum_{i=1}^m U_i - \mathbb{E}U_i| \geq 2M + D\Big] &\leq \mathbb{P}_0\Big(\,|\, \frac{1}{m}\sum_{i=1}^m U_i \,|\geq M + D\Big) \\ &\leq 2\inf_{t\geq 0}\Big\{\mathbb{E}\exp\Big(\frac{t}{m}\sum_{i=1}^m U_i - Mt - Dt\Big)\Big\}, \\ &\leq 2\exp\Big(-\frac{mD^2}{8\sigma^2\|\mu\|_2^2}\Big) \leq \frac{\delta}{3\tilde{k}}. \end{aligned}$$

By combining the bounds on T_1 and T_2, some algebra shows that our choices of $t, \tilde{k}$ yield the claimed bound—namely, that $\mathbb{P}\big[e_n(X) \geq t/(2\sigma^2)\big] \leq \delta$.

Proofs of technical lemmas

In this section, we collect the proofs of various technical lemmas cited in the previous sections.

Proof of Lemma 8

We use the following concentration theorem (e.g., [51]): suppose that the function $f : \mathbb{R}^n \to \mathbb{R}$ is coordinate-wise convex and L-Lipschitz with respect to the Euclidean norm. Then for any i.i.d. sequence of variables $\{X_i\}_{i=1}^n$ taking values in the interval $[a, b]$, we have

$$\mathbb{P}\big[f(X) \geq \mathbb{E}f(X) + \delta\big] \leq \mathrm{e}^{-\frac{\delta^2}{4L^2(b-a)^2}} \tag{2.66}$$

We consider the process without absolute values (which introduces the factor of two in the lemma) and see that $\epsilon := (\epsilon_1, \ldots, \epsilon_n)$ is a random vector with bounded entries and that the supremum over affine functions is convex.

It remains to show that the function $\epsilon \mapsto V_m(\widetilde{X}, u)$ is Lipschitz with $L_m(\widetilde{X}; u)$ as follows

$$\begin{aligned}
&\big|\sup_\theta \frac{1}{m}\sum_{i=1}^m \epsilon_i f_\theta(\widetilde{X}_{i;2k}) - \sup_\theta \frac{1}{m}\sum_{i=1}^m \epsilon_i' f_\theta(\widetilde{X}_{i;2k})\big| \\
&\leq \frac{1}{m}|\sum_{i=1}^m (\epsilon_i - \epsilon_i') f_\theta(\widetilde{X}_{i;2k})| \\
&\leq \frac{1}{m}\sqrt{\sum_{i=1}^m (2w_{\tilde{\theta}}(\widetilde{X}_{i;2k}) - 1)^2 \langle X_{i,k}, u\rangle^2}\|\epsilon - \epsilon'\|_2 \\
&\leq L_m(\widetilde{X}; u)\|\epsilon - \epsilon'\|_2
\end{aligned}$$

where $\tilde{\theta} = \arg\max_{\theta\in\Omega} \sum_i \epsilon_i f_\theta(\widetilde{X}_{i;2k})$ in the last line and we use that $|2w_\theta(\widetilde{X}_{i;2k}) - 1| \leq 1$.

Proof of Lemma 9

The proof consists of three steps. First, we observe that the Rademacher complexity is upper bounded by the Gaussian complexity. Then we use Gaussian comparison inequalities to reduce the process to a simpler one, followed by a final step to convert it back to a Rademacher process.

Relating the Gaussian and Rademacher complexity: Let $g_i \sim \mathcal{N}(0, 1)$. It is easy to see that using Jensen's inequality and the fact that $\epsilon_i|g_i| \overset{d}{=} g_i$

$$\begin{aligned}
\mathbb{E}_\epsilon \sup_\theta \frac{1}{m}\sum_{i=1}^m \epsilon_i f_\theta(\widetilde{X}_{i;2k}) &= \sqrt{\frac{2}{\pi}}\mathbb{E}_\epsilon \sup_\theta \frac{1}{m}\sum_{i=1}^m \epsilon_i \mathbb{E}_g[|g_i|] f_\theta(\widetilde{X}_{i;2k}) \\
&\leq \sqrt{\frac{2}{\pi}}\mathbb{E}_g \sup_\theta \frac{1}{m}\sum_{i=1}^m g_i f_\theta(\widetilde{X}_{i;2k}).
\end{aligned}$$

Lipschitz continuity: For $\theta = (\mu, \beta)$ define the corresponding effective parameter that is obtained by treating the observed variables X as fixed

$$\tilde{\theta}_i := (\gamma_i, \beta) = (\frac{\langle \mu, X_{i,1}\rangle}{\sigma^2}, \ldots, \frac{\langle \mu, X_{i,2k}\rangle}{\sigma^2}, \beta). \tag{2.67}$$

Now we can use results in the proof of Corollary 1 to see that $\tilde{\theta}_i \mapsto F(\tilde{\theta}_i; \widetilde{X}_{i;2k}) := f_\theta(\widetilde{X}_{i;2k})$ is Lipschitz in the Euclidean norm, i.e. there exists an L, only dependent on ρ_{mix} such that

$$|F(\tilde{\theta}_i; \widetilde{X}_{i;2k}) - F(\tilde{\theta}_i'; \widetilde{X}_{i;2k})| \leq L\|\tilde{\theta}_i - \tilde{\theta}_i'\|_2 |\langle X_{i,k}, u\rangle| \tag{2.68}$$

For this we directly use results (exponential family representation) that were used to show Corollary 1. We overload notation and write $X_\ell := X_{1,\ell}$ and analyze Lipschitz continuity for

the first block. First note that $F(\tilde{\theta}_i, X_{1;2k}) = (2\mathbb{E}_{Z_k|X_1^{2k},\theta} Z_k - 1)X_{i,k}$. By Taylor's theorem, we then have

$$\begin{aligned}
|F(\tilde{\theta}_i; \widetilde{X}_{i;2k}) - F(\tilde{\theta}'_i; \widetilde{X}_{i;2k})| &= |\langle X_{i,k}, u\rangle||\mathbb{E}_{Z_k|\widetilde{X}_{i;2k},\theta} Z_k - \mathbb{E}_{Z_k|\widetilde{X}_{i;2k},\theta'} Z_k| \\
&\leq |\langle X_{i,k}, u\rangle||\mathbb{E}_{Z_k|\widetilde{X}_{i;2k},(\mu,\beta)} Z_k - \mathbb{E}_{Z_k|\widetilde{X}_{i;2k},(\mu',\beta)} Z_k| \\
&+ |\langle X_{i,k}, u\rangle||\mathbb{E}_{Z_k|\widetilde{X}_{i;2k},(\mu',\beta)} Z_k - \mathbb{E}_{Z_k|\widetilde{X}_{i;2k},(\mu',\beta')} Z_k|
\end{aligned}$$

Let us examine each of the summands separately. By the Cauchy-Schwartz inequality and Lemma 6, we have

$$\begin{aligned}
|\mathbb{E}_{Z_k|\widetilde{X}_{i;2k},(\mu,\beta)} Z_k - \mathbb{E}_{Z_k|\widetilde{X}_{i;2k},(\mu',\beta)} Z_k| &= \frac{1}{\sigma}|\sum_{\ell=1}^{2k} \frac{\partial^2 \Phi}{\partial\gamma_\ell \partial\gamma_0}\Big|_{\theta=\tilde{\theta}}(\gamma_\ell - \gamma'_\ell)| \\
&= |\sum_{\ell=1}^{2k} \mathrm{cov}(Z_0, Z_\ell \mid \widetilde{X}_{i;2k}, \tilde{\theta})(\langle \mu, X_\ell\rangle - \langle \mu', X_\ell\rangle)| \\
&\leq \sqrt{(\sum_{\ell=1}^{2k} 4\rho_{\mathrm{mix}}^{2\ell}) \sum_{\ell=1}^{2k} (\gamma_\ell - \gamma'_\ell)^2},
\end{aligned}$$

as well as

$$\begin{aligned}
|\mathbb{E}_{Z_k|\widetilde{X}_{i;2k},(\mu',\beta)} Z_k - \mathbb{E}_{Z_k|\widetilde{X}_{i;2k},(\mu',\beta')} Z_k| &= |\sum_{\ell=1}^{2k} \frac{\partial^2 \Phi}{\partial\beta_\ell \partial\gamma_0}\Big|_{\theta=\tilde{\theta}}(\beta - \beta')| \\
&= |\sum_{\ell=1}^{2k} \mathrm{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \widetilde{X}_{i;2k}, \tilde{\theta})(\beta - \beta')| \\
&\leq \frac{2}{1-\rho_{\mathrm{mix}}}|\beta - \beta'|.
\end{aligned}$$

Combining these two bounds yields

$$\begin{aligned}
|F(\tilde{\theta}_i; \widetilde{X}_{i;2k}) - F(\tilde{\theta}'_i; \widetilde{X}_{i;2k})|^2 &\leq \langle X_{i,k}, u\rangle^2 L\big(\sum_{\ell=1}^{2k} (\gamma_\ell - \gamma'_\ell)^2 + (\beta - \beta')^2\big) \\
&= \langle X_{i,k}, u\rangle^2 L^2 \|\tilde{\theta}_i - \tilde{\theta}'_i\|_2^2
\end{aligned}$$

with $L^2 = \frac{8}{(1-\rho_{\mathrm{mix}})^2}$.

Applying the Sudakov-Fernique Gaussian comparison: Let us introduce the shorthands $X_\theta = \frac{1}{m}\sum_i g_i f_\theta(\widetilde{X}_{i;2k})$, and

$$Y_\theta = \frac{1}{m} L \sum_i \big(\sum_{\ell=1}^{2k} g_{i\ell} \frac{\langle \mu, X_{i,\ell}\rangle}{\sigma^2} + g_{i,2k+1}\beta\big)\langle X_{i,k}, u\rangle.$$

By construction, the random variable $X_\theta - X'_\theta$ is a zero-mean Gaussian variable with variance

$$\begin{aligned}\mathbb{E}_g(X_\theta - X_{\theta'})^2 &= \sum_i (F(\tilde{\theta}; \widetilde{X}_{i;2k}) - F(\tilde{\theta}'; \widetilde{X}_{i;2k}))^2 \\ &\leq L^2 \sum_i \langle X_{i,k}, u\rangle^2 \big(\sum_{\ell=1}^{2k} (\gamma_{i,\ell} - \gamma'_{i,\ell})^2 + (\beta - \beta')^2 \big) \\ &= \mathbb{E}_g(Y_\theta - Y_{\theta'})^2 \end{aligned} \tag{2.69}$$

By the Sudakov-Fernique comparison [53],we are then guaranteed that $\mathbb{E}\sup_\theta X_\theta \leq \mathbb{E}\sup_\theta Y_\theta$. Therefore, it is sufficient to bound

$$\mathbb{E}_g \sup_{\theta\in\Omega} Y_\theta = \underbrace{\mathbb{E}_g \sup_\theta \frac{L}{\sigma^2 n} \sum_{i=1}^m \sum_{\ell=1}^{2k} g_{i\ell} \langle \mu, X_{i,\ell}\rangle \langle X_{i,k}, u\rangle}_{T_1} + \underbrace{\mathbb{E}_g \sup_\theta \frac{L}{n} \sum_{i=1}^m g_{i,2k+1} \beta \langle X_{i,k}, u\rangle}_{T_2}.$$

Converting back to a Rademacher process: We now convert the term T_1 back to a Rademacher process, which allows us to use sub-exponential tail bounds. We do so by re-introducing additional Rademacher variables, and then removing the term $\max_i |g_i|$ via the Ledoux-Talagrand contraction theorem [53]. Given a Rademacher variable ϵ_{il} independent of g, note the distributional equivalence $\epsilon_{il} g_{il} \stackrel{d}{=} g_{i\ell}$. Then consider the function $\phi_{i\ell}(g_{i\ell}) := g_{i\ell} h_{i\ell}$ with $h_{i\ell} := \langle \mu, X_{i,\ell}\rangle \langle X_{i,k}, u\rangle$ for which it is easy to see that

$$|\phi_{i\ell}(g_{i\ell}, h_{i\ell}) - \phi_{i\ell}(g_{i\ell}, h'_{i\ell})| \leq |g_{i\ell}||h_{i\ell} - h'_{i\ell}| \tag{2.70}$$

Applying Theorem 4.12. in Ledoux and Talgrand [53] yields

$$\mathbb{E}\sup_\theta \frac{1}{m} \sum_{i=1}^m \sum_{\ell=1}^{2k} \epsilon_{i\ell} g_{i\ell} \langle \mu, X_{i\ell}\rangle \langle X_{i,k}, u\rangle \leq \mathbb{E}_g \|g\|_\infty \mathbb{E}_\epsilon \sup_\theta \frac{1}{m} \sum_{i=1}^m \sum_{\ell=1}^{2k} \epsilon_{i\ell} \langle \mu, X_{i,\ell}\rangle \langle X_{i,k}, u\rangle.$$

Putting together the pieces yields the claim (2.62).

Proof of Lemma 10

We prove that the probability of each of the events corresponding to the inequalities is smaller than $\frac{1}{3} \mathrm{e}^{-\tilde{c} d \log(\frac{k}{\delta})}$.

Bounding L_m: We start by bounding $L_m(\widetilde{X}; u)$. Note that

$$\frac{1}{m} \sum_{i=1}^m \langle X_{i,k}, u\rangle^2 \leq \|\mu^*\|_2^2 + \frac{1}{m} \sum_{i=1}^m \langle n_{i,k}, u\rangle^2 + \frac{1}{m} \sum_{i=1}^m \langle \mu^*, u\rangle \langle n_{i,k}, u\rangle$$

where the sum $\sum_{i=1}^m \langle n_{i,k}, u\rangle^2$ is sub-exponential random variable with parameters $(2\sqrt{m}\sigma^2, 4)$ so that

$$\mathbb{P}(\frac{1}{m}\sum_{i=1}^m \langle n_{i,k}, u\rangle^2 - \sigma^2 \geq \tilde{c}^2\sigma^2 \log(1/\delta)) \leq \mathrm{e}^{-c' m \log\frac{1}{\delta}} \leq \mathrm{e}^{-c' d \log\frac{1}{\delta}}$$

where the last inequality follows since $m \geq d$ by assumption. Since $\langle n_{i,k}, u\rangle$ can be readily bounded by a sub-Gaussian tailbound it then follows directly that $L_m^2(\widetilde{X}; u) \leq \tilde{c}^2 \frac{2(\|\mu^*\|_2^2 + \sigma^2)\log\frac{1}{\delta}}{m}$ with probability at least $1 - \frac{1}{3}\mathrm{e}^{-c' d \log(\frac{k}{\delta})}$ for c' large enough.

Bounding N_m: In order to bound $N_m(\widetilde{X}; u)$, we first introduce an extra Rademacher random variable into its definition; doing so does not change its value (now defined by an expectation over both g and the Rademacher variables). We now require a result for a product of the form ϵgh where g, h are independent Gaussian random variables.

Lemma 11. *Let (ϵ, g, h) be independent random variables, with ϵ Rademacher, $g \sim \mathcal{N}(0, \sigma_g^2)$, and $h \sim \mathcal{N}(0, \sigma_h^2)$. Then the random variable $Z = \epsilon gh$ is a zero-mean sub-exponential random variable with parameters $(\frac{\sigma_g^2\sigma_h^2}{2}, \frac{1}{4})$.*

Proof. Note that $g' = \alpha h$ with $\alpha = \frac{\sigma_g}{\sigma_h}$ is identically distributed as g. Therefore, we have

$$gh = \frac{1}{\alpha} g g' = \frac{1}{4\alpha}[(g - g')^2 + (g + g')^2]$$

The random variables $g - g'$ and $g + g'$ are independent and therefore $(g - g')^2, (g + g')^2$ are sub-exponential with parameters $\nu^2 = 4\sigma_g^4$, $b = \frac{1}{4}$. This directly yields

$$\mathbb{E}\mathrm{e}^{\lambda\epsilon[(g+g')^2 - (g-g')^2]} \leq \mathrm{e}^{4\lambda^2\sigma_g^4}$$

for $|\lambda| \leq \frac{1}{b}$. Therefore $\mathbb{E}\mathrm{e}^{\lambda\epsilon gh} \leq \mathrm{e}^{\frac{\lambda^2\sigma_g^2\sigma_h^2}{4}}$, which shows that ϵgh is sub-exponential with parameters $(\frac{\sigma_g^2\sigma_h^2}{2}, \frac{1}{4})$. □

Returning to the random variable $N_m(\widetilde{X}; u)$, each term $\epsilon_i g_{i,2k+1}\langle X_{i,k}, u\rangle$ is a sub-exponential random variable with mean zero and parameter $\nu^2 = \|\mu^*\|_2^2 + \frac{\sigma^2}{2}$. Consequently, there are universal constants such that $N_m(\widetilde{X}; u) \leq cL\nu\sqrt{\frac{d\log\frac{k}{\delta}}{m}}$ with probability at least $1 - \frac{1}{3}\mathrm{e}^{-c' d\log(\frac{k}{\delta})}$.

Bounding M_m: Our next claim is that with probability at least $1 - \frac{1}{3}\mathrm{e}^{-c' d\log(\frac{k}{\delta})}$, we have

$$\mathbb{E}_\epsilon \|\frac{1}{m}\sum_{i=1}^m \epsilon_{i\ell} X_{i,\ell}\langle X_{i,k}, u\rangle\|_2 \leq (\|\mu^*\|_2^2 + \sigma^2)\sqrt{\frac{d\log\frac{k}{\delta}}{m}}, \tag{2.71}$$

which then implies that $M_m(\widetilde{X}; u) \leq c\|\mu^*\|_2\big(\frac{\|\mu^*\|_2}{\sigma^2} + 1\big)k\sqrt{\frac{d \log m \log \frac{k}{\delta}}{m}}$. In order to establish this claim, we first observe that by Lemma 11, the random variable $\epsilon_{i,\ell}\langle X_{i,\ell}, u\rangle\langle X_{i,k}, u\rangle$ is zero mean, sub-exponential with parameter at most $\nu^2 = (\|\mu^*\|_2^2 + \sigma^2)^2$. The bound then follows by the same argument used to bound the quantity N_m.

2.10 Mixing related results

In the following we use the shorthand notation $\pi_k^\theta := p(z_k \mid x_0^k, \theta)$ which we refer to the filtering distribution which is tied to some distribution μ on z_0.

Introducing the shorthand notation $p_\mu(x_k) := \sum_{z_k}\sum_{z_{k-1}} p(x_k \mid z_k)p(z_k \mid z_{k-1})\mu(z_{k-1})$, we define the filter operator

$$F_i\nu(z_i) := \frac{\sum_{z_{i-1}} p(x_i \mid z_i)p(z_i \mid z_{i-1})\nu(z_{i-1})}{\sum_{z_i}\sum_{z_{i-1}} p(x_i \mid z_i)p(z_i \mid z_{i-1})\nu(z_{i-1})} \;=\; \sum_{z_{i-1}} \frac{p(x_i \mid z_i)p(z_i \mid z_{i-1})}{p_\nu(x_i)}\nu(z_{i-1}).$$

where the observations x are fix. Using this notation, the filtering distribution can then be rewritten in the more compact form $\pi_k^\theta = p(z_k \mid x_0^k, \theta) = F_k \dots F_1\mu$. Similarly, we define

$$K_{j|i}(z_j \mid z_{j-1}) := \frac{p(z_j \mid z_{j-1})p(x_j \mid z_j)p(x_{j+1}^i \mid z_j)}{\sum_{z_j} p(z_j \mid z_{j-1})p(x_j \mid z_j)p(x_{j+1}^i|z_j)}, \quad \text{and} \quad \nu_{\ell|i} := \frac{p(x_{\ell+1}^i \mid z_\ell)\nu(z_\ell)}{\sum_{z_\ell} p(x_{\ell+1}^i \mid z_\ell)\nu(z_\ell)}$$

Note that $\epsilon_{\text{mix}}C_0 \leq p(x_{\ell+1}^i \mid z_\ell) \leq \epsilon_{\text{mix}}^{-1}C_0$ where

$$C_0 = \sum_{z_i \dots z_{\ell+1}} p(x_i \mid z_i)p(z_i \mid z_{i-1}) \dots p(x_{\ell+1} \mid z_{\ell+1})\overline{\pi}(z_{\ell+1})$$

and therefore by definition of ϵ_{mix} (2.3)

$$\sup_x \frac{\sup_z p(x_{\ell+1}^i \mid z_\ell)}{\inf_z p(x_{\ell+1}^i \mid z_\ell)} \leq \epsilon_{\text{mix}}^{-2}. \tag{2.72}$$

With these definitions, it can be verified (e.g., see Chapter 5 of [75]) that $F_i \dots F_{\ell+1}\nu = \nu_{\ell+1|i}^T K_{\ell+1|i} \dots K_{i|i}$, where $\nu^T K := \int \nu(x')K(x|x')\mathrm{d}x'$. In the discrete setting, this relation can be written as the row vector ν being right multiplied by the kernel matrix K.

Consequences of mixing

In this technical section we derive several useful consequences of the geometric mixing condition on the stochastic process Z_i.

Lemma 12. *For any geometrically ρ_{mix}-mixing and time reversible Markov chain $\{Z_i\}$ with s states, there is a universal constant c such that*

$$\sup_{z_0} \big|p(z_0 \mid x_k^n) - p(z_0)\big| < \frac{c(s+1)}{\overline{\pi}_{\min}^3 \epsilon_{\text{mix}}^3}\rho_{\text{mix}}^k. \tag{2.73}$$

Proof. We first prove the following relation

$$\sup_x |p(z_i \mid x_{i+k}) - p(z_i)| \leq c_0 \frac{\rho_{\text{mix}}^k}{\overline{\pi}_{\min}}. \tag{2.74}$$

Using time reversibility and the definition of mixing (2.4) we obtain

$$\begin{aligned}
\max_x (p(z_0 \mid x_k) - \overline{\pi}(z_0)) &= \sum_{z_k} (p(z_0 \mid z_k) - \overline{\pi}(z_0)) p(z_k \mid x_k) \\
&\leq \max_{z_k} |(p(z_0 \mid z_k) - \overline{\pi}(z_0))| \sum_{z_k} p(z_k \mid x_k) \\
&\leq \max_{z_k} \Big| \frac{p(z_k \mid z_0)\overline{\pi}(z_0)}{\overline{\pi}(z_k)} - \frac{\overline{\pi}(z_0)\overline{\pi}(z_k)}{\overline{\pi}(z_k)} \Big| \\
&\leq \frac{\overline{\pi}(z_0)}{\overline{\pi}(z_k)} \max_{z_k} |p(z_k \mid z_0) - \overline{\pi}(z_0)| \leq \frac{c_0 \rho_{\text{mix}}^k}{\overline{\pi}_{\min}}
\end{aligned}$$

where $p(z_k \mid z_0) = P(Z_k = z_k \mid Z_0 = z_0)$ and $p(z_0 \mid z_k) = P(Z_0 = z_0 \mid Z_k = z_k)$.

Using this result we can now prove inequality (2.73). By definition, we have

$$p(z_0) = \frac{p(x_{k+1}^n \mid x_k) p(x_k) p(z_0)}{p(x_{k+1}^n, x_k)}, \quad \text{and} \quad p(z_0 \mid x_k, x_{k+1}^n) = \frac{p(x_{k+1}^n \mid x_k, z_0) p(x_k \mid z_0) p(z_0)}{p(x_{k+1}^n, x_k)}$$

and therefore

$$\begin{aligned}
|p(z_0) - p(z_0 \mid x_k^n)| \leq \frac{p(x_k)p(z_0)}{p(x_k^n)} &|p(x_{k+1}^n \mid x_k) - p(x_{k+1}^n \mid x_k, z_0)| \\
&+ \frac{p(x_{k+1}^n \mid x_k, z_0) p(z_0)}{p(x_{k+1}^n \mid x_k)} |p(x_k) - p(x_k \mid z_0)|
\end{aligned} \tag{2.75}$$

In the following we bound each of the two differences. Note that

$$\begin{aligned}
|p(x_{k+1}^n \mid x_k, z_0) - p(x_{k+1}^n \mid x_k)| &= \sum_{z_k} \sum_{z_{k+1}} p(x_{k+1}^n \mid z_{k+1}) p(z_{k+1} \mid z_k) |p(z_k \mid x_k, z_0) - p(z_k \mid x_k)| \\
&\leq \sup_{z_k, x_k} |p(z_k \mid x_k, z_0) - p(z_k \mid x_k)| \sum_{z_k} p(x_{k+1}^n \mid z_k)
\end{aligned} \tag{2.76}$$

The last term $\sum_{z_k} p(x_{k+1}^n \mid z_k)$ is bounded by s for s-state models. Using the bound (2.74), we obtain

$$|p(x_k \mid z_0) - p(x_k)| = \frac{|p(z_0 \mid x_k) - \overline{\pi}(z_0)| p(x_k)}{\overline{\pi}(z_0)} \leq \frac{p(x_k)}{\overline{\pi}_{\min}^2} \rho_{\text{mix}}^k \tag{2.77}$$

which yields

$$\begin{aligned}
|p(z_k \mid x_k, z_0) - p(z_k \mid x_k)| &= p(x_k \mid z_k) \left| \frac{p(z_k \mid z_0)}{p(x_k \mid z_0)} - \frac{\overline{\pi}(z_k)}{p(x_k)} \right| \\
&\leq \frac{p(x_k \mid z_k)}{p(x_k \mid z_0)} \big(|p(z_k \mid z_0) - \overline{\pi}(z_k)| + \frac{\overline{\pi}(z_k)}{p(x_k)} |p(x_k \mid z_0) - p(x_k)| \big) \\
&\leq \frac{p(x_k \mid z_k)}{p(x_k \mid z_0)} \left(\rho_{\text{mix}}^k + \frac{1}{\overline{\pi}_{\text{min}}^2} \rho_{\text{mix}}^k \right) \\
&\leq \frac{2\rho_{\text{mix}}^k}{p(z_k \mid z_0)\overline{\pi}_{\text{min}}^2} \leq \frac{2}{\overline{\pi}_{\text{min}}^3 \epsilon_{\text{mix}}} \rho_{\text{mix}}^k. \qquad (2.78)
\end{aligned}$$

The last statement is true because one can check that for all $t \in \mathbb{N}$ we have

$$\min_{z_k, z_0} p(z_k \mid z_0) = \min_{ij} (A^t)_{ij} \geq \min_{ij} (A)_{ij} \geq \epsilon_{\text{mix}} \overline{\pi}_{\text{min}}$$

for any stochastic matrix A which satisfies the mixing condition (2.3).

Substituting (2.76) with (2.78) and (2.77) into (2.75), we obtain

$$\begin{aligned}
|p(z_0) - p(z_0 \mid x_k^n)| &\leq \frac{\sum_{z_k} p(x_{k+1}^n \mid z_k) p(z_0)}{\sum_{z_k} p(x_{k+1}^n \mid z_k) p(z_k \mid x_k)} \frac{2\rho_{\text{mix}}^k}{\overline{\pi}_{\text{min}}^3 \epsilon_{\text{mix}}} + \frac{p(x_{k+1}^n \mid x_k, z_0) p(z_0)}{p(x_{k+1}^n \mid x_k)} \frac{\rho_{\text{mix}}^k}{\overline{\pi}_{\text{min}}} \\
&\leq \left(\frac{2s}{\overline{\pi}_{\text{min}}^3 \epsilon_{\text{mix}}^3} + \frac{1}{\epsilon_{\text{mix}}^2 \overline{\pi}_{\text{min}}} \right) \rho_{\text{mix}}^k \leq \frac{2s+1}{\overline{\pi}_{\text{min}}^3 \epsilon_{\text{mix}}^3} \rho_{\text{mix}}^k
\end{aligned}$$

where we use (2.72) to see that

$$\frac{\sum_{z_k} p(x_{k+1}^n \mid z_k) p(z_k \mid x_k, z_0)}{\sum_{z_k} p(x_{k+1}^n \mid z_k) p(z_k \mid x_k)} \leq \frac{\max_{z_k} p(x_{k+1}^n \mid z_k)}{\min_{z_k} p(x_{k+1}^n \mid z_k)} \leq \epsilon_{\text{mix}}^{-2}$$

and similarly for the first term. □

Lemma 13 (Filter stability)**.** *For any mixing Markov chain which fulfills condition* (2.3)*, the following holds*

$$\|F_i \dots F_1(\nu - \nu')\|_\infty \leq \epsilon_{\text{mix}}^{-2} \widetilde{\rho}_{\text{mix}}^i \|\nu - \nu'\|_1$$

where $\widetilde{\rho}_{\text{mix}} = 1 - \epsilon_{\text{mix}} \overline{\pi}_{\text{min}}$*. In particular we have*

$$\sup_{z_i} |p(z_i \mid x_1^i) - p(z_i \mid x_{-n}^i)| \leq 2\epsilon_{\text{mix}}^{-2} \widetilde{\rho}_{\text{mix}}^i. \qquad (2.79)$$

Proof. Given the mixing assumption (2.3) we can show that $K_{j|i}(x|y) \geq \epsilon p_{j|i}(x)$ with $\epsilon = \epsilon_{\text{mix}} \overline{\pi}_{\text{min}}$ for some probability distribution $p_{j|i}(\cdot)$. This is because we can lower bound

$$\begin{aligned}
K_{j|i}(z_j \mid z_{j-1}) &= \frac{p(z_j \mid z_{j-1}) p(x_j \mid z_j) p(x_{j+1}^i \mid z_j)}{\sum_{z_j} p(z_j \mid z_{j-1}) p(x_j \mid z_j) p(x_{j+1}^i \mid z_j)} \\
&\geq \underbrace{\frac{\epsilon_{\text{mix}} \overline{\pi}(z_j) p(x_j \mid z_j) p(x_{j+1}^i \mid z_j)}{\sum_{z_j} \frac{\overline{\pi}(z_j)}{\overline{\pi}_{\text{min}}} p(x_j \mid z_j) p(x_{j+1}^i \mid z_j)}}_{=: \epsilon p_{j|i}(z_j)}
\end{aligned}$$

with $\epsilon = \epsilon_{\text{mix}}\overline{\pi}_{\text{min}}$. This allows us to define the stochastic matrix

$$Q_{j|i} = \frac{1}{1-\epsilon}(K_{j|i} - \epsilon P_{j|i}) \text{ or } K_{j|i} = \epsilon P_{j|i} + (1-\epsilon)Q_{j|i}.$$

where $(P_{j|i})_{k\ell} = p_{j|i}(\ell)$ and for any two probability distributions ν_1, ν_2 we have $(\nu_1 - \nu_2)^T P_{j|i} = 0$. Using $\widetilde{\rho}_{\text{mix}} = 1 - \epsilon$ we then obtain by induction, Hoelder's inequality and inequality (2.72)

$$\begin{aligned}
&\|(\nu_{1|i} - \nu'_{1|i})^T K_{1|i} \dots K_{i|i}\|_\infty \\
&\leq \prod_{j=1}^{i}(1-\epsilon)\|(\nu_{1|i} - \nu'_{1|i})^T \otimes_{j=1}^{i} Q_{j|i}\|_2 \\
&\leq \widetilde{\rho}^i_{\text{mix}}\|\nu_{1|i} - \nu'_{1|i}\|_2 \prod_{j=1}^{i}\|Q^T_{j|i}\|_{op} \leq \widetilde{\rho}^i_{\text{mix}}\|\nu_{1|i} - \nu'_{1|i}\|_2 \\
&\leq \widetilde{\rho}^i_{\text{mix}}\left\|\frac{p(x_2^i \mid \cdot)\nu(\cdot)}{\sum_{z_1} p(x_2^i \mid z_1)\nu(z_1)} - \frac{p(x_2^i \mid \cdot)\nu'(\cdot)}{\sum_{z_1} p(x_2^i \mid z_1)\nu'(z_1)}\right\|_2 \\
&\leq \widetilde{\rho}^i_{\text{mix}}\Big[\Big\|\frac{p(x_2^i \mid \cdot)}{\sum_{z_1} p(x_2^i \mid z_1)\nu(z_1)}(\nu(\cdot) - \nu'(\cdot))\Big\|_2 \\
&+ \Big|\sup_{z_1} p(x_2^i \mid z_1)\Big(\frac{1}{\sum_{z_1} p(x_2^i \mid z_1)\nu(z_1)} - \frac{1}{\sum_{z_1} p(x_2^i \mid z_1)\nu'(z_1)}\Big)\Big|\|\nu'(\cdot)\|_1\Big] \\
&\leq \widetilde{\rho}^i_{\text{mix}}\Big(\frac{\sup_{z_1} p(x_2^i \mid z_1)}{\inf_{z_1} p(x_2^i \mid z_1)}\Big)^2\|\nu - \nu'\|_1 \leq \epsilon^{-2}_{\text{mix}}\widetilde{\rho}^i_{\text{mix}}\|\nu - \nu'\|_1,
\end{aligned}$$

since $Q_{j|i}$ are stochastic matrices and $\|\nu\|_2 \leq \|\nu\|_1 \leq 1$ for probability vectors. The second statement is readily derived by substituting $\nu(z_1) = p(z_1)$ and $\nu'(z_1) = p(z_1 \mid x^1_{-n})$. □

Proof of Lemma 3

Recall the shorthand $\widetilde{\rho}_{\text{mix}} = 1 - \epsilon_{\text{mix}}\overline{\pi}_{\text{min}}$. First observe that

$$\begin{aligned}
\sup_{z_i}|p(z_i \mid x_1^n) - p(z_i \mid x_{i-k}^{i+k})| \leq |p(z_i|x_{i+1}^n)p(z_i|x_1^i) - p(z_i|x_{i+1}^{i+k})p(z_i|x_{i-k+1}^i)|\frac{p(x_{i+1}^n)}{p(x_{i+1}^n|x_1^i)p(z_i)} \\
+ |A-1|\frac{p(x_{i+1}^{i+k})}{p(x_{i+1}^{i+k}|x_{i-k+1}^i)}\frac{1}{p(z_i)}
\end{aligned}$$

where $A = \frac{p(x_{i+1}^n)}{p(x_{i+1}^n|x_1^i)}\left(\frac{p(x_{i+1}^{i+k})}{p(x_{i+1}^{i+k}|x_{i-k+1}^i)}\right)^{-1}$. We bound the two terms in the sum separately.

From Lemma 13 we directly obtain the following upper bounds

$$\begin{aligned}
\sup_{z,x}|p(z_i \mid x_1^i) - p(z_i \mid x_{i-k+1}^i)| &\leq \epsilon^{-2}_{\text{mix}}\widetilde{\rho}^{\min\{i,k\}}_{\text{mix}} \\
\sup_{z,x}|p(z_i \mid x_{i+1}^n) - p(z_i \mid x_{i+1}^{i+k})| &\leq \epsilon^{-2}_{\text{mix}}\widetilde{\rho}^{\max\{n-i,k\}}_{\text{mix}}
\end{aligned}$$

where the latter follows because of reversibility assumption (2.2) of the Markov chain. Inequality (2.72) can also be used to show that $\frac{p(x_{i+1}^n)}{p(x_{i+1}^n|x_1^i)}, \frac{p(x_{i+1}^{i+k})}{p(x_{i+1}^{i+k}|x_{i-k+1}^i)} \leq \epsilon_{\text{mix}}^{-2}$. A proof for a similar statement is given after inequality (2.80). The first term of the sum is therefore bounded above by $2\frac{\tilde{\rho}_{\text{mix}}^{\min\{i,n-i,k\}}}{\pi_{\min}\epsilon_{\text{mix}}^4}$.

For the second term, we mainly need to bound $|A-1|$. In order to simplify the notation in the proof, we divide the sequence of values all observed variables in the window $i-k, i+k$ around index i, i.e. $x_{\min\{i-k,1\}}^{\max\{i+k,n\}}$, into four disjoint chunks and call them a, b, c, d in chronological order, explicitly defined as

$$a := x_{\min\{i-k,1\}}^{\max\{i-k,1\}} \quad b := x_{\max\{i-k,1\}+1}^{i} \quad c := x_{i+1}^{\min\{i+k,n\}} \quad d := x_{\min\{i+k,n\}+1}^{\max\{i+k,n\}}.$$

Note that the definition depends on whether $i-k>1$ or $i+k<n$. Depending on the combination of $i+k<n$ and $i-k>1$ being true or false,

$$A = \begin{cases} \frac{p(d|c)p(a|b)}{p(d|a,b,c)p(a|b,c)} & \text{if } i-k>1,\ i+k<n \\ \frac{p(a|b)p(d|a,b,c)}{p(a|b,c,d)p(d|c)} & \text{if } i-k>1,\ i+k>n \end{cases}.$$

For the other two possible cases, A is an inverse of the above. We demonstrate the main argument by looking into these two cases in more detail. Observe that the following inequality holds for the first case

$$|A-1| \leq \frac{|p(d \mid c) - p(d \mid a,b,c)|}{p(d \mid a,b,c)} \frac{p(a \mid b)}{p(a \mid b,c)} + \frac{|p(a \mid b) - p(a \mid b,c)|}{p(a \mid b,c)}$$

holds for all x. For the second case there is only an additional conditioning on d for the second term on the right hand side. In the inverse case that $i-k<1, i+k>n$ we have

$$|A-1| \leq \frac{|p(d \mid c) - p(d \mid a,b,c)|}{p(d \mid c)} \frac{p(a \mid b,c)}{p(a \mid b)} + \frac{|p(a \mid b) - p(a \mid b,c)|}{p(a \mid b)}$$

and equivalently with an additional conditioning on d for $i-k<1, i+k<n$. It is thus sufficient to consider $\sup_x |p(d \mid c) - p(d \mid a,b,c)|$ and $\sup_x |p(a \mid b) - p(a \mid b,c,d)|$. We see later that this is also the critical quantity to bound for the inverses.

First note that

$$\max\left\{\frac{p(a \mid b)}{p(a \mid b,c)}, \frac{p(a \mid b,c)}{p(a \mid b)}\right\} \leq \epsilon_{\text{mix}}^{-2}. \tag{2.80}$$

For the first term we see that for all x we have

$$\begin{aligned} \frac{p(a \mid b)}{p(a \mid b,c)} = \frac{p(x_\alpha^\beta \mid x_{\beta+1}^i)}{p(x_\alpha^\beta \mid x_{\beta+1}^i, x_{i+1}^\gamma)} &= \frac{\sum_{z_{\beta+1}} p(x_\alpha^\beta \mid z_{\beta+1}) p(z_{\beta+1} \mid x_{\beta+1}^i)}{\sum_{z_{\beta+1}} p(x_\alpha^\beta \mid z_{\beta+1}) p(z_{\beta+1} \mid x_{\beta+1}^\gamma)} \\ &\leq \frac{\sup_z p(x_\alpha^\beta \mid z_{\beta+1})}{\inf_z p(x_\alpha^\beta \mid z_{\beta+1})} \leq \epsilon_{\text{mix}}^{-2} \end{aligned}$$

where the second inequality holds because of conditional independence of $x_{\beta+1}^i$ and x_α^β given $z_{\beta+1}$ in an HMM, and the last line holds because of inequality (2.72) and the fact that the Markov chain is invertible. Observe that the same arguments goes through for the inverse as well so that inequality (2.80) holds.

Let us now look at the rest of the terms involving differences. For the sake of simplification, let us introduce the shorthand notation

$$\alpha := \min\{i-k, 1\}, \quad \beta := \max\{i-k, 1\}, \quad \gamma := \min\{i+k, n\}, \quad \text{and} \quad \delta := \max\{i+k, n\}.$$

Then we can write $x_\alpha^\beta = a; x_{\beta+1}^i = b; x_{i+1}^\gamma = c; x_{\gamma+1}^\delta = d$ and bound $|p(d \mid c) - p(d \mid a, b, c)|$. Using Lemma 13 and inequality (2.72), we have the bound

$$\begin{aligned}
\frac{p(d \mid c) - p(d \mid a, b, c)}{p(d \mid a, b, c)} &= \frac{|p(x_{\gamma+1}^\delta | x_{i+1}^\gamma) - p(x_{\gamma+1}^\delta | x_\alpha^i, x_{i+1}^\gamma)|}{p(x_{\gamma+1}^\delta | x_\alpha^i, x_{i+1}^\gamma)} \\
&\leq \frac{\sum_{z_\gamma} p(x_{\gamma+1}^\delta \mid z_\gamma) |p(z_\gamma | x_{i+1}^\gamma) - p(z_\gamma | x_\alpha^i, x_{i+1}^\gamma)|}{\sum_{z_\gamma} p(x_{\gamma+1}^\delta | z_\gamma) p(z_\gamma | x_\alpha^\gamma)} \\
&\leq \frac{\sup_z p(x_{\gamma+1}^\delta | z_\gamma)}{\inf_z p(x_{\gamma+1}^\delta | z_\gamma)} \sum_{z_\gamma} |p(z_\gamma | x_{i+1}^\gamma) - p(z_\gamma | x_\alpha^\gamma)| \\
&\leq Cs\epsilon_{\text{mix}}^{-4} \widetilde{\rho}_{\text{mix}}^{\gamma - i} = Cs\epsilon_{\text{mix}}^{-4} \widetilde{\rho}_{\text{mix}}^{\min\{n-i, k\}}.
\end{aligned}$$

The same argument applies if the denominator is $p(d \mid c)$. Analogously, we have that

$$\frac{|p(a \mid b) - p(a \mid b, c)|}{p(a \mid b, c)} \leq Cs\epsilon_{\text{mix}}^{-4} \widetilde{\rho}_{\text{mix}}^{i-\beta+1} = Cs\epsilon_{\text{mix}}^{-4} \widetilde{\rho}_{\text{mix}}^{\min\{k, i\}}$$

and the same holds for the case when the denominator is $p(a \mid b)$ by inequality (2.80).

Note that the additional conditioning on d, does not change the result. Also, considering the inverses we see that the inequalities still hold.

Putting everything together now yields

$$|A - 1| \leq C' s\epsilon_{\text{mix}}^{-6} \widetilde{\rho}_{\text{mix}}^{\min\{n-i, i, k\}},$$

where C' is a generic constant and thus

$$\sup_{z_i} |p(z_i \mid x_1^n) - p(z_i \mid x_{i-k}^{i+k})| \leq C \frac{s \widetilde{\rho}_{\text{mix}}^{\min\{i, n-i, k\}}}{\overline{\pi}_{\min} \epsilon_{\text{mix}}^8}.$$

Proof of Lemma 6

The latter inequality is valid in our particular case because

$$\begin{aligned}|\operatorname{cov}(z_0, z_\ell \mid x_0, \ldots, x_k)| &= |\sum_{z_0, z_\ell} z_0 z_\ell p(z_\ell \mid z_0, x) p(z_0|x) - \sum_{z_0} z_0 p(z_0|x) \sum_{z_\ell} z_\ell p(z_\ell|x)| \\ &= |\sum_{z_0, z_\ell} z_0 z_\ell p(z_0|x)(p(z_\ell|z_0, x) - p(z_\ell \mid x))| \\ &\leq \sup_{z_\ell, z_0} |p(z_\ell|z_0, x) - p(z_\ell \mid x)| \sum_{z_0} \sum_{z_\ell} |z_0 z_\ell| p(z_0 \mid x)\end{aligned}$$

Let us now show that $\sup_{z_\ell, z_0} \left|p(z_\ell \mid z_0, x) - p(z_\ell \mid x)\right| \leq \rho_{\text{mix}}^\ell$. Introducing the shorthand $\Delta(\ell) = p(z_\ell = 1 \mid z_0 = 1, x) - p(z_\ell = 1 \mid z_0 = -1, x)$, we first claim that

$$|\Delta(1)| \leq \rho_{\text{mix}} \tag{2.81}$$

To establish this fact, note that

$$\begin{aligned}\Delta(1) &= \left| \frac{p(x \mid z_\ell = 1)}{p(x \mid z_{\ell-1} = 1)} p(z_\ell = 1 \mid z_{\ell-1} = 1) - \frac{p(x \mid z_\ell = 1)}{p(x \mid z_{\ell-1} = -1)} p(z_\ell = 1 \mid z_{\ell-1} = -1) \right| \\ &= \frac{ap}{ap + b(1-p)} - \frac{a(1-p)}{a(1-p) + bp} \\ &= \frac{ab}{(ap + b(1-p))(a(1-p) + bp)}(2p - 1)\end{aligned}$$

where we write $a = p(x \mid z_\ell = 1)$ and $b = p(x \mid z_\ell = -1)$. The denominator is minimized at $p = 1$ so that inequality (2.81) is shown. The same argument shows that $|\Delta(-1)| \leq \rho_{\text{mix}}$.

Induction step: Assume that $\Delta(\ell - 1) \leq \rho_{\text{mix}}^{\ell-1}$. It then follows that

$$\begin{aligned}&|p(z_\ell = 1 \mid z_0 = 1, x) - p(z_\ell = 1 \mid z_0 = -1, x)| \\ &= |\sum_{z_{\ell-1}} p(z_\ell = 1 \mid z_{\ell-1}, x) p(z_{\ell-1} \mid z_0 = 1, x) - p(z_\ell = 1 \mid z_{\ell-1}, x) p(z_{\ell-1}|z_0 = -1, x)| \\ &= \Delta(1)\Delta(\ell - 1) \leq \rho_{\text{mix}}^\ell\end{aligned}$$

Since

$$p(z_\ell = 1 \mid z_0 = -1, x) - p(z_\ell = 1 \mid z_0 = 1, x) = -p(z_\ell = -1 \mid z_0 = -1, x) + p(z_\ell = -1 \mid z_0 = 1, x)$$

we use the shorthand $s = p(z_0 = 1 \mid x)$ to obtain

$$\begin{aligned}&\sup_{z_\ell, z_0} |\, p(z_\ell \mid z_0, x) - p(z_\ell \mid x)| \\ &= \sup_{b_\ell, b_0} p(z_\ell = b_\ell \mid z_0 = b_0, x) - [(p(z_\ell = b_\ell \mid z_0 = 1, x)s + p(z_\ell = b_\ell \mid z_0 = -1, x)(1 - s)] \\ &\leq (1 - s)|\Delta(\ell)| \leq \rho_{\text{mix}}\end{aligned}$$

which proves the bound for $\mathrm{cov}(Z_0, Z_1 \mid \gamma)$.

For the two state mixing we define $\widetilde{\Delta}(\ell) = p(z_\ell = 1 \mid z_1 z_0 = 1, x) - p(z_\ell = 1 \mid z_1 z_0 = -1, x)$ and can readily see that $|\widetilde{\Delta}(1)| \leq \rho_{\mathrm{mix}}$ and

$$\begin{aligned} &|p(z_{\ell+1} z_{\ell+2} = 1 \mid z_\ell z_{\ell-1} = 1, x) - p(z_{\ell+1} z_{\ell+2} = 1 \mid z_\ell z_{\ell-1} = -1, x)| \\ &= [p(z_{\ell+2} = 1 \mid z_{\ell+1} = 1, x) - p(z_{\ell+2} = -1 \mid z_{\ell+1} = -1, x)]\tilde{\Delta}(2) \end{aligned}$$

Using equation (2.81), we obtain

$$|\widetilde{\Delta}(2)| = |p(z_1 = 1 \mid z_0 = 1, x) - p(z_1 = -1 \mid z_0 = -1, x)|\widetilde{\Delta}(1) \leq \rho_{\mathrm{mix}} \tag{2.82}$$

from which it directly follows that

$$|p(z_{\ell+1} z_{\ell+2} = 1 \mid z_\ell z_{\ell-1} = 1, x) - p(z_{\ell+1} z_{\ell+2} = 1 \mid z_\ell z_{\ell-1} = -1, x)| \leq \rho_{\mathrm{mix}}$$

The rest follows the same arguments as above and the bound for $\mathrm{cov}(Z_0 Z_1, Z_\ell Z_{\ell+1} \mid \gamma)$ in inequality (2.55) is shown.

Finally, the bound for $\mathrm{cov}(Z_0, Z_\ell Z_{\ell+1} \mid \gamma)$ in inequality (2.55) follows in a straightforward way using the relation (2.82) and induction with equation (2.81), as above.

Chapter 3

Early stopping of kernel boosting algorithms

3.1 Introduction

While non-parametric models offer great flexibility, they can also lead to overfitting, and thus poor generalization performance. For this reason, it is well-understood that procedures for fitting non-parametric models must involve some form of regularization. When models are fit via a form of empirical risk minimization, the most classical form of regularization is based on adding some type of penalty to the objective function. An alternative form of regularization is based on the principle of *early stopping*, in which an iterative algorithm is run for a pre-specified number of steps, and terminated prior to convergence.

While the basic idea of early stopping is fairly old (e.g., [71, 2, 79]), recent years have witnessed renewed interests in its properties, especially in the context of boosting algorithms and neural network training (e.g., [60, 23]). Over the past decade, a line of work has yielded some theoretical insight into early stopping, including works on classification error for boosting algorithms [7, 30, 42, 55, 89, 92], L^2-boosting algorithms for regression [18, 17], and similar gradient algorithms in reproducing kernel Hilbert spaces (e.g. [21, 20, 78, 89, 64]). A number of these papers establish consistency results for particular forms of early stopping, guaranteeing that the procedure outputs a function with statistical error that converges to zero as the sample size increases. On the other hand, there are relatively few results that actually establish *rate optimality* of an early stopping procedure, meaning that the achieved error matches known statistical minimax lower bounds. To the best of our knowledge, Bühlmann and Yu [18] were the first to prove optimality for early stopping of L^2-boosting as applied to spline classes, albeit with a rule that was not computable from the data. Subsequent work by Raskutti et al. [64] refined this analysis of L^2-boosting for kernel classes and first established an important connection to the localized Rademacher complexity; see also the related work [89, 65, 19] with rates for particular kernel classes.

More broadly, relative to our rich and detailed understanding of regularization via penal-

ization (e.g., see the books [34, 74, 73, 81] and papers [6, 48] for details), our understanding of early stopping regularization is not as well developed. Intuitively, early stopping should depend on the same bias-variance tradeoffs that control estimators based on penalization. In particular, for penalized estimators, it is now well-understood that complexity measures such as the *localized Gaussian width*, or its Rademacher analogue, can be used to characterize their achievable rates [6, 48, 73, 81]. Is such a general and sharp characterization also possible in the context of early stopping?

The main contribution of this chapter is to answer this question in the affirmative for the early stopping of boosting algorithms for a certain class of regression and classification problems involving functions in reproducing kernel Hilbert spaces (RKHS). A standard way to obtain a good estimator or classifier is through minimizing some penalized form of loss functions of which the method of kernel ridge regression [80] is a popular choice. Instead, we consider an iterative update involving the kernel that is derived from a greedy update. Borrowing tools from empirical process theory, we are able to characterize the "size" of the effective function space explored by taking T steps, and then to connect the resulting estimation error naturally to the notion of localized Gaussian width defined with respect to this effective function space. This leads to a principled analysis for a broad class of loss functions used in practice, including the loss functions that underlie the L^2-boost, LogitBoost and AdaBoost algorithms, among other procedures.

The remainder of this chapter is organized as follows. In Section 3.2, we provide background on boosting methods and reproducing kernel Hilbert spaces, and then introduce the updates studied in this chapter. Section 3.3 is devoted to statements of our main results, followed by a discussion of their consequences for particular function classes in Section 3.4. We provide simulations that confirm the practical effectiveness of our stopping rules, and show close agreement with our theoretical predictions. In Section 3.5, we provide the proofs of our main results, with certain more technical aspects deferred to the appendices.

3.2 Background and problem formulation

The goal of prediction is to learn a function that maps *covariates* $x \in \mathcal{X}$ to *responses* $y \in \mathcal{Y}$. In a regression problem, the responses are typically real-valued, whereas in a classification problem, the responses take values in a finite set. In this chapter, we study both regression ($\mathcal{Y} = \mathbb{R}$) and classification problems (e.g., $\mathcal{Y} = \{-1, +1\}$ in the binary case). Our primary focus is on the case of *fixed design*, in which we observe a collection of n pairs of the form $\{(x_i, Y_i)\}_{i=1}^n$, where each $x_i \in \mathcal{X}$ is a fixed covariate, whereas $Y_i \in \mathcal{Y}$ is a random response drawn independently from a distribution $\mathbb{P}_{Y|x_i}$ which depends on x_i. Later in the chapter, we also discuss the consequences of our results for the case of random design, where the (X_i, Y_i) pairs are drawn in an i.i.d. fashion from the joint distribution $\mathbb{P} = \mathbb{P}_X \mathbb{P}_{Y|X}$ for some distribution $\mathbb{P}_X$ on the covariates.

In this section, we provide some necessary background on a gradient-type algorithm which is often referred to as *boosting* algorithm. We also discuss briefly about the reproducing

kernel Hilbert spaces before turning to a precise formulation of the problem that is studied in this chapter.

Boosting and early stopping

Consider a cost function $\phi : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, where the non-negative scalar $\phi(y, \theta)$ denotes the cost associated with predicting θ when the true response is y. Some common examples of loss functions ϕ that we consider in later sections include:

- the *least-squares loss* $\phi(y, \theta) := \frac{1}{2}(y - \theta)^2$ that underlies L^2-boosting [18],
- the *logistic regression loss* $\phi(y, \theta) = \ln(1+e^{-y\theta})$ that underlies the LogitBoost algorithm [31, 32], and
- the *exponential loss* $\phi(y, \theta) = \exp(-y\theta)$ that underlies the AdaBoost algorithm [30].

The least-squares loss is typically used for regression problems (e.g., [18, 21, 20, 78, 89, 64]), whereas the latter two losses are frequently used in the setting of binary classification (e.g., [30, 55, 32]).

Given some loss function ϕ, we define the *population cost functional* $f \mapsto \mathcal{L}(f)$ via

$$\mathcal{L}(f) := \mathbb{E}_{Y_1^n}\Big[\frac{1}{n}\sum_{i=1}^{n}\phi\big(Y_i, f(x_i)\big)\Big]. \tag{3.1}$$

Note that with the covariates $\{x_i\}_{i=1}^n$ fixed, the functional $\mathcal{L}$ is a non-random object. Given some function space $\mathscr{F}$, the optimal function* minimizes the population cost functional—that is

$$f^* := \arg\min_{f \in \mathscr{F}} \mathcal{L}(f). \tag{3.2}$$

As a standard example, when we adopt the least-squares loss $\phi(y, \theta) = \frac{1}{2}(y - \theta)^2$, the population minimizer f^* corresponds to the conditional expectation $x \mapsto \mathbb{E}[Y \mid x]$.

Since we do not have access to the population distribution of the responses however, the computation of f^* is impossible. Given our samples $\{Y_i\}_{i=1}^n$, we consider instead some procedure applied to the *empirical loss*

$$\mathcal{L}_n(f) := \frac{1}{n}\sum_{i=1}^{n}\phi(Y_i, f(x_i)), \tag{3.3}$$

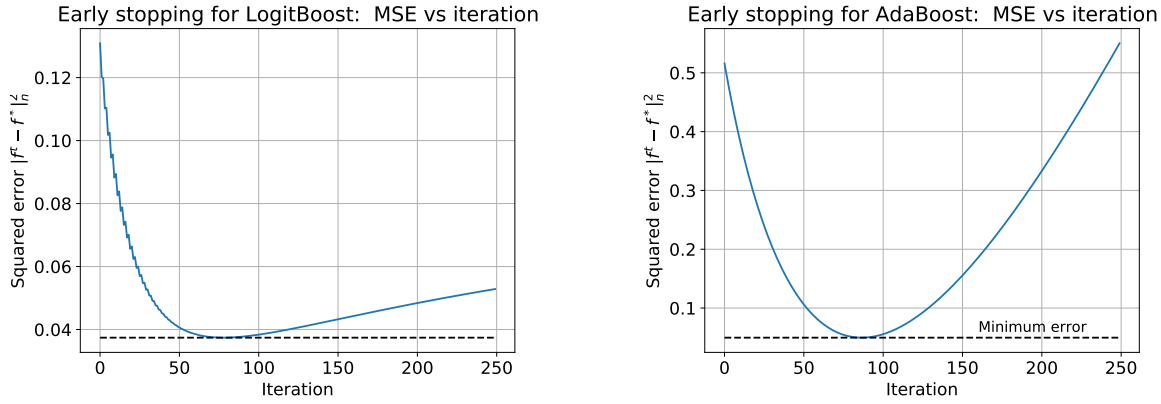
where the population expectation has been replaced by an empirical expectation. For example, when $\mathcal{L}_n$ corresponds to the log likelihood of the samples with $\phi(Y_i, f(x_i)) = \log[\mathbb{P}(Y_i; f(x_i))]$, direct unconstrained minimization of $\mathcal{L}_n$ would yield the maximum likelihood estimator.

It is well-known that direct minimization of $\mathcal{L}_n$ over a sufficiently rich function class $\mathscr{F}$ may lead to overfitting. There are various ways to mitigate this phenomenon, among

*As clarified in the sequel, our assumptions guarantee uniqueness of f^*.

which the most classical method is to minimize the sum of the empirical loss with a penalty regularization term. Adjusting the weight on the regularization term allows for trade-off between fit to the data, and some form of regularity or smoothness in the fit. The behavior of such penalized of regularized estimation methods is now quite well understood (for instance, see the books [34, 74, 73, 81] and papers [6, 48] for more details).

In this chapter, we study a form of *algorithmic regularization*, based on applying a gradient-type algorithm to $\mathcal{L}_n$ but then stopping it "early"—that is, after some fixed number of steps. Such methods are often referred to as *boosting algorithms*, since they involve "boosting" or improve the fit of a function via a sequence of additive updates (see e.g. [66, 30, 16, 15, 67]). Many boosting algorithms, among them AdaBoost [30], L^2-boosting [18] and LogitBoost [31, 32], can be understood as forms of functional gradient methods [55, 32]; see the survey paper [17] for further background on boosting. The way in which the number of steps is chosen is referred to as a stopping rule, and the overall procedure is referred to as *early stopping* of a boosting algorithm.



(a) (b)

Figure 3.1: Plots of the squared error $\|f^t - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f^t(x_i) - f^*(x_i))^2$ versus the iteration number t for (a) LogitBoost using a first-order Sobolev kernel (b) AdaBoost using the same first-order Sobolev kernel $\mathbb{K}(x, x') = 1 + \min(x, x')$ which generates a class of Lipschitz functions (splines of order one). Both plots correspond to a sample size $n = 100$.

In more detail, a broad class of boosting algorithms [55] generate a sequence $\{f^t\}_{t=0}^{\infty}$ via updates of the form

$$f^{t+1} = f^t - \alpha^t g^t \quad \text{with} \quad g^t \propto \arg\max_{\|d\|_{\mathscr{F}} \leq 1} \langle \nabla \mathcal{L}_n(f^t),\, d(x_1^n) \rangle, \tag{3.4}$$

where the scalar $\{\alpha^t\}_{t=0}^{\infty}$ is a sequence of step sizes chosen by the user, the constraint $\|d\|_{\mathscr{F}} \leq 1$ defines the unit ball in a given function class $\mathscr{F}$, $\nabla \mathcal{L}_n(f) \in \mathbb{R}^n$ denotes the gradient taken at the vector $\big(f(x_1), \ldots, f(x_n)\big)$, and $\langle h,\, g \rangle$ is the usual inner product between vectors

$h, g \in \mathbb{R}^n$. For non-decaying step sizes and a convex objective $\mathcal{L}_n$, running this procedure for an infinite number of iterations will lead to a minimizer of the empirical loss, thus causing overfitting. In order to illustrate this phenomenon, Figure 3.1 provides plots of the squared error $\|f^t - f^*\|_n^2 := \frac{1}{n}\sum_{i=1}^n \big(f^t(x_i) - f^*(x_i)\big)^2$ versus the iteration number, for LogitBoost in panel (a) and AdaBoost in panel (b). See Section 3.4 for more details on exactly how these experiments were conducted.

In the plots in Figure 3.1, the dotted line indicates the minimum mean-squared error ρ_n^2 over all iterates of that particular run of the algorithm. Both plots are qualitatively similar, illustrating the existence of a "good" number of iterations to take, after which the MSE greatly increases. Hence a natural problem is to decide at what iteration T to stop such that the iterate f^T satisfies bounds of the form

$$\mathcal{L}(f^T) - \mathcal{L}(f^*) \precsim \rho_n^2 \quad \text{and} \quad \|f^T - f^*\|_n^2 \precsim \rho_n^2 \tag{3.5}$$

with high probability. The main results of this chapter provide a stopping rule T for which bounds of the form (3.5) do in fact hold with high probability over the randomness in the observed responses.

Moreover, as shown by our later results, under suitable regularity conditions, the expectation of the minimum squared error ρ_n^2 is proportional to the *statistical minimax risk* $\inf_{\widehat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}[\mathcal{L}(\widehat{f}) - \mathcal{L}(f)]$, where the infimum is taken over all possible estimators $\widehat{f}$. Note that the minimax risk provides a fundamental lower bound on the performance of any estimator uniformly over the function space $\mathcal{F}$. Coupled with our stopping time guarantee (3.5), we are guaranteed that our estimate achieves the minimax risk up to constant factors. As a result, our bounds are unimprovable in general (see Corollary 4).

Reproducing Kernel Hilbert Spaces

The analysis of this chapter focuses on algorithms with the update (3.4) when the function class $\mathscr{F}$ is a reproducing kernel Hilbert space $\mathscr{H}$ (RKHS, see standard sources [80, 33, 68, 12]), consisting of functions mapping a domain $\mathcal{X}$ to the real line $\mathbb{R}$. Any RKHS is defined by a bivariate symmetric *kernel function* $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which is required to be positive semidefinite, i.e. for any integer $N \geq 1$ and a collection of points $\{x_j\}_{j=1}^N$ in $\mathcal{X}$, the matrix $[\mathbb{K}(x_i, x_j)]_{ij} \in \mathbb{R}^{N \times N}$ is positive semidefinite.

The associated RKHS is the closure of linear span of the form $f(\cdot) = \sum_{j \geq 1} \omega_j \mathbb{K}(\cdot, x_j)$, where $\{x_j\}_{j=1}^\infty$ is some collection of points in $\mathcal{X}$, and $\{\omega_j\}_{j=1}^\infty$ is a real-valued sequence. For two functions $f_1, f_2 \in \mathscr{H}$ which can be expressed as a finite sum $f_1(\cdot) = \sum_{i=1}^{\ell_1} \alpha_i \mathbb{K}(\cdot, x_i)$ and $f_2(\cdot) = \sum_{j=1}^{\ell_2} \beta_j \mathbb{K}(\cdot, x_j)$, the inner product is defined as $\langle f_1, f_2 \rangle_{\mathscr{H}} = \sum_{i=1}^{\ell_1} \sum_{j=1}^{\ell_2} \alpha_i \beta_j \mathbb{K}(x_i, x_j)$ with induced norm $\|f_1\|_{\mathscr{H}}^2 = \sum_{i=1}^{\ell_1} \alpha_i^2 \mathbb{K}(x_i, x_i)$. For each $x \in \mathcal{X}$, the function $\mathbb{K}(\cdot, x)$ belongs to $\mathscr{H}$, and satisfies the reproducing relation

$$\langle f, \mathbb{K}(\cdot, x) \rangle_{\mathscr{H}} = f(x) \quad \text{for all } f \in \mathscr{H}. \tag{3.6}$$

Moreover, when the covariates X_i are drawn i.i.d. from a distribution $\mathbb{P}_X$ with domain $\mathcal{X}$ we can invoke Mercer's theorem which states that any function in $\mathscr{H}$ can be represented as

$$\mathbb{K}(x, x') = \sum_{k=1}^{\infty} \mu_k \phi_k(x) \phi_k(x'), \tag{3.7}$$

where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are the *eigenvalues* of the kernel function $\mathbb{K}$ and $\{\phi_k\}_{k=1}^{\infty}$ are eigenfunctions of $\mathbb{K}$ which form an orthonormal basis of $L^2(\mathcal{X}, \mathbb{P}_X)$ with the inner product $\langle f, g \rangle := \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\mathbb{P}_X(x)$. We refer the reader to the standard sources [80, 33, 68, 12] for more details on RKHSs and their properties.

Throughout this chapter, we assume that the kernel function is uniformly bounded, meaning that there is a constant L such that $\sup_{x \in \mathcal{X}} \mathbb{K}(x, x) \leq L$. Such a boundedness condition holds for many kernels used in practice, including the Gaussian, Laplacian, Sobolev, other types of spline kernels, as well as any trace class kernel with trigonometric eigenfunctions. By rescaling the kernel as necessary, we may assume without loss of generality that $L = 1$. As a consequence, for any function f such that $\|f\|_{\mathscr{H}} \leq r$, we have by the reproducing relation that

$$\|f\|_{\infty} = \sup_x \langle f, \mathbb{K}(\cdot, x) \rangle_{\mathscr{H}} \leq \|f\|_{\mathscr{H}} \sup_x \|\mathbb{K}(\cdot, x)\|_{\mathscr{H}} \leq r.$$

Given samples $\{(x_i, y_i)\}_{i=1}^n$, by the representer theorem [47], it is sufficient to restrict ourselves to the linear subspace $\mathscr{H}_n = \overline{\text{span}}\{\mathbb{K}(\cdot, x_i)\}_{i=1}^n$, for which all $f \in \mathscr{H}_n$ can be expressed as

$$f = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \omega_i \mathbb{K}(\cdot, x_i) \tag{3.8}$$

for some coefficient vector $\omega \in \mathbb{R}^n$. Among those functions which achieve the infimum in expression (3.1), let us define f^* as the one with the minimum Hilbert norm. This definition is equivalent to restricting f^* to be in the linear subspace $\mathscr{H}_n$.

Boosting in kernel spaces

For a finite number of covariates x_i from $i = 1 \ldots n$, let us define the *normalized kernel matrix* $K \in \mathbb{R}^{n \times n}$ with entries $K_{ij} = \mathbb{K}(x_i, x_j)/n$. Since we can restrict the minimization of $\mathcal{L}_n$ and $\mathcal{L}$ from $\mathscr{H}$ to the subspace $\mathscr{H}_n$ w.l.o.g., using expression (3.8) we can then write the function value vectors $f(x_1^n) := (f(x_1), \ldots, f(x_n))$ as $f(x_1^n) = \sqrt{n} K\omega$. As there is a one-to-one correspondence between the n-dimensional vectors $f(x_1^n) \in \mathbb{R}^n$ and the corresponding function $f \in \mathscr{H}_n$ in $\mathscr{H}$ by the representer theorem, minimization of an empirical loss in the subspace $\mathscr{H}_n$ essentially becomes the n-dimensional problem of fitting a response vector y over the set range(K). In the sequel, all updates will thus be performed on the function value vectors $f(x_1^n)$.

With a change of variable $d(x_1^n) = \sqrt{n}\sqrt{K}z$ we then have

$$d^t(x_1^n) := \underset{\substack{\|d\|_{\mathscr{H}} \leq 1 \\ d \in \text{range}(K)}}{\arg\max} \langle \nabla \mathcal{L}_n(f^t),\, d(x_1^n) \rangle = \frac{\sqrt{n} K \nabla \mathcal{L}_n(f^t)}{\sqrt{\nabla \mathcal{L}_n(f^t) K \nabla \mathcal{L}_n(f^t)}}.$$

In this chapter, we study the choice $g^t = \langle \nabla \mathcal{L}_n(f^t),\, d^t(x_1^n) \rangle d^t$ in the boosting update (3.4), so that the function value iterates take the form

$$f^{t+1}(x_1^n) = f^t(x_1^n) - \alpha n K \nabla \mathcal{L}_n(f^t), \tag{3.9}$$

where $\alpha > 0$ is a constant stepsize choice. Choosing $f^0(x_1^n) = 0$ ensures that all iterates $f^t(x_1^n)$ remain in the range space of K.

In this chapter we consider the following three error measures for an estimator $\widehat{f}$:

$$\begin{aligned} L^2(\mathbb{P}_n) \text{ norm:} \quad & \|\widehat{f} - f^*\|_n^2 = \frac{1}{n} \sum_{i=1}^n \big(\widehat{f}(x_i) - f^*(x_i)\big)^2, \\ L^2(\mathbb{P}_X) \text{ norm:} \quad & \|\widehat{f} - f^*\|_2^2 := \mathbb{E}\big(\widehat{f}(X) - f^*(X)\big)^2, \\ \text{Excess risk:} \quad & \mathcal{L}(\widehat{f}) - \mathcal{L}(f^*), \end{aligned}$$

where the expectation in the $L^2(\mathbb{P}_X)$-norm is taken over random covariates X which are independent of the samples (X_i, Y_i) used to form the estimate $\widehat{f}$. Our goal is to propose a stopping time T such that the averaged function $\widehat{f} = \frac{1}{T}\sum_{t=1}^T f^t$ satisfies bounds of the type (3.5). We begin our analysis by focusing on the empirical $L^2(\mathbb{P}_n)$ error, but as we will see in Corollary 3, bounds on the empirical error are easily transformed to bounds on the population $L^2(\mathbb{P}_X)$ error. Importantly, we exhibit such bounds with a statistical error term δ_n that is specified by the *localized Gaussian complexity* of the kernel class.

3.3 Main results

We now turn to the statement of our main results, beginning with the introduction of some regularity assumptions.

Assumptions

Recall from our earlier set-up that we differentiate between the empirical loss function $\mathcal{L}_n$ in expression (3.3), and the population loss $\mathcal{L}$ in expression (3.1). Apart from assuming differentiability of both functions, all of our remaining conditions are imposed on the population loss. Such conditions at the population level are weaker than their analogues at the empirical level.

For a given radius $r > 0$, let us define the Hilbert ball around the optimal function f^* as

$$\mathbb{B}_{\mathscr{H}}(f^*, r) := \{f \in \mathscr{H} \mid \|f - f^*\|_{\mathscr{H}} \leq r\}. \tag{3.10}$$

Our analysis makes particular use of this ball defined for the radius $C_{\mathscr{H}}^2 := 2\max\{\|f^*\|_{\mathscr{H}}^2,\, 32, \sigma^2\}$ where the *effective noise level* is defined by

$$\sigma := \begin{cases} \min\left\{t \mid \max_{i=1,\ldots,n} \mathbb{E}[e^{((Y_i - f^*(x_i))^2/t^2)}] < \infty\right\} & \text{for least squares} \\ 4\,(2M+1)(1+2C_{\mathscr{H}}) & \text{for } \phi'\text{-bounded losses.} \end{cases} \tag{3.11}$$

We assume that the population loss is m-strongly convex and M-smooth over $\mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$, meaning that the sandwich inequality

$$m\text{-}M\textbf{-condition} \quad \frac{m}{2}\|f-g\|_n^2 \;\le\; \mathcal{L}(f) - \mathcal{L}(g) - \langle \nabla\mathcal{L}(g),\, f(x_1^n) - g(x_1^n)\rangle \;\le\; \frac{M}{2}\|f-g\|_n^2$$

holds for all $f, g \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$ and all design points $\{x_i\}_{i=1}^n$. In addition, we assume that the function ϕ is M-Lipschitz in its second argument over the interval $\theta \in [\min_{i\in[n]} f^*(x_i) - 2C_{\mathscr{H}}, \max_{i\in[n]} f^*(x_i) + 2C_{\mathscr{H}}]$. To be clear, here $\nabla\mathcal{L}(g)$ denotes the vector in $\mathbb{R}^n$ obtained by taking the gradient of $\mathcal{L}$ with respect to the vector $g(x_1^n)$. It can be verified by a straightforward computation that when $\mathcal{L}$ is induced by the least-squares cost $\phi(y,\theta) = \frac{1}{2}(y-\theta)^2$, the m-M-condition holds for $m = M = 1$. The logistic and exponential loss satisfy this condition (see supp. material), where it is key that we have imposed the condition *only locally* on the ball $\mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$.

In addition to the least-squares cost, our theory also applies to losses $\mathcal{L}$ induced by scalar functions ϕ that satisfy the following condition:

$$\phi'\textbf{-boundedness} \quad \max_{i=1,\ldots,n} \left|\frac{\partial\phi(y,\theta)}{\partial\theta}\right|_{\theta=f(x_i)} \le B \qquad \text{for all } f \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}}) \text{ and } y \in \mathcal{Y}.$$

This condition holds with $B = 1$ for the logistic loss for all $\mathcal{Y}$, and $B = \exp(2.5C_{\mathscr{H}})$ for the exponential loss for binary classification with $\mathcal{Y} = \{-1, 1\}$, using our kernel boundedness condition. Note that whenever this condition holds with some finite B, we can always rescale the scalar loss ϕ by $1/B$ so that it holds with $B = 1$, and we do so in order to simplify the statement of our results.

Upper bound in terms of localized Gaussian width

Our upper bounds involve a complexity measure known as the localized Gaussian width. In general, Gaussian widths are widely used to obtain risk bounds for least-squares and other types of M-estimators. In our case, we consider Gaussian complexities for "localized" sets of the form

$$\mathcal{E}_n(\delta, 1) := \Big\{f - g \mid f, g \in \mathscr{H}, \quad \|f-g\|_{\mathscr{H}} \le 1, \quad \|f-g\|_n \le \delta\Big\}. \tag{3.12}$$

The Gaussian complexity localized at scale δ is given by

$$\mathcal{G}_n\big(\mathcal{E}_n(\delta,1)\big) := \mathbb{E}\Big[\sup_{g\in\mathcal{E}_n(\delta,1)} \frac{1}{n}\sum_{i=1}^n w_i g(x_i)\Big], \tag{3.13}$$

where $(w_1,\ldots,w_n)$ denotes an i.i.d. sequence of standard Gaussian variables.

An essential quantity in our theory is specified by a certain fixed point equation that is now standard in empirical process theory [73, 6, 48, 64]. Let us define the *effective noise level*

$$\sigma := \begin{cases} \min\Big\{t \mid \max\limits_{i=1,\ldots,n} \mathbb{E}[e^{((Y_i-f^*(x_i))^2/t^2)}] < \infty\Big\} & \text{for least squares} \\ 4\,(2M+1)(1+2C_{\mathscr{H}}) & \text{for } \phi'\text{-bounded losses.} \end{cases} \tag{3.14}$$

The *critical radius* δ_n is the smallest positive scalar such that

$$\frac{\mathcal{G}_n(\mathcal{E}_n(\delta,1))}{\delta} \;\leq\; \frac{\delta}{\sigma}. \tag{3.15}$$

We note that past work on localized Rademacher and Gaussian complexity [56, 6] guarantees that there exists a unique $\delta_n > 0$ that satisfies this condition, so that our definition is sensible.

Upper bounds on excess risk and empirical $L^2(\mathbb{P}_n)$-error

With this set-up, we are now equipped to state our main theorem. It provides high-probability bounds on the excess risk and $L^2(\mathbb{P}_n)$-error of the estimator $\bar{f}^T := \frac{1}{T}\sum_{t=1}^T f^t$ defined by averaging the T iterates of the algorithm. It applies to both the least-squares cost function, and more generally, to any loss function satisfying the m-M-condition and the ϕ'-boundedness condition.

Theorem 3. *Suppose that the sample size n large enough such that $\delta_n \leq \frac{M}{m}$, and we compute the sequence $\{f^t\}_{t=0}^\infty$ using the update (3.9) with initialization $f^0 = 0$ and any step size $\alpha \in (0, \min\{\frac{1}{M}, M\}]$. Then for any iteration $T \in \big\{0,1,\ldots\lfloor\frac{m}{8M\delta_n^2}\rfloor\big\}$, the averaged function estimate $\bar{f}^T$ satisfies the bounds*

$$\mathcal{L}(\bar{f}^T) - \mathcal{L}(f^*) \leq CM\Big(\frac{1}{\alpha m T} + \frac{\delta_n^2}{m^2}\Big), \quad \text{and} \tag{3.16a}$$

$$\|\bar{f}^T - f^*\|_n^2 \leq C\Big(\frac{1}{\alpha m T} + \frac{\delta_n^2}{m^2}\Big), \tag{3.16b}$$

where both inequalities hold with probability at least $1 - c_1\exp(-C_2\frac{m^2 n\delta_n^2}{\sigma^2})$.

We prove Theorem 3 in Section 3.5.

A few comments about the constants in our statement: in all cases, constants of the form c_j are universal, whereas the capital C_j may depend on parameters of the joint distribution

and population loss $\mathcal{L}$. In Theorem 3, we have the explicit value $C_2 = \{\frac{m^2}{\sigma^2}, 1\}$ and C^2 is proportional to the quantity $2\max\{\|f^*\|_{\mathscr{H}}^2,\ 32,\ \sigma^2\}$. While inequalities (3.16a) and (3.16b) are stated as high probability results, similar bounds for expected loss (over the response y_i, with the design fixed) can be obtained by a simple integration argument.

In order to gain intuition for the claims in the theorem, note that apart from factors depending on (m, M), the first term $\frac{1}{\alpha m T}$ dominates the second term $\frac{\delta_n^2}{m^2}$ whenever $T \lesssim 1/\delta_n^2$. Consequently, up to this point, taking further iterations reduces the upper bound on the error. This reduction continues until we have taken of the order $1/\delta_n^2$ many steps, at which point the upper bound is of the order δ_n^2.

More precisely, suppose that we perform the updates with step size $\alpha = \frac{m}{M}$; then, after a total number of $\tau := \frac{1}{\delta_n^2 \max\{8, M\}}$ many iterations, the extension of Theorem 3 to expectations guarantees that the mean squared error is bounded as

$$\mathbb{E}\|\bar{f}^{\tau} - f^*\|_n^2 \leq C' \frac{\delta_n^2}{m^2}, \tag{3.17}$$

where C' is another constant depending on $C_{\mathscr{H}}$. Here we have used the fact that $M \geq m$ in simplifying the expression. It is worth noting that guarantee (3.17) matches the best known upper bounds for kernel ridge regression (KRR)—indeed, this must be the case, since a sharp analysis of KRR is based on the same notion of localized Gaussian complexity (e.g. [5, 6]) . Thus, our results establish a strong parallel between the *algorithmic regularization* of early stopping, and the *penalized regularization* of kernel ridge regression. Moreover, as will be clarified in Section 3.3, under suitable regularity conditions on the RKHS, the critical squared radius δ_n^2 also acts as a lower bound for the expected risk, meaning that our upper bounds are not improvable in general.

Note that the critical radius δ_n^2 only depends on our observations $\{(x_i, y_i)\}_{i=1}^n$ through the solution of inequality (3.15). In many cases, it is possible to compute and/or upper bound this critical radius, so that a concrete and valid stopping rule can indeed by calculated in advance. In Section 3.4, we provide a number of settings in which this can be done in terms of the eigenvalues $\{\mu_j\}_{j=1}^n$ of the normalized kernel matrix.

Consequences for random design regression

Thus far, our analysis has focused purely on the case of fixed design, in which the sequence of covariates $\{x_i\}_{i=1}^n$ is viewed as fixed. If we instead view the covariates as being sampled in an i.i.d. manner from some distribution $\mathbb{P}_X$ over $\mathcal{X}$, then the empirical error $\|\widehat{f} - f^*\|_n^2 = \frac{1}{n}\sum_{i=1}^n \big(f(x_i) - f^*(x_i)\big)^2$ of a given estimate $\widehat{f}$ is a random quantity, and it is interesting to relate it to the squared population $L^2(\mathbb{P}_X)$-norm $\|\widehat{f} - f^*\|_2^2 = \mathbb{E}\big[(\widehat{f}(X) - f^*(X))^2\big]$.

In order to state an upper bound on this error, we introduce a population analogue of the critical radius δ_n, which we denote by $\bar{\delta}_n$. Consider the set

$$\bar{\mathcal{E}}(\delta, 1) := \Big\{ f - g \ \mid \ f, g \in \mathscr{H}, \quad \|f - g\|_{\mathscr{H}} \leq 1, \quad \|f - g\|_2 \leq \delta \Big\}. \tag{3.18}$$

It is analogous to the previously defined set $\mathcal{E}(\delta, 1)$, except that the empirical norm $\|\cdot\|_n$ has been replaced by the population version. The population Gaussian complexity localized at scale δ is given by

$$\bar{\mathcal{G}}_n\big(\bar{\mathcal{E}}(\delta, 1)\big) := \mathbb{E}_{w,X}\Big[\sup_{g \in \bar{\mathcal{E}}(\delta,1)} \frac{1}{n}\sum_{i=1}^{n} w_i g(X_i)\Big], \tag{3.19}$$

where $\{w_i\}_{i=1}^n$ are an i.i.d. sequence of standard normal variates, and $\{X_i\}_{i=1}^n$ is a second i.i.d. sequence, independent of the normal variates, drawn according to $\mathbb{P}_X$. Finally, the population critical radius $\bar{\delta}_n$ is defined by equation (3.19), in which $\mathcal{G}_n$ is replaced by $\bar{\mathcal{G}}_n$.

Corollary 3. *In addition to the conditions of Theorem 3, suppose that the sequence $\{(X_i, Y_i)\}_{i=1}^n$ of covariate-response pairs are drawn i.i.d. from some joint distribution $\mathbb{P}$, and we compute the boosting updates with step size $\alpha = \frac{m}{M}$ and initialization $f^0 = 0$. Then the averaged function estimate $\bar{f}^T$ at time $T := \lfloor \frac{1}{\delta_n^2 \max\{8, M\}} \rfloor$ satisfies the bound*

$$\mathbb{E}_X\big(\bar{f}^T(X) - f^*(X)\big)^2 \;=\; \|\bar{f}^T - f^*\|_2^2 \leq \tilde{c}\, \bar{\delta}_n^2$$

with probability at least $1 - c_1 \exp(-C_2 \frac{m^2 n \delta_n^2}{\sigma^2})$ over the random samples.

The proof of Corollary 3 follows directly from standard empirical process theory bounds [6, 64] on the difference between empirical risk $\|\bar{f}^T - f^*\|_n^2$ and population risk $\|\bar{f}^T - f^*\|_2^2$. In particular, it can be shown that $\|\cdot\|_2$ and $\|\cdot\|_n$ norms differ only by a factor proportion to $\bar{\delta}_n$. Furthermore, one can show that the empirical critical quantity δ_n is bounded by the population $\bar{\delta}_n$. By combining both arguments the corollary follows. We refer the reader to the papers [6, 64] for further details on such equivalences.

It is worth comparing this guarantee with the past work of Raskutti et al. [64], who analyzed the kernel boosting iterates of the form (3.9), but with attention restricted to the special case of the least-squares loss. Their analysis was based on first decomposing the squared error into bias and variance terms, then carefully relating the combination of these terms to a particular bound on the localized Gaussian complexity (see equation (3.20) below). In contrast, our theory more directly analyzes the effective function class that is explored by taking T steps, so that the localized Gaussian width (3.19) appears more naturally. In addition, our analysis applies to a broader class of loss functions.

In the case of reproducing kernel Hilbert spaces, it is possible to sandwich the localized Gaussian complexity by a function of the eigenvalues of the kernel matrix. Mendelson [56] provides this argument in the case of the localized Rademacher complexity, but similar arguments apply to the localized Gaussian complexity. Letting $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ denote the ordered eigenvalues of the normalized kernel matrix K, define the function

$$\mathcal{R}(\delta) = \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{n} \min\{\delta^2, \mu_j\}}. \tag{3.20}$$

Up to a universal constant, this function is an upper bound on the Gaussian width $\mathcal{G}_n\big(\mathcal{E}(\delta, 1)\big)$ for all $\delta \geq 0$, and up to another universal constant, it is also a lower bound for all $\delta \geq \frac{1}{\sqrt{n}}$.

Achieving minimax lower bounds

In this section, we show that the upper bound (3.17) matches known minimax lower bounds on the error, so that our results are unimprovable in general. We establish this result for the class of *regular kernels*, as previously defined by Yang et al. [88], which includes the Gaussian and Sobolev kernels as special cases.

The class of regular kernels is defined as follows. Let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ denote the ordered eigenvalues of the normalized kernel matrix K, and define the quantity $d_n := \operatorname{argmin}_{j=1,\ldots,n}\{\mu_j \leq \delta_n^2\}$. A kernel is called *regular* whenever there is a universal constant c such that the tail sum satisfies $\sum_{j=d_n+1}^{n} \mu_j \leq c\, d_n \delta_n^2$. In words, the tail sum of the eigenvalues for regular kernels is roughly on the same or smaller scale as the sum of the eigenvalues bigger than δ_n^2.

For such kernels and under the Gaussian observation model ($Y_i \sim N(f^*(x_i), \sigma^2)$), Yang et al. [88] prove a minimax lower bound involving δ_n. In particular, they show that the minimax risk over the unit ball of the Hilbert space is lower bounded as

$$\inf_{\widehat{f}} \sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \geq c_\ell \delta_n^2. \tag{3.21}$$

Comparing the lower bound (3.21) with upper bound (3.17) for our estimator $\bar{f}^T$ stopped after $O(1/\delta_n^2)$ many steps, it follows that the bounds proven in Theorem 3 are unimprovable apart from constant factors.

We now state a generalization of this minimax lower bound, one which applies to a sub-class of *generalized linear models*, or GLM for short. In these models, the conditional distribution of the observed vector $Y = (Y_1, \ldots, Y_n)$ given $\big(f^*(x_1), \ldots, f^*(x_n)\big)$ takes the form

$$\mathbb{P}_\theta(y) = \prod_{i=1}^{n} \Big[h(y_i) \exp\big(\frac{y_i f^*(x_i) - \Phi(f^*(x_i))}{s(\sigma)}\big)\Big], \tag{3.22}$$

where $s(\sigma)$ is a known scale factor and $\Phi : \mathbb{R} \to \mathbb{R}$ is the cumulant function of the generalized linear model. As some concrete examples:

- The linear Gaussian model is recovered by setting $s(\sigma) = \sigma^2$ and $\Phi(t) = t^2/2$.
- The logistic model for binary responses $y \in \{-1, 1\}$ is recovered by setting $s(\sigma) = 1$ and $\Phi(t) = \log(1 + \exp(t))$.

Our minimax lower bound applies to the class of GLMs for which the cumulant function Φ is differentiable and has uniformly bounded second derivative $|\Phi''| \leq L$. This class includes the linear, logistic, multinomial families, among others, but excludes (for instance) the Poisson family. Under this condition, we have the following:

Corollary 4. *Suppose that we are given i.i.d. samples* $\{y_i\}_{i=1}^n$ *from a GLM* (3.22) *for some function* f^* *in a regular kernel class with* $\|f^*\|_{\mathscr{H}} \leq 1$. *Then running* $T := \lfloor \frac{1}{\delta_n^2 \max\{8,M\}} \rfloor$ *iterations with step size* $\alpha = \frac{m}{M}$ *and* $f^0 = 0$ *yields an estimate* $\bar{f}^T$ *such that*

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \asymp \inf_{\widehat{f}} \sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2. \tag{3.23}$$

As always, in the minimax claim (3.23), the infimum is taken over all measurable functions of the input data and the expectation is taken over the randomness of the response variables $\{Y_i\}_{i=1}^n$. Since we know that $\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \precsim \delta_n^2$, an equivalent way to interpret the the bound (3.23) is that it asserts that $\inf_{\widehat{f}} \sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \succsim \delta_n^2$. See Section 3.5 for the proof of this result.

At a high level, the statement in Corollary 4 shows that early stopping prevents us from overfitting to the data; in particular, using the stopping time T yields an estimate that attains the optimal balance between bias and variance.

3.4 Consequences for various kernel classes

In this section, we apply Theorem 3 to derive some concrete rates for different kernel spaces and then illustrate them with some numerical experiments. It is known that the complexity of an RKHS in association with a distribution over the covariates $\mathbb{P}_X$ can be characterized by the decay rate (3.7) of the eigenvalues of the kernel function. In the finite sample setting, the analogous quantities are the eigenvalues $\{\mu_j\}_{j=1}^n$ of the normalized kernel matrix K. The representation power of a kernel class is directly correlated with the eigen-decay: the faster the decay, the smaller the function class. When the covariates are drawn from the distribution $\mathbb{P}_X$, empirical process theory guarantees that the empirical and population eigenvalues are close.

Theoretical predictions as a function of decay

In this section, let us consider two broad types of eigen-decay:

- **γ-exponential decay**: For some $\gamma > 0$, the kernel matrix eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 \exp(-c_2 j^\gamma)$, where c_1, c_2 are universal constants. Examples of kernels in this class include the Gaussian kernel, which for the Lebesgue measure satisfies such a bound with $\gamma = 2$ (real line) or $\gamma = 1$ (compact domain).
- **β-polynomial decay**: For some $\beta > 1/2$, the kernel matrix eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 j^{-2\beta}$, where c_1 is a universal constant. Examples of kernels in this class include the k^{th}-order Sobolev spaces for some fixed integer $k \geq 1$ with Lebesgue measure on a bounded domain. We consider Sobolev spaces that consist of functions that have k^{th}-order weak derivatives $f^{(k)}$ being Lebesgue integrable and $f(0) = f^{(1)}(0) = \cdots = f^{(k-1)}(0) = 0$. For such classes, the β-polynomial decay condition holds with $\beta = k$.

Given eigendecay conditions of these types, it is possible to compute an upper bound on the critical radius δ_n. In particular, using the fact that the function $\mathcal{R}$ from equation (3.20) is an upper bound on the function $\mathcal{G}_n\big(\mathcal{E}(\delta, 1)\big)$, we can show that for γ-exponentially decaying kernels, we have $\delta_n^2 \precsim \frac{(\log n)^{1/\gamma}}{n}$, whereas for β-polynomial kernels, we have $\delta_n^2 \precsim n^{-\frac{2\beta}{2\beta+1}}$ up to universal constants. Combining with our Theorem 3, we obtain the following result:

Corollary 5 (Bounds based on eigendecay)**.** *Under the conditions of Theorem 3:*

(a) For kernels with γ-exponential eigen-decay, we

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \leq c\,\frac{\log^{1/\gamma} n}{n} \qquad \text{at} \quad T \asymp \tfrac{n}{\log^{1/\gamma} n} \text{ steps.} \tag{3.24a}$$

(b) For kernels with β-polynomial eigen-decay, we have

$$\mathbb{E}\|\bar{f}^T - f^*\|_n^2 \precsim n^{-2\beta/(2\beta+1)} \qquad \text{at} \quad T \asymp n^{2\beta/(2\beta+1)} \text{ steps..} \tag{3.24b}$$

See Section 3.5 for the proof of Corollary 5.

In particular, these bounds hold for LogitBoost and AdaBoost. We note that similar bounds can also be derived with regard to risk in $L^2(\mathbb{P}_n)$ norm as well as the excess risk $\mathcal{L}(f^T) - \mathcal{L}(f^*)$.

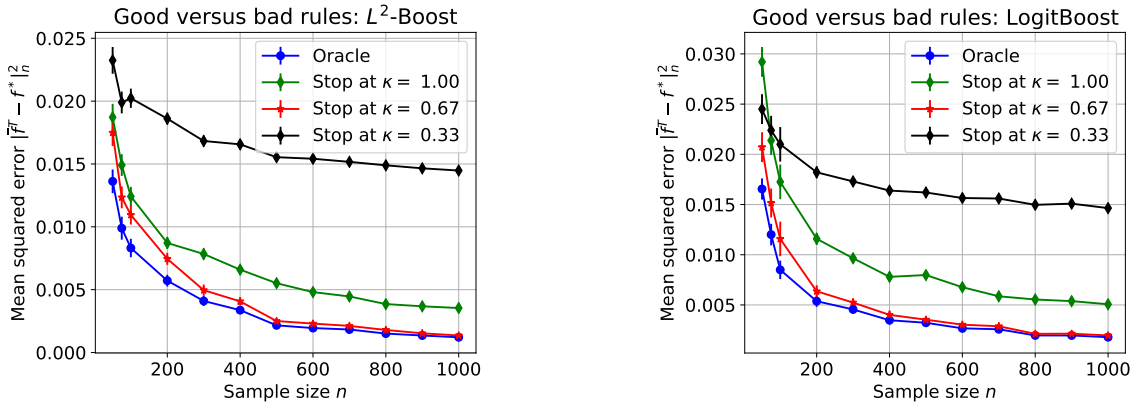
To the best of our knowledge, this result is the first to show non-asymptotic and optimal statistical rates for the $\|\cdot\|_n^2$-error when using early stopping LogitBoost or AdaBoost with an explicit dependence of the stopping rule on n. Our results also yield similar guarantees for L^2-boosting, as has been established in past work [64]. Note that we can observe a similar trade-off between computational efficiency and statistical accuracy as in the case of kernel least-squares regression [89, 64]: although larger kernel classes (e.g. Sobolev classes) yield higher estimation errors, boosting updates reach the optimum faster than for a smaller kernel class (e.g. Gaussian kernels).

Numerical experiments

We now describe some numerical experiments that provide illustrative confirmations of our theoretical predictions. While we have applied our methods to various kernel classes, in this section, we present numerical results for the first-order Sobolev kernel as two typical examples for exponential and polynomial eigen-decay kernel classes.

Let us start with the first-order Sobolev space of Lipschitz functions on the unit interval $[0, 1]$, defined by the kernel $\mathbb{K}(x, x') = 1 + \min(x, x')$, and with the design points $\{x_i\}_{i=1}^n$ set equidistantly over $[0, 1]$. Note that the equidistant design yields β-polynomial decay of the eigenvalues of K with $\beta = 1$ as in the case when x_i are drawn i.i.d. from the uniform measure on $[0, 1]$. Consequently we have that $\delta_n^2 \asymp n^{-2/3}$. Accordingly, our theory predicts that the stopping time $T = (cn)^{2/3}$ should lead to an estimate $\bar{f}^T$ such that $\|\bar{f}^T - f^*\|_n^2 \precsim n^{-2/3}$.

In our experiments for L^2-Boost, we sampled Y_i according to $Y_i = f^*(x_i) + w_i$ with $w_i \sim \mathcal{N}(0, 0.5)$, which corresponds to the probability distribution $\mathbb{P}(Y \mid x_i) = \mathcal{N}(f^*(x_i); 0.5)$, where $f^*(x) = |x - \frac{1}{2}| - \frac{1}{4}$ is defined on the unit interval $[0, 1]$. By construction, the function f^* belongs to the first-order Sobolev space with $\|f^*\|_{\mathscr{H}} = 1$. For LogitBoost, we sampled Y_i according to $\text{Bin}(p(x_i), 5)$ where $p(x) = \frac{\exp(f^*(x))}{1+\exp(f^*(x))}$. In all cases, we fixed the initialization $f^0 = 0$, and ran the updates (3.9) for L^2-Boost and LogitBoost with the constant step size $\alpha = 0.75$. We compared various stopping rules to the *oracle gold standard* G, meaning the procedure that examines all iterates $\{f^t\}$, and chooses the stopping time $G = \arg\min_{t \geq 1} \|f^t - f^*\|_n^2$ that yields the minimum prediction error. Of course, this procedure is unimplementable in practice, but it serves as a convenient lower bound with which to compare.



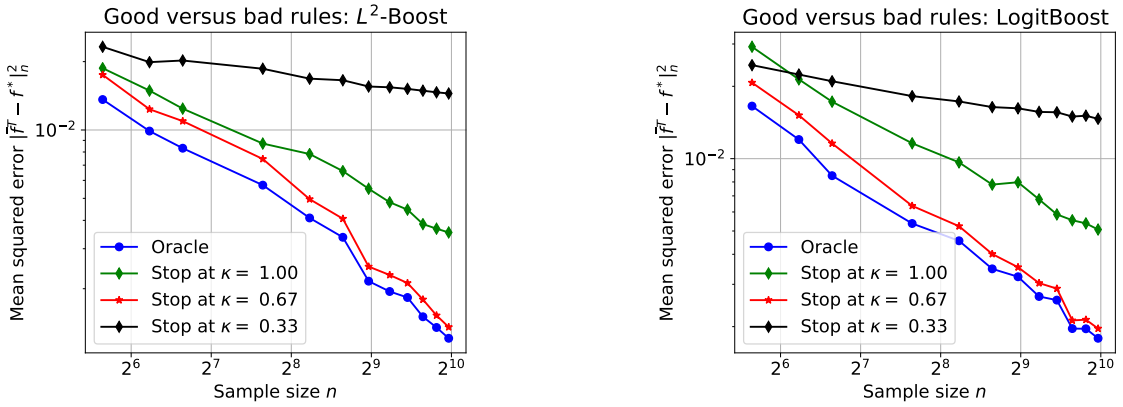
(a) (b)

Figure 3.2: The mean-squared errors for the stopped iterates $\bar{f}^T$ at the Gold standard, i.e. iterate with the minimum error among all unstopped updates (blue) and at $T = (7n)^\kappa$ (with the theoretically optimal $\kappa = 0.67$ in red, $\kappa = 0.33$ in black and $\kappa = 1$ in green) for (a) L^2-Boost and (b) LogitBoost.

Figure 3.2 shows plots of the mean-squared error $\|\bar{f}^T - f^*\|_n^2$ over the sample size n averaged over 40 trials, for the gold standard $T = G$ and stopping rules based on $T = (7n)^\kappa$ for different choices of κ. Error bars correspond to the standard errors computed from our simulations. Panel (a) shows the behavior for L^2-boosting, whereas panel (b) shows the behavior for LogitBoost.

Note that both plots are qualitatively similar and that the theoretically derived stopping rule $T = (7n)^\kappa$ with $\kappa^* = 2/3 = 0.67$, while slightly worse than the Gold standard, tracks its performance closely. We also performed simulations for some "bad" stopping rules, in particular for an exponent κ *not equal* to $\kappa^* = 2/3$, indicated by the green and black curves. In the log scale plots in Figure 3.3 we can clearly see that for $\kappa \in \{0.33, 1\}$ the performance is indeed much worse, with the difference in slope even suggesting a different scaling of the error

with the number of observations n. Recalling our discussion for Figure 3.1, this phenomenon likely occurs due to underfitting and overfitting effects. These qualitative shifts are consistent with our theory.



(a) (b)

Figure 3.3: Logarithmic plots of the mean-squared errors at the Gold standard in blue and at $T = (7n)^\kappa$ (with the theoretically optimal rule for $\kappa = 0.67$ in red, $\kappa = 0.33$ in black and $\kappa = 1$ in green) for (a) L^2-Boost and (b) LogitBoost.

3.5 Proof of main results

In this section, we present the proofs of our main results. The technical details are deferred to section 3.7.

In the following, recalling the discussion in Section 3.2, we denote the vector of function values of a function $f \in \mathscr{H}$ evaluated at $(x_1, x_2, \ldots, x_n)$ as $\theta_f := f(x_1^n) = (f(x_1), f(x_2), \ldots f(x_n)) \in \mathbb{R}^n$, where we omit the subscript f when it is clear from the context. As mentioned in the main text, updates on the function value vectors $\theta^t \in \mathbb{R}^n$ correspond uniquely to updates of the functions $f^t \in \mathscr{H}$. In the following we repeatedly abuse notation by defining the Hilbert norm and empirical norm on vectors in $\Delta \in \text{range}(K)$ as

$$\|\Delta\|_{\mathscr{H}}^2 = \frac{1}{n}\Delta^T K^\dagger \Delta \quad \text{and} \quad \|\Delta\|_n^2 = \frac{1}{n}\|\Delta\|_2^2,$$

where $K^\dagger$ is the pseudoinverse of K. We also use $\mathbb{B}_{\mathscr{H}}(\theta, r)$ to denote the ball with respect to the $\|\cdot\|_{\mathscr{H}}$-norm in range(K).

Proof of Theorem 3

The proof of our main theorem is based on a sequence of lemmas, all of which are stated with the assumptions of Theorem 3 in force. The first lemma establishes a bound on the

empirical norm $\|\cdot\|_n$ of the error $\Delta^{t+1} := \theta^{t+1} - \theta^*$, provided that its Hilbert norm is suitably controlled.

Lemma 14. *For any stepsize $\alpha \in (0, \frac{1}{M}]$ and any iteration t we have*

$$\frac{m}{2}\|\Delta^{t+1}\|_n^2 \leq \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} + \langle \nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1}\rangle. \quad (3.25)$$

See Section 3.7 for the proof of this claim.

The second term on the right-hand side of the bound (3.25) involves the difference between the population and empirical gradient operators. Since this difference is being evaluated at the random points Δ^t and Δ^{t+1}, the following lemma establishes a form of uniform control on this term.

Let us define the set

$$\mathbb{S} := \Big\{\Delta, \widetilde{\Delta} \in \mathbb{R}^n \;\mid\; \|\Delta\|_{\mathscr{H}} \geq 1, \quad \text{and} \quad \theta^\star + \Delta,\ \theta^\star + \widetilde{\Delta} \in \mathbb{B}_{\mathscr{H}}(\theta^*, 2C_{\mathscr{H}})\Big\}, \quad (3.26)$$

and consider the uniform bound

$$\langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\, \Delta\rangle \leq 2\delta_n\|\Delta\|_n + 2\delta_n^2\|\Delta\|_{\mathscr{H}} + \frac{m}{c_3}\|\Delta\|_n^2 \quad \text{for all } \Delta, \widetilde{\Delta} \in \mathbb{S}. \quad (3.27)$$

Lemma 15. *Let $\mathcal{E}$ be the event that bound* (3.27) *holds. There are universal constants (c_1, c_2) such that $\mathbb{P}[\mathcal{E}] \geq 1 - c_1\exp(-c_2\frac{m^2 n\delta_n^2}{\sigma^2})$.*

See Section 3.7 for the proof of Lemma 15.

Note that Lemma 14 applies only to error iterates with a bounded Hilbert norm. Our last lemma provides this control for some number of iterations:

Lemma 16. *There are constants (C_1, C_2) independent of n such that for any step size $\alpha \in \big(0, \min\{M, \frac{1}{M}\}\big]$, we have*

$$\|\Delta^t\|_{\mathscr{H}} \leq C_{\mathscr{H}} \qquad \text{for all iterations } t \leq \tfrac{m}{8M\delta_n^2} \quad (3.28)$$

with probability at least $1 - C_1\exp(-C_2 n\delta_n^2)$, where $C_2 = \max\{\frac{m^2}{\sigma^2}, 1\}$.

See Section 3.7 for the proof of this lemma which also uses Lemma 15.

Taking these lemmas as given, we now complete the proof of the theorem. We first condition on the event $\mathcal{E}$ from Lemma 15, so that we may apply the bound (3.27). We then fix some iterate t such that $t < \frac{m}{8M\delta_n^2} - 1$, and condition on the event that the bound (3.28) in Lemma 16 holds, so that we are guaranteed that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq C_{\mathscr{H}}$. We then split the analysis into two cases:

Case 1: First, suppose that $\|\Delta^{t+1}\|_n \leq \delta_n C_{\mathscr{H}}$. In this case, inequality (3.16b) holds directly.

Case 2: Otherwise, we may assume that $\|\Delta^{t+1}\|_n > \delta_n \|\Delta^{t+1}\|_{\mathscr{H}}$. Applying the bound (3.27) with the choice $(\widetilde{\Delta}, \Delta) = (\Delta^t, \Delta^{t+1})$ yields

$$\langle \nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1} \rangle \leq 4\delta_n \|\Delta^{t+1}\|_n + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2. \tag{3.29}$$

Substituting inequality (3.29) back into equation (3.25) yields

$$\frac{m}{2}\|\Delta^{t+1}\|_n^2 \leq \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} + 4\delta_n\|\Delta^{t+1}\|_n + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2.$$

Re-arranging terms yields the bound

$$\gamma m \|\Delta^{t+1}\|_n^2 \leq D^t + 4\delta_n \|\Delta^{t+1}\|_n, \tag{3.30}$$

where we have introduced the shorthand notation $D^t := \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\}$, as well as $\gamma = \frac{1}{2} - \frac{1}{c_3}$

Equation (3.30) defines a quadratic inequality with respect to $\|\Delta^{t+1}\|_n$; solving it and making use of the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ yields the bound

$$\|\Delta^{t+1}\|_n^2 \leq \frac{c\delta_n^2}{\gamma^2 m^2} + \frac{2D^t}{\gamma m}, \tag{3.31}$$

for some universal constant c. By telescoping inequality (3.31), we find that

$$\frac{1}{T}\sum_{t=1}^{T}\|\Delta^t\|_n^2 \leq \frac{c\delta_n^2}{\gamma^2 m^2} + \frac{1}{T}\sum_{t=1}^{T}\frac{2D^t}{\gamma m} \tag{3.32}$$

$$\leq \frac{c\delta_n^2}{\gamma^2 m^2} + \frac{1}{\alpha\gamma m T}[\|\Delta^0\|_{\mathscr{H}}^2 - \|\Delta^T\|_{\mathscr{H}}^2]. \tag{3.33}$$

By Jensen's inequality, we have

$$\|\bar{f}^T - f^*\|_n^2 = \|\frac{1}{T}\sum_{t=1}^{T}\Delta^t\|_n^2 \;\leq\; \frac{1}{T}\sum_{t=1}^{T}\|\Delta^t\|_n^2,$$

so that inequality (3.16b) follows from the bound (3.32).

On the other hand, by the smoothness assumption, we have

$$\mathcal{L}(\bar{f}^T) - \mathcal{L}(f^*) \leq \frac{M}{2}\|\bar{f}^T - f^*\|_n^2,$$

from which inequality (3.16a) follows.

Proof of Corollary 4

Similar to the proof of Theorem 1 in Yang et al. [88], a generalization can be shown using a standard argument of Fanos inequality. By definition of the transformed parameter $\theta = DU\alpha$ with $K = U^T DU$, we have for any estimator $\widehat{f} = \sqrt{n}U^T\theta$ that $\|\widehat{f} - f^*\|_n^2 = \|\theta - \theta^\star\|_2^2$. Therefore our goal is to lower bound the Euclidean error $\|\theta - \theta^\star\|_2$ of any estimator of $\theta^\star$. Borrowing Lemma 4 in Yang et al. [88], there exists $\delta/2$-packing of the set $B = \{\theta \in \mathbb{R}^n \mid \|D^{-1/2}\theta\|_2 \leq 1\}$ of cardinality $M = e^{d_n/64}$ with $d_n := \arg\min_{j=1,\ldots,n}\{\mu_j \leq \delta_n^2\}$. This is done through packing the following subset of B

$$\mathcal{E}(\delta) := \Big\{\theta \in \mathbb{R}^n \;\; | \;\; \sum_{j=1}^{n} \frac{\theta_j^2}{\min\{\delta^2, \mu_j\}} \leq 1\Big\}.$$

Let us denote the packing set by $\{\theta^1, \ldots, \theta^M\}$. Since $\theta \in \mathcal{E}(\delta)$, by simple calculation, we have $\|\theta^i\|_2 \leq \delta$.

By considering the random ensemble of regression problem in which we first draw at index Z at random from the index set $[M]$ and then condition on $Z = z$, we observe n i.i.d samples $y_1^n := \{y_1, \ldots, y_n\}$ from $\mathbb{P}_{\theta^z}$, Fano's inequality implies that

$$\mathbb{P}(\|\widehat{\theta} - \theta^\star\|_2 \geq \frac{\delta^2}{4}) \geq 1 - \frac{I(y_1^n; Z) + \log 2}{\log M}.$$

where $I(y_1^n; Z)$ is the mutual information between the samples Y and the random index Z.

So it is only left for us to control the mutual information $I(y_1^n; Z)$. Using the mixture representation, $\bar{\mathbb{P}} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{P}_{\theta^i}$ and the convexity of the KullbackLeibler divergence, we have

$$I(y_1^n; Z) = \frac{1}{M}\sum_{j=1}^{M} \|\mathbb{P}_{\theta^j},\ \bar{\mathbb{P}}\|_{\mathrm{KL}} \leq \frac{1}{M^2}\sum_{i,j} \|\mathbb{P}_{\theta^i},\ \mathbb{P}_{\theta^j}\|_{\mathrm{KL}}.$$

We now claim that

$$\|\mathbb{P}_\theta(y),\ \mathbb{P}_{\theta'}(y)\|_{\mathrm{KL}} \leq \frac{nL\|\theta - \theta'\|_2^2}{s(\sigma)}. \tag{3.34}$$

Since each $\|\theta^i\|_2 \leq \delta$, triangle inequality yields $\|\theta_i - \theta_j\|_2 \leq 2\delta$ for all $i \neq j$. It is therefore guaranteed that

$$I(y_1^n; Z) \leq \frac{4nL\delta^2}{s(\sigma)}.$$

Therefore, similar to Yang et al. [88], following by the fact that the kernel is regular and hence $s(\sigma)d_n \geq cn\delta_n^2$, any estimator $\widehat{f}$ has prediction error lower bounded as

$$\sup_{\|f^*\|_{\mathscr{H}} \leq 1} \mathbb{E}\|\widehat{f} - f^*\|_n^2 \geq c_l\delta_n^2.$$

Corollary 4 thus follows using the upper bound in Theorem 3.

Proof of inequality (3.34)**:** Direct calculations of the KL-divergence yield

$$\begin{aligned}\|\mathbb{P}_\theta(y),\ \mathbb{P}_{\theta'}(y)\|_{\mathrm{KL}} &= \int \log(\frac{\mathbb{P}_\theta(y)}{\mathbb{P}_{\theta'}(y)})\mathbb{P}_\theta(y)dy \\ &= \frac{1}{s(\sigma)}\int \sum_{i=1}^n \Big[\sqrt{n}y_i\langle u_i,\, \theta-\theta'\rangle + \Phi(\sqrt{n}\langle u_i,\, \theta'\rangle) - \Phi(\sqrt{n}\langle u_i,\, \theta\rangle)\Big]\mathbb{P}_\theta dy \\ &= \sum_{i=1}^n \frac{\Phi(\sqrt{n}\langle u_i,\, \theta'\rangle) - \Phi(\sqrt{n}\langle u_i,\, \theta\rangle)}{s(\sigma)} + \frac{\sqrt{n}}{s(\sigma)}\int \sum_{i=1}^n \big[y_i\langle u_i,\, \theta-\theta'\rangle\big]\mathbb{P}_\theta dy.\end{aligned} \tag{3.35}$$

To further control the right hand side of expression (3.35), we concentrate on expressing $\int \sum_{i=1}^n y_i u_i \mathbb{P}_\theta dy$ differently. Leibniz's rule allow us to inter-change the order of integral and derivative, so that

$$\int \frac{dP_\theta}{d\theta}dy = \frac{d}{d\theta}\int P_\theta dy = 0. \tag{3.36}$$

Observe that

$$\int \frac{dP_\theta}{d\theta}dy = \frac{\sqrt{n}}{s(\sigma)}\int P_\theta \cdot \sum_{i=1}^n u_i\big(y_i - \Phi'(\sqrt{n}\langle u_i,\, \theta'\rangle)\big)dy$$

so that equality (3.36) yields

$$\int \sum_{i=1}^n y_i u_i \mathbb{P}_\theta dy = \sum_{i=1}^n u_i \Phi'(\sqrt{n}\langle u_i,\, \theta\rangle).$$

Combining the above inequality with expression (3.35), the KL divergence between two generalized linear models $\mathbb{P}_\theta, \mathbb{P}_{\theta'}$ can thus be written as

$$\|\mathbb{P}_\theta(y),\ \mathbb{P}_{\theta'}(y)\|_{\mathrm{KL}} = \frac{1}{s(\sigma)}\sum_{i=1}^n \Phi(\sqrt{n}\langle u_i,\, \theta'\rangle) - \Phi(\sqrt{n}\langle u_i,\, \theta\rangle) - \sqrt{n}\langle u_i,\, \theta'-\theta\rangle\Phi'(\sqrt{n}\langle u_i,\, \theta\rangle). \tag{3.37}$$

Together with the fact that

$$|\Phi(\sqrt{n}\langle u_i,\, \theta'\rangle) - \Phi(\sqrt{n}\langle u_i,\, \theta\rangle) - \sqrt{n}\langle u_i,\, \theta'-\theta\rangle\Phi'(\sqrt{n}\langle u_i,\, \theta\rangle)| \leq nL\|\theta-\theta'\|_2^2.$$

which follows by assumption on Φ having a uniformly bounded second derivative. Putting the above inequality with inequality (3.37) establishes our claim (3.34).

Proof of Corollary 5

The general statement follows directly from Theorem 3. In order to invoke Theorem 3 for the particular cases of LogitBoost and AdaBoost, we need to verify the conditions, i.e. that the m-M-condition and ϕ'-boundedness conditions hold for the respective loss function over the ball $\mathbb{B}_{\mathscr{H}}(\theta^*, 2C_{\mathscr{H}})$. The following lemma provides such a guarantee:

Lemma 17. *With $D := C_{\mathscr{H}} + \|\theta^*\|_{\mathscr{H}}$, the logistic regression cost function satisfies the m-M-condition with parameters*

$$m = \frac{1}{e^{-D} + e^{D} + 2}, \quad M = \frac{1}{4}, \quad \textit{and} \quad B = 1.$$

The AdaBoost cost function satisfies the m-M-condition with parameters

$$m = \mathrm{e}^{-D}, \quad M = \mathrm{e}^{D}, \quad \textit{and} \quad B = \mathrm{e}^{D}.$$

See Section 3.7 for the proof of Lemma 17.

γ-exponential decay: If the kernel eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 \exp(-c_2 j^{\gamma})$, where c_1, c_2 are universal constants, the function $\mathcal{R}$ from equation (3.20) can be upper bounded as

$$\mathcal{R}(\delta) = \sqrt{\frac{2}{n}} \sqrt{\sum_{i=1}^{n} \min\{\delta^2, \mu_j\}} \leq \sqrt{\frac{2}{n}} \sqrt{k\delta^2 + \sum_{j=k+1}^{n} c_1 e^{-c_2 j^2}},$$

where k is the smallest integer such that $c_1 \exp(-c_2 k^{\gamma}) < \delta^2$. Since the localized Gaussian width $\mathcal{G}_n\big(\mathcal{E}_n(\delta, 1)\big)$ can be sandwiched above and below by multiples of $\mathcal{R}(\delta)$, some algebra shows that the critical radius scales as $\delta_n^2 \asymp \frac{n}{\log(n)^{1/\gamma}\sigma^2}$.

Consequently, if we take $T \asymp \frac{\log(n)^{1/\gamma}\sigma^2}{n}$ steps, then Theorem 3 guarantees that the averaged estimator $\bar{\theta}^T$ satisfies the bound

$$\|\bar{\theta}^T - \theta^*\|_n^2 \lesssim \left(\frac{1}{\alpha m} + \frac{1}{m^2}\right) \frac{\log^{1/\gamma} n}{n} \sigma^2,$$

with probability $1 - c_1 \exp(-c_2 m^2 \log^{1/\gamma} n)$.

β-polynomial decay: Now suppose that the kernel eigenvalues satisfy a decay condition of the form $\mu_j \leq c_1 j^{-2\beta}$ for some $\beta > 1/2$ and constant c_1. In this case, a direct calculation yields the bound

$$\mathcal{R}(\delta) \leq \sqrt{\frac{2}{n}} \sqrt{k\delta^2 + c_2 \sum_{j=k+1}^{n} j^{-2}},$$

where k is the smallest integer such that $c_2 k^{-2} < \delta^2$. Combined with upper bound $c_2 \sum_{j=k+1}^{n} j^{-2} \leq c_2 \int_{k+1} j^{-2} \leq k\delta^2$, we find that the critical radius scales as $\delta_n^2 \asymp n^{-2\beta/(1+2\beta)}$.

Consequently, if we take $T \asymp n^{-2\beta/(1+2\beta)}$ many steps, then Theorem 3 guarantees that the averaged estimator $\bar{\theta}^T$ satisfies the bound

$$\|\bar{\theta}^T - \theta^*\|_n^2 \;\leq\; \left(\frac{1}{\alpha m} + \frac{1}{m^2}\right)\left(\frac{\sigma^2}{n}\right)^{2\beta/(2\beta+1)},$$

with probability at least $1 - c_1 \exp(-c_2 m^2 (\frac{n}{\sigma^2})^{1/(2\beta+1)})$.

3.6 Discussion

In this chapter, we have proven non-asymptotic bounds for early stopping of kernel boosting for a relatively broad class of loss functions. These bounds allowed us to propose simple stopping rules which, for the class of regular kernel functions [88], yield minimax optimal rates of estimation. Although the connection between early stopping and regularization has long been studied and explored in the theoretical literature and applications alike, to the best of our knowledge, this paper is the first one to establish a general relationship between the statistical optimality of stopped iterates and the localized Gaussian complexity. This connection is important, because this localized Gaussian complexity measure, as well as its Rademacher analogue, are now well-understood to play a central role in controlling the behavior of estimators based on regularization [73, 6, 48, 81].

There are various open questions suggested by our results. The stopping rules in this paper depend on the eigenvalues of the empirical kernel matrix; for this reason, they are data-dependent and computable given the data. However, in practice, it would be desirable to avoid the cost of computing all the empirical eigenvalues. Can fast approximation techniques for kernels be used to approximately compute our optimal stopping rules? Second, our current theoretical results apply to the averaged estimator $\bar{f}^T$. We strongly suspect that the same results apply to the stopped estimator f^T, but some new ingredients are required to extend our proofs.

3.7 Proof of technical lemmas

Proof of Lemma 14

Recalling that $K^\dagger$ denotes the pseudoinverse of K, our proof is based on the linear transformation

$$z := n^{-1/2}(K^\dagger)^{1/2}\theta \quad \Longleftrightarrow \quad \theta = \sqrt{n} K^{1/2} z.$$

as well as the new function $\mathcal{J}_n(z) := \mathcal{L}_n(\sqrt{n}\sqrt{K}z)$ and its population equivalent $\mathcal{J}(z) := \mathbb{E}\mathcal{J}_n(z)$. Ordinary gradient descent on $\mathcal{J}_n$ with stepsize α takes the form

$$z^{t+1} = z^t - \alpha\nabla\mathcal{J}_n(z^t) \;=\; z^t - \alpha\sqrt{n}\sqrt{K}\nabla\mathcal{L}_n(\sqrt{n}\sqrt{K}z^t). \tag{3.38}$$

If we transform this update on z back to an equivalent one on θ by multiplying both sides by $\sqrt{n}\sqrt{K}$, we see that ordinary gradient descent on $\mathcal{J}_n$ is equivalent to the kernel boosting update $\theta^{t+1} = \theta^t - \alpha nK\nabla\mathcal{L}_n(\theta^t)$.

Our goal is to analyze the behavior of the update (3.38) in terms of the population cost $\mathcal{J}(z^t)$. Thus, our problem is one of analyzing a noisy form of gradient descent on the function $\mathcal{J}$, where the noise is induced by the difference between the empirical gradient operator $\nabla\mathcal{J}_n$ and the population gradient operator $\nabla\mathcal{J}$.

Recall that the $\mathcal{L}$ is M-smooth by assumption. Since the kernel matrix K has been normalized to have largest eigenvalue at most one, the function $\mathcal{J}$ is also M-smooth, whence

$$\mathcal{J}(z^{t+1}) \leq \mathcal{J}(z^t) + \langle\nabla\mathcal{J}(z^t),\, d^t\rangle + \frac{M}{2}\|d^t\|_2^2, \qquad \text{where} \quad d^t := z^{t+1} - z^t = -\alpha\nabla\mathcal{J}_n(z^t).$$

Morever, since the function $\mathcal{J}$ is convex, we have $\mathcal{J}(z^*) \geq \mathcal{J}(z^t) + \langle\nabla\mathcal{J}(z^t),\, z^* - z^t\rangle$, whence

$$\begin{aligned}\mathcal{J}(z^{t+1}) - \mathcal{J}(z^*) &\leq \langle\nabla\mathcal{J}(z^t),\, d^t + z^t - z^*\rangle + \frac{M}{2}\|d^t\|_2^2 \\ &= \langle\nabla\mathcal{J}(z^t),\, z^{t+1} - z^*\rangle + \frac{M}{2}\|d^t\|_2^2. \end{aligned} \tag{3.39}$$

Now define the difference of the squared errors $V^t := \frac{1}{2}\Big\{\|z^t - z^*\|_2^2 - \|z^{t+1} - z^*\|_2^2\Big\}$. By some simple algebra, we have

$$\begin{aligned} V^t \;=\; \frac{1}{2}\Big\{\|z^t - z^*\|_2^2 - \|d^t + z^t - z^*\|_2^2\Big\} &= -\langle d^t,\, z^t - z^*\rangle - \frac{1}{2}\|d^t\|_2^2 \\ &= -\langle d^t,\, -d^t + z^{t+1} - z^*\rangle - \frac{1}{2}\|d^t\|_2^2 \\ &= -\langle d^t,\, z^{t+1} - z^*\rangle + \frac{1}{2}\|d^t\|_2^2. \end{aligned}$$

Substituting back into equation (3.39) yields

$$\begin{aligned}\mathcal{J}(z^{t+1}) - \mathcal{J}(z^*) &\leq \frac{1}{\alpha}V^t + \langle\nabla\mathcal{J}(z^t) + \frac{d^t}{\alpha},\, z^{t+1} - z^*\rangle \\ &= \frac{1}{\alpha}V^t + \langle\nabla\mathcal{J}(z^t) - \nabla\mathcal{J}_n(z^t),\, z^{t+1} - z^*\rangle, \end{aligned}$$

where we have used the fact that $\frac{1}{\alpha} \geq M$ by our choice of stepsize α.

Finally, we transform back to the original variables $\theta = \sqrt{n}\sqrt{K}z$, using the relation $\nabla\mathcal{J}(z) = \sqrt{n}\sqrt{K}\nabla\mathcal{L}(\theta)$, so as to obtain the bound

$$\mathcal{L}(\theta^{t+1}) - \mathcal{L}(\theta^*) \leq \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} + \langle\nabla\mathcal{L}(\theta^t) - \nabla\mathcal{L}_n(\theta^t),\, \theta^{t+1} - \theta^*\rangle.$$

Note that the optimality of θ^* implies that $\nabla\mathcal{L}(\theta^*) = 0$. Combined with m-strong convexity, we are guaranteed that $\frac{m}{2}\|\Delta^{t+1}\|_n^2 \leq \mathcal{L}(\theta^{t+1}) - \mathcal{L}(\theta^*)$, and hence

$$\frac{m}{2}\|\Delta^{t+1}\|_n^2 \leq \frac{1}{2\alpha}\Big\{\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2\Big\} + \langle \nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1}\rangle,$$

as claimed.

Proof of Lemma 15

We split our proof into two cases, depending on whether we are dealing with the least-squares loss $\phi(y, \theta) = \frac{1}{2}(y - \theta)^2$, or a classification loss with uniformly bounded gradient ($\|\phi'\|_\infty \leq 1$).

Least-squares case

The least-squares loss is m-strongly convex with $m = M = 1$. Moreover, the difference between the population and empirical gradients can be written as $\nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}) = \frac{\sigma}{n}(w_1, \ldots, w_n)$, where the random variables $\{w_i\}_{i=1}^n$ are i.i.d. and sub-Gaussian with parameter 1. Consequently, we have

$$|\langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\, \Delta\rangle| = \left|\frac{\sigma}{n}\sum_{i=1}^n w_i\Delta(x_i)\right|.$$

Under these conditions, one can show (see [81] for reference) that

$$\left|\frac{\sigma}{n}\sum_{i=1}^n w_i\Delta(x_i)\right| \leq 2\delta_n\|\Delta\|_n + 2\delta_n^2\|\Delta\|_{\mathscr{H}} + \frac{1}{16}\|\Delta\|_n^2, \tag{3.40}$$

which implies that Lemma 15 holds with $c_3 = 16$.

Gradient-bounded ϕ-functions

We now turn to the proof of Lemma 15 for gradient bounded ϕ-functions. First, we claim that it suffices to prove the bound (3.27) for functions $g \in \partial\mathscr{H}$ and $\|g\|_{\mathscr{H}} = 1$ where $\partial\mathscr{H} := \{f - g \mid f, g \in \mathscr{H}\}$. Indeed, suppose that it holds for all such functions, and that we are given a function Δ with $\|\Delta\|_{\mathscr{H}} > 1$. By assumption, we can apply the inequality (3.27) to the new function $g := \Delta/\|\Delta\|_{\mathscr{H}}$, which belongs to $\partial\mathscr{H}$ by nature of the subspace $\mathscr{H} = \overline{\text{span}}\{\mathbb{K}(\cdot, x_i)\}_{i=1}^n$.

Applying the bound (3.27) to g and then multiplying both sides by $\|\Delta\|_{\mathscr{H}}$, we obtain

$$\begin{aligned}\langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\, \Delta\rangle &\leq 2\delta_n\|\Delta\|_n + 2\delta_n^2\|\Delta\|_{\mathscr{H}} + \frac{m}{c_3}\frac{\|\Delta\|_n^2}{\|\Delta\|_{\mathscr{H}}} \\ &\leq 2\delta_n\|\Delta\|_n + 2\delta_n^2\|\Delta\|_{\mathscr{H}} + \frac{m}{c_3}\|\Delta\|_n^2,\end{aligned}$$

where the second inequality uses the fact that $\|\Delta\|_{\mathscr{H}} > 1$ by assumption.

In order to establish the bound (3.27) for functions with $\|g\|_{\mathscr{H}} = 1$, we first prove it uniformly over the set $\{g \mid \|g\|_{\mathscr{H}} = 1, \quad \|g\|_n \leq t\}$, where $t > 1$ is a fixed radius (of course, we restrict our attention to those radii t for which this set is non-empty.) We then extend the argument to one that is also uniform over the choice of t by a "peeling" argument.

Define the random variable

$$\mathcal{Z}_n(t) := \sup_{\Delta, \widetilde{\Delta} \in \mathcal{E}(t,1)} \langle \nabla \mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla \mathcal{L}_n(\theta^* + \widetilde{\Delta}), \, \Delta \rangle. \tag{3.41}$$

The following two lemmas, respectively, bound the mean of this random variable, and its deviations above the mean:

Lemma 18. *For any $t > 0$, the mean is upper bounded as*

$$\mathbb{E}\mathcal{Z}_n(t) \leq \sigma \mathcal{G}_n(\mathcal{E}(t,1)), \tag{3.42}$$

where $\sigma := 2M + 4C_{\mathscr{H}}$.

Lemma 19. *There are universal constants (c_1, c_2) such that*

$$\mathbb{P}\Big[\mathcal{Z}_n(t) \geq \mathbb{E}\mathcal{Z}_n(t) + \alpha\Big] \leq c_1 \exp\Big(-\frac{c_2 n \alpha^2}{t^2}\Big). \tag{3.43}$$

See Appendices 3.7 and 3.7 for the proofs of these two claims.

Equipped with Lemmas 18 and 19, we now prove inequality (3.27). We divide our argument into two cases:

Case $t = \delta_n$: We first prove inequality (3.27) for $t = \delta_n$. From Lemma 18, we have

$$\mathbb{E}\mathcal{Z}_n(\delta_n) \leq \sigma \mathcal{G}_n(\mathcal{E}(\delta_n, 1)) \overset{(i)}{\leq} \delta_n^2, \tag{3.44}$$

where inequality (i) follows from the definition of δ_n in inequality (3.15). Setting $\alpha = \delta_n^2$ in expression (3.43) yields

$$\mathbb{P}\Big[\mathcal{Z}_n(\delta_n) \geq 2\delta_n^2\Big] \leq c_1 \exp\left(-c_2 n \delta_n^2\right), \tag{3.45}$$

which establishes the claim for $t = \delta_n$.

Case $t > \delta_n$: On the other hand, for any $t > \delta_n$, we have

$$\mathbb{E}\mathcal{Z}_n(t) \overset{(i)}{\leq} \sigma \mathcal{G}_n(\mathcal{E}(t,1)) \overset{(ii)}{\leq} t\sigma \frac{\mathcal{G}_n(\mathcal{E}(t,1))}{t} \leq t\delta_n,$$

where step (i) follows from Lemma 18, and step (ii) follows because the function $u \mapsto \frac{\mathcal{G}_n(\mathcal{E}(u,1))}{u}$ is non-increasing on the positive real line. (This non-increasing property is a direct consequence of the star-shaped nature of $\partial\mathscr{H}$.) Finally, using this upper bound on expression $\mathbb{E}\mathcal{Z}_n(\delta_n)$ and setting $\alpha = t^2 m/(4c_3)$ in the tail bound (3.43) yields

$$\mathbb{P}\Big[\mathcal{Z}_n(t) \geq t\delta_n + \frac{t^2 m}{4c_3}\Big] \leq c_1 \exp\left(-c_2 n m^2 t^2\right). \tag{3.46}$$

Note that the precise values of the universal constants c_2 may change from line to line throughout this section.

Peeling argument Equipped with the tail bounds (3.45) and (3.46), we are now ready to complete the peeling argument. Let $\mathcal{A}$ denote the event that the bound (3.27) is violated for some function $g \in \partial\mathscr{H}$ with $\|g\|_{\mathscr{H}} = 1$. For real numbers $0 \leq a < b$, let $\mathcal{A}(a,b)$ denote the event that it is violated for some function such that $\|g\|_n \in [a,b]$, and $\|g\|_{\mathscr{H}} = 1$. For $k = 0, 1, 2, \ldots$, define $t_k = 2^k \delta_n$. We then have the decomposition $\mathcal{E} = (0, t_0) \cup (\bigcup_{k=0}^{\infty} \mathcal{A}(t_k, t_{k+1}))$ and hence by union bound,

$$\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{A}(0, \delta_n)] + \sum_{k=1}^{\infty} \mathbb{P}[\mathcal{A}(t_k, t_{k+1})]. \tag{3.47}$$

From the bound (3.45), we have $\mathbb{P}[\mathcal{A}(0,\delta_n)] \leq c_1 \exp(-c_2 n \delta_n^2)$. On the other hand, suppose that $\mathcal{A}(t_k, t_{k+1})$ holds, meaning that there exists some function g with $\|g\|_{\mathscr{H}} = 1$ and $\|g\|_n \in [t_k, t_{k+1}]$ such that

$$\begin{aligned}
\langle \nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\, g\rangle &> 2\delta_n \|g\|_n + 2\delta_n^2 + \frac{m}{c_3}\|g\|_n^2 \\
&\overset{(i)}{\geq} 2\delta_n t_k + 2\delta_n^2 + \frac{m}{c_3} t_k^2 \\
&\overset{(ii)}{\geq} \delta_n t_{k+1} + 2\delta_n^2 + \frac{m}{4c_3} t_{k+1}^2,
\end{aligned}$$

where step (i) uses the $\|g\|_n \geq t_k$ and step (ii) uses the fact that $t_{k+1} = 2t_k$. This lower bound implies that $\mathcal{Z}_n(t_{k+1}) > t_{k+1}\delta_n + \frac{t_{k+1}^2 m}{4c_3}$ and applying the tail bound (3.46) yields

$$\mathbb{P}(\mathcal{A}(t_k, t_{k+1})) \leq \mathbb{P}(\mathcal{Z}_n(t_{k+1}) > t_{k+1}\delta_n + \frac{t_{k+1}^2 m}{4c_3}) \leq \exp\left(-c_2 n m^2 2^{2k+2} \delta_n^2\right).$$

Substituting this inequality and our earlier bound (3.45) into equation (3.47) yields

$$\mathbb{P}(\mathcal{E}) \leq c_1 \exp(-c_2 n m^2 \delta_n^2),$$

where the reader should recall that the precise values of universal constants may change from line-to-line. This concludes the proof of Lemma 15.

Proof of Lemma 18

Recalling the definitions (3.1) and (3.3) of $\mathcal{L}$ and $\mathcal{L}_n$, we can write

$$\mathcal{Z}_n(t) = \sup_{\Delta, \widetilde{\Delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^{n} (\phi'(y_i, \theta_i^* + \widetilde{\Delta}_i) - \mathbb{E}\phi'(y_i, \theta_i^* + \widetilde{\Delta}_i))\Delta_i$$

Note that the vectors Δ and $\widetilde{\Delta}$ contain function values of the form $f(x_i) - f^*(x_i)$ for functions $f \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$. Recall that the kernel function is bounded uniformly by one. Consequently, for any function $f \in \mathbb{B}_{\mathscr{H}}(f^*, 2C_{\mathscr{H}})$, we have

$$|f(x) - f^*(x)| = |\langle f - f^*, \mathbb{K}(\cdot, x)\rangle_{\mathscr{H}}| \leq \|f - f^*\|_{\mathscr{H}} \|\mathbb{K}(\cdot, x)\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}.$$

Thus, we can restrict our attention to vectors $\Delta, \widetilde{\Delta}$ with $\|\Delta\|_\infty, \|\widetilde{\Delta}\|_\infty \leq 2C_{\mathscr{H}}$ from hereonwards.

Letting $\{\varepsilon_i\}_{i=1}^n$ denote an i.i.d. sequence of Rademacher variables, define the symmetrized variable

$$\tilde{\mathcal{Z}}_n(t) := \sup_{\Delta, \widetilde{\Delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi'(y_i, \theta_i^* + \widetilde{\Delta}_i)\, \Delta_i. \tag{3.48}$$

By a standard symmetrization argument [74], we have $\mathbb{E}_y[\mathcal{Z}_n(t)] \leq 2\mathbb{E}_{y,\epsilon}[\tilde{\mathcal{Z}}_n(t)]$. Moreover, since

$$\phi'(y_i, \theta_i^* + \widetilde{\Delta}_i)\, \Delta_i \leq \frac{1}{2}\Big(\phi'(y_i, \theta_i^* + \widetilde{\Delta}_i)\Big)^2 + \frac{1}{2}\Delta_i^2$$

we have

$$\begin{aligned}
\mathbb{E}\mathcal{Z}_n(t) &\leq \mathbb{E} \sup_{\widetilde{\Delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \big(\phi'(y_i, \theta_i^* + \widetilde{\Delta}_i)\big)^2 \;+\; \mathbb{E} \sup_{\Delta \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \Delta_i^2 \\
&\leq \underbrace{2\,\mathbb{E} \sup_{\widetilde{\Delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi'(y_i, \theta_i^* + \widetilde{\Delta}_i)}_{T_1} \;+\; 4C_{\mathscr{H}} \underbrace{\mathbb{E} \sup_{\Delta \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \Delta_i}_{T_2},
\end{aligned}$$

where the second inequality follows by applying the Rademacher contraction inequality [53], using the fact that $\|\phi'\|_\infty \leq 1$ for the first term, and $\|\Delta\|_\infty \leq 2C_{\mathscr{H}}$ for the second term.

Focusing first on the term T_1, since $\mathbb{E}[\varepsilon_i \phi'(y_i, \theta_i^*)] = 0$, we have

$$\begin{aligned}
T_1 &= \mathbb{E} \sup_{\widetilde{\Delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \underbrace{\Big(\phi'(y_i, \theta_i^* + \widetilde{\Delta}_i) - \phi'(y_i; \theta_i^*)\Big)}_{\varphi_i(\widetilde{\Delta}_i)} \\
&\overset{(i)}{\leq} M\mathbb{E} \sup_{\widetilde{\Delta} \in \mathcal{E}(t,1)} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \widetilde{\Delta}_i \\
&\overset{(ii)}{\leq} \sqrt{\frac{\pi}{2}} M \mathcal{G}_n(\mathcal{E}(t,1)),
\end{aligned}$$

where step (i) follows since each function φ_i is M-Lipschitz by assumption; and step (ii) follows since the Gaussian complexity upper bounds the Rademacher complexity up to a factor of $\sqrt{\frac{\pi}{2}}$. Similarly, we have

$$T_2 \leq \sqrt{\frac{\pi}{2}}\, \mathcal{G}_n(\mathcal{E}(t,1)),$$

and putting together the pieces yields the claim.

Proof of Lemma 19

Recall the definition (3.48) of the symmetrized variable $\tilde{\mathcal{Z}}_n$. By a standard symmetrization argument [74], there are universal constants c_1, c_2 such that

$$\mathbb{P}\Big[\mathcal{Z}_n(t) \geq \mathbb{E}\mathcal{Z}_n[t] + c_1\alpha\Big] \leq c_2\mathbb{P}\Big[\tilde{\mathcal{Z}}_n(t) \geq \mathbb{E}\tilde{\mathcal{Z}}_n[t] + \alpha\Big].$$

Since $\{\varepsilon_i\}_{i=1}^n$ are $\{y_i\}_{i=1}^n$ are independent, we can study $\tilde{\mathcal{Z}}_n(t)$ conditionally on $\{y_i\}_{i=1}^n$. Viewed as a function of $\{\varepsilon_i\}_{i=1}^n$, the function $\tilde{\mathcal{Z}}_n(t)$ is convex and Lipschitz with respect to the Euclidean norm with parameter

$$L^2 := \sup_{\Delta,\widetilde{\Delta}\in\mathcal{E}(t,1)} \frac{1}{n^2}\sum_{i=1}^n \Big(\phi'(y_i, \theta_i^* + \widetilde{\Delta}_i)\, \Delta_i\Big)^2 \;\leq\; \frac{t^2}{n},$$

where we have used the facts that $\|\phi'\|_\infty \leq 1$ and $\|\Delta\|_n \leq t$. By Ledoux's concentration for convex and Lipschitz functions [52], we have

$$\mathbb{P}\Big[\tilde{\mathcal{Z}}_n(t) \geq \mathbb{E}\tilde{\mathcal{Z}}_n[t] + \alpha \;\mid\; \{y_i\}_{i=1}^n\Big] \leq c_3 \exp\Big(-c_4\frac{n\alpha^2}{t^2}\Big).$$

Since the right-hand side does not involve $\{y_i\}_{i=1}^n$, the same bound holds unconditionally over the randomness in both the Rademacher variables and the sequence $\{y_i\}_{i=1}^n$. Consequently, the claimed bound (3.43) follows, with suitable redefinitions of the universal constants.

Proof of Lemma 16

We first require an auxiliary lemma, which we state and prove in the following section. We then prove Lemma 16 in Section 3.7.

An auxiliary lemma

The following result relates the Hilbert norm of the error to the difference between the empirical and population gradients:

Lemma 20. *For any convex and differentiable loss function $\mathcal{L}$, the kernel boosting error $\Delta^{t+1} := \theta^{t+1} - \theta^*$ satisfies the bound*

$$\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq \|\Delta^t\|_{\mathscr{H}} \|\Delta^{t+1}\|_{\mathscr{H}} + \alpha \langle \nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t), \, \Delta^{t+1} \rangle. \tag{3.49}$$

Proof. Recall that $\|\Delta^t\|_{\mathscr{H}}^2 = \|\theta^t - \theta^*\|_{\mathscr{H}}^2 = \|z^t - z^*\|_2^2$ by definition of the Hilbert norm. Let us define the population update operator G on the population function $\mathcal{J}$ and the empirical update operator G_n on $\mathcal{J}_n$ as

$$G(z^t) := z^t - \alpha \nabla \mathcal{J}(\sqrt{n}\sqrt{K} z^t), \quad \text{and} \quad z^{t+1} := G_n(z^t) = z^t - \alpha \nabla \mathcal{J}_n(\sqrt{n}\sqrt{K} z^t). \tag{3.50}$$

Since $\mathcal{J}$ is convex and smooth, it follows from standard arguments in convex optimization that G is a non-expansive operator—viz.

$$\|G(x) - G(y)\|_2 \leq \|x - y\|_2 \qquad \text{for all } x, y \in \mathcal{C}. \tag{3.51}$$

In addition, we note that the vector z^* is a fixed point of G—that is, $G(z^*) = z^*$. From these ingredients, we have

$$\begin{aligned}
\|\Delta^{t+1}\|_{\mathscr{H}}^2 &= \langle z^{t+1} - z^*, \, G_n(z^t) - G(z^t) + G(z^t) - z^* \rangle \\
&\overset{(i)}{\leq} \|z^{t+1} - z^*\|_2 \|G(z^t) - G(z^*)\|_2 + \\
&\qquad \alpha \langle \sqrt{n}\sqrt{K} [\nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t)], \, z^{t+1} - z^* \rangle \\
&\overset{(ii)}{\leq} \|\Delta^{t+1}\|_{\mathscr{H}} \|\Delta^t\|_{\mathscr{H}} + \alpha \langle \nabla \mathcal{L}(\theta^* + \Delta^t) - \nabla \mathcal{L}_n(\theta^* + \Delta^t), \, \Delta^{t+1} \rangle
\end{aligned}$$

where step (i) follows by applying the Cauchy-Schwarz to control the inner product, and step (ii) follows since $\Delta^{t+1} = \sqrt{n}\sqrt{K}(z^{t+1} - z^*)$, and the square root kernel matrix $\sqrt{K}$ is symmetric. □

Proof of Lemma 16

We now prove Lemma 16. The argument makes use of Lemmas 14 and 15 combined with Lemma 20.

In order to prove inequality (3.28), we follow an inductive argument. Instead of proving (3.28) directly, we prove a slightly stronger relation which implies it, namely

$$\max\{1, \|\Delta^t\|_{\mathscr{H}}^2\} \leq \max\{1, \|\Delta^0\|_{\mathscr{H}}^2\} + t\delta_n^2 \frac{4M}{\widetilde{\gamma} m}. \tag{3.52}$$

Here $\widetilde{\gamma}$ and c_3 are constants linked by the relation

$$\widetilde{\gamma} := \frac{1}{32} - \frac{1}{4c_3} = 1/C_{\mathscr{H}}^2. \tag{3.53}$$

We claim that it suffices to prove that the error iterates Δ^{t+1} satisfy the inequality (3.52). Indeed, if we take inequality (3.52) as given, then we have

$$\|\Delta^t\|_{\mathscr{H}}^2 \leq \max\{1, \|\Delta^0\|_{\mathscr{H}}^2\} + \frac{1}{2\widetilde{\gamma}} \leq C_{\mathscr{H}}^2,$$

where we used the definition $C_{\mathscr{H}}^2 = 2\max\{\|\theta^*\|_{\mathscr{H}}^2, 32\}$. Thus, it suffices to focus our attention on proving inequality (3.52).

For $t = 0$, it is trivially true. Now let us assume inequality (3.52) holds for some $t \leq \frac{m}{8M\delta_n^2}$, and then prove that it also holds for step $t+1$.

If $\|\Delta^{t+1}\|_{\mathscr{H}} < 1$, then inequality (3.52) follows directly. Therefore, we can assume without loss of generality that $\|\Delta^{t+1}\|_{\mathscr{H}} \geq 1$.

We break down the proof of this induction into two steps:

- First, we show that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}$ so that Lemma 15 is applicable.
- Second, we show that the bound (3.52) holds and thus in fact $\|\Delta^{t+1}\|_{\mathscr{H}} \leq C_{\mathscr{H}}$.

Throughout the proof, we condition on the event $\mathcal{E}$ and $\mathcal{E}_0 := \{\frac{1}{\sqrt{n}}\|y - \mathbb{E}[y \mid x]\|_2 \leq \sqrt{2}\sigma\}$. Lemma 15 guarantees that $\mathbb{P}(\mathcal{E}^c) \leq c_1 \exp(-c_2 \frac{m^2 n \delta_n^2}{\sigma^2})$ whereas $\mathbb{P}(\mathcal{E}_0) \geq 1 - \mathrm{e}^{-n}$ follows from the fact that Y^2 is sub-exponential with parameter $\sigma^2 n$ and applying Hoeffding's inequality. Putting things together yields an upper bound on the probability of the complementary event, namely

$$\mathbb{P}(\mathcal{E}^c \cup \mathcal{E}_0^c) \leq 2c_1 \exp(-C_2 n \delta_n^2)$$

with $C_2 = \max\{\frac{m^2}{\sigma^2}, 1\}$.

Showing that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}$ In this step, we assume that inequality (3.52) holds at step t, and show that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}$. Recalling that $z := \frac{(K^\dagger)^{1/2}}{\sqrt{n}}\theta$, our update can be written as

$$z^{t+1} - z^* = z^t - \alpha\sqrt{n}\sqrt{K}\nabla\mathcal{L}(\theta^t) - z^* + \alpha\sqrt{n}\sqrt{K}(\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t)).$$

Applying the triangle inequality yields the bound

$$\|z^{t+1} - z^*\|_2 \leq \|\underbrace{z^t - \alpha\sqrt{n}\sqrt{K}\nabla\mathcal{L}(\theta^t)}_{G(z^t)} - z^*\|_2 + \|\alpha\sqrt{n}\sqrt{K}(\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t))\|_2$$

where the population update operator G was previously defined (3.50), and observed to be non-expansive (3.51). From this non-expansiveness, we find that

$$\|z^{t+1} - z^*\|_2 \leq \|z^t - z^*\|_2 + \|\alpha\sqrt{n}\sqrt{K}(\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t))\|_2,$$

Note that the ℓ_2 norm of z corresponds to the Hilbert norm of θ. This implies

$$\|\Delta^{t+1}\|_{\mathscr{H}} \leq \|\Delta^t\|_{\mathscr{H}} + \underbrace{\|\alpha\sqrt{n}\sqrt{K}(\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t))\|_2}_{:=T}$$

Observe that because of uniform boundedness of the kernel by one, the quantity T can be bounded as

$$T \leq \alpha\sqrt{n}\|\nabla\mathcal{L}_n(\theta^t) - \nabla\mathcal{L}(\theta^t))\|_2 = \alpha\sqrt{n}\frac{1}{n}\|v - \mathbb{E}v\|_2,$$

where we have define the vector $v \in \mathbb{R}^n$ with coordinates $v_i := \phi'(y_i, \theta_i^t)$. For functions ϕ satisfying the gradient boundedness and $m - M$ condition, since $\theta^t \in \mathbb{B}_{\mathscr{H}}(\theta^*, C_{\mathscr{H}})$, each coordinate of the vectors v and $\mathbb{E}v$ is bounded by 1 in absolute value. We consequently have

$$T \leq \alpha \leq C_{\mathscr{H}},$$

where we have used the fact that $\alpha \leq m/M < 1 \leq \frac{C_{\mathscr{H}}}{2}$. For least-squares ϕ we instead have

$$T \leq \alpha\frac{\sqrt{n}}{n}\|y - \mathbb{E}[y \mid x]\|_2 =: \frac{\alpha}{\sqrt{n}}Y \leq \sqrt{2}\sigma \leq C_{\mathscr{H}}$$

conditioned on the event $\mathcal{E}_0 := \{\frac{1}{\sqrt{n}}\|y - \mathbb{E}[y \mid x]\|_2 \leq \sqrt{2}\sigma\}$. Since Y^2 is sub-exponential with parameter $\sigma^2 n$ it follows by Hoeffding's inequality that $\mathbb{P}(\mathcal{E}_0) \geq 1 - \mathrm{e}^{-n}$.

Putting together the pieces yields that $\|\Delta^{t+1}\|_{\mathscr{H}} \leq 2C_{\mathscr{H}}$, as claimed.

Completing the induction step We are now ready to complete the induction step for proving inequality (3.52) using Lemma 14 and Lemma 15 since $\|\Delta^{t+1}\|_{\mathscr{H}} \geq 1$. We split the argument into two cases separately depending on whether or not $\|\Delta^{t+1}\|_{\mathscr{H}}\delta_n \geq \|\Delta^{t+1}\|_n$. In general we can assume that $\|\Delta^{t+1}\|_{\mathscr{H}} > \|\Delta^t\|_{\mathscr{H}}$, otherwise the induction inequality (3.52) satisfies trivially.

Case 1: When $\|\Delta^{t+1}\|_{\mathscr{H}}\delta_n \geq \|\Delta^{t+1}\|_n$, inequality (3.27) implies that

$$\langle\nabla\mathcal{L}(\theta^* + \widetilde{\Delta}) - \nabla\mathcal{L}_n(\theta^* + \widetilde{\Delta}),\, \Delta^{t+1}\rangle \leq 4\delta_n^2\|\Delta^{t+1}\|_{\mathscr{H}} + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2, \tag{3.54}$$

Combining Lemma 20 and inequality (3.54), we obtain

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 &\leq \|\Delta^t\|_{\mathscr{H}}\|\Delta^{t+1}\|_{\mathscr{H}} + 4\alpha\delta_n^2\|\Delta^{t+1}\|_{\mathscr{H}} + \alpha\frac{m}{c_3}\|\Delta^{t+1}\|_n^2 \\ &\implies \|\Delta^{t+1}\|_{\mathscr{H}} \leq \frac{1}{1 - \alpha\delta_n^2\frac{m}{c_3}}\big[\|\Delta^t\|_{\mathscr{H}} + 4\alpha\delta_n^2\big],\end{aligned} \tag{3.55}$$

where the last inequality uses the fact that $\|\Delta^{t+1}\|_n \leq \delta_n\|\Delta^{t+1}\|_{\mathscr{H}}$.

Case 2: When $\|\Delta^{t+1}\|_{\mathscr{H}}\delta_n < \|\Delta^{t+1}\|_n$, we use our assumption $\|\Delta^{t+1}\|_{\mathscr{H}} \geq \|\Delta^t\|_{\mathscr{H}}$ together with Lemma 20 and inequality (3.27) which guarantee that

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 &\leq \|\Delta^t\|_{\mathscr{H}}^2 + 2\alpha\langle\nabla\mathcal{L}(\theta^* + \Delta^t) - \nabla\mathcal{L}_n(\theta^* + \Delta^t),\, \Delta^{t+1}\rangle \\ &\leq \|\Delta^t\|_{\mathscr{H}}^2 + 8\alpha\delta_n\|\Delta^{t+1}\|_n + 2\alpha\frac{m}{c_3}\|\Delta^{t+1}\|_n^2.\end{aligned}$$

Using the elementary inequality $2ab \leq a^2 + b^2$, we find that

$$\begin{aligned}\|\Delta^{t+1}\|_{\mathscr{H}}^2 &\leq \|\Delta^t\|_{\mathscr{H}}^2 + 8\alpha\left[m\widetilde{\gamma}\|\Delta^{t+1}\|_n^2 + \frac{1}{4\widetilde{\gamma}m}\delta_n^2\right] + 2\alpha\frac{m}{c_3}\|\Delta^{t+1}\|_n^2 \\ &\leq \|\Delta^t\|_{\mathscr{H}}^2 + \alpha\frac{m}{4}\|\Delta^{t+1}\|_n^2 + \frac{2\alpha\delta_n^2}{\widetilde{\gamma}m}, \qquad (3.56)\end{aligned}$$

where in the final step, we plug in the constants $\widetilde{\gamma}, c_3$ which satisfy equation (3.53).

Now Lemma 14 implies that

$$\begin{aligned}\frac{m}{2}\|\Delta^{t+1}\|_n^2 &\leq D^t + 4\|\Delta^{t+1}\|_n\delta_n + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2 \\ &\overset{(i)}{\leq} D^t + 4\left[\widetilde{\gamma}m\|\Delta^{t+1}\|_n^2 + \frac{1}{4\widetilde{\gamma}m}\delta_n^2\right] + \frac{m}{c_3}\|\Delta^{t+1}\|_n^2,\end{aligned}$$

where step (i) again uses $2ab \leq a^2 + b^2$. Thus, we have $\frac{m}{4}\|\Delta^{t+1}\|_n^2 \leq D^t + \frac{1}{\widetilde{\gamma}m}\delta_n^2$. Together with expression (3.56), we find that

$$\begin{aligned}&\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq \|\Delta^t\|_{\mathscr{H}}^2 + \frac{1}{2}(\|\Delta^t\|_{\mathscr{H}}^2 - \|\Delta^{t+1}\|_{\mathscr{H}}^2) + \frac{4\alpha}{\widetilde{\gamma}m}\delta_n^2 \\ \implies &\|\Delta^{t+1}\|_{\mathscr{H}}^2 \leq \|\Delta^t\|_{\mathscr{H}}^2 + \frac{4\alpha}{\widetilde{\gamma}m}\delta_n^2. \qquad (3.57)\end{aligned}$$

Combining the pieces: By combining the two previous cases, we arrive at the bound

$$\max\left\{1, \|\Delta^{t+1}\|_{\mathscr{H}}^2\right\} \leq \max\left\{1, \kappa^2(\|\Delta^t\|_{\mathscr{H}} + 4\alpha\delta_n^2)^2, \|\Delta^t\|_{\mathscr{H}}^2 + \frac{4M}{\widetilde{\gamma}m}\delta_n^2\right\}, \qquad (3.58)$$

where $\kappa := \frac{1}{(1-\alpha\delta_n^2\frac{m}{c_3})}$ and we used that $\alpha \leq \min\{\frac{1}{M}, M\}$.

Now it is only left for us to show that with the constant c_3 chosen such that $\widetilde{\gamma} = \frac{1}{32} - \frac{1}{4c_3} = 1/C_{\mathscr{H}}^2$, we have

$$\kappa^2(\|\Delta^t\|_{\mathscr{H}} + 4\alpha\delta_n^2)^2 \leq \|\Delta^t\|_{\mathscr{H}}^2 + \frac{4M}{\widetilde{\gamma}m}\delta_n^2.$$

Define the function $f : (0, C_{\mathscr{H}}] \to \mathbb{R}$ via $f(\xi) := \kappa^2(\xi + 4\alpha\delta_n^2)^2 - \xi^2 - \frac{4M}{\widetilde{\gamma}m}\delta_n^2$. Since $\kappa \geq 1$, in order to conclude that $f(\xi) < 0$ for all $\xi \in (0, C_{\mathscr{H}}]$, it suffices to show that $\operatorname{argmin}_{x\in\mathbb{R}} f(x) < 0$

and $f(C_{\mathscr{H}}) < 0$. The former is obtained by basic algebra and follows directly from $\kappa \geq 1$. For the latter, since $\widetilde{\gamma} = \frac{1}{32} - \frac{1}{4c_3} = 1/C_{\mathscr{H}}^2$, $\alpha < \frac{1}{M}$ and $\delta_n^2 \leq \frac{M^2}{m^2}$ it thus suffices to show

$$\frac{1}{(1 - \frac{M}{8m})^2} \leq \frac{4M}{m} + 1$$

Since $(4x+1)(1-\frac{x}{8})^2 \geq 1$ for all $x \leq 1$ and $\frac{m}{M} \leq 1$, we conclude that $f(C_{\mathscr{H}}) < 0$.

Now that we have established $\max\{1, \|\Delta^{t+1}\|_{\mathscr{H}}^2\} \leq \max\{1, \|\Delta^t\|_{\mathscr{H}}^2\} + \frac{4M}{\widetilde{\gamma} m}\delta_n^2$, the induction step (3.52) follows. which completes the proof of Lemma 16.

Proof of Lemma 17

Recall that the LogitBoost algorithm is based on logistic loss $\phi(y, \theta) = \ln(1 + e^{-y\theta})$, whereas the AdaBoost algorithm is based on the exponential loss $\phi(y, \theta) = \exp(-y\theta)$. We now verify the m-M-condition for these two losses with the corresponding parameters specified in Lemma 17.

m-M-condition for logistic loss

The first and second derivatives are given by

$$\frac{\partial \phi(y,\theta)}{\partial \theta} = \frac{-ye^{-y\theta}}{1 + e^{-y\theta}}, \qquad \text{and} \qquad \frac{\partial^2 \phi(y,\theta)}{(\partial\theta)^2} = \frac{y^2}{(e^{-y\theta/2} + e^{y\theta/2})^2}.$$

It is easy to check that $|\frac{\partial \phi(y,\theta)}{\partial \theta}|$ is uniformly bounded by $B = 1$.

Turning to the second derivative, recalling that $y \in \{-1, +1\}$, it is straightforward to show that

$$\max_{y \in \{-1,+1\}} \sup_{\theta} \frac{y^2}{(e^{-y\theta/2} + e^{y\theta/2})^2} \leq \frac{1}{4},$$

which implies that $\frac{\partial \phi(y,\theta)}{\partial \theta}$ is a $1/4$-Lipschitz function of θ, i.e. with $M = 1/4$.

Our final step is to compute a value for m by deriving a uniform lower bound on the Hessian. For this step, we need to exploit the fact that $\theta = f(x)$ must arise from a function f such that $\|f\|_{\mathscr{H}} \leq D := C_{\mathscr{H}} + \|\theta^*\|_{\mathscr{H}}$. Since $\sup_x \mathbb{K}(x,x) \leq 1$ by assumption, the reproducing relation for RKHS then implies that $|f(x)| \leq D$. Combining this inequality with the fact that $y \in \{-1, 1\}$, it suffices to lower the bound the quantity

$$\min_{y \in \{-1,+1\}} \min_{|\theta| \leq D} \left| \frac{\partial^2 \phi(y,\theta)}{(\partial\theta)^2} \right| = \min_{|y| \leq 1} \min_{|\theta| \leq D} \frac{y^2}{(e^{-y\theta/2} + e^{y\theta/2})^2} \geq \underbrace{\frac{1}{e^{-D} + e^{D} + 2}}_{m},$$

which completes the proof for the logistic loss.

m-M-condition for AdaBoost

The AdaBoost algorithm is based on the cost function $\phi(y, \theta) = e^{-y\theta}$, which has first and second derivatives (with respect to its second argument) given by

$$\frac{\partial \phi(y,\theta)}{\partial \theta} = -ye^{-y\theta}, \qquad \text{and} \quad \frac{\partial^2 \phi(y,\theta)}{(\partial \theta)^2} = e^{-y\theta}.$$

As in the preceding argument for logistic loss, we have the bound $|y| \leq 1$ and $|\theta| \leq D$. By inspection, the absolute value of the first derivative is uniformly bounded $B := e^D$, whereas the second derivative always lies in the interval $[m, M]$ with $M := e^D$ and $m := e^{-D}$, as claimed.

Part II

Testing

Chapter 4

Adaptive Sampling for multiple testing

4.1 Introduction

For most modern internet companies, wherever there is a metric that can be measured (e.g., time spent on a page, click-through rates, conversion of curiousity to a sale), there is almost always a randomized trial behind the scenes, with the goal of identifying an alternative website design that provides improvements over the default design. The use of such data-driven decisions for perpetual improvement is colloquially known as *A/B testing* in the case of two alternatives, or *A/B/n testing* for several alternatives. Given a default configuration and several alternatives (e.g., color schemes of a website), the standard practice is to divert a small amount of scientist-traffic to a randomized trial over these alternatives and record the desired metric for each of them. If an alternative appears to be significantly better, it is implemented; otherwise, the default setting is maintained. The idea is illustrated in Figure 4.1 in a setting that could be representative for a multiple testing application by a tech company.

At first glance, this procedure seems intuitive and simple. However, in cases where the aim is to optimize over one particular metric, this common tool suffers from several downsides. (1) First, whereas some alternatives may be clearly worse than the default, others may only have a slight edge. If one wishes to minimize the amount of time and resources spent on this randomized trial the more promising alternatives should intuitively get a larger share of the traffic than the clearly-worse alternatives. Yet typical A/B/n testing frameworks allocate traffic uniformly over alternatives. (2) Second, companies often desire to continuously monitor an ongoing A/B test as they may adjust their termination criteria as time goes by and possibly stop earlier or later than originally intended. However, just as if you flip a coin long enough, a long string of heads is eventually inevitable, the practice of continuous monitoring (without mathematically correcting for it) can easily fool the tester to believe that a result is statistically significant, when in reality it is not. This is one of the reasons for the

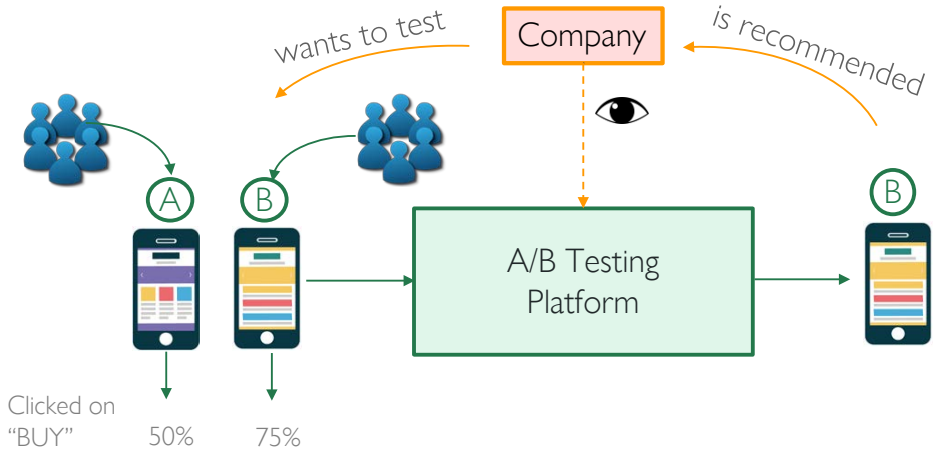


Figure 4.1: A typical example for an A/B testing framework as used by many tech companies, where the reward is a purchase and the treatment is an alternative app design

lack of reproducibility of scientific results, an issue recently receiving increased attention from the public media. (3) Third, the lack of sufficient evidence or an insignificant improvement of the metric may make it undesirable from a practical or financial perspective to replace the default. Therefore, when a company runs hundreds to thousands of A/B tests within a year as illustrated in Figure 4.2 (a), ideally the number of statistically insignificant changes that it made should be small compared to the total number of changes made. Controlling the false alarm rate of each individual test at a desired level α however does *not* achieve this type of control, also known as controlling the false discovery rate. Of course, it is also desirable to detect better alternatives (when they exist), and to do so as quickly as possible.

In this chapter, we provide a novel framework that addresses the above shortcomings of A/B or A/B/n testing. The first concern is tackled by employing recent advances in adaptive sampling like the pure-exploration multi-armed bandit (MAB) algorithm. For the second concern, we adopt the notion of any-time p-values for guilt-free continuous monitoring, and we make the advantages and risks of early-stopping transparent. Finally, we handle the third issue using recent advances in online false discovery rate (FDR) control. Hence the combined framework can be described as doubly-sequential (sequences of MAB tests, each of which is itself sequential) as illustrated in Figure 4.2. Although each of those problems has been studied in hitherto disparate communities, how to leverage the best of all worlds, if at all possible, has remained an open problem. The main contributions of this chapter are in merging these ideas in a combined framework and presenting the conditions under which it can be shown to yield near-optimal sample complexity, near-optimal best-alternative discovery rate, as well as FDR control.

While the above concerns raised about A/B/n testing were discussed using the example of modern internet companies, the same concerns carry forward qualitatively to other domains, like pharmaceutical companies running sequential clinical trials with a control (often placebo)

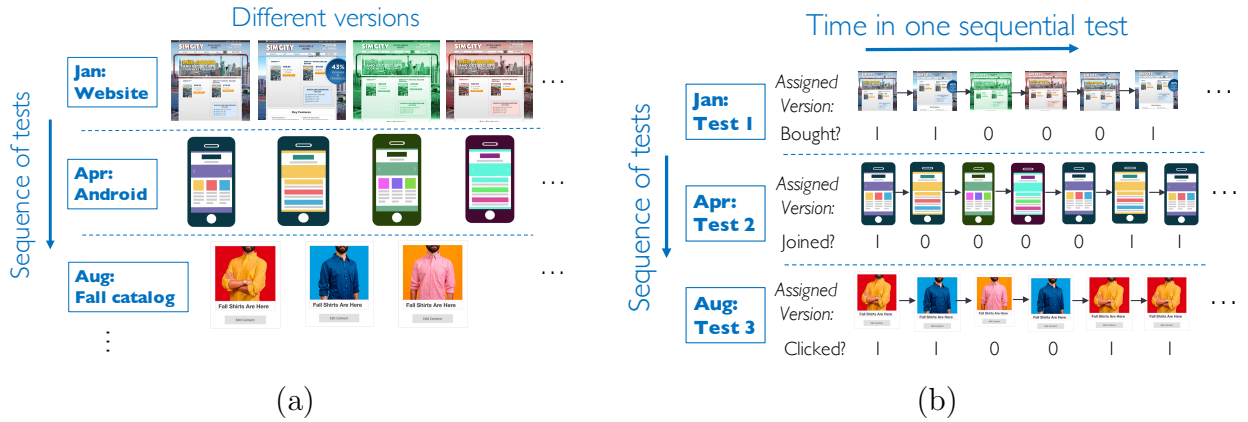


Figure 4.2: (a) A company possibly conducts many experiments with multiple treatments over the year. In this example, the "website" experiment aims at finding a version that leads to more purchases, the "android" experiment is conducted to determine the design with the highest percentage of new memberships and the "fall catalog" should be represented by the picture which leads to the biggest number of clicks (b) Our proposed doubly sequential procedure, where the version that is assigned to the next person depends on previous outcomes, achieves false discovery control and true discovery guarantees. The outcomes are binary variables which indicate whether the user took the company preferred action (e.g. "buy")

and a few treatments (like different doses or drug substances). In a manufacturing or food production setting, one may be interested in identifying (perhaps cheaper) substitutes for individual materials without compromising the quality of a product too much. In a government setting, pilot programs are funded in search of improvements over current programs and it is desirable from a social welfare standpoint and cost to limit the adoption of ineffective policies.

The remainder of this chapter is organized as follows. In Section 4.2, we lay out the primary goals of the chapter, and describe a meta-algorithm that combines adaptive sampling strategies with FDR control procedures. Section 4.3 is devoted to the description of a concrete procedure, along with some theoretical guarantees on its properties. In Section 4.5, we describe the results of our extensive experiments on both simulated and real-world data sets that are available to us, before we conclude with a discussion in Section 4.7.

4.2 Formal experimental setup and a meta-algorithm

In this section we first formalize the setup of a typical A/B/n test and provide a high-level overview of our proposed combined framework aimed at addressing the shortcomings mentioned in the introduction. A specific instantiation of this meta-algorithm along with detailed theoretical guarantees are specified in Section 4.3.

For concreteness, we refer to the system designer, whether a tech company or a pharmaceutical company, as a (data) scientist. We assume that the scientist needs to possibly

conduct an infinite number of experiments sequentially, indexed by j. Each experiment has one default setting, referred to as the *control*, and $K = K(j)$ alternative settings, called the *treatments* or *alternatives.* The scientist must return one of the $K + 1$ options that is the "best" according to some predefined metric, before the next experiment is started. Such a setup is a simple mathematical model both for clinical trials run by pharmaceutical labs, and A/B/n testing used at scale by tech companies.

One full experiment consists of steps of the following kind: In each step, the scientist assigns a new person—who arrives at the website or who enrolls in the clinical trial—to one of the $K + 1$ options and obtains a measurable outcome. In practice, the role of the scientist could be taken by an adaptive algorithm, which determines the assignment at time step j by careful consideration of all previous outcomes. Borrowing terminology from the multi-armed bandit (MAB) literature, we refer to each of the $K + 1$ options as an *arm*, and each assignment to arm i is termed "pulling arm i". For concreteness, we assign the index 0 to the default or control arm, and note that this index is known to the algorithm.

We assume that the observable metric from each pull of arm $i = 0, 1, \ldots, K$ corresponds to an independent draw from an unknown probability distribution with expectation μ_i. Ideally, if the means were known, we would use them as scores to compare the arms where higher is better. In the sequel we use $\mu_{i_\star} := \max_{i=1,\ldots,K} \mu_i$ to denote the mean of the best arm. We refer the reader to Table 4.1 for a glossary of the notation used throughout this chapter.

Some desiderata and difficulties

Given the setup above, how can we mathematically describe the guarantees that the companies might desire from an improved multiple-A/B/n testing framework? Which parts of the puzzle can be directly transferred from known results, and what challenges remain?

In order to answer the first question, let us adopt terminology from the hypothesis testing literature and view each experiment as a test of a *null hypothesis*. Any claim that an alternative arm is the best is called a *discovery*, and if such a claim is erroneous then it is called a false discovery. When multiple hypotheses need to be tested, the scientist needs to define the quantity it wants to control. While we may desire that the probability of even a single false discovery—called the family-wise error rate—is small, this is usually far too stringent for a large and unknown number of tests. For this reason, [11] proposed that it may be more interesting to control the expected ratio of false discoveries to the total number of discoveries (called the False Discovery Rate, or *FDR* for short) or ratio of expected number of false discoveries to the expected number of total discoveries (called the modified FDR or *mFDR* for short). Over the past decades, the FDR and its variants like mFDR have become standard quantities for multiple testing applications. In the following, if not otherwise specified, we use the term FDR to denote both measures in order to simplify the presentation. In Section 4.3, we show that both mFDR and FDR can be controlled for different choices of procedures.

Challenges in viewing an MAB instance as a hypothesis test

In our setup, we want to be able to control the FDR at any time in an online manner. Online FDR procedures were first introduced by Foster and Stine [29], and have since been studied by other authors (e.g., [1, 41]). A typical online FDR procedure is based on comparing a valid p-value P^j with carefully-chosen levels α_j for each hypothesis test*. We reject the null hypothesis, represented as $R_j = 1$, when $P^j \leq \alpha_j$ and we set $R_j = 0$ otherwise.

As mentioned, we want to use adaptive MAB algorithms in each experiment to test each hypothesis, since they can find a best arm among $K+1$ with near-optimal sample complexity. However the traditional MAB setup does not account for the asymmetry between the arms as is the case in a testing setup, with one being the default (control) and others being alternatives (treatments). This is the standard scenario in A/B/n testing applications, as for example a company might prefer wrong claims that the control is the best (false negative), rather than wrong claims that an alternative is the best (false positive), simply because new system-wide adoption of selected alternatives might involve high costs. What would be a suitable null hypothesis in this hybrid setting? To allow continuous monitoring, is it possible to define and compute always-valid p-values that are super-uniformly distributed under the null hypothesis when computed at any time t? (This could be especially challenging given that the number of samples from each the arm is random, and different for each arm.)

In addition to asymmetry, the practical scientist might have a different incentive than the ideal outcome for MAB algorithms. In particular, he/she might not want to find the best alternative if it is not *substantially* better than the control. Indeed, if the net gain made by adopting a new alternative is small, it might be offset by the cost of implementing the change from the existing default choice. By similar reasoning, we may not require identifying the single best arm if there is a *set* of arms with similar means that are all larger than the rest.

We propose a sensible null-hypothesis for each experiment which incorporates the approximation and improvement notions as described above and provide an always valid p-value which can be easily calculated at each time step in the experiment. We show that a slight modification of the usual LUCB algorithm caters to this specific null-hypothesis while still maintaining near-optimal sample complexity.

Interaction between MAB and FDR

In order to take advantage of the sample efficiency of best-arm bandit algorithms, it is crucial to set the confidence levels close to what is needed. Given a user-defined level α, at each hypothesis j, online FDR procedures automatically output the significance level α_j which are "needed" to guarantee FDR control, based on past decisions.

Can we directly set the MAB confidence levels to the output levels α_j from the online FDR procedure? If we do, our p-values are not independent across different hypotheses anymore: P^j directly depends on the FDR levels α_j and each α_j in turn depends on past

*A valid P^j must be stochastically dominated by a uniform distribution on $[0, 1]$, which we henceforth refer to as *super-uniformly distributed*.

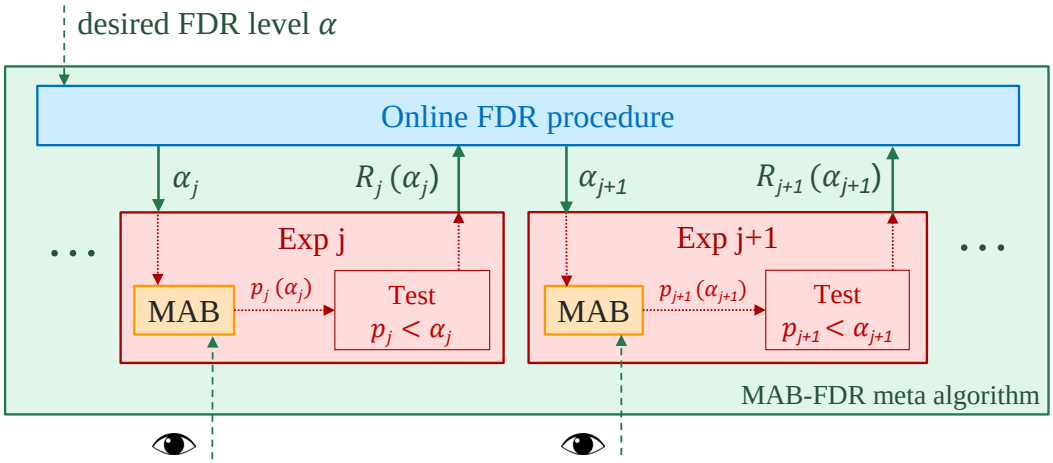


Figure 4.3: Diagram of the MAB-FDR meta algorithm designed to achieve online FDR control along with near-optimal sample complexity. The green arrows symbolize interaction between the MAB and FDR procedures via the FDR test levels α_j and rejection indicator variables R_j. Notice that the P^j-values are now dependent as each α_j depends on $R_1, \ldots, R_{j-1}$. The eyes represent possible continuous monitoring by the scientist.

MAB rejections, thus on past MAB p-values (see Figure 4.3). Does the new interaction compromise FDR guarantees?

Although known online FDR procedures [29, 41] guarantee FDR control for independent p-values, this does not hold for dependent p-values in general. Hence FDR control guarantees cannot simply be obtained out of the box. In particular, it is not a priori obvious that the introduced dependence between the p-values does not cause problems, i.e. violates necessary conditions for FDR control type theorems. A key insight that emerges from our analysis is that an appropriate bandit algorithm actually shapes the p-value distribution under the null in a good way that allows us to control FDR.

A meta-algorithm

Procedure 1 summarizes our doubly-sequential procedure, with a corresponding flowchart in Figure 4.3. We will prove theoretical guarantees after instantiating the separate modules. Note that our framework allows the scientist to plug in their favorite best-arm MAB algorithm or online FDR procedure. The choice for each of them determines which guarantees can be proven for the entire setup. Any independent improvement in either of the two parts would immediately lead to an overall performance boost of the overall framework.

Procedure 1 MAB-FDR Meta algorithm skeleton

1. The scientist sets a desired FDR control rate α.
2. For each $j = 1, 2, \ldots$:
   - Experiment j receives a designated control arm and some number of alternative arms.
   - An *online-FDR procedure* returns an α_j that is some function of the past values $\{P^\ell\}_{\ell=1}^{j-1}$.
   - An *MAB procedure* with inputs (a) the control arm and $K(j)$ alternative arms, (b) confidence level α_j, and (c) (optional) a precision $\epsilon \geq 0$, is executed and if the procedure self-terminates, returns a recommended arm.
   - Throughout the MAB procedure, an *always valid p-value* is constructed continuously for each time t using only the samples collected up to that time from the j-th experiment: for any t, it is a random variable $P_t^j \in [0, 1]$ that is super-uniformly distributed whenever the control-arm is best.
   - When the MAB procedure is terminated at time t (either by itself or by a user-defined stopping criterion that may depend on P_t^j), if the arm with the highest empirical mean is *not* the control arm and $P_t^j \leq \alpha_j$, then we return $P^j := P_t^j$, and the control arm is rejected in favor of this empirically best arm.

4.3 A concrete procedure with guarantees

We now take the high-level road map given in Procedure 1, and show that we can obtain a concrete, practically implementable framework with FDR control and power guarantees. We first discuss the key modeling decisions we have to make in order to seamlessly embed MAB algorithms into an online FDR framework. We then outline a modified version of a commonly used best-arm algorithm, before we finally prove FDR and power guarantees for the concrete combined procedure.

Defining null hypotheses and constructing p-values

Our first task is to define a null hypothesis for each experiment. As mentioned before, the choice of the null is not immediately obvious, since we sample from *multiple* distributions *adaptively* instead of independently. In particular, we will generally not have the same number of samples for all arms. Given a distribution with default mean μ_0 and alternative distributions with means $\{\mu_i\}_{i=1}^K$, we propose that the null hypothesis for the j-th experiment should be defined as

$$H_0^j : \mu_0 \geq \mu_i - \epsilon \quad \text{for all } i = 1, \ldots, K. \tag{4.1}$$

In words, the null corresponds to there being no alternative arm that is ϵ-better than the control arm.

It remains to define a p-value for each experiment that is stochastically dominated by a uniform random variable under the null; such a p-value is said to be *superuniform*. In order to simplify notation below, we omit the index j for the experiment and retain only the index i for the choice of arms. In order to be able to use a p-value at arbitrary times in the testing procedure and to allow scientists to monitor the algorithm's progress in real time, it is helpful to define an *always valid p-value*, as previously defined by Johari et al. [43]. An always valid p-value is a stochastic process $\{P_t\}_{t=1}^{\infty}$ such that for all fixed and random stopping times T, under any distribution $\mathbb{P}_0$ over the arm rewards such that the null hypothesis is true, we have

$$\mathbb{P}_0(P_T \leq \alpha) \leq \alpha. \tag{4.2}$$

When all arms are drawn independently an equal number of times, by linearity of expectation one can regard the distance of each pair of samples as a random variable drawn i.i.d. from a distribution with mean $\tilde{\mu}_i := \mu_0 - \mu_i$. We can then view the problem as testing the standard hypothesis $H_0 : \tilde{\mu}_i > -\epsilon$. However, when the arms are pulled adaptively, a different solution needs to be found—indeed, in this case, the sample means are *not unbiased estimators* of the true means, since the number of times an arm was pulled now depends on the empirical means of all the arms.

Our strategy is to construct always valid p-values by using the fact that p-values can be obtained by inverting confidence intervals. To construct always-valid confidence bounds, we resort to the fundamental concept of the law of the iterated logarithm (LIL), for which non-asymptotic versions have been recently derived and used for both bandits and testing problems (see [39], [4]).

To elaborate, define the function

$$\varphi_n(\delta) = \sqrt{\frac{\log(\frac{1}{\delta}) + 3\log(\log(\frac{1}{\delta})) + \frac{3}{2}\log(\log(\mathrm{e}n))}{n}}. \tag{4.3}$$

If $\widehat{\mu}_{i,n}$ is the empirical average of independent samples from a sub-Gaussian distribution, then it is known (see, for instance, [Theorem 8 45]) that for all $\delta \in (0,1)$, we have

$$\max\Big\{\mathbb{P}\Big(\bigcup_{n=1}^{\infty}\{\widehat{\mu}_{i,n} - \mu_i > \varphi_n(\delta \wedge 0.1)\}\Big), \quad \mathbb{P}\Big(\bigcup_{n=1}^{\infty}\{\widehat{\mu}_{i,n} - \mu_i < -\varphi_n(\delta \wedge 0.1)\}\Big)\Big\} \leq \delta, \tag{4.4}$$

where $\delta \wedge 0.1 := \min\{\delta, 0.1\}$.

We are now ready to propose single arm p-values of the form

$$\begin{aligned} P_{i,t} :&= \sup\Big\{\gamma \in [0,1] \;\mid\; \widehat{\mu}_{i,n_i(t)} - \varphi_{n_i(t)}(\tfrac{\gamma}{2K}) \leq \widehat{\mu}_{0,n_0(t)} + \varphi_{n_0(t)}(\tfrac{\gamma}{2}) + \epsilon\Big\} \\ &= \sup\Big\{\gamma \in [0,1] \;\mid\; \mathrm{LCB}_i(t) \leq \mathrm{UCB}_0(t) + \epsilon\Big\} \end{aligned} \tag{4.5}$$

Here we set $P_{i,t} = 1$ if the supremum is taken over an empty set. Given these single arm p-values, the always-valid p-value for the experiment is defined as

$$P_t := \min_{s \le t} \min_{i=1,\dots,K} P_{i,s}. \tag{4.6}$$

We claim that this procedure leads to an always valid p-value (with proof in Section 4.6).

Proposition 1. *The sequence $\{P_t\}_{t=1}^{\infty}$ defined via equation (4.6) is an always valid p-value.*

See Section 4.6 for the proof of this proposition.

Adaptive sampling for best-arm identification

In the traditional A/B testing setting described in the introduction, samples are allocated uniformly to the different alternatives. But by allocating different numbers of samples to the alternatives, decisions can be made with the same statistical significance using far fewer samples. Suppose moreover that there is a unique maximizer $i_\star := \arg\max_{i=0,1,\dots,K} \mu_i$, so that

$$\Delta_i := \mu_{i_\star} - \mu_i > 0 \qquad \text{for all } i \neq i_\star.$$

Then for any $\delta \in (0,1)$, best-arm identification algorithms for the multi-armed bandit problem can identify $i_\star$ with probability at least $1-\delta$ based on at most[†] $\sum_{i \neq i_\star} \Delta_i^{-2} \log(1/\delta)$ total samples (see the paper [38] for a brief survey and [77] for an application to clinical trials). In contrast, if samples are allocated *uniformly* to the alternatives under the same conditions, then the most natural procedures require $K \max_{i \neq i_\star} \Delta_i^{-2} \log(K/\delta)$ samples before returning $i_\star$ with probability at least $1-\delta$.

However, standard best-arm bandit algorithms do not incorporate asymmetry as induced by null-hypotheses as in definition (4.1) by default. Furthermore, recall that a practical scientist might desire the ability to incorporate approximation and a minimum improvement requirement. More precisely, it is natural to consider the requirement that the returned arm i_b satisfies the bounds $\mu_{i_b} \ge \mu_0 + \epsilon$ and $\mu_{i_b} \ge \mu_{i_\star} - \epsilon$ for some $\epsilon > 0$. For those readers unfamiliar with best-arm MAB algorithms, it is likely helpful to first grasp the entire framework in the special $\epsilon = 0$ throughout, before understanding it in full generality with the complications introduced by setting $\epsilon > 0$. In the following we present a modified MAB algorithm based on the common LUCB algorithm (see [44, 70] and a high-level illustration in Figure 4.4).

Inside the loop of Algorithm 1, we use $h_t \in \{0, 1, \dots, K\}$ to denote the current empirically-best arm, ℓ_t to denote the most promising contender among the other arms that has not yet been sampled enough to be ruled out. The parameter $\epsilon \ge 0$ is a slack variable, and the algorithm is easiest to first understand when $\epsilon = 0$. We provide a visualization of how ϵ affects the stopping condition in Figure 4.5. Step (a) checks if h_t is within ϵ of the true highest mean, and if it is also at least ϵ greater than the true mean of the control arm (or

[†]Here we have ignored some doubly-logarithmic factors.

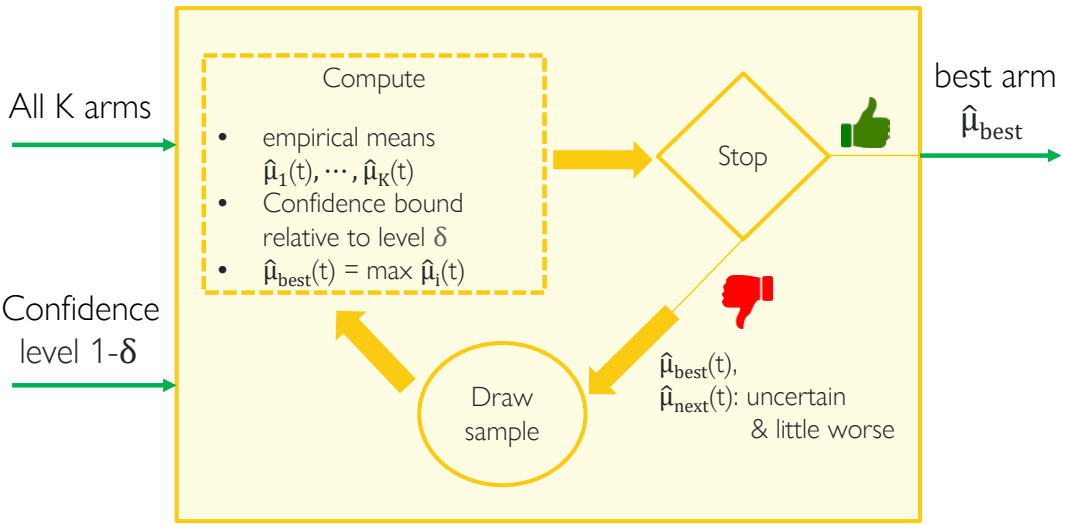


Figure 4.4: Cartoon of traditional best-arm identification algorithms relying on confidence bounds: when the stopping criterion is not fulfilled, the best empirical arm at the time and the runner-up are sampled. The latter is either drawn fewer times than the best empirical arm (and is thus more uncertain) or has an empirical mean that is only a little bit worse.

Algorithm 1 Best-arm identification with a control arm for confidence δ and precision $\epsilon \geq 0$

For all t let $n_i(t)$ be the number of times arm i has been pulled up to time t. In addition, for each arm i let $\widehat{\mu}_i(t) = \frac{1}{n_i(t)} \sum_{\tau=1}^{n_i(t)} r_i(\tau)$, define

$$\mathrm{LCB}_i(t) := \widehat{\mu}_{i,n_i(t)} - \varphi_{n_i(t)}(\tfrac{\delta}{2K}) \qquad \text{and} \qquad \mathrm{UCB}_i(t) := \widehat{\mu}_{i,n_i(t)} + \varphi_{n_i(t)}(\tfrac{\delta}{2}).$$

1. Set $t = 1$ and sample every arm once.
2. Repeat: Compute $h_t = \arg\max_{i=0,1,\ldots,K} \widehat{\mu}_i(t)$, and $\ell_t = \arg\max_{i=0,1,\ldots,K, i \neq h_t} \mathrm{UCB}_i(t)$

   (a) If $\mathrm{LCB}_0(t) > \mathrm{UCB}_i(t) - \epsilon$, for all $i \neq 0$, then output 0 and terminate.
   Else if $\mathrm{LCB}_{h_t}(t) > \mathrm{UCB}_{\ell_t}(t) - \epsilon$ and $\mathrm{LCB}_{h_t}(t) > \mathrm{UCB}_0(t) + \epsilon$, then output h_t and terminate.

   (b) If $\epsilon > 0$, let $u_t = \arg\max_{i \neq 0} \mathrm{UCB}_i(t)$ and pull all distinct arms in $\{0, u_t, h_t, \ell_t\}$ once.
   If $\epsilon = 0$, pull arms h_t and ℓ_t and set $t = t + 1$.

---

is the control arm), terminates with this arm h_t. Step (b) ensures that the control arm is sufficiently sampled when $\epsilon > 0$. Step (c) pulls h_t and ℓ_t, reducing the overall uncertainty in the difference between their two means.

The following proposition applies to Algorithm 1 run with a control arm indexed by $i = 0$ with mean μ_0 and alternative arms indexed by $i = 1, \ldots, K$ with means μ_i, respectively. Let i_b denote the random arm returned by the algorithm assuming that it exits, and define the set

$$\mathcal{S}^\star := \{i_\star \neq 0 \mid \mu_{i_\star} \geq \max_{i=1,\ldots,K} \mu_i - \epsilon \quad \text{and} \quad \mu_{i_\star} > \mu_0 + \epsilon\}. \tag{4.7}$$

Note that the mean associated with any index $i_\star \in \mathcal{S}^\star$, assuming that the set is non-empty, is guaranteed to be ϵ-superior to the control mean, and at most ϵ-inferior to the maximum mean over all arms.

Proposition 2. *The algorithm 1 terminates in finite time with probability one. Furthermore, suppose that the samples from each arm are independent and sub-Gaussian with scale* 1*. Then for any* $\delta \in (0, 1)$ *and* $\epsilon \geq 0$*, Algorithm 1 has the following guarantees:*

(a) Suppose that $\mu_0 > \max_{i=1,\ldots,K} \mu_i - \epsilon$*. Then with probability at least* $1 - \delta$*, the algorithm exits with* $i_b = 0$ *after taking at most* $O\left(\sum_{i=0}^{K} \widetilde{\Delta}_i^{-2} \log(K \log(\widetilde{\Delta}_i^{-2})/\delta)\right)$ *time steps with effective gaps*

$$\widetilde{\Delta}_0 = (\mu_0 + \epsilon) - \max_{j=1,\ldots,K} \mu_j \quad \textit{and}$$
$$\widetilde{\Delta}_i = (\mu_0 + \epsilon) - \mu_i.$$

(b) Otherwise, suppose that the set $\mathcal{S}^\star$ *as defined in equation* (4.7) *is non-empty. Then with probability at least* $1 - \delta$*, the algorithm exits with* $i_b \in \mathcal{S}^\star$ *after taking at most*

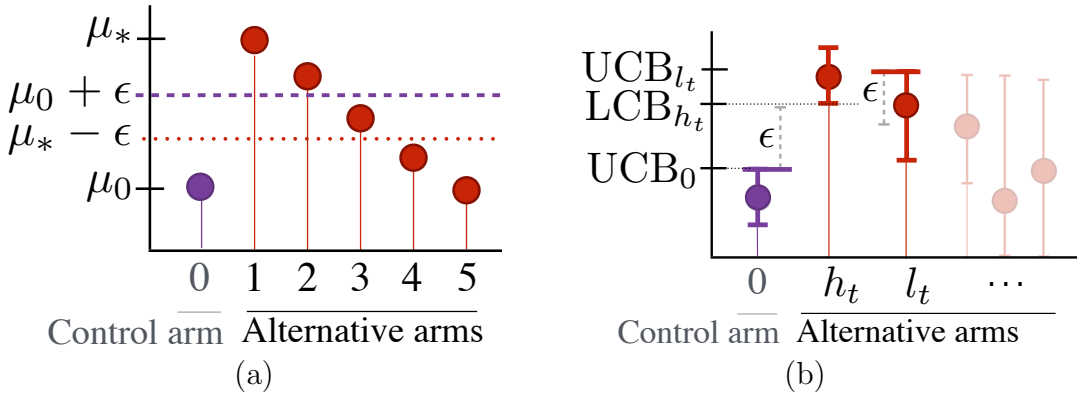


Figure 4.5: (a) The means of arms $\{1, 2, 3\}$ are within ϵ of the best arm, but only arms $\{1, 2\}$ are at least ϵ better than the control arm 0. Thus, returning any of arms $\{3, 4, 5\}$ would result in a false discovery when $\epsilon > 0$. (b) An example of the stopping condition being critically met and returning a non-control arm h_t. While $\text{LCB}_{h_t} > \text{UCB}_{\ell_t} - \epsilon$ is satisfied with some slack, $\text{LCB}_{h_t} > \text{UCB}_0 + \epsilon$ is just barely satisfied.

$O\left(\sum_{i=0}^{K} \widetilde{\Delta}_i^{-2} \log(K \log(\widetilde{\Delta}_i^{-2})/\delta)\right)$ *time steps with effective gaps*

$$\widetilde{\Delta}_0 = \min\left\{\max_{j=1,\dots,K} \mu_j - (\mu_0 + \epsilon), \max\{\Delta_0, \epsilon\}\right\} \quad \textit{and}$$
$$\widetilde{\Delta}_i = \max\left\{\Delta_i, \min\left\{\max_{j=1,\dots,K} \mu_j - (\mu_0 + \epsilon), \epsilon\right\}\right\}.$$

See Section 4.6 for the proof of this claim. Part (a) of Proposition 2 guarantees that when no alternative arm is ϵ-superior to the control arm (i.e. under the null hypothesis), the algorithm stops and returns the control arm after a certain number of samples with probability at least $1 - \delta$, where the sample complexity depends on ϵ-modified gaps between the means μ_0 and μ_i. Part (b) guarantees that if there is in fact at least one alternative that is ϵ-superior to the control arm (i.e. under the alternative), then the algorithm will find at least one of them that is at most ϵ-inferior to the best of all possible arms with the same sample complexity and probability.

Note that the required number of samples $O\left(\sum_{i=0}^{K} \widetilde{\Delta}_i^{-2} \log(K \log(\widetilde{\Delta}_i^{-2})/\delta)\right)$ in Proposition 2 is comparable, up to log factors, with the well-known results in [44, 70] for the case $\epsilon = 0$, with the modified gaps $\widetilde{\Delta}_i$ replacing $\Delta_i = \mu_{i_\star} - \mu_i$. Indeed, the nearly optimal sample complexity result of [70] implies that the algorithm terminates under settings (a) and (b) after at most $O(\max_{j \neq i_\star} \Delta_j^{-2} \log(K \log(\Delta_j^{-2})/\delta) + \sum_{i \neq i_\star} \Delta_i^{-2} \log(\log(\Delta_i^{-2})/\delta))$ samples are taken.

In our development to follow, we now bring back the index for experiment j, in particular using P^j to denote the quantity P_T^j at any stopping time T. Here the stopping time can either be defined by the scientist, or in an algorithmic manner.

Best-arm MAB interacting with online FDR

After having established null hypotheses and p-values in the context of best-arm MAB algorithms, we are now ready to embed them into an online FDR procedure. In the following, we consider p-values for the j-th experiment $P^j := P_{T_j}^j$ which is just the p-value as defined in equation (4.6) at the stopping time T_j, which depends on α_j.

We denote the set of true null and false null hypotheses up to experiment J as $\mathcal{H}_0(J)$ and $\mathcal{H}_1(J)$ respectively, where we drop the argument whenever it's clear from the context. The variable $R_j = \mathbb{1}_{P^j \leq \alpha_j}$ indicates whether a the null hypothesis of experiment j has been rejected, where $R_j = 1$ denotes a claimed discovery that an alternative was better than the control. The false discovery rate (FDR) and modified FDR *up to experiment* J are then defined as

$$\mathrm{FDR}(J) := \mathbb{E}\frac{\sum_{j \in \mathcal{H}_0} R_j}{\sum_{i=1}^{J} R_i \vee 1} \qquad \text{and} \qquad \mathrm{mFDR}(J) := \frac{\mathbb{E}\sum_{j \in \mathcal{H}_0} R_j}{\mathbb{E}\sum_{i=1}^{J} R_i + 1}. \tag{4.8}$$

Notation	Terminology and explanation
MAB	(pure exploration for best-arm identification in) multi-armed bandits
FDR(J)	expected ratio of # false disc. to # disc. up to experiment J
mFDR(J)	the ratio of expected # false discoveries to expected # discoveries
α	target for FDR or mFDR control after any number of experiments
BDR(J)	the best arm discovery rate (generalization of test power)
ϵBDR(J)	the ϵ-best arm discovery rate (softer metric than BDR)
LCB, UCB	the lower and upper confidence bounds used in the best-arm algorithms
$j \in \mathbb{N}$	experiment counter (number of MAB instances)
$T_j \in \mathbb{N}$	stopping time for the j-th experiment
$P_t^j, P_t \in [0,1]$	always valid p-value after time t (in experiment j, explicit or implicit)
P^j	always valid p-value for experiment j at its stopping time T_j
$\alpha_j \in [0,1]$	threshold set by the online FDR algorithm for P^j, using $\{p_i\}_{i=1}^{j-1}$
$T(\alpha_j) \in \mathbb{N}$	stopping time for the j-th experiment, when experiment uses α_j
0	the control or default arm
$\{1, \ldots, K\}$	$K = K(j)$ alternatives or treatment arms (experiment j implicit)
$i \in \{0, \ldots, K\}$	$K+1$ options or "all arms"
$i_\star, i_b$	the best of all arms, and the arm returned by MAB
μ_i, μ_*	the mean of the i-th arm, and the mean of the best arm
$t, n_i(t) \in \mathbb{N}$	total number of pulls, number of times arm i is pulled up to time t

Table 4.1: A summary of the notation used in this chapter.

Here the expectations are taken with respect to distributions of the arm pulls and the respective sampling algorithm. In general, it is not true that control of one quantity implies control of the other. Nevertheless, in the long run (when the law of large numbers is a good approximation), one does not expect a major difference between the two quantities in practice.

The set of true nulls $\mathcal{H}_0$ thus includes all experiments where H_0^j is true, and the FDR and mFDR are well-defined for any number of experiments J, since we often desire to control FDR(J) or mFDR(J) for all $J \in \mathbb{N}$. In order to measure power, we define the *ϵ-best-arm discovery rate* as

$$\epsilon\text{BDR}(J) := \frac{\mathbb{E} \sum_{j \in \mathcal{H}_1} R_j \mathbb{1}_{\mu_{i_b} \geq \mu_{i_\star} - \epsilon} \mathbb{1}_{\mu_{i_b} \geq \mu_0 + \epsilon}}{|\mathcal{H}_1(J)|} \tag{4.9}$$

We provide a concrete procedure 2 for our doubly sequential framework, where we use a particular online FDR algorithm due to Javanmard and Montanari [41] known as LORD (now called LORD 3 in the updated version); the reader should note that other online FDR procedure could be used to obtain essentially the same set of guarantees. Given a desired level α, the LORD procedure starts off with an initial "α-wealth" of $W(0) < \alpha$. Based on an infinite sequence $\{\gamma_i\}_{i=1}^{\infty}$ that sums to one, and the time of the most recent discovery τ_j, it

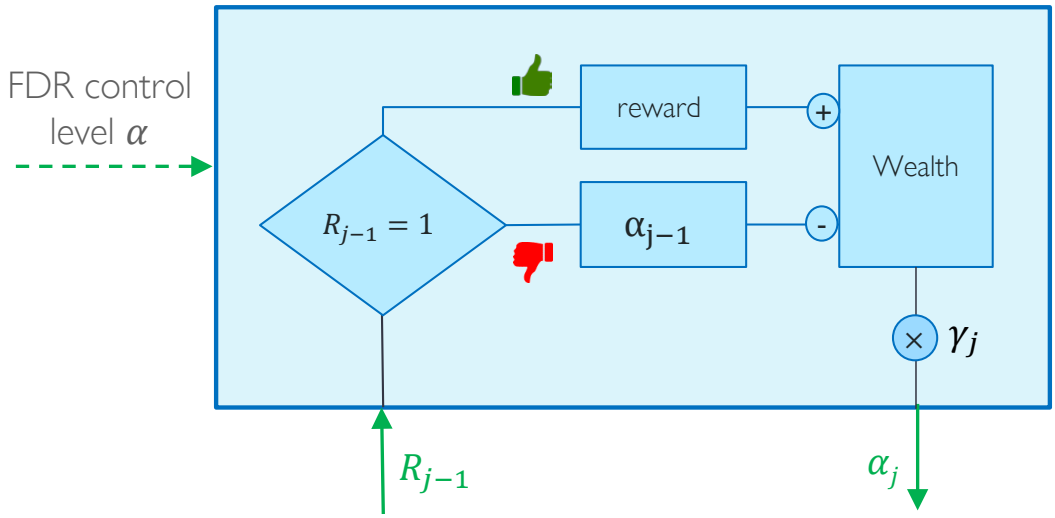


Figure 4.6: High-level cartoon of a LORD procedure as in [41] based on wealth to provide some intuition. Note that here, γ_j represents the fact that α_j corresponds to some fraction of the wealth which depends on j, and is not directly equivalent with the γ_j in the main text.

uses up a fraction $\gamma_{j-\tau_j}$ of the remaining α-wealth to test. Whenever there is a rejection, we increase the α-wealth by $\alpha - W(0)$. Figure 4.6 provides a cartoon to illustrate wealth-based online FDR procedures like LORD introduced in [41]. A feasible choice for a stopping time in practice is $T_j := \min\{T(\alpha_j), M\}$, where M is a maximal number of samples the scientist wants to pull and $T(\alpha_j)$ is the stopping time of the best-arm MAB algorithm run at confidence α_j.

Procedure 2 MAB-LORD: best-arm identification with online FDR control

1. Initialize $W(0) < \alpha$, set $\tau_0 = 0$, and choose a sequence $\{\gamma_i\}$ s.t. $\sum_{i=1}^{\infty} \gamma_i = 1$
2. At each step j, compute $\alpha_j = \gamma_{j-\tau_j} W(\tau_j)$ and $W(j+1) = W(j) - \alpha_j + R_j(\alpha - W(0))$
3. Output α_j and run Algorithm 1 using α_j-confidence and stop at a stopping time T_j.
4. Algorithm 1 returns P^j and we reject the null hypothesis if $P^j \leq \alpha_j$.
5. Set $R_j = \mathbb{1}_{P^j \leq \alpha_j}, \tau_j = \tau_{j-1} \vee jR_j$, update $j = j+1$ and go back to step 2.

The following theorem provides guarantees on mFDR and power for the MAB-LORD procedure.

Theorem 4 (Online mFDR control for MAB-LORD)**.**

(a) Procedure 2 achieves mFDR control at level α for stopping times $T_j = \min\{T(\alpha_j), M\}$.

(b) Furthermore, if we set $M = \infty$, Procedure 2 satisfies

$$\epsilon BDR(J) \geq \frac{\sum_{j=1}^{J} \mathbb{1}_{j \in \mathcal{H}_1}(1 - \alpha_j)}{|\mathcal{H}_1(J)|}. \tag{4.10}$$

The proof of this theorem can be found in Section 4.6. Note that by the arguments in the proof of Theorem 4, mFDR control itself is actually guaranteed for any generalized α-investing procedure [1] combined with any best-arm MAB algorithm. In fact we could use any adaptive stopping time T_j which depend on the history only via the rejections $R_1, \ldots, R_{j-1}$. Furthermore, using a modified LORD proposed by Javanmard and Montanari [40], we can also guarantee FDR control– which can be found in Section 4.4.

It is noteworthy that small values of α do not only guarantee smaller FDR error but also higher BDR. However, there is no free lunch — a smaller α implies a smaller α_j at each experiment, which in turn causes the best-arm MAB algorithm to employ a larger number of pulls in each experiment.

4.4 Notes on FDR control

Apart from mFDR for general online FDR procedures, it turns out that we can in fact prove FDR control for our framework using the specific online FDR procedure called LORD '15 introduced in [40]. When used in Procedure 2, the only adjustment that needs to be made is to reset $W(j+1)$ to α in step 2 after every rejection, yielding $\alpha_j = \alpha\gamma_{j-\tau_j}$ for any sequence $\{\gamma_j\}_{j=1}^{\infty}$ such that $\sum_{j=1}^{\infty} \gamma_j = 1$. We call the adjusted procedure MAB-LORD' for short.

Theorem 5 (Online FDR control for MAB-LORD)**.** *(a) MAB-LORD' achieves mFDR and FDR control at a specified level α for stopping times $T_j = \min\{T(\alpha_j), M\}$.*

(b) Furthermore, if we set $M = \infty$, MAB-LORD' satisfies

$$\epsilon BDR(J) \geq \frac{(1 - \alpha)}{|\mathcal{H}_1(J)|}. \tag{4.11}$$

Note that LORD as in [40] is less powerful than in [41] since the values α_j in the former can be much smaller than those in [41], which could in fact exceed the level α. Therefore, for FDR control we currently do have to sacrifice some power.

Proof. We leverage the proposition that can be obtained from a slightly more careful analysis of the procedure than in [40].

Proposition 3. *If $\mathbb{P}_0(P^j \leq \alpha_j \mid \tau_j) \leq \alpha_j$, i.e. the distribution of the p−values under the null are superuniform conditioned on the last rejection, using the online LORD'15 procedure controls the FDR at each t.*

Note that this proposition allows online FDR control for any, possibly dependent, p-values which are conditionally superuniform. This condition is not equivalent to (4.14) in general, it is in fact less restrictive since the probability is conditioned only on a function $\widetilde{\tau}_j = \max\{k \leq j : R_k = 1\}$ of all past rejections. Formally, the sigma algebra induced by τ_{j-1} is contained in $\mathcal{F}^{j-1}$ and hence $\mathbb{P}_0(P^j \leq \alpha_j \mid \tau_{j-1}) \leq \mathbb{P}_0(P^j \leq \alpha_j \mid R_1, \ldots, R_j)$ by the tower property. Finally, utilizing the fact that our p-values are conditionally super-uniform as proven in Section 4.6, i.e. inequality (4.14) holds, the condition for Proposition 3 is fulfilled and the proof is complete. □

Proof of Proposition 3

Let $\widetilde{\tau}_i$ denote the time of the i-th rejection with $\widetilde{\tau}_0 = 0$ (note that this is different from τ_j). and define $k(t) = \sum_{j=1}^{t} R_j$. Let H_j be the $j-$th hypothesis that was rejected. We adjust an argument from [40].

First observe that $\{k(t) = \ell\} = \{\widetilde{\tau}_\ell \leq t, \widetilde{\tau}_{\ell+1} > t\}$ and $FDP(t) = FDP(\widetilde{\tau}_{k(t)})$ so that

$$\begin{aligned}
\mathbb{E}FDP(t) = \mathbb{E}FDP(\tau_{k(t)}) &= \sum_{\ell=1}^{t} \mathbb{E}\big[\frac{\sum_{j\in\mathcal{H}_0} R_j}{\ell} \mid k(t) = \ell\big] P(k(t) = \ell) \\
&= \sum_{\ell=1}^{t} P(k(t) = \ell) \sum_{i=1}^{\ell} \mathbb{E}\big[\frac{\mathbb{1}_{H_i\in\mathcal{H}_0}}{\ell} \mid k(t) = \ell\big] \\
&= \sum_{\ell=1}^{t} P(k(t) = \ell) \sum_{i=1}^{\ell} \mathbb{E}\big[\mathbb{E}\big(\frac{\sum_{j=\widetilde{\tau}_{i-1}+1}^{\widetilde{\tau}_i} R_j \mathbb{1}_{j\in\mathcal{H}_0}}{\ell} \mid \widetilde{\tau}_0, \ldots, \widetilde{\tau}_{i-1}\big) \mid \widetilde{\tau}_\ell \leq t, \widetilde{\tau}_{\ell+1} > t\big]
\end{aligned}$$

Since for the LORD '15 procedure, we have $\alpha_t = \gamma_{t-\tau_t}$, and thus for all positive integers i, the random variables R_j with $j \geq \widetilde{\tau}_{i-1}$ are conditionally independent of $\widetilde{\tau}_0, \ldots, \widetilde{\tau}_{i-2}$ given $\widetilde{\tau}_{i-1}$. Additionally noting that $\widetilde{\tau}_{i-1} = \tau_j$ for all $j \geq \widetilde{\tau}_{i-1}$ by definition of $\widetilde{\tau}$ and τ, using $\mathbb{E}_0(\mathbb{1}_{p_j\leq\alpha_j} \mid \tau_j) \leq \alpha_j$ we obtain

$$\begin{aligned}
\mathbb{E}\big(\frac{\sum_{j\in(\widetilde{\tau}_{i-1},\widetilde{\tau}_i]\bigcap j\in\mathcal{H}_0} R_j}{\ell} \mid \widetilde{\tau}_0, \ldots, \widetilde{\tau}_{i-1}\big) &= \mathbb{E}\big(\frac{\sum_{j=\widetilde{\tau}_{i-1}+1}^{\widetilde{\tau}_i} R_j \mathbb{1}_{j\in\mathcal{H}_0}}{\ell} \mid \widetilde{\tau}_{i-1}\big) \\
&\leq \frac{\sum_{j=\tau_{i-1}+1}^{\tau_i} \mathbb{1}_{j\in H_0} \mathbb{E}[R_j \mid \tau_j]}{\ell} \\
&\leq \frac{\sum_{j=\tau_{i-1}+1}^{\tau_i} \alpha_j}{\ell} \leq \frac{\alpha}{\ell}.
\end{aligned}$$

The last inequality follows since between any two rejection times τ_k, τ_{k+1}, we have

$$\sum_{i=\tau_k}^{\tau_{k+1}} \alpha_i \leq \alpha \sum_{i=1}^{\infty} \gamma_i \leq \alpha.$$

Since $\sum_{\ell=1}^{t} P(k(t) = \ell) = 1$ it follows that FDR control is obtained.

4.5 Experimental results

In the following, we describe the results of experiments [‡] on both simulated and real-world data sets to illustrate the properties and guarantees of our procedure described in Section 4.3. In particular, we show that the mFDR is indeed controlled over time and that MAB-FDR (used interchangeably with MAB-LORD here) is highly advantageous in terms of sample complexity and power compared to a straightforward extension of A/B testing that is embedded in online FDR procedures. Unless otherwise noted, we set $\epsilon = 0$ in all of our simulations to focus on the main ideas and keep the discussion concise.

There are two natural frameworks to compare against MAB-FDR. The first, called AB-FDR or AB-LORD, swaps the MAB part for an A/B (i.e. A/B/n) test (uniformly sampling all alternatives until termination). The second comparator swaps the online FDR control for independent testing at α for all hypotheses – we call this MAB-IND. Formally, AB-FDR swaps step 3 in Procedure 2 with "*Output α_j and uniformly sample each arm until stopping time T_j.*" while MAB-IND swaps step 4 in Procedure 2 with "*The algorithm returns P^j and we reject the null hypothesis if $P^j \leq \alpha$.*". In order to compare the performances of these procedures, we ran three sets of simulations using Procedure 2 with $\epsilon = 0$ and $\gamma_j = 0.07 \frac{\log(j \vee 2)}{j e^{\sqrt{\log j}}}$ as in [41]. The first two sets are on artificial data (Gaussian and Bernoulli draws from sets of randomly drawn means μ_i), while the third is based on data from the New Yorker Cartoon Caption Contest (Bernoulli draws).

Our experiments are run on artificial data with Gaussian/Bernoulli draws and real-world Bernoulli draws from the New Yorker Cartoon Caption Contest. Recall that the sample complexity of the best-arm MAB algorithm is determined by the gaps $\Delta_j = \mu_{i_\star} - \mu_j$. One of the main relevant differences to consider between an experiment of artificial or real-world nature is thus the distribution of the means μ_i for $i = 1, \ldots, K$. The artificial data simulations are run with a fixed gap between the mean of the best arm $\mu_{i_\star}$ and second best arm μ_2, which we denote by $\Delta = \mu_{i_\star} - \mu_2$. In each experiment (hypothesis), the means of the other arms are set uniformly in $[0, \mu_2]$. For our real-world simulations with the cartoon contest, the means for the arms in each experiment are not arbitrary but correspond to empirical means from the caption contest. In addition, the contests actually follow a natural chronological order (see details below), which makes this dataset highly relevant to our purposes. In all simulations, 60% of all the hypotheses are true nulls, and their indices are chosen uniformly.

Power and sample complexity

The first set of simulations compares MAB-FDR against AB-FDR. They confirm that the total number of necessary pulls to determine significance (which we refer to as *sample complexity*) is much smaller for MAB-FDR than for AB-FDR. In the MAB-FDR framework, this also effectively leads to higher power given a fixed truncation time.

[‡]The code for reproducing all experiments and plots in this chapter is publicly available at `https://github.com/fanny-yang/MABFDR`

Two types of plots are used to demonstrate the superiority of our procedure: for one we fix the number of arms and plot the ϵBDR with $\epsilon = 0$ (which we call BDR for short) for both procedures over different choices of truncation times M. For the other we fix M and show how the sample complexity varies with the number of arms. Note that low BDR means that the bandit algorithm often reaches truncation time before it could stop.

Simulated Gaussian and Bernoulli trials

For the Gaussian draws, we set $\mu_{i_\star} = 8$. The gap to the second best arm is $\Delta = 3$ so that all means $\mu_{i \neq i_\star}$ are drawn uniformly between $Unif \sim [0, 5]$. The number of hypotheses is fixed to be 500. For Bernoulli draws we choose the maximum mean to be $\mu_{i_\star} = 0.4$, $\Delta = 0.3$ so that all means $\mu_{i \neq i_\star}$ are drawn uniformly between $Unif \sim [0, 0.1]$. The number of hypotheses is fixed at 50. We display the empirical average over 100 runs where each run uses the same hypothesis sequence (indicating which hypotheses are true and false) and sequence of means μ_i for each hypothesis. The only randomness we average over comes from the random Gaussian/Bernoulli draws which cause different rejections R_j and α_j, so that the randomness in each draw propagates through the online FDR procedure. The results can be seen in Figures 4.7 and 4.8.

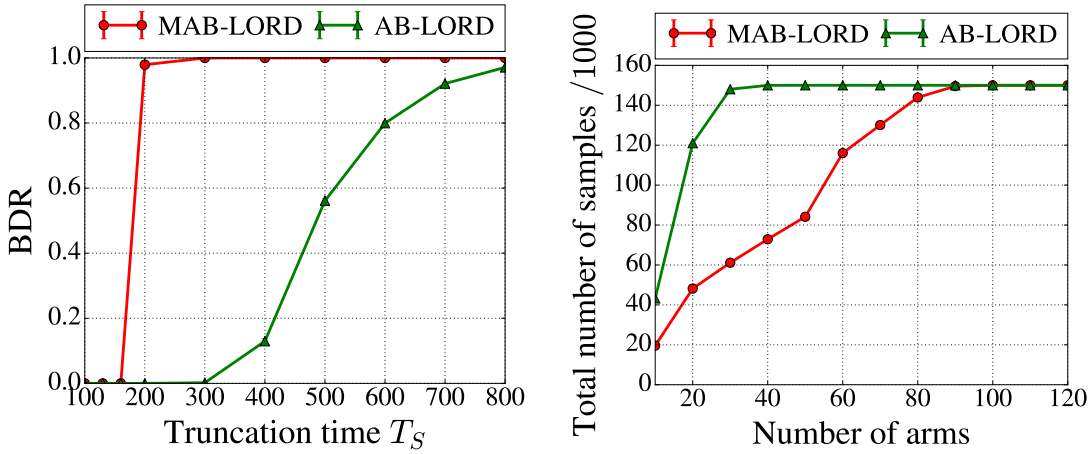


Figure 4.7: (a) Power vs. truncation time T_S (per hypothesis) for 50 arms and (b) Sample complexity vs. # arms for truncation time $M = 300$ for Gaussian draws with fixed $\mu_{i_\star} = 8$, $\Delta = 3$ over 500 hypotheses with 200 non-nulls, averaged over 100 runs.

The power at any given truncation time is much higher for MAB-FDR than AB-FDR. This is because the best-arm MAB is more likely to satisfy the stopping criterion before any given truncation time than the uniform sampling algorithm. The plot in Fig. 4.7(a) suggests that the actual stopping time of the algorithm is concentrated between 160 and 200 while it is much more spread out for the uniform algorithm.

The sample complexity plot in Fig. 4.7(b) qualitatively shows how the total number of necessary arm pulls for AB-FDR increases much faster with the number of arms than for the

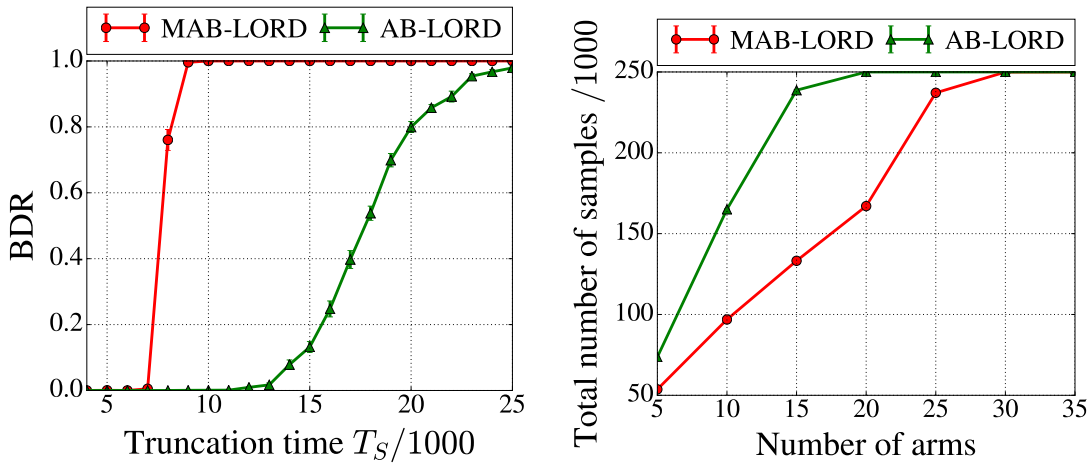


Figure 4.8: (a) Power over truncation time T_S (per hypothesis) for 50 arms and (b) Sample complexity over number of arms for truncation time $M = 5000$ for Bernoulli draws with fixed $\mu_{i_\star} = 0.7$, $\Delta = 0.3$ over 50 hypotheses with 20 non-nulls, averaged over 100 runs.

MAB-FDR, before it plateaus at the truncation time multiplied by the number of hypotheses. Recall that whenever the best-arm MAB stops before the truncation time in each hypothesis, the stopping criterion is met, i.e. the best arm is identified with probability at least $1 - \alpha_j$, so that the power is bound to be close to one whenever $T_j = T(\alpha_j)$.

For Bernoulli draws we choose the maximum mean to be $\mu_{i_\star} = 0.4$, $\Delta = 0.3$ so that all means $\mu_{i \neq i_\star}$ are drawn uniformly between $Unif \sim [0, 0.1]$. The number of hypotheses is fixed at 50. Otherwise the experimental setup is identical to those discussed in the main text for Gaussians. The plots for Bernoulli data can be found in Fig. 4.8.

The behavior for both Gaussian and Bernoullis are comparable, which is not surprising due to the choice of the subGaussian LIL bound. However one may notice that the choice of the gap of $\Delta = 3$ vs. $\Delta = 0.3$ drastically increases sample complexity so that the phase transition for power is shifted to very large T_S.

Application to New Yorker captions

In the simulations with real data we consider the crowd-sourced data collected for the *New Yorker Magazine's* Cartoon Caption contest: for a fixed cartoon, captions are shown to individuals online one at a time and they are asked to rate them as 'unfunny', 'somewhat funny', or 'funny'. We considered 30 contests[§] where for each contest, we computed the fraction of times each caption was rated as either 'somewhat funny' or 'funny'. We treat each caption as an arm, but because each caption was only shown a finite number of times in the dataset, we simulate draws from a Bernoulli distribution with the observed empirical mean

[§]Contest numbers 520-551, excluding 525 and 540 as they were not present. Full dataset and its description is available at `https://github.com/nextml/NEXT-data/`.

computed from the dataset. When considering subsets of the arms in any given experiment, we always use the captions with the highest empirical means (i.e. if $n = 10$ then we use the 10 captions that had the highest empirical means in that contest).

Although MAB-FDR still outperforms AB-FDR by a large margin, the plots in Figure 4.9 also show how the power and sample complexity notably differ from our toy simulation, where we seem to have chosen a rather benign distribution of means - in this setting, the gap Δ is much lower, often around ~ 0.01.

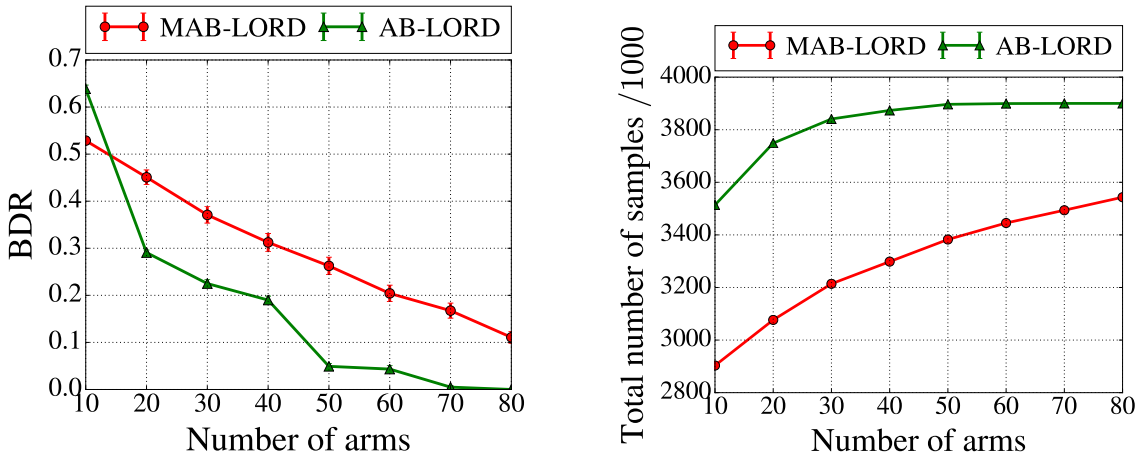
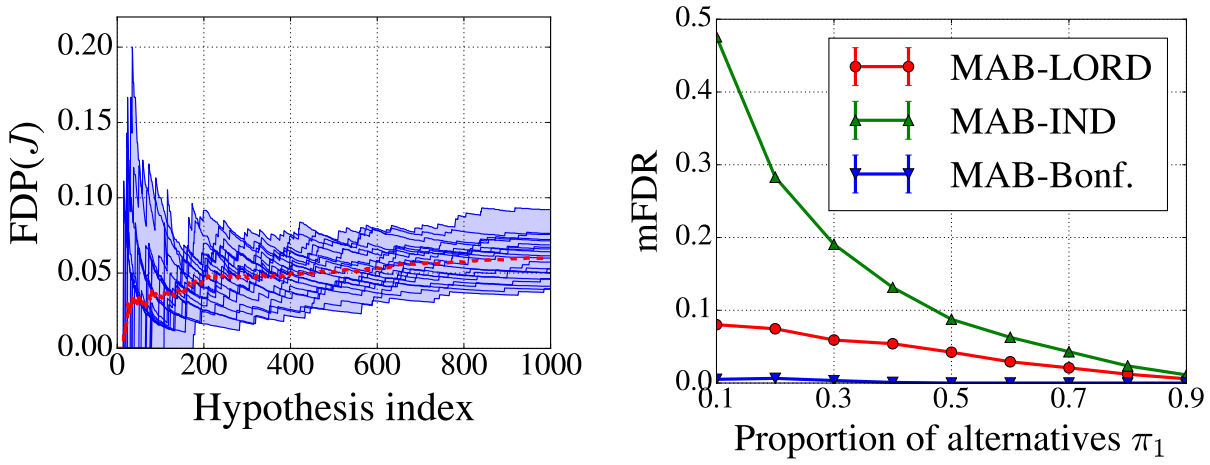


Figure 4.9: (a) BDR over number of arms, i.e. truncation time per hypothesis for 10 arms and (b) Sample complexity over number of arms for truncation time $M = 130000$ for Bernoulli draws, 30 hypotheses with 12 non-nulls and averaged over 100 runs.

mFDR and FDR control

In this section we use simulations to demonstrate the second part of our meta algorithm which deals with the control of the false discovery rate or its modified version. Since bandit algorithms have a very high best-arm discovery guarantee which in practice even exceeds its theoretical guarantee of at least $1 - \alpha_j$, mFDR and FDR plots on MAB-FDR directly do not lead to very insightful plots - namely the constant 0 line. However, we can demonstrate that even under adversarial conditions, i.e. when the P-value under the null is much less concentrated around one than obtained via the best arm bandit algorithm, mFDR or the false discovery proportion (FDP) in each run are still controlled *at any time* t as Theorem 4 guarantees. Albeit not exactly reflecting mFDR control in the case of MAB-FDR but in fact in an even harder setting, results from these experiments can be regarded as valuable on their own - it emphasizes the fact that Theorem 4 guarantees mFDR control independent of the adaptive sampling algorithm and specific choice of p-value as long as it is always valid.

For Figure 4.10, we again consider Gaussian draws with the same settings as described in 4.5. This time however, for each true null hypothesis we skip the bandit experiment and



(a) (b)

Figure 4.10: (a) Single runs of MAB-LORD (blue) and their average (red) with uniformly drawn p-values for null hypotheses and Gaussian draws for non-nulls with $\mu_{i_\star} = 8$, $\Delta = 3$ and $T_S = 200$, 500 hypotheses with 200 true nulls and 30 arms, the desired mFDR level is $\alpha = 0.1$ (b) mFDR over different proportions of non-nulls π_1, with same settings, averaged over 80 runs.

directly draw $P^j \sim [0, 1]$ to compare with the significance levels α_j from our online FDR procedure 2. As mentioned above, by Theorem 4, mFDR should still be controlled as it only requires the p-values to be super-uniform. In Figure 4.10(a) we plot the instantaneous false discovery proportion (number of false discoveries over total discoveries) $\mathrm{FDP}(J) = \frac{\sum_{j \in \mathcal{H}_0 J} R_j}{\sum_{j=1}^{T} R_j}$ over the hypothesis index for different runs with the same settings. Apart from fluctuations in the beginning due to the relatively small denominator, we can observe how the guarantee for the $\mathrm{FDR}(J) = \mathbb{E}\ \mathrm{FDP}(J)$, with its empirical value depicted by the red line, transfers to the control of each individual run (blue lines).

In Figure 4.10, we compare the mFDR (which in fact coincides with the FDR in this plot) of MAB-FDR using different multiple testing procedures, including MAB-IND and a Bonferroni type correction. The latter uses a simple union bound and chooses α_j such that $\sum_{j=1}^{\infty} \alpha_j \leq \alpha$ and thus trivially allows for any time FWER, and thus FDR control. In our simulations we use $\alpha_j = \frac{6\alpha}{\pi^2 j^2}$. As expected, Bonferroni is too conservative and barely makes any rejections whereas the naive MAB-IND approach does not control FDR. LORD avoids both extremes and controls FDR while having reasonable power.

4.6 Proofs

In this section we provide the proofs of the main results in the chapter.

Proof of Proposition 1

For any fixed $\gamma \in (0,1)$, we have the equivalence

$$\widehat{\mu}_{i,n_i(t)} - \varphi_{n_i(t)}(\tfrac{\gamma}{2K}) > \widehat{\mu}_{0,n_0(t)} + \varphi_{n_0(t)}(\tfrac{\gamma}{2}) + \epsilon \quad \Longleftrightarrow \quad p_{i,t} \leq \gamma.$$

If $\max_{i=1,\dots,K} \mu_i \leq \mu_0 + \epsilon$, then we have

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{i=1}^{K}\bigcup_{t=1}^{\infty}\left\{\widehat{\mu}_{i,n_i(t)} - \varphi_{n_i(t)}(\tfrac{\gamma}{2K}) > \widehat{\mu}_{0,n_0(t)} + \varphi_{n_0(t)}(\tfrac{\gamma}{2}) + \epsilon\right\}\right) \\
&\qquad = 1 - \mathbb{P}\left(\bigcap_{i=1}^{K}\bigcap_{t=1}^{\infty}\left\{\widehat{\mu}_{i,n_i(t)} - \varphi_{n_i(t)}(\tfrac{\gamma}{2K}) \leq \widehat{\mu}_{0,n_0(t)} + \varphi_{n_0(t)}(\tfrac{\gamma}{2}) + \epsilon\right\}\right) \\
&\qquad \leq 1 - \mathbb{P}\left(\bigcap_{t=1}^{\infty}\left\{\mu_0 \leq \widehat{\mu}_{0,t} + \varphi_t(\tfrac{\gamma}{2})\right\} \cap \bigcap_{i=1}^{K}\bigcap_{t=1}^{\infty}\left\{\widehat{\mu}_{i,n_i(t)} - \varphi_{n_i(t)}(\tfrac{\gamma}{2K}) \leq \mu_i\right\}\right) \\
&\qquad \leq \mathbb{P}\left(\bigcup_{t=1}^{\infty}\left\{\mu_0 > \widehat{\mu}_{0,t} + \varphi_t(\tfrac{\gamma}{2})\right\}\right) + \sum_{i=1}^{K}\mathbb{P}\left(\bigcup_{t=1}^{\infty}\left\{\widehat{\mu}_{i,n_i(t)} - \varphi_{n_i(t)}(\tfrac{\gamma}{2K}) > \mu_i\right\}\right) \\
&\qquad \leq \tfrac{\gamma}{2} + K\tfrac{\gamma}{2K} = \gamma
\end{aligned}$$

by equation (4.4). Thus, we have $\mathbb{P}\left(\bigcup_{i=1}^{K}\bigcup_{t=1}^{\infty}\left\{p_{i,t} \leq \gamma\right\}\right) \leq \gamma$, which completes the proof.

Proof of Proposition 2

Here we prove that the algorithm 1 terminates in finite time. The technical proof for sample complexity is moved to Section 4.8. It suffices to argue for $\delta/2 \leq 0.1$ and we discuss the other case at the end.

Proof of termination in finite time First we prove by contradiction that the algorithm terminates in finite time with probability one for the case $\mu_0 \geq \max_{i=1,\dots,K} \mu_i - \epsilon$.

Assuming that there exist runs for which the algorithm does not terminate, the set of arms defined by

$$S := \{i : \mathrm{LCB}_0(t) \leq \mathrm{UCB}_i(t) - \epsilon \text{ infinitely often (i.o.)}\}$$

is necessarily non-empty for these runs. We now show that this assumption yields a contradiction so that

$$\mathbb{P}(\text{Algorithm does not terminate}) \leq \mathbb{P}(\mathrm{LCB}_0(t) \leq \max_{i=1,\dots,K} \mathrm{UCB}_i(t) - \epsilon \text{ i.o.}) = 0 \qquad (4.12)$$

First take note that by definition of the algorithm, if an arm i is drawn infinitely often (i.o.), then so is the control arm 0 and we have $\mathrm{LCB}_0(t) \to \mu_0$ as well as $\mathrm{UCB}_i(t) \to \mu_i$ as

$t \to \infty$. This follows by the law of large numbers combined with the fact that $\varphi_{n_i(t)}, \varphi_{n_0(t)} \to 0$ as $t \to \infty$, since $\varphi_n \to 0$ as $n \to \infty$. Since for the null hypothesis we have $\mu_0 > \mu_i - \epsilon$, it follows that $\mathrm{LCB}_0(t) > \mathrm{UCB}_i(t) - \epsilon$ for all $t \geq t'$ for some t'.

This argument implies that all arms $i \in S$ can only be drawn a finite number of times, i.e. $n_i(t) < \infty$ for all $i \in S$. However, the fact that they are not drawn i.o. implies that $h_t \neq i$ and $\ell_t \neq i$ i.o. for all $i \in S$, so that there exists $i' \notin S$ such that $\max_{i \in S} \mathrm{UCB}_i(t) \leq \mathrm{UCB}_{i'}(t)$ i.o. By definition of S we then obtain

$$\mathrm{LCB}_0(t) \leq \mathrm{UCB}_{i'}(t) - \epsilon \text{ i.o.} \tag{4.13}$$

However, since $i' \notin S$, inequality (4.13) cannot hold and equation (4.12) is proved.

A nearly identical argument to the above shows that the stopping condition is met in finite time.

Proof of Theorem 4

We now turn to the proof of Theorem 4, splitting our argument into parts (a) and (b), respectively.

Proof of part (a)

In order for generalized alpha-investing procedures such as LORD to successfully control the mFDR, it is sufficient that p-values under the null be *conditionally super-uniform*, meaning that for all $j \in \mathcal{H}_0$, we have

$$\mathbb{P}_0(P^j \leq \alpha_j | \mathcal{F}^{j-1}) \leq \alpha_j(R_1, \ldots, R_{j-1}) \tag{4.14}$$

where $\mathcal{F}^{j-1}$ is the σ-field induced by $R_1, \ldots, R_{j-1}$. Note that as long as condition (4.14) is satisfied, T_j and thus P^j could potentially depend on α_j, i.e. the rejection indicator variables $R_1, \ldots, R_{j-1}$ and potentially $P^1, \ldots, P^{j-1}$. See Aharoni and Rosset [1] for further details.

It thus suffices to show that condition (4.14) holds for our definition of p-value in our framework. We know that by Proposition 1 we have for any random stopping time, thus any fixed truncation time M, that $\mathbb{P}_0(P^j_T \leq \alpha_j) \leq \alpha_j$. We now show that the same bound also holds for the (α_j-dependent) bandit stopping time $T(\alpha_j)$, i.e. that $\mathbb{P}_0(P^j_{T(\alpha_j)} \leq \alpha_j) \leq \alpha_j$.

Under the null hypothesis, the best arm is at most ϵ better than the control arm, i.e. $\mu_0 > \mu_i - \epsilon$, so that by Proposition 2 we have that with probability $\geq 1 - \alpha_j$, $i_b = 0$, i.e. $\mathrm{LCB}_0(t) > \mathrm{UCB}_i(t) - \epsilon$ for all $i \neq 0$. Hence, $\mathrm{LCB}_i(t) - \mathrm{UCB}_0(t) < \epsilon$, and thus, by the definition of the p-values, $P^j_{i,T(\alpha_j)} = 1$ for all i with probability $\geq 1 - \alpha_j$. It finally follows that $\mathbb{P}_0(P^j_{T(\alpha_j)} \leq \alpha_j) \leq \alpha_j$.

Putting things together, under the true null hypothesis (omitting the index $j \in \mathcal{H}_0$ to simplify notation) we directly have that for any α_j

$$\begin{aligned}\mathbb{P}_0(P^j_{T_j}(\alpha_j) \leq \alpha_j) &= \mathbb{P}_0\big(P^j_{T(\alpha_j)} \leq \alpha_j \big| T(\alpha_j) \leq M\big)\mathbb{P}_0(T(\alpha_j) \leq M) \\ &\quad + \mathbb{P}_0\big(P^j_M \leq \alpha_j \big| T(\alpha_j) > M\big)\mathbb{P}_0(T(\alpha_j) > M) \\ &\leq \alpha_j(\mathbb{P}_0(T(\alpha_j) \leq M) + \mathbb{P}_0(T(\alpha_j) > M)) = \alpha_j\end{aligned}$$

for all fixed α_j even when the stopping time $T(\alpha_j)$ is dependent on α_j. This is equivalent to stating that for any sequence $R_1, \ldots, R_{j-1}$ we have

$$\begin{aligned}\mathbb{P}_0(P^j \leq \alpha_j(R_1, \ldots, R_{j-1}) | \mathcal{F}^{j-1}) &= \mathbb{P}_0(P^j_{T(\alpha_j(R_1,\ldots,R_{j-1}))} \leq \alpha_j(R_1, \ldots, R_{j-1})) \\ &\leq \alpha_j(R_1, \ldots, R_{j-1})\end{aligned}$$

and the proof is complete.

Proof of part (b)

It suffices to prove that for a single experiment j and $M = \infty$, we have $\mathbb{P}_1(P^j_{T(\alpha_j)} \leq \alpha_j) \geq 1 - \alpha_j$ where $\mathbb{P}_1$ is the distribution of a non-null experiment j. First observe that at stopping time $T(\alpha_j)$ of Algorithm 1, either $P^j_{i,T(\alpha_j)} \leq \alpha_j$ or $P^j_{i,T(\alpha_j)} = 1$ for all i. The former event happens whenever the algorithm exits with $i_b \in \mathcal{S}^\star$, i.e. when $\mathrm{LCB}_{i_b}(t) \geq \mathrm{UCB}_{\ell_t}(t) - \epsilon$ holds. Then, by definition of the p-value in equation (4.6) and ℓ_t we must have that $P^j_{i_b,T(\alpha_j)} \leq \alpha_j$. As a consequence, by Proposition 2, we have

$$\begin{aligned}\mathbb{P}_1(P^j_{T(\alpha_j)} \leq \alpha_j) &\geq \mathbb{P}(P^j_{T(\alpha_j)} \leq \alpha_j) \\ &\geq \mathbb{P}_1(\text{Algorithm 1 exits with } i_b \in \mathcal{S}^\star) \\ &\geq 1 - \alpha_j\end{aligned}$$

and the proof is complete.

4.7 Discussion

The recent focus in popular media about the lack of reproducibility of scientific results erodes the public's confidence in published scientific research. To maintain high standards of published results and claimed discoveries, simply increasing the statistical significance standards of each individual experimental work (e.g., reject at level 0.001 rather than 0.05) would drastically hurt power. We take the alternative approach of controlling the ratio of false discoveries to claimed discoveries at some desired value (e.g., 0.05) over many sequential experiments. This means that the statistical significance for validating a discovery changes from experiment to experiment, and could be larger or smaller than 0.05, requiring less or more data to be collected. Unlike earlier works on online FDR control, our framework synchronously interacts with adaptive sampling methods like MABs over uniform sampling to

make the overall sampling procedure as efficient as possible. We do not know of other works in the literature combining the benefits of adaptive sampling and FDR control. It should be clear that any improvement, theoretical or practical, to either online FDR algorithms or best-arm identification in MAB (or their variants), immediately results in a corresponding improvement for our MAB-FDR framework.

More general notions of FDR with corresponding online procedures have recently been developed by Ramdas et al [63]. In particular, they incorporate the notion of memory and a priori importance of each hypothesis. This could prove to be a valuable extension for our setting, especially in cases when only the percentage of wrong rejections in the recent past matters. It would be useful to establish FDR control for these generalized notions of FDR as well.

There are several directions that could be explored in future work. First, it would be interesting to extend the MAB aspect (in which each arm is univariate) of our framework to more general settings. Balasubramani and Ramdas [4] show how to construct sequential tests for many multivariate nonparametric testing problems, using LIL confidence intervals, which can again be inverted to provide always valid p-values. It might be of interest to marry the ideas in our chapter with theirs. For example, the null hypothesis might be that the control arm has the same (multivariate) mean as other arms (K-sample testing), and under the alternative, we would like to pick the arm whose mean is furthest away from the control. A more complicated example could involve dependence, where we observe pairs of arms, and the null hypothesis is that the rewards in the control arm are independent of the alternatives, and if the null is false we may want to pick the most correlated arm. The work by Zhao et al. [93] on tightening LIL-bounds could be practically relevant. Recent work on sequential p-values by Malek et al. [54] also naturally fit into our framework. Lastly, in this work we treat samples or pulls from arms as identical from a statistical perspective; it might be of interest in subsequent work to extend our framework to the contextual bandit setting, in which the samples are associated with features to aid exploration.

4.8 Proof of sample complexity for Proposition 2

In the sequel we use $\gtrsim, \sim$ for inequality and equality up to constant factors.

Define $i_\star = \arg\max_{i=0,1,\ldots,K} \mu_i$ (breaking ties arbitrarily) and $n_i(t)$ to be the number of times sample i was drawn until time t. For any $i \in \{0, 1, \ldots, K\}$ and $\eta \in \mathbb{R}$ we define the following key quantity

$$\begin{aligned} \tau_i(\eta, \xi) &:= \min\{n \in \mathbb{N} : 2\varphi_n(\tfrac{\delta}{2K}) < \max\{|\eta - \mu_i|, \xi\}\} \\ &\lesssim \min\left\{(\eta - \mu_i)^{-2} \log(K \log(\eta - \mu_i)^{-2})/\delta), \xi^{-2} \log(K \log(\xi^{-2})/\delta)\right\} \end{aligned} \tag{4.15}$$

where we set $\tau_i(\mu_i, 0) = \infty$, but this case does not arise in our analysis.

Let us define the events

$$\mathcal{E}_i = \bigcap_{n=1}^{\infty} \{|\widehat{\mu}_{i,n} - \mu_i| \le \varphi_n(\tfrac{\delta}{2K})\}.$$

By a union bound and the LIL bound in (4.4), we have for $\delta/2K < 0.1$ that

$$\mathbb{P}\left(\bigcup_{i=0}^{K} \mathcal{E}_i^c\right) \leq \frac{K+1}{2K}\delta \leq \delta$$

for $K \geq 2$. For $\frac{\delta}{2K} > 0.1$, note that for all $\delta' < \delta$ we have $\varphi_n(\delta') \leq \varphi_n(\delta)$ so that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_i^c) &= \mathbb{P}(\varphi_n(\tfrac{\delta}{2K}) < \widehat{\mu}_{i,n} - \mu_i) \\ &\leq \mathbb{P}(\varphi_n(0.1) < \widehat{\mu}_{i,n} - \mu_i) \leq \tfrac{\delta}{2K} \qquad \forall i = 1, \ldots, K\end{aligned}$$

Throughout the rest of the proof we assume the events $\mathcal{E}_i$ hold.

The following simple lemma regarding the key quantity τ_i will be used throughout the proof.

Lemma 21. *Fix $i \in \{0, 1, \ldots, K\}$ and $\eta > 0$. For any $t \in \mathbb{N}$, whenever $n_i(t) \geq \tau_i(\eta, \xi)$ we have that under the event $\bigcap_{i=0,\ldots,K} \mathcal{E}_i$, we have*

$$\begin{aligned}UCB_i(t) &\leq \max\{\eta, \mu_i + \xi\} \text{ if } \eta \geq \mu_i \\ LCB_i(t) &\geq \min\{\eta, \mu_i - \xi\} \text{ if } \eta \leq \mu_i\end{aligned}$$

Proof. Assume $n_i(t) \geq \tau_i(\eta, \xi)$. If $\eta \geq \mu_i$ we have by definition of $\mathcal{E}_i$ that

$$\mathrm{UCB}_i(t) = \widehat{\mu}_{i,n_i(t)} + \varphi_{n_i(t)}(\tfrac{\delta}{2}) \leq \mu_i + 2\varphi_{n_i(t)}(\tfrac{\delta}{2K}) < \mu_i + \max\{\eta - \mu_i, \xi\}$$

and if $\eta \leq \mu_i$

$$\begin{aligned}\mathrm{LCB}_i(t) &= \widehat{\mu}_{i,n_i(t)} - \varphi_{n_i(t)}(\tfrac{\delta}{2K}) \geq \mu_i - 2\varphi_{n_i(t)}(\tfrac{\delta}{2K}) \\ &> \mu_i - \max\{\mu_i - \eta, \xi\} = \mu_i + \min\{\eta - \mu_i, -\xi\}.\end{aligned}$$

□

Proof of Proposition 2 (a) $\mu_0 > \max_{i=1,\ldots,K} \mu_i - \epsilon$

At each time t which does not satisfy the stopping condition, arm 0 and $\arg\max_{i=1,\ldots,K} \mathrm{UCB}_i(t)$ are pulled. Note that by Lemma 21

$$\{n_0(t) \geq \tau_0(\tfrac{\mu_0 + (\max_{i=1,\ldots,K} \mu_i - \epsilon)}{2}, 0)\} \implies \mathrm{LCB}_0(t) \geq \min\{\tfrac{\mu_0 + (\max_{i=1,\ldots,K} \mu_i - \epsilon)}{2}, \mu_0\} \geq \tfrac{\mu_0 + (\max_{i=1,\ldots,K} \mu_i - \epsilon)}{2} \tag{4.16}$$

so that $t > n_0(t)$ makes sure that there were enough draws for the particular arm 0 (since it's drawn every time). For $i \neq 0$ we have

$$\{n_i(t) \geq \tau_i(\tfrac{(\mu_0 + \epsilon) + \max_{i=1,\ldots,K} \mu_i}{2}, 0)\} \implies \mathrm{UCB}_i(t) \leq \max\{\tfrac{(\mu_0 + \epsilon) + \max_{i=1,\ldots,K} \mu_i}{2}, \mu_i\} \leq \tfrac{(\mu_0 + \epsilon) + \max_{i=1,\ldots,K} \mu_i}{2}. \tag{4.17}$$

which makes $t > \sum_{i=0}^{K} n_i(t)$ a necessary condition.

Reversely whenever $t > \sum_{i=0}^{K} n_i(t)$, for all arms $i \neq 0$ we have $\mathrm{UCB}_i(t) \leq \frac{(\mu_0+\epsilon)+\max_{i=1,\ldots,K}\mu_i}{2}$. In essence, once arm i has been sampled $n_i(t)$ times, because of (4.17), it will not be sampled again - either, because all of the other $UCB_i(t)$ satisfy the same upper bound, the algorithm will have stopped, or, if for some i we have $UCB_i(t) > \frac{(\mu_0+\epsilon)+\max_{i=1,\ldots,K}\mu_i}{2}$ that will be the arm that is drawn. Thus,

$$\begin{aligned}&\{t \geq B_1(\mu,\delta) := \tau_0(\tfrac{\mu_0+(\max_{i=1,\ldots,K}\mu_i-\epsilon)}{2},0) + \sum_{i=1}^{K}\tau_i(\tfrac{(\mu_0+\epsilon)+\max_{i=1,\ldots,K}\mu_i}{2},0)\}\\ &\implies \{\mathrm{LCB}_0(t) - \mathrm{UCB}_i(t) \geq -\epsilon \quad \forall i \neq 0\},\end{aligned}$$

i.e., the stopping condition is met, where the first term accounts for satisfying (4.16), the second term accounts for satisfying (4.17) for all $i \neq 0$, and the third term accounts for satisfying Equation (4.18). Denoting $T(\delta)$ as the stopping time of the algorithm, this implies that with probability at least $1-\delta$, we have $T(\delta) \leq B_1(\mu,\delta)$ and arm 0 is returned.

Let us now simplify the expression to make it more accessible to the reader and arrive at the theorem statement. Defining $\widetilde{\Delta}_i := \max\{|\eta-\mu_i|,\xi\}$ as the *effective gap* in the definition of $\tau_i(\eta,\xi)$ in Equation (4.15), it is straightforward to verify that the effective gap associated with arm 0 is equal to

$$\widetilde{\Delta}_0 \sim (\mu_0+\epsilon) - \max_{j=1,\ldots,K}\mu_j,$$

and the effective gap for any other arm i is equal to

$$\widetilde{\Delta}_i \gtrsim (\mu_0+\epsilon) - \mu_i.$$

Using these quantities, we can see that the upper bound $B_1(\mu,\delta)$ scales like

$$\sum_{i=0}^{K} \widetilde{\Delta}_i^{-2}\log(K\log(\widetilde{\Delta}_i^{-2})/\delta).$$

Proof of Proposition 2 (b) $\max_{i=1,\ldots,K}\mu_i = \mu_{i_\star} > \mu_0+\epsilon$

At each time t which does not satisfy the stopping condition, arm 0 is pulled. Note again that by Lemma 21

$$\{n_0(t) \geq \tau_0(\tfrac{(\mu_{i_\star}-\epsilon)+\mu_0}{2},0)\} \implies \mathrm{UCB}_0(t) \leq \max\{\tfrac{(\mu_{i_\star}-\epsilon)+\mu_0}{2},\mu_0\} \leq \frac{(\mu_{i_\star}-\epsilon)+\mu_0}{2}.$$

The following claim is key to proving this case (where $u \in (0,1)$ be an absolute constant to be defined later).

Claim 1. *Under the event* $\bigcap_{i=0,\dots,K} \mathcal{E}_i$*, for any* $u \leq \frac{2}{7}$ *and* $\bar{\mu} \in [\max_{j \neq i_\star} \mu_j, \mu_{i_\star}]$*, we have*

$$|\{s \geq 2\sum_{i=0}^{K} \tau_i(\bar{\mu}, u\epsilon) : LCB_{h_s}(s) \leq \mu_{i_\star} - \tfrac{5}{2}u\epsilon \text{ or } UCB_{\ell_s}(s) \geq \mu_{i_\star} + u\epsilon\}| < \sum_{i=0}^{K} \tau_i(\bar{\mu}, u\epsilon) \tag{4.18}$$

The proof of this claim can be found in Section 4.8. Note that for all s we have that

$$\mathrm{LCB}_{h_s}(s) \geq \mu_{i_\star} - \tfrac{5}{2}u\epsilon \text{ and } \mathrm{UCB}_{\ell_s}(s) \leq \mu_{i_\star} + u\epsilon \implies \mathrm{LCB}_{h_s}(s) \geq \mathrm{UCB}_{\ell_s}(s) - \epsilon.$$

Intuitively the inequality (4.18) thus limits the number of times that for $t \geq 2\sum_{i=0}^{K} \tau_i(\bar{\mu}, u\epsilon)$, the criterion $\mathrm{LCB}_{h_s}(s) \geq \mathrm{UCB}_{\ell_s}(s) - \epsilon$ is not fulfilled. We refer to the times when the condition on the left hand side of inequality (4.18) is fulfilled, as "good" times.

Applying Claim 1 with $\bar{\mu} = \max_{j \neq i_\star} \frac{\mu_{i_\star} + \mu_j}{2}$ and $u = \frac{\mu_{i_\star} - (\mu_0 + \epsilon)}{5\epsilon}$ we then observe that on the "good" times, we have

$$\mathrm{LCB}_{h_t} \geq \mu_{i_\star} - \tfrac{5}{2}u\epsilon = \frac{\mu_{i_\star} + (\mu_0 + \epsilon)}{2} = \frac{(\mu_{i_\star} - \epsilon) + \mu_0}{2} + \epsilon,$$

so that we directly obtain that with probability at least $1 - \delta$,

$$T(\delta) \leq B_2(\mu, \delta) := \tau_0(\tfrac{(\mu_{i_\star} - \epsilon) + \mu_0}{2}, 0) + 3\sum_{i=0}^{K} \tau_i(\max_{j \neq i_\star} \tfrac{\mu_{i_\star} + \mu_j}{2}, \min\{\tfrac{2}{7}\epsilon, \tfrac{\mu_{i_\star} - (\mu_0 + \epsilon)}{5}\}).$$

Let us now simplify the expression. It is straightforward to verify that the effective gap associated with arm 0 is equal to

$$\begin{aligned}\widetilde{\Delta}_0 &\gtrsim \min\left\{\tfrac{\mu_{i_\star} - (\mu_0 + \epsilon)}{2}, \max\left\{\max_{j \neq i_\star} \tfrac{\mu_{i_\star} + \mu_j}{2} - \mu_0, \tfrac{2}{7}\epsilon\right\}\right\} \\ &\gtrsim \min\left\{\mu_{i_\star} - (\mu_0 + \epsilon), \max\{\Delta_0, \frac{4}{7}\epsilon\}\right\}\end{aligned}$$

and the effective gap for any other arm i is equal to

$$\begin{aligned}\widetilde{\Delta}_i &= \max\left\{|\max_{j \neq i_\star} \tfrac{\mu_{i_\star} + \mu_j}{2} - \mu_i|, \min\{\tfrac{2}{7}\epsilon, \tfrac{\mu_{i_\star} - (\mu_0 + \epsilon)}{5}\}\right\} \\ &\gtrsim \max\left\{\Delta_i, \min\left\{\mu_{i_\star} - (\mu_0 + \epsilon), \epsilon\right\}\right\}\end{aligned}$$

where we recall that $\Delta_i = \mu_{i_\star} - \mu_i$ if $i \neq i_\star$, and $\Delta_{i_\star} = \mu_{i_\star} - \max_{j \neq i_\star} \mu_j$ otherwise. Using these quantities, the upper bound $B_2(\mu, \delta)$ on the stopping time $T(\delta)$ scales like $\sum_{i=0}^{K} \widetilde{\Delta}_i^{-2} \log(K \log(\widetilde{\Delta}_i^{-2})/\delta)$. This concludes the proof of the proposition.

Proof of Claim 1

Let $\bar{\mu} \in [\max_{j \neq i_\star} \mu_j, \mu_{i_\star}]$ and $\tau_i := \tau_i(\bar{\mu}, u\epsilon)$. The following result is a a key ingredient for the proof of the claim.

Proposition 4. *For any time t and $u \leq 1/2$,*

$$\begin{aligned} &\Big\{|\{s \leq t : h_s = i_\star\}| \geq \sum_{i=0}^{K} \tau_i\Big\} \\ &\implies \{UCB_{\ell_t}(t) \leq \bar{\mu} + u\epsilon\} \cap \{LCB_{h_t}(t) \geq \bar{\mu} - u\epsilon\} \\ &\implies \{LCB_{h_t}(t) - UCB_{\ell_t}(t) \geq -\epsilon\}. \end{aligned}$$

Proof. If $h_s = i_\star$ then *some* $i \neq i_\star$ is assigned to ℓ_s and $\mathrm{UCB}_i(s) \leq \max\{\bar{\mu}, \mu_i + u\epsilon\} \leq \bar{\mu} + u\epsilon$ whenever $n_i(s) \geq \tau_i(\bar{\mu}, u\epsilon)$. Because ℓ_s is the highest upper confidence bound, the sum over all τ_i represents exhausting all arms (i.e., pigeonhole principle). An analogous result holds for $\mathrm{LCB}_{i_\star}(t)$. □

A direct consequence of Proposition 4 is that even though we don't know which arm will be assigned to h_t at any given time t, we do know that if $h_t = i_\star$ for a sufficient number of times, namely $\sum_{i=0}^{K} \tau_i$ times, the termination criteria will be met. Thus, assume $h_t \neq i_\star$ and note that

$$\begin{aligned} \{h_t = i,\, \mu_i < \mu_{i_\star} - \tfrac{5}{2}u\epsilon,\, \widehat{\mu}_{i,n_i(t)} \geq \min\{\bar{\mu}, \mu_{i_\star} - \tfrac{3}{2}u\epsilon\}\} \\ \implies \min\{\bar{\mu}, \mu_{i_\star} - \tfrac{3}{2}u\epsilon\} \leq \widehat{\mu}_{i,n_i(t)} \leq \mu_i + \varphi_{n_i(t)}(\tfrac{\delta}{2K}) \\ \implies \{n_i(t) < \tau_i\} \end{aligned}$$

where the last line follows from $\mu_i + \varphi_{n_i(t)}(\frac{\delta}{2K}) < \min\{\bar{\mu}, \mu_i + u\epsilon\} \leq \min\{\bar{\mu}, \mu_{i_\star} - \frac{3}{2}u\epsilon\}$ whenever $n_i(t) \geq \tau_i$. Furthermore, the following Proposition 5, says for $t \geq 2\sum_{i=0}^{K} \tau_i$ we have that $\widehat{\mu}_{h_t, n_{h_t}(t)} \geq \min\{\bar{\mu}, \mu_{i_\star} - \frac{3}{2}u\epsilon\}$.

Proposition 5. *For any time t,*

$$\{t \geq 2\sum_{i=0}^{K} \tau_i\} \implies \{\widehat{\mu}_{h_t, n_{h_t}(t)} \geq \min\{\bar{\mu}, \mu_{i_\star} - \tfrac{3}{2}u\epsilon\}\}.$$

The proof of the proposition can be found in Section 4.8.

Combining this fact with the display immediately above and the observation that some $i = h_t$, we have that $|\{s \geq 2\sum_{i=0}^{K} \tau_i : \mu_{i_\star} - \mu_{h_s} \geq \frac{5}{2}u\epsilon\}| < \sum_{i=0}^{K} \tau_i$. Now, on one of these times t such that $\{h_t = i, n_i(t) \geq \tau_i, \mu_{i_\star} - \mu_i < \frac{5}{2}u\epsilon\}$, we have

$$\mathrm{LCB}_i(t) = \widehat{\mu}_{i,n_i(t)} - \varphi_{n_i(t)}(\tfrac{\delta}{2K}) \geq \mu_i - 2\varphi_{n_i(t)}(\tfrac{\delta}{2K}) \geq \min\{\bar{\mu}, \mu_i - u\epsilon\} \geq \mu_{i_\star} - \tfrac{5}{2}u\epsilon.$$

The above display with the next proposition completes the proof of Equation 4.18.

Proposition 6. *For any time t,*

$$\{t \geq \sum_{i=0}^{K} \tau_i\} \implies \{\max_{i=0,1,\ldots,K} UCB_i(t) \leq \mu_{i_\star} + u\epsilon\}.$$

Proof. Note that

$$\begin{aligned}\{\mathrm{UCB}_i(t) \geq \mu_{i_\star} + u\epsilon\} &\implies \{\mu_{i_\star} + u\epsilon \leq \mathrm{UCB}_i(t) = \widehat{\mu}_{i,n_i(t)} + \varphi_{n_i(t)}(\tfrac{\delta}{2}) \leq \mu_i + 2\varphi_{n_i(t)}(\tfrac{\delta}{2K})\} \\ &\implies \{n_i(t) < \tau_i\}\end{aligned}$$

since $\mu_i + 2\varphi_{n_i(t)}(\frac{\delta}{2K}) < \max\{\bar{\mu}, \mu_i + u\epsilon\} \leq \mu_{i_\star} + u\epsilon$ whenever $n_i(t) \geq \tau_i$. Now, because at each time t, the arm $\arg\max_{j=0,1,\ldots,K} \mathrm{UCB}_j(t)$ is pulled because it is either h_t or ℓ_t, we conclude that this arm can only be pulled τ_i times before satisfying $\mathrm{UCB}_i(t) \leq \mu_{i_\star} + u\epsilon$. □

Proof of Proposition 5

The above proposition implies,

$$\{t \geq 2\sum_{i=0}^{K} \tau_i\} \implies \left\{|\{s \leq t : h_s \neq i_\star\}| \geq \sum_{i=0}^{K} \tau_i\right\}.$$

Now consider the event

$$\begin{aligned}\{h_t \neq i_\star, \ell_t = i\} &\implies \mu_{i_\star} \leq \widehat{\mu}_{i_\star,n_{i_\star}(t)} + \varphi_{n_{i_\star}(t)}(\tfrac{\delta}{2}) \leq \widehat{\mu}_{i,n_i(t)} + \varphi_{n_i(t)}(\tfrac{\delta}{2}) \leq \mu_i + 2\varphi_{n_i(t)}(\tfrac{\delta}{2K}) \\ &\implies \{\mu_{i_\star} - \mu_i \leq 2\varphi_{n_i(t)}(\tfrac{\delta}{2K})\} \\ &\implies \{n_i(t) < \tau_i\} \cup \{n_i(t) \geq \tau_i, \mu_{i_\star} - \mu_i \leq 2\varphi_{n_i(t)}(\tfrac{\delta}{2K})\} \\ &\implies \{n_i(t) < \tau_i\} \cup \{n_i(t) \geq \tau_i, \mu_{i_\star} - \mu_i \leq \max\{|\bar{\mu} - \mu_i|, u\epsilon\}\} \\ &\implies \{n_i(t) < \tau_i\} \cup \{n_i(t) \geq \tau_i, \mu_{i_\star} - \mu_i < u\epsilon\} \cup \{n_i(t) \geq \tau_i, i = i_\star\}\end{aligned}$$

by the definition of τ_i. Because at each time $s \leq t$ we have that *some* $i = \ell_s$, if $|\{s \leq t : h_s \neq i_\star\}| \geq \sum_{i=0}^{K} \tau_i$, we have that

$$\{t \geq 2\sum_{i=0}^{K} \tau_i\} \implies \{\exists i : n_i(t) \geq \tau_i \text{ and } \mu_{i_\star} - \mu_i < u\epsilon\} \cup \{n_i(t) \geq \tau_i \text{ and } i = i_\star\}.$$

We use the fact that such an $\ell_t = i \neq i_\star$ exists that satisfies $\mu_{i_\star} - \mu_i < u\epsilon$ to say

$$\exists i \neq i_\star : \widehat{\mu}_{i,n_i(t)} \geq \mu_i - \varphi_{n_i(t)}(\tfrac{\delta}{2K}) \geq \mu_i - \max\{\mu_{i_\star} - \mu_i, u\epsilon\}/2 \geq \mu_{i_\star} - \tfrac{3}{2}u\epsilon$$

or $\ell_t = i_\star$ and

$$\widehat{\mu}_{i_\star,n_{i_\star}(t)} \geq \mu_{i_\star} - \varphi_{n_{i_\star}(t)}(\tfrac{\delta}{2K}) \geq \mu_{i_\star} - \max\{\mu_{i_\star} - \bar{\mu}, u\epsilon\}/2 = \min\{\bar{\mu}, \mu_{i_\star} - \tfrac{1}{2}u\epsilon\}.$$

Because $\widehat{\mu}_{h_t,n_{h_t}(t)} \geq \max_{i=0,1,\ldots,K} \widehat{\mu}_{i,n_i(t)}$, the proof of the claim is complete.

Bibliography

[1] E. Aharoni and S. Rosset. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794, 2014.

[2] R. S. Anderssen and P. M. Prenter. A formal comparison of methods proposed for the numerical solution of first kind integral equations. *Jour. Australian Math. Soc. (Ser. B)*, 22:488–500, 1981.

[3] S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

[4] A. Balsubramani and A. Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 42–51. AUAI Press, 2016.

[5] P. Bartlett and S. Mendelson. Gaussian and Rademacher complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[6] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[7] P. L. Bartlett and M. Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8(Oct):2347–2368, 2007.

[8] L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, pages 1554–1563, 1966.

[9] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, pages 164–171, 1970.

[10] M. Belkin and K. Sinha. Toward learning Gaussian mixtures with arbitrary separation. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 407–419, 2010.

[11] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[12] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, Norwell, MA, 2004.

[13] P. Bickel, Y. Ritov, and T. Rydén. Asymptotic normality of the maximum-likelihood estimator for general Hidden Markov Models. *The Annals of Statistics*, 26(4):1614–1635, 08 1998.

[14] P. J. Bickel, Y. Ritov, and T. Ryden. Asymptotic normality of the maximum-likelihood estimator for general Hidden Markov Models. *The Annals of Statistics*, pages 1614–1635, 1998.

[15] L. Breiman. Prediction games and arcing algorithms. *Neural computation*, 11(7):1493–1517, 1999.

[16] L. Breiman et al. Arcing classifier (with discussion and a rejoinder by the author). *Annals of Statistics*, 26(3):801–849, 1998.

[17] P. Bühlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, pages 477–505, 2007.

[18] P. Bühlmann and B. Yu. Boosting with L^2 loss: Regression and classification. *Journal of American Statistical Association*, 98:324–340, 2003.

[19] R. Camoriano, T. Angles, A. Rudi, and L. Rosasco. Nytro: When subsampling meets early stopping. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1403–1411, 2016.

[20] A. Caponetto and Y. Yao. Adaptation for regularization operators in learning theory. Technical Report CBCL Paper #265/AI Technical Report #063, Massachusetts Institute of Technology, September 2006.

[21] A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical Report CBCL Paper #264/AI Technical Report #062, Massachusetts Institute of Technology, September 2006.

[22] O. Cappé, E. Moulines, and T. Rydén. Hidden Markov Models, 2004.

[23] R. Caruana, S. Lawrence, and C. L. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems*, pages 402–408, 2001.

[24] A. Chaganty and P. Liang. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning*, pages 1040–1048, 2013.

[25] S. Dasgupta. Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science, FOCS '99, 17-18 October, 1999, New York, NY, USA*, pages 634–644, 1999.

[26] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, pages 1–38, 1977.

[27] R. Durbin. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[28] R. Elliott, L. Aggoun, and J. Moore. *Hidden Markov Models: Estimation and Control*. Applications of Mathematics. Springer, 1995.

[29] D. P. Foster and R. A. Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444, 2008.

[30] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[31] J. Friedman, T. Hastie, R. Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of statistics*, 28(2):337–407, 2000.

[32] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[33] C. Gu. *Smoothing spline ANOVA models*. Springer Series in Statistics. Springer, New York, NY, 2002.

[34] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002.

[35] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[36] D. Hsu and S. Kakade. Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 11–20, New York, NY, USA, 2013. ACM.

[37] D. Hsu, S. Kakade, and T. Zhang. A spectral algorithm for learning Hidden Markov Models. *J. Comput. Syst. Sci.*, 78(5):1460–1480, 2012.

[38] K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pages 1–6. IEEE, 2014.

[39] K. G. Jamieson, M. Malloy, R. D. Nowak, and S. Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *COLT*, volume 35, pages 423–439, 2014.

[40] A. Javanmard and A. Montanari. On online control of false discovery rate. *arXiv preprint arXiv:1502.06197*, 2015.

[41] A. Javanmard and A. Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of Statistics*, 2017.

[42] W. Jiang. Process consistency for adaboost. *Annals of Statistics*, 21:13–29, 2004.

[43] R. Johari, L. Pekelis, and D. J. Walsh. Always valid inference: Bringing sequential analysis to A/B testing. *arXiv preprint arXiv:1512.04922*, 2015.

[44] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 655–662, 2012.

[45] E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 2015.

[46] C. Kim and C. Nelson. *State-space Models with Regime Switching: Classical and Gibbs-sampling Approaches with Applications*. MIT Press, 1999.

[47] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Jour. Math. Anal. Appl.*, 33:82–95, 1971.

[48] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[49] L. A. Kontorovich, B. Nadler, and R. Weiss. On learning parametric-output HMMs. In *Proc. 30th International Conference Machine Learning*, pages 702–710, June 2013.

[50] F. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Info. Theory*, 47(2):498–519, February 2001.

[51] M. Ledoux. On Talagrand's deviation inequalities for product measures. *ESAIM: Probability and statistics*, 1:63–87, 1997.

[52] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

[53] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[54] A. Malek, Y. Chow, M. Ghavamzadeh, and S. Katariya. Sequential multiple hypothesis testing with type I error control. In *The 20th International Conference on Artificial Intelligence and Statistics, 2017*, pages 1343–1351, 2017.

[55] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean. Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems 12*, pages 512–518, 1999.

[56] S. Mendelson. Geometric parameters of kernel machines. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 29–43, 2002.

[57] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 93–102, Washington, DC, USA, 2010. IEEE Computer Society.

[58] E. Mossel and S. Roch. Learning nonsingular phylogenies and Hidden Markov Models. *The Annals of Applied Probability*, 16(2):583–614, 05 2006.

[59] A. Nobel and A. Dembo. A note on uniform laws of averages for dependent processes. *Statistics & Probability Letters*, 17(3):169–172, 1993.

[60] L. Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[61] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Pearson Education Signal Processing Series. Pearson Education, 1993.

[62] A. Ramdas, F. Yang, M. J. Wainwright, and M. I. Jordan. Online control of the false discovery rate with decaying memory. In *(Oral) Advances in Neural Information Processing Systems (NIPS)*, pages 5655–5664, 2017.

[63] A. Ramdas, F. Yang, M. J. Wainwright, and M. I. Jordan. Online control of the false discovery rate with decaying memory. In *Advances in Neural Information Processing Systems (NIPS) 2017*, pages 5655–5664, 2017.

[64] G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014.

[65] L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[66] R. E. Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[67] R. E. Schapire. The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer, 2003.

[68] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[69] S. M. Siddiqi, B. Boots, and G. J. Gordon. Reduced-rank Hidden Markov Models. In *Proc. 13th International Conference on Artificial Intelligence and Statistics*, 2010.

[70] M. Simchowitz, K. Jamieson, and B. Recht. The simulator: Understanding adaptive sampling in the moderate-confidence regime. In *Conference on Learning Theory*, pages 1794–1834, 2017.

[71] O. N. Strand. Theory and methods related to the singular value expansion and Landweber's iteration for integral equations of the first kind. *SIAM J. Numer. Anal.*, 11:798–825, 1974.

[72] S. Terwijn. On the learnability of Hidden Markov Models. In *Proceedings of the 6th International Colloquium on Grammatical Inference: Algorithms and Applications*, ICGI '02, pages 261–268, London, UK, UK, 2002. Springer-Verlag.

[73] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[74] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

[75] R. van Handel. Hidden Markov Models. *Unpublished lecture notes*, 2008.

[76] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860, June 2004.

[77] S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[78] E. D. Vito, S. Pereverzyev, and L. Rosasco. Adaptive kernel methods using the balancing principle. *Foundations of Computational Mathematics*, 10(4):455–479, 2010.

[79] G. Wahba. Three topics in ill-posed problems. In M. Engl and G. Groetsch, editors, *Inverse and ill-posed problems*, pages 37–50. Academic Press, 1987.

[80] G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PN, 1990.

[81] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2017.

[82] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1—305, December 2008.

[83] Z. Wang, Q. Gu, Y. Ning, and H. Liu. High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*, 2014.

[84] J. C. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, pages 95–103, 1983.

[85] F. Yang, S. Balakrishnan, and M. J. Wainwright. Statistical and computational guarantees for the Baum-Welch algorithm. *Journal for Machine Learning Research*, 18(1):1–53, 2017.

[86] F. Yang, A. Ramdas, K. Jamieson, and M. J. Wainwright. A framework for Multi-A(rmed)/B(andit) testing with online FDR control. In *(Spotlight) Advances in Neural Information Processing Systems (NIPS)*, pages 5959–5968, 2017.

[87] F. Yang, Y. Wei, and M. J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities (first two authors contributed equally). In *(Spotlight) Advances in Neural Information Processing Systems (NIPS)*, pages 6067–6077, 2017.

[88] Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal non-parametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017.

[89] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[90] X. Yi and C. Caramanis. Regularized EM algorithms: A unified framework and provable statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 1567–1575, 2015.

[91] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

[92] T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005.

[93] S. Zhao, E. Zhou, A. Sabharwal, and S. Ermon. Adaptive concentration inequalities for sequential decision problems. In *Advances In Neural Information Processing Systems*, pages 1343–1351, 2016.