Safety, Robustness, and Interpretability in Machine Learning

Samuel Pfrommer

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-67
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-67.html

May 15, 2025

Copyright © 2025, 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.

Safety, Robustness, and Interpretability in Machine Learning

by

Samuel Ian Pfrommer

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Somayeh Sojoudi, Chair
Professor Javad Lavaei
Professor Venkatachalam Anantharam

Spring 2025

Safety, Robustness, and Interpretability in Machine Learning

Copyright 2025
by
Samuel Ian Pfrommer

Abstract

Safety, Robustness, and Interpretability in Machine Learning

by

Samuel Ian Pfrommer

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Somayeh Sojoudi, Chair

Machine learning is poised to have a dramatic impact across many scientific, industrial, and social domains. While current Artificial Intelligence (AI) systems generally involve human supervision, future applications will demand significantly more autonomy. Such a transition will require us to trust the behavior of increasingly large models. This dissertation addresses three critical research areas towards this goal: safety, robustness, and interpretability.

We first address safety concerns in Reinforcement Learning (RL) and Imitation Learning (IL). While learned policies have achieved impressive performance, they often exhibit unsafe behavior due to training-time exploration and test-time environmental shifts. We introduce a model predictive control-based safety guide which refines the actions of a base RL policy, conditioned on user-provided constraints. With an appropriate optimization formulation and loss function, we show theoretically that the final base policy is provably safe at optimality. IL suffers from a distinct *causal confusion* safety concern, where spurious correlations between observations and expert actions can lead to unsafe behavior upon deployment. We leverage tools from Structural Causal Models (SCMs) to identify and mask problematic observations. Whereas previous work requires access to a queryable expert or an expert reward function, our approach uses the typical ability of an experimenter to intervene on the initial state of an episode.

The second part of this dissertation concerns robustifying machine learning classifiers against adversarial inputs. Classifiers are a critical component of many AI systems and have been shown to be highly sensitive to small input perturbations. We first extend randomized smoothing beyond traditional isotropic certification by projecting inputs into a data-manifold subspace, resulting in orders-of-magnitude improvements in certified volume. We then revisit the fundamental robustness problem by proposing asymmetric certification. This binary classification setting requires only certified robustness for one class, reflecting the fact that many real-world adversaries are strictly interested in producing false negatives. This more focused problem admits an interesting class of feature-convex architectures, which we leverage to provide efficient, deterministic, and closed-form certified radii.

The third part of this dissertation discusses two distinct aspects of interpretability: how Large Language Models (LLMs) decide what to recommend to human users, and how we can build learned models which obey human-interpretable structures. We first analyze conversational search engines, in which we use LLMs to rank consumer products for a user query. Our results show that LLMs vary widely in prioritizing product names, associated website content, and input context position. Finally, we propose a new family of interpretable models in domains where latent embeddings carry mathematical structure: structural transport nets. Via a learned bijection to a carefully-designed mirrored algebra, we produce interpretable latent-space operations which respect the laws of the original input space. We demonstrate that respecting underlying algebraic laws is crucial for learning accurate and self-consistent operations.

To my family for starting me on this journey.

To Vera, for her love and support along the way.

Contents

Acknowledgements

I would like to especially thank my advisor, Professor Somayeh Sojoudi, for her guidance, encouragement, and understanding throughout my PhD. Somayeh has always championed my development as a researcher, lifted me up when I've stumbled, and sincerely advocated for me inside and outside of Berkeley. I am also deeply grateful to Professor Javad Lavaei for his support in my academic pursuits and to my qualifying and dissertation committee members, Professor Kameshwar Poolla and Professor Venkat Anantharam, for their insightful feedback.

Similarly important to my research journey has been the influence of my academic collaborators at Berkeley: Brendon, Tanmay, Yatong, Hyunin, George, Yixiao, Julien, and Alec. Brendon in particular was the first to engage with me on adversarial robustness work, and every one of the six papers we've written together has been a true pleasure.

I would not have started a PhD without inspiration from my time at Penn. Thank you to Matt and Fernando for guiding me in some of my earliest research experiences. I'm also grateful for my first academic mentor, Professor Michael Posa. Michael taught me how to break down a research problem and helped me develop persistence when things inevitably don't work.

There have been many who filled my life outside of work with meaning. Thank you to Mathias, Justin K., Justin P., Matt, Alok, Tanmay, Jake, Daniel, Sam, Bhaskar, Devon, Berkeley Veritas, and the salsa community for all the adventures we've shared. Thank you to my family for loving me unconditionally, for raising me with a sense of independence and curiosity, and for making a thousand sacrifices, big and small, to make this moment possible.

Most importantly, thank you to my fiancée Vera for making these few years the happiest of my life.

Introduction

The capabilities of machine learning models have outpaced our ability to understand and control their behavior. Modern architectures are large and relatively unstructured, with characteristics which are diffusely distributed across potentially billions of parameters. This renders safety difficult to enforce, robustness difficult to ensure, and intent difficult to interpret.

What does it mean for models to be safe, robust, and interpretable? We succinctly define these terms as follows:

> *Safety*: can a model be constrained?
> *Robustness*: can a model be defended?
> *Interpretability*: can a model be understood?

Autonomous driving serves as an illustrative example. Relevant *safety* questions might include: can we guarantee (mathematically or empirically) that the autonomous vehicle will not enter an intersection with pedestrians? Can we verify that our learned controller will not accelerate at a red light? *Robustness* concerns the model's handling of adversarial edge cases. Can the autonomous vehicle handle foggy conditions that it has not been trained on? Is it vulnerable to a malicious adversary tampering with road signs? *Interpretability* addresses deeper questions of intent. Why did the vehicle briefly swerve left – was it avoiding a pothole, or did the neighboring car stray from its lane?

These areas have deep interconnections and arguably run from most specific to most foundational. If a model's behavior changes radically under adversarial pressures, how can we possibly claim that it is safe? And how can we fully ascertain a model's adversarial robustness if we do not understand the basic principles of its behavior? Nevertheless, this hierarchy is not strict, and each field features distinct challenges and methodologies. We can address the former settings without claiming to fully resolve the latter.

Safety

We consider safety to be an agentic problem involving learned policies in a sequential setting. Agency in Reinforcement Learning (RL) has typically been restricted to sandboxed environments such as video games [Berner et al., 2019, Mnih et al., 2015]. This limited scope contrasts with the proliferation of agents in real-world settings, including industrial control [Dalal et al., 2018] and autonomous driving [Bojarski et al., 2016]. Real-world agency involves operating alongside humans in a shared environment and poses unique challenges for safety. Chief among these is the satisfaction of behavioral constraints in accordance with human expectations. A factory robot arm collaborating with a worker must be constrained from performing dangerous motions.

This dissertation focuses on the safety of reinforcement and imitation learning agents in a classical Markov Decision Process (MDP) setting. While we do not explicitly consider language model agency, we note that many of the same safety concerns are also relevant to LLMs and can likewise be considered as constraints. An LLM email assistant should not leak privileged information to a scammer [Cohen et al., 2024]. An LLM software generation tool should not write vulnerable code [Wu et al., 2023a]. These constraints are more challenging to formalize than those for a robot arm. But we expect that some techniques and insights may transfer between these two settings.

How can we formulate policies that satisfy constraints? Even in an MDP setting with a well-trained model, nondeterministic policies and environmental uncertainty can lead to unsafe behavior. One class of approaches rely on better-understood optimal control techniques such as Model Predictive Control (MPC) [Qin and Badgwell, 2003, Rawlings and Mayne, 2009]. A representative approach is that of Wabersich and Zeilinger [2019], which samples actions from an RL base policy and uses an MPC controller as a *safety filter* to correct unsafe behaviors. In Cheng et al. [2019], the authors propose a method that combines model-free RL algorithms with control barrier functions to guarantee safety during training. Generally, these approaches require an accurate model of the environment, which can be learned with some difficulty [Koller et al., 2018].

Training-time behavior adds additional complexity as RL algorithms engage in trial-and-error exploration of the policy space. Attempts to address this by training in simulation are handicapped by limited simulator realism [Ray et al., 2019]. Constrained Reinforcement Learning (CRL) instead aims to formalize safety requirements as constraints within the RL optimization problem [Achiam et al., 2017, Dalal et al., 2018, Tessler et al., 2018]. Overarching drawbacks to these approaches include limited training-time safety guarantees and the potential for unsafe policies after training.

While Imitation Learning (IL) agents do not perform training-time exploration, they feature distinct safety challenges which arise when transitioning from training to deployment. One key issue is *causal confusion*, where the learned policy mistakes observations which are correlated with expert actions as being causally related [De Haan et al., 2019, Kaddour et al., 2022]. Consider an IL agent trained on a human-perspective driving dataset containing a dashboard light that activates upon braking. Since the light and the braking

action correlate strongly in the training data, the policy will learn to brake only when the light appears [De Haan et al., 2019]. This means that the policy might appear to brake normally during training only to be completely unsafe at deployment. Existing approaches for removing such *nuisance variables* make strong assumptions, typically requiring either a queryable expert or an expert reward function [De Haan et al., 2019, Ortega et al., 2021, Ross et al., 2011a]. Other works introduce regularization techniques which mitigate, but do not eliminate, the problem [Park et al., 2021].

Robustness

Adversarial robustness is arguably a precursor for safety. Even if a non-robust model behaves safely under standard conditions, malicious actors could exploit its vulnerabilities to cause serious harm. Despite state-of-the-art performance on a range of tasks, ML models are shockingly sensitive to *adversarial examples*—inputs with small (often human-imperceptible) perturbations that are maliciously crafted to induce failure [Biggio et al., 2013, Nguyen et al., 2015, Szegedy et al., 2014]. This is particularly problematic in safety-critical applications, such as autonomous driving [Bojarski et al., 2016, Wu et al., 2017a], power system operation [Kong et al., 2017], and medical diagnostics [Amato et al., 2013, Yadav and Jadhav, 2019]. Eykholt et al. [2018] established that just a carefully-placed physical patch can cause an image classifier to completely misclassify a traffic sign. The robsutness problem is also present in language models, where a malicious adversary can "jailbreak" a model to generate harmful text [Zou et al., 2023].

The canonical task of adversarial robustness is to ensure that a model's correct classification of an input is invariant under some set of bounded perturbations—typically characterized as having a small ℓ_p norm. Several empirical defenses have claimed to provide heuristic robustness guarantees along these lines, only to be subsequently broken by stronger attacks [Athalye et al., 2018, Carlini and Wagner, 2017, Kurakin et al., 2017, Madry et al., 2018, Uesato et al., 2018]. This has inspired research interest in *certifiable robustness*, which provides provable robustness guarantees under arbitrary attacks of a bounded norm. The final ℓ_p-ball *certified radii* are tightly coupled with the model architecture, with off-the-shelf models generally featuring large Lipschitz constants and thus weak certificates [Fazlyab et al., 2019, Hein and Andriushchenko, 2017, Yang et al., 2020b]. Various lines of work have addressed this by introducing model families which admit tractable certification procedures [Cohen et al., 2019a, Li et al., 2019, Trockman and Kolter, 2021, Wong and Kolter, 2018, Zhang et al., 2021a]. Of particular note for this dissertation is the family of *randomized smoothing* methods, which provide high-probability robustness certificates by aggregating predictions over random corruptions of the input [Cohen et al., 2019a]. These approaches in turn suffer from a range of drawbacks, including only certifying against one specific norm or requiring prohibitively expensive computations.

Interpretability

Model interpretability is perhaps the most foundational and poorly-understood of the three considered problem settings. A modern LLM's capabilities are scattered across an

enormous number of parameters, frustrating attempts to localize behaviors to a specific set of weights or develop clean intuitions about overall functionality.

While this dissertation does not directly explore *mechanistic interpretability*—attempting to reverse-engineer distinct internal mechanisms within a model—this subfield is useful for understanding the broader challenges of interpretability. We consider the well-explored problem of *lie detection* in LLMs as a specific case study. Early lie-detection approaches leveraged *activation probing* to identify whether the last-token activations of a model captured the truthfulness of an input statement [Azaria and Mitchell, 2023]. At a high level, this involves curating a dataset of true and false statements and training a binary classifier over the last-token final-layer activation. Azaria and Mitchell [2023] found that such a probe achieved up to a 83% lie-detection accuracy. But this lie detector was soon shown to be faulty, along with related approaches such as Burns et al. [2022]. Levinstein and Herrmann [2024] discovered that these probes achieved effectively random performance on a dataset consisting of Boolean negations of the original test statements; in effect, the probes were not identifying truthfulness but rather the presence of "negation words" that were spuriously correlated with truthfulness.

A separate family of interpretability approaches sidesteps the challenges of analyzing internal activations by instead studying the input-output behavior of LLMs. For instance, work on evaluating natural language reasoning has shown that LLMs are sensitive to paraphrastic input variations [Srikanth et al., 2024]. Behavioral studies have revealed that even when models are designed to appear unbiased, they can still replicate societal stereotypes by consistently assigning gendered roles in occupational contexts [Kotek et al., 2023]. Benchmarks such as TruthfulQA examine how models manage misinformation by analyzing truthfulness for commonly-misanswered questions [Lin et al., 2022]. Additional studies have analyzed models' self-reported confidence levels to better understand uncertainty in output predictions [Kadavath et al., 2022].

A distinct and underemphasized aspect of interpretability considers the alignment of model structure with human expectations. This line of work involves designing model architectures that inherently respect known mathematical, physical, or domain-specific properties. Early work on enforcing structure involved designing networks with monotonicity properties [Sill, 1997], ensuring that model predictions in business applications are non-decreasing with respect to relevant financial metrics. The Hamiltonian Neural Networks of Greydanus et al. [2019] guarantee that learned dynamics functions obey physical conservation laws. Physics-Informed Neural Networks incorporate differential equations directly into the training process, resulting in networks that extrapolate beyond training data in a human-interpretable manner [Raissi et al., 2019]. Pawlowski et al. [2020] combines deep learning with structural causal models, both enhancing interpretability and enabling counterfactual inference.

Summary of contributions

These three problem settings form the three major parts of this dissertation. Each part in turn consists of two chapters derived from previously published work. We briefly highlight the major contributions of this research in the context of our established framework.

Part I: safety

Chapter 1 addresses the safety of RL agents both during training and after deployment. We introduce a Model Predictive Control (MPC)-based *safety guide* which contains two main innovations over previous safety filter work. The first is a chance-constrained problem formulation which permits optimization over action distributions, aligning with the stochastic formulation of RL policies. This enables the second innovation: a safety penalty in the policy gradient objective that encourages the policy to imitate the guide in safety-critical situations. While the safety guide permits high-probability safety guarantees during training, we show theoretically that the safety penalty ensures that the optimal RL policy is provably safe at deployment.

Chapter 2 focuses on the causal confusion problem in imitation learning, which has serious implications for the safe and predictable operation of IL agents after deployment. Specifically, we build upon a family of approaches which mask confounders in a disentangled representation of the observation space. Existing methods in this family require either a queryable expert, an expert reward function, or a manually specified causal graph. We instead propose a method which leverages the typical ability of an experimenter to specify the initial state of an episode. Our algorithm uses tools from Structural Causal Models (SCMs) to mask spuriously correlated latent variables. We prove that this method is *conservative* in the sense that it does not mask observations that causally affect the expert's behavior, and empirically demonstrate its effectiveness in illustrative controls tasks.

Part I is based on the following published works:

Samuel Pfrommer, Tanmay Gautam, Alec Zhou, and Somayeh Sojoudi. Safe reinforcement learning with chance-constrained model predictive control. In *Learning for Dynamics and Control Conference*, pages 291–303. PMLR, 2022.

Samuel Pfrommer, Yatong Bai, Hyunin Lee, and Somayeh Sojoudi. Initial state interventions for deconfounded imitation learning. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 2312–2319. IEEE, 2023d.

Part II: robustness

Chapter 3 extends randomized smoothing beyond ℓ_p norm ball certified regions, whose volume decays factorially fast in the dimensionality of the input space. We propose a classifier architecture which projects inputs into a principal component subspace and applies randomized smoothing in this lower-dimensional space. The resulting certified regions are characterized as a subspace-perpendicular extrusion of a low-dimensional sphere

and are shown empirically to contain meaningful adversarial vulnerabilities. Leveraging mathematical results regarding high-dimensional cube-subspace intersections, we derive a tractable lower bound on the volume of this certified region. In accordance with the manifold hypothesis, we show that the factorial volume decay in the much lower-dimensional projected dimension—as opposed to the original input dimension—results in orders-of-magnitude improvements in certified volume.

Chapter 4 reframes certified robustness as an *asymmetric* binary classification problem, where certificates are only required for inputs from one class. This reflects real-world settings where an adversary is only concerned with producing false negatives (e.g. spam email classification). In this more focused domain, we introduces feature-convex neural networks, which compose a Lipschitz-continuous feature map with a learned convex classifier. This architecture admits closed-form, deterministic certified radii for any ℓ_p norm. We experimentally show that these radii outperform existing methods while being orders of magnitude faster to compute than competitive baselines.

Part II is based on the following published works:

Samuel Pfrommer, Brendon G. Anderson, and Somayeh Sojoudi. Projected randomized smoothing for certified adversarial robustness. *Transactions on Machine Learning Research*, 2023b.

Samuel Pfrommer, Brendon Anderson, Julien Piet, and Somayeh Sojoudi. Asymmetric certified robustness via feature-convex neural networks. *Advances in Neural Information Processing Systems*, 36:52365–52400, 2023a.

Part III: interpretability

Chapter 5 investigates conversational search engines from both an interpretability and robustness perspective. Conversational search engines, such as `perplexity.ai` and Google AI overview, operate by loading both user queries and website content into an LLM's context window. We introduce a focused dataset of consumer product websites and characterize the LLM's "ranking" of these products as the order in which they are mentioned in the LLM's response. Our experiments reveal that different LLMs exhibit distinct patterns in their ranking behavior, with varying emphasis on product names, document content, and position in the context window. We then present a tree-of-attacks prompt injection technique which allows a website operator to artificially boost their product's ranking.

Chapter 6 explores imposing human-interpretable algebraic structure onto the learned embeddings of objects that lie in a larger mathematical space. For example, in 3D modeling applications subsets of Euclidean space can be embedded as vectors using implicit neural representations. These subsets feature a natural algebraic structure consisting of operations (e.g., union) and corresponding laws (e.g., associativity). This chapter proposes *structural transport nets* to learn operations which provably respect algebraic laws by construction. The core architectural innovation is a learned bijection

from the latent space to a Euclidean-space "mirrored algebra" which is constructed in accordance with desired laws. We evaluate structural transport nets against naive baselines and show that respecting underlying algebraic structure is key for learning accurate and self-consistent operations.

Part III is based on the following published works:

Samuel Pfrommer, Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. Ranking manipulation for conversational search engines. *Empirical Methods in Natural Language Processing*, 2023c.

Samuel Pfrommer, Brendon G Anderson, and Somayeh Sojoudi. Transport of algebraic structure to latent embeddings. *International Conference on Machine Learning*, 2024.

Part I

Safety

Chapter 1

Safe Reinforcement Learning via Chance-Constrained Model Predictive Control

Real-world reinforcement learning (RL) problems often demand that agents behave safely by obeying a set of designed constraints. We address the challenge of safe RL by coupling a *safety guide* based on model predictive control (MPC) with a modified policy gradient framework in a linear setting with continuous actions. The guide enforces safe operation of the system by embedding safety requirements as chance constraints in the MPC formulation. The policy gradient training step then includes a safety penalty which trains the base policy to behave safely. We show theoretically that this penalty allows for a provably safe optimal base policy and illustrate our method with a simulated linearized quadrotor experiment.

This chapter is based on the following published work:

Samuel Pfrommer, Tanmay Gautam, Alec Zhou, and Somayeh Sojoudi. Safe reinforcement learning with chance-constrained model predictive control. In *Learning for Dynamics and Control Conference*, pages 291–303. PMLR, 2022.

1.1 Introduction

Reinforcement learning has been extensively studied in the context of closed environments, where it has gained popularity for its success in mastering games such as Atari and Go [Mnih et al., 2015, Silver et al., 2017, Sutton and Barto, 1998]. A pressing need to deploy autonomous agents in the physical world has introduced a new challenge: agents must be able to interact with their environments in a safe and comprehensible manner. This is especially critical in industrial settings [Dalal et al., 2018].

For safety-critical tasks, the trial-and-error nature of exploration in RL often prevents

agent deployment in the real world during training, motivating the use of simulators. However, when dealing with complex environments, simulators may fail to sufficiently model the complexity of the environment [Ray et al., 2019]. Furthermore, reward functions may be unknown a priori, making learning in simulation impossible. This is where methods that guarantee safe exploration during training offer a substantial advantage.

Our work employs policy gradients and model predictive control (MPC) as its primary building blocks to address the safe RL problem. Policy gradient methods learn a parameterized policy to maximize long-term expected rewards using gradient ascent and play a central role in reinforcement learning due to their ability to handle stochasticity, superior convergence properties and training stability, and efficacy in high-dimensional action spaces [Sutton and Barto, 1998]. This family of algorithms is also *model-free*, relying solely on reward signals from the environment without modeling any dynamics. Policy gradient variations have since proliferated under the deep learning paradigm, notably including "natural" policy gradients and actor-critic methods in addition to techniques such as experience replay and importance sampling for better sample efficiency [Peters and Schaal, 2008, Wang et al., 2017].

Model predictive control is a flexible optimal control framework that has seen successes across a wide variety of settings, including process control in chemical plants and oil refineries, power electronics and power system balancing, autonomous vehicles and drones, and building control [Qin and Badgwell, 2003, Rawlings and Mayne, 2009]. It is *model-based*, requiring the system dynamics to be identified either a priori or through learning [Koller et al., 2018]. Its interpretability lends itself to robust extensions, where system uncertainties and disturbances can be incorporated to probabilistically guarantee agent safety [Koller et al., 2018].

1.1.1 Related work

Safety filters are the closest line of work to our proposed algorithm [Wabersich and Zeilinger, 2019, 2021]. This is a decoupled method that takes sampled actions from any base policy and uses an MPC controller as the "safety filter" to correct unsafe behaviors. However, these two components function independently, which may lead to conflicting and potentially oscillatory behaviour between the MPC and RL objectives. The computationally taxing safety filter must also be used at both training and test times, making the technique ill-suited for real-world deployment on constrained hardware. Cheng et al. [2019] proposes a related framework that combines model-free RL algorithms with control barrier functions to guarantee safety during training. While this approach accommodates model uncertainty and learns the dynamics online, it is decoupled in a manner similar to safety filters and retains the same drawbacks. Wagener et al. [2021] describes SAILR - an alternative intervention-based approach that utilizes advantage functions to learn a safe policy during training. While empirically the authors demonstrate that this algorithm outperforms other safe RL methods, it is still shown to occasionally violate safety constraints during training.

Constrained reinforcement learning (CRL) aims to formalize the reliability and safety requirements of an agent by encoding these explicitly as constraints within the RL optimization problem. Achiam et al. [2017] proposes a trust-region based policy search algorithm for CRL with guarantees, under some policy regularity assumptions, that the policy stays within the constraints in expectation. This approach cannot be used in applications where safety must be ensured at all visited states. Dalal et al. [2018] addresses the CRL problem by adding a safety layer to the policy that analytically solves an action correction formulation for each state. While this approach guarantees constraint satisfaction, it does not yield a safe policy at the end of training. In [Tessler et al., 2018], the constraints are embedded as a penalty signal into the reward function, guiding the policy towards a constraint satisfying solution. Similar to [Achiam et al., 2017], safety is not ensured at each state.

Model-based RL methods generally offer higher sample efficiency than their model-free counterparts and can be applied in safety-critical settings with more interpretable safety constraints. This area of work includes learning-based robust MPC [Koller et al., 2018]. Berkenkamp et al. [2017] proposes an algorithm that considers safety in terms of Lyapunov stability guarantees. More specifically, the approach demonstrates how, starting from an initial safe policy, the safe region of attraction can be expanded by collecting data within the safe region and adapting the policy.

Imitation learning attempts to learn a policy by direct supervision from expert demonstration. This approach is frequently plagued by distribution mismatch and compounding errors. Dataset Aggregation (DAgger) is an iterative method used to mitigate these drawbacks by reducing the distribution mismatch [Ross et al., 2011b]. In Menda et al. [2019], the authors extend DAgger to EnsembleDAgger, which addresses the challenge of safe exploration by quantifying the confidence of the learned policy. It does this by using an ensemble of neural networks to estimate the variance of the action proposed by the learned policy at a particular state. While showing solid empirical performance, EnsembleDAgger lacks formal safety guarantees.

1.1.2 Paper contributions

Our approach wraps a policy gradient *base policy* with an MPC-based *safety guide* that corrects any potentially unsafe actions. The base policy learns to optimize the agent's long-term behaviour, while the MPC component accounts for state-space safety constraints. By optimizing over an action distribution in the safety guide, we show that adding a safety penalty to the policy gradient loss allows for a provably safe optimal base policy. This resolves tension between the base policy and the safety guide and permits the removal of the computationally expensive safety guide after training.

1.2 Background

1.2.1 Notation

Throughout this work, we let $s_t \in \mathcal{X}$, $a_t \in \mathcal{A}$, and $r(s_t, a_t) \in \mathbb{R}$ refer to the state, action, and reward at time t. A sequence of states and actions is termed a trajectory and denoted by τ, and the sum of rewards over a trajectory is denoted $r(\tau)$. We focus on the setting where $\mathcal{X} \subseteq \mathbb{R}^n$ and $\mathcal{A} \subseteq \mathbb{R}^m$. Since our action space is continuous, we represent a stochastic policy as $\pi : \mathcal{X} \to \mathcal{N}(\mathcal{A})$, where $\mathcal{N}(\mathcal{A})$ is a Gaussian distribution over actions. More specifically, we can write $\pi(\,\cdot \mid s) = \mathcal{N}(\mu(s), \Sigma(s))$ for some Gaussian mean $\mu(s)$ and covariance $\Sigma(s)$. The space of such policies is denoted as Π. When such a policy is parameterized by a vector θ, we use the notation π_θ. With some abuse of notation, we write $\tau \sim \pi$ to denote sampling a trajectory from the policy π; similarly, $(s, a) \sim \pi$ denotes sampling a state s from the stationary distribution induced by π and then sampling a from $\pi(\,\cdot \mid s)$. Furthermore, $\|\cdot\|_p$ denotes the ℓ_p-norm within $\mathbb{R}^n$. The symbol $\mathbf{1}_n$ defines an n-dimensional column vector of ones, and $\mathrm{Tr}(A)$ denotes the trace of the matrix A. $\mathbb{E}_{p(x)}[\cdot]$ is the expectation operator with respect to the probability distribution $p(x)$.

1.2.2 Policy gradient

Policy gradient methods attempt to find the optimal parameters θ^* for the objective

$$\max_\theta J(\pi_\theta), \qquad J(\pi_\theta) = \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{M} \gamma^t r(s_t, a_t)\right]. \tag{1.1}$$

The vanilla policy gradient approach performs gradient ascent to maximize this objective [Williams, 1992]. The gradient can be approximated with the Monte-Carlo estimator

$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{N}\sum_{j=1}^{N}\sum_{t=0}^{M} \nabla_\theta \log \pi_\theta(a_t^j \mid s_t^j) \sum_{t'=t}^{M} \gamma^{t'-t} r(s_{t'}^j, a_{t'}^j), \tag{1.2}$$

with $0 \leq \gamma < 1$ a discount factor. While many variance-reduction techniques can be used to improve (1.2), for simplicity of exposition we employ this basic formulation.

1.2.3 Model predictive control

Model predictive control is a purely optimization-based planning framework. Given a dynamics model and a set of state and action constraints (safety requirements, physical limitations, etc.), the finite-horizon MPC problem computes the near-optimal open-loop action sequence that minimizes a specified cumulative cost function. The first of these actions is executed, and the entire optimization repeats on the next time step. While the MPC framework offers concreteness in its constraints, it requires a pre-specified reward function and is incapable of forming reward-maximizing plans beyond its horizon.

1.2.4 Problem setting

We represent the environment dynamics as a known linear time-invariant system

$$s_{t+1} = As_t + Ba_t, \tag{1.3}$$

with initial state s_0, dynamics matrix $A \in \mathbb{R}^{n\times n}$, and input matrix $B \in \mathbb{R}^{n\times m}$. The safety requirements are captured by a polyhedral state safe set $\mathcal{S} \subset \mathcal{X}$. The goal is to learn a policy which maximizes the cumulative reward signal r while ensuring that the exploration during training is safe at all times, i.e. $s_t \in \mathcal{S}$ for all t.

1.3 Method

A high-level overview of our method combining policy gradient learning and model predictive control is displayed in Figure 1.1. We first outline the construction of the *safety guide*, which solves a chance-constrained MPC optimization to enforce the safety of actions proposed by the underlying base policy. This allows for guaranteed safety during training time with arbitrarily high probability. Section 1.3.2 discusses how the safety guide is incorporated into the overarching policy optimization.

1.3.1 Safety guide design

The safety guide solves a convex MPC problem for each time step during training to ensure system safety. This safety guide is not needed later at test time, which is justified theoretically in Section 1.4. We begin by making the following assumption.

Assumption 1.1. There exists a polyhedral terminal safe set $\mathcal{S}_T \subset \mathcal{S} \subset \mathcal{X}$ that is invariant, meaning that for any state $s \in \mathcal{S}_T$, there exists a sequence of control inputs that keep the system in $\mathcal{S}_T$ for all subsequent time steps.

The construction of invariant sets has a well-established theory due to its applications in systems and control. For linear systems, several recursive algorithms have been proposed to construct polyhedral invariant sets [Gilbert and Tan, 1991, Pluymers et al., 2005], with nonlinear systems considered in [Bravo et al., 2003, Korda et al., 2013].

Algorithm 1.1 specifies the safety guide optimization problem. Intuitively, the safety guide attempts to find an action distribution that is as close as possible to that outputted by the base policy, subject to safety constraints. Taking inspiration from techniques in the obstacle avoidance literature [Blackmore et al., 2011], we formulate this in a chance-constrained model predictive fashion.

Variables. The optimization variables consist of a sequence of state means μ_t^s and open-loop control actions μ_t^a over a planning horizon of length H, with the first action containing some uncertainty represented by $\overline{\Sigma}_0^a$. The bar over any Σ denotes that this matrix is related to the relevant covariance matrix via $\Sigma = \overline{\Sigma}\ \overline{\Sigma}^\intercal$. This decomposition allows for subsequent chance constraints to be expressed as closed-form convex constraints.

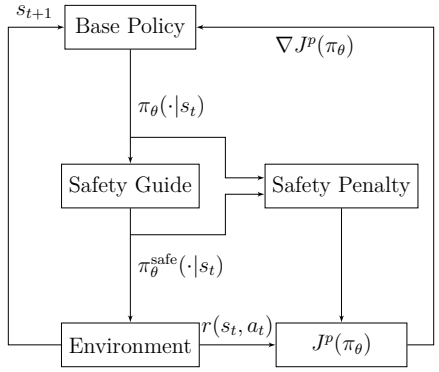


Figure 1.1: The training scheme. The base policy π_θ suggests a distribution over actions given s_t. The safety guide potentially shifts this distribution to ensure safety and outputs the distribution $\pi_\theta^{\text{safe}}(\ \cdot\ |\ s_t)$, from which the next action is sampled. The environment reward and a safety penalty on the distance between these two distributions are combined in the objective, whose gradient is approximated using Monte Carlo rollouts.

Algorithm 1.1 Safety guide

Input: starting state s', base policy mean $\mu_\theta^a(s')$ and covariance $\Sigma_\theta^a(s')$

Parameters: Planning horizon H, safety tolerance ϵ, system matrices A and B, state safe set $\mathcal{S}$, safe terminal set $\mathcal{S}_T$, feasible action set $\mathcal{A}$

Solve the convex optimization problem

$$
\begin{aligned}
\underset{\substack{\overline{\Sigma}_0^a \\ \mu_0^a,\dots,\mu_H^a \\ \mu_0^s,\dots,\mu_H^s}}{\arg\min}\quad & \mathrm{KL}\left(\mathcal{N}\left(\mu_0^a, \overline{\Sigma}_0^a \overline{\Sigma}_0^{a\intercal}\right) \parallel \mathcal{N}\left(\mu_\theta^a(s'), \Sigma_\theta^a(s')\right)\right) \\
& \mu_{t+1}^s = A\mu_t^s + B\mu_t^a, \quad 0 \le t < H \\
& \overline{\Sigma}_t^s \coloneqq A^t B \overline{\Sigma}_0^a, \quad 0 \le t < H \\
& \Pr\left[\mathcal{N}(\mu_t^s, \overline{\Sigma}_t^s \overline{\Sigma}_t^{s\intercal}) \notin \mathcal{S}\right] < \epsilon, \quad 0 \le t < H \\
& \Pr\left[\mathcal{N}(\mu_H^s, \overline{\Sigma}_H^s \overline{\Sigma}_H^{s\intercal}) \notin \mathcal{S}_T\right] < \epsilon, \quad t = H \\
& \mu_t^a \in \mathcal{A}, \quad 0 < t \le H \\
& \mu_0^s = s'
\end{aligned}
$$

If infeasible **then** relax constraints and resolve

Return $\pi_\theta^{\text{safe}}(\ \cdot\ |\ s') = \mathcal{N}\left(\mu_0^{a*}, \overline{\Sigma}_0^{a*} \overline{\Sigma}_0^{a*\intercal}\right)$

Since we are interested in allowing the base policy to have as much freedom as possible, we avoid the additional conservatism that would result from incorporating uncertainty over future actions and allow these to be chosen deterministically.

Objective. The safety guide objective minimizes the divergence between the base policy action distribution and the distribution of the MPC's first action. If the base policy distribution allows for subsequent actions that maintain safety, the objective vanishes and the returned safe distribution is the original distribution specified by the base policy. KL divergence is not symmetric; we choose this argument order to make the objective convex in the variables μ_0^a and $\overline{\Sigma}_0^a$. To see this, consider the following form for the KL divergence, dropping references to s' for notational convenience:

$$
\begin{aligned}
& \mathrm{KL}\left(\mathcal{N}\left(\mu_0^a, \overline{\Sigma}_0^a \overline{\Sigma}_0^{a\intercal}\right) \parallel \mathcal{N}\left(\mu_\theta^a, \Sigma_\theta^a\right)\right) \\
= \ & \log\det \Sigma_\theta^a - \log\det \overline{\Sigma}_0^a \overline{\Sigma}_0^{a\intercal} - n + \mathrm{Tr}\left((\Sigma_\theta^a)^{-1} \overline{\Sigma}_0^a \overline{\Sigma}_0^{a\intercal}\right) + (\mu_\theta^a - \mu_0^a)^\intercal \left(\Sigma_\theta^a\right)^{-1} (\mu_\theta^a - \mu_0^a).
\end{aligned}
$$

Recall that symbols subscripted by θ are constants in the optimization, while symbols subscripted by 0 are optimization variables. Therefore we disregard the constant terms $\log\det \Sigma_\theta^a$ and $-n$. Convexity of $-\log\det \overline{\Sigma}_0^a \overline{\Sigma}_0^{a\intercal}$ follows from multiplicative properties of the determinant and concavity of the log det operator. For the remaining terms, we assume that Σ_θ^a is positive definite, a practically satisfied assumption. The fourth term

$\mathrm{Tr}\Big((\Sigma_\theta^a)^{-1}\overline{\Sigma}_0^a\overline{\Sigma}_0^{a\intercal}\Big)$ can then be rewritten as $\mathrm{Tr}(XX^T)$ with $X = \sqrt{(\Sigma_\theta^a)^{-1}}\ \overline{\Sigma}_0^a$, which is a convex function composed with a linear function and is therefore convex. Finally, the last term is a positive definite quadratic form and is therefore convex in μ_0^a.

Dynamics. The state propagation equations follow from known properties of linear transformations of Gaussian random variables [Liu, 2019]. Since actions after index 0 are entirely deterministic, we can express state uncertainty at future time steps $\overline{\Sigma}_t^s$ directly as linear functions of the initial action uncertainty $\overline{\Sigma}_0^a$. This parallels results in the chance-constrained path planning literature [Blackmore et al., 2011].

Safety Constraints. The safety constraints for $\mathcal{S}$ and $\mathcal{S}_T$ can be handled similarly. Consider the chance constraint $\Pr\Big[s \notin \mathcal{S}\Big] < \epsilon$, with $s \sim \mathcal{N}(\mu_t^s, \overline{\Sigma}_t^s\overline{\Sigma}_t^{s\intercal})$. Evaluating such a constraint using sampling would require prohibitively many samples for small ϵ and result in a nonconvex optimization problem. We instead leverage techniques from chance constrained optimization to represent this constraint deterministically. Let the polyhedral safe set be defined by r linear inequalities as

$$\mathcal{S} = \bigcap_{i=1}^{r} \{s \mid u_i^\intercal s \leq v_i\}.$$

Deriving a tight closed-form expression for a joint constraint over multiple linear inequalities is a nontrivial problem that is typically handled by an approximation scheme [Cheng and Lisser, 2012]. We conservatively bound the probability of violating each inequality by ϵ/r, noting that this implies

$$\Pr\ [s \notin S] \leq \sum_{i=1}^{r} \Pr\ [u_i^\intercal s > v_i] \leq \sum_{i=1}^{r} \frac{\epsilon}{r} = \epsilon.$$

We now aim to derive a closed-form counterpart for r constraints of the form

$$\Pr[u_i^T s > v_i] \leq \frac{\epsilon}{r}, \qquad s \sim \mathcal{N}(\mu_t^s, \overline{\Sigma}_t^s\overline{\Sigma}_t^{s\intercal}).$$

Since s is normally distributed, this constraint is equivalent to the deterministic constraint

$$v_i - \mu_t^{s\intercal} u_i \geq \Phi^{-1}\left(1 - \frac{\epsilon}{r}\right) \|\overline{\Sigma}_t^s u_i\|_2,$$

where Φ is the standard Gaussian CDF [Duchi, 2021]. Each of our r constraints now becomes a second-order cone constraint and can be handled by conventional convex optimization solvers. If the original problem is infeasible, we relax these constraints with slack variables which we linearly penalize in the objective.

1.3.2 Policy gradient with safety penalty

We now modify the standard policy gradient formulation (1.1) to include a term penalizing corrections by the safety guide, effectively training the base policy to behave safely. Our

objective becomes

$$\max_{\theta} J^p(\pi_\theta), \qquad J^p(\pi_\theta) = \mathop{\mathbb{E}}_{(s,a)\sim\pi_\theta^{\text{safe}}}\Big[r(s,a) - \beta\; d\Big(\pi_\theta^{\text{safe}}(\,\cdot \mid s), \pi_\theta(\,\cdot \mid s)\Big)\Big], \tag{1.4}$$

where d is a positive definite statistical distance which is continuous in s for $\pi_\theta, \pi_\theta^{\text{safe}} \in \Pi$ and $\beta > 0$ is a regularization parameter. For notational convenience, the expectation draws from the stationary state distribution induced by π_θ^{safe} and the associated action distribution. We show in Section 1.4.2 that any positive definite, continuous d results in a safe base policy after training. We choose the squared l_2 parameter distance for its numerical properties:

$$d\Big(\pi_\theta^{\text{safe}}(\,\cdot \mid s), \pi_\theta(\,\cdot \mid s)\Big) \coloneqq \|\mu^a_{\text{safe}} - \mu^a_\theta\|_2^2 + \|\Sigma^a_{\text{safe}} - \Sigma^a_\theta\|_2^2.$$

We can now obtain our optimal parameters θ^* using gradient ascent on a Monte Carlo estimator similar to (1.2) with an added term for the safety penalty.

1.4 Theoretical analysis

We show that our policy leads to safe exploration at training time with arbitrarily high probability. We then prove that coupling reward maximization with a safety penalty in (1.4) leads to a safe optimal base policy. This is highly desirable as it eliminates conflict between the base policy and the safety guide, mitigates distributional shift, and reduces the computational burden on the agent at test time.

1.4.1 Training time safety

Consider a standard episodic training setting where an episode terminates after a set number of time steps or upon violation of the state safety constraints.

Proposition 1.2. Consider an arbitrary natural number T and safety tolerance $\epsilon > 0$ from Algorithm 1.1. Then over T training steps, the expected number of states s_t such that $s_t \notin \mathcal{S}$ is at most ϵT.

This follows directly from the constraints on the optimization problem in Algorithm 1.1. Specifically, there is an ϵ chance of sampling an action from the safe distribution that leads to an unsafe state, in which case the episode ends in at most H time steps. Assumption 1.1 guarantees that with probability $1 - \epsilon$ the action sampled will be safe and subsequent optimizations will remain feasible.

Since ϵ is a design parameter, this expectation can be driven to be arbitrarily small, at the cost of imposing additional conservatism in the exploration process. In practice, this quantity can be effectively set to zero by a small concession on the size of the safe sets $\mathcal{S}$ and $\mathcal{S}_T$. Shrinking these by some factor $1 - \delta$ gives the safe policy a buffer to the true unsafe region, allowing it to recover from unsafe actions by softening the chance inequality constraints in Algorithm 1.1. Our experiments in Section 1.5 use this technique to maintain *perfect safety* over the course of a million training steps.

1.4.2 Base policy safety

In order to derive theoretical guarantees for the optimal policy of (1.4), we introduce two assumptions.

Assumption 1.3. The parameterized base policy class π_θ is a *universal approximator*. Namely, for every policy $\pi^* \in \Pi$ and desired ϵ, there exists a parameterized $\pi_\theta \in \Pi$ such that

$$\sup_{s,a} |\pi^*(a \mid s) - \pi_\theta(a \mid s)| < \epsilon.$$

Assumption 1.4. The reward $r(\tau)$ is bounded over all trajectories τ.

Assumption 1.3 parallels a standard assumption in the deep learning literature that a richly parameterized network is arbitrarily expressive. Assumption 1.4 is similarly benign, and is immediately satisfied in a typical setting where rewards are bounded and trajectories are finite.

Lemma 1.5. For every $\pi^* \in \Pi$ and $\epsilon_J > 0$, there exists a learned parameterization π_θ such that

$$J(\pi^*) - J(\pi_\theta) < \epsilon_J,$$

where $J(\pi) = \mathbb{E}_{\tau\sim\pi}\, r(\tau)$ is the standard reinforcement learning objective.

Proof. Let $p(s)$ be the initial state distribution and $p(s_{t+1} \mid s_t, a_t)$ represent the environment transition dynamics. For notational simplicity, we define $\Delta J := J(\pi^*) - J(\pi_\theta)$. Then we can write

$$\Delta J = \int r(\tau) p(s_0) \delta\pi(\tau) \prod_{t=0}^{M} p(s_{t+1} \mid s_t, a_t) d\tau, \tag{1.5}$$

where

$$\delta\pi(\tau) = \prod_{t=0}^{M} \pi^*(a_t \mid s_t) - \prod_{t=0}^{M} \pi_\theta(a_t \mid s_t).$$

Assumption 1.3 implies that there exists a parameter vector θ such that $\delta\pi(\tau)$ can be bounded for all τ by an arbitrarily small quantity. Since $r(\tau)$ in (1.5) is bounded by Assumption 1.4 and probability distributions integrate to 1, ΔJ can be driven arbitrarily close to zero. □

Lemma 1.5 relates the universal approximation properties from Assumption 1.3 to the reward incurred by the policy. We now proceed with the main theoretical result.

Theorem 1.6. An optimal parameter vector θ^* which maximizes (1.4) is such that the base policy $\pi_{\theta*}$ is safe; i.e., the equality $\pi_{\theta*}(\,\cdot \mid s_t) = \pi^{\text{safe}}_{\theta^*}(\,\cdot \mid s_t)$ holds except on a set of measure zero with respect to the stationary state density function induced by $\pi_{\theta*}$ in a Radon-Nikodym sense.

Proof. To prove by contradiction, assume that $\pi_{\theta*}$ is not safe. Define the set of states where the policy diverges from its safe representation as

$$U = \{s : d(\pi^{\text{safe}}_{\theta^*}(\,\cdot \mid s), \pi_{\theta*}(\,\cdot \mid s)) > 0\} = \{s : d(s) > 0\}.$$

with some abuse of notation. Now, consider the measure μ^* induced by the stationary state density function of $\pi_{\theta*}$. Since probability distributions integrate to one, μ^* is finite on compact sets; Euclidian space is also locally compact Hausdorff and second countable, and hence we have μ^* regular (Theorem 7.8 in [Folland, 1999]).

Since d is continuous in s by assumption, U is the inverse image of an open set under a continuous function and is therefore open. Regularity of μ^* and $\mu^*(U) > 0$ (by assumption) implies that there exists a compact set $\bar{U} \subset U$ such that $\mu^*(\bar{U}) > 0$. Since continuous functions attain their minimum over compact sets, we have that $d(s) > \delta$ for all $s \in \bar{U}$ for some $\delta > 0$.

We now show that the difference in objectives between the safe and base policies is given by

$$J^p(\pi^{\text{safe}}_{\theta^*}) - J^p(\pi_{\theta*}) \geq \beta\delta\mu^*(\bar{U}) > 0.$$

Observe that the state action marginal in the expectation (1.4) is always taken with respect to $\pi^{\text{safe}}_{\theta^*}$; therefore, the reward terms vanish and the safety penalty is the only remaining term. By the previous discussion, this is at least $\delta\mu^*(\bar{U})$, providing the desired expression.

Finally, we invoke Lemma 1.5 to construct a policy $\pi_{\theta'} \in \Pi$ such that $J^p(\pi^{\text{safe}}_{\theta^*}) - J^p(\pi_{\theta'}) < \beta\delta\mu^*(\bar{U})/2$, noting that the safety penalty in (1.4) can be driven arbitrarily close to zero by Assumption 1.3 and continuity of d. This implies $J^p(\pi_{\theta'}) > J^p(\pi_{\theta*})$, which is a contradiction. □

Theorem 1.6 shows that the optimal parameters θ^* for our objective (1.4) produce a *safe base policy* $\pi_{\theta*}$. Provided that gradient ascent effectively maximizes (1.4), we can be confident that the policy has learned to behave safely and no longer requires the safety guide. This has three key advantages.

1. *Harmony between the base policy and safety guide.* Without a safety penalty, there is limited incentive for the base policy to learn to correct its own unsafe actions; the executed actions and ensuing rewards are always drawn from the action distribution of the safety guide. As noted in [Koller et al., 2018], this decoupling can lead to a perpetual conflict between the base policy and the safety guide, with the base policy constantly approaching the boundaries of the safe set and the guide constantly correcting. Theorem 1.6 shows that our method resolves this issue.
2. *Mitigation of distributional shift.* One potential concern with this method involves distributional shift; our policy gradient step updates the base policy, while rewards are sampled using the safe policy. Theorem 1.6 implies that as training progresses, the distributional shift between these two policies decays to zero.

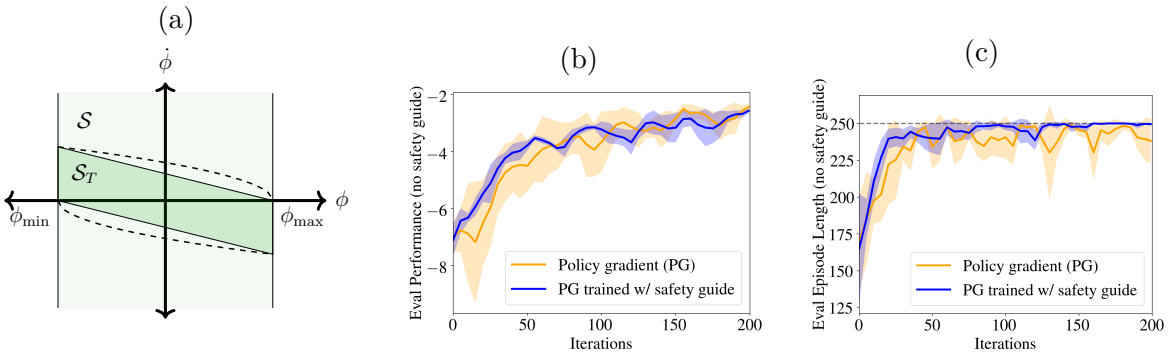


Figure 1.2: Experimental setup and results for the quadrotor setting. (a) The ϕ-$\dot{\phi}$ plane of the quadrotor system safety sets. The dashed lines represent the true bounds of the terminal safety set $\mathcal{S}_T$; we inner approximate this by a polytope. In practice, we also slightly shrink $\phi_{\min}$ and $\phi_{\max}$ by some factor $1-\delta$. (b) Test-time performances of a policy gradient agent trained with and without the safety guide on the double integrator task. The thick line indicates mean performance over five runs, with the shaded area representing the standard deviation. (c) Test-time average episode length. The policy trained with the safety guide achieves the maximum episode length of 250 even when the safety guide is removed, indicating that the base policy has learned to behave safely.

3. *Reduction of computational burden.* Solving the safety guide optimization problem requires significant computational effort. Theorem 1.6 shows that the safety guide can be removed at test time without compromising safety. This can free up agent resources for other tasks.

We note that in the setting where $J^p(\pi_\theta)$ is not completely maximized, the safety penalty $d(s)$ can still be concretely evaluated in any region of the state space. This provides the designer of the system with a quantitative measure of the level of safety of the base policy as well as insights into which regions of the state space are most dangerous.

1.5 Numerical experiments

Consider a two-dimensional quadrotor with state $s_t = [x_t, \dot{x}_t, y_t, \dot{y}_t, \phi_t, \dot{\phi}_t]$, where (x, y) is the quadrotor position and ϕ is the counter-clockwise angle to the vertical. The episode terminates if the quadrotor hits the ground or tilts more than 0.5 radians. For early termination, the reward penalizes impact speed for hitting the ground ($r(s_t) = -1 - 2|\dot{y}_t|$) or rotational speed for excessive tilt ($r(s_t) = -1 - 5|\dot{\phi}_t|$). Otherwise, the quadrotor is incentivized to hover close to the ground while remaining centered horizontally ($r(s_t) = -0.01y_t - 0.01|x_t|$). The control inputs are $a_t = [f_t, \tau_t]$, with $f_t \in [-2, 2]$ the vertical thrust and $\tau_t \in [-2, 2]$ the torque. Using the time step $\Delta t = 0.02$, we simulate the system

using the following linearized dynamics about the hovering equilibrium

$$\begin{bmatrix} x_{t+1} \\ \dot{x}_{t+1} \\ y_{t+1} \\ \dot{y}_{t+1} \\ \phi_{t+1} \\ \dot{\phi}_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 & \Delta t & 0 \\ 0 & 0 & 1 & 0 & 0 & \Delta t \\ 0 & 0 & -g\Delta t & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ \dot{x}_t \\ y_t \\ \dot{y}_t \\ \phi_t \\ \dot{\phi}_t \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \Delta t/m & 0 \\ 0 & \Delta t/I \end{bmatrix} \begin{bmatrix} f_t \\ \tau_t \end{bmatrix}$$

for mass $m = 1$, inertia $I = 1$, and gravity $g = 1$. We design our safety set $\mathcal{S}$ to have the constraints $y \geq 0.05$ and $-0.45 \leq \phi \leq 0.45$. Our terminal safe set $\mathcal{S}_T$ consists of the same position bounds as well as a position-dependent velocity bound that captures the maximum velocity that can be brought to zero by the end of the corresponding safe set interval. Since this curve scales with the square root of distance, we inner approximate this by a polytope (Figure 1.2a). The safety tolerance, planning horizon, and safety penalty are set as $\epsilon = 0.01$, $H = 15$, and $\beta = 1.5$. Our network consists of two hidden layers of size 64 with tanh nonlinearities. We collected 5000 steps per batch with an episode length of 250 steps. The learning rate is 0.002 and discount factor is $\gamma = 0.95$. Our reported results include the top 5 of 10 seeds by average eval performance—a common approach for mitigating policy initialization variance [Wu et al., 2017b]. The safety guide optimization problem is solved using MOSEK [ApS, 2019].

Our training approach achieved perfect safety over a training corpus of a million steps without compromising performance (Figure 1.2b). Furthermore, Figure 1.2c shows that safety guide-trained policy rapidly achieves the optimal average episode length of 250 steps even when the safety guide is removed. This suggests that the safety penalty effectively induces the base policy to behave safely without having to try unsafe actions.

1.6 Conclusion

This work addresses the challenge of safe RL using a novel approach that combines a policy gradient agent with a chance-constrained MPC safety guide. The safety guide receives as input the proposed action distribution from the base policy and imposes additional safety requirements. By design, the safety guide intervenes minimally and modifies the base policy's proposed action distribution only if it inevitably leads towards an unsafe region of the state space. An additional safety penalty on these corrections in the overall objective allows us to provide theoretical guarantees that our base policy learns to behave safely without having to explore unsafe actions. We empirically justify our proposed method through numerical experiments on a linearized quadrotor control task.

Chapter 2

Initial State Interventions for Deconfounded Imitation Learning

Imitation learning suffers from *causal confusion*. This phenomenon occurs when learned policies attend to features that do not causally influence the expert actions but are instead spuriously correlated. Causally confused agents produce low open-loop supervised loss but poor closed-loop performance upon deployment. We consider the problem of masking observed confounders in a disentangled representation of the observation space. Our novel masking algorithm leverages the usual ability to intervene in the initial system state, avoiding any requirement involving expert querying, expert reward functions, or causal graph specification. Under certain assumptions, we theoretically prove that this algorithm is *conservative* in the sense that it does not incorrectly mask observations that causally influence the expert; furthermore, intervening on the initial state serves to strictly reduce excess conservatism. The masking algorithm is applied to behavior cloning for two illustrative control systems: CartPole and Reacher.

This chapter is based on the following published work:

Samuel Pfrommer, Yatong Bai, Hyunin Lee, and Somayeh Sojoudi. Initial state interventions for deconfounded imitation learning. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 2312–2319. IEEE, 2023d.

2.1 Introduction

Imitation learning aims to train an intelligent agent to mimic expert demonstrations for a particular task. Various imitation learning instantiations, such as behavior cloning and inverse reinforcement learning, have been widely applied to fields including robotics [Calinon and Billard, 2007, Krishnan et al., 2018], autonomous driving [Kuefler et al., 2017, Wang et al., 2021c], and optimal navigation [Hussein et al., 2018, Shou et al., 2020]. Imitation learning enables agents to learn from high-quality samples instead of exploring from scratch, leading to significantly higher learning efficiency when compared

with reinforcement learning methods [Bojarski et al., 2016]. This is especially important in safety-critical settings where reinforcement learning are difficult to execute [Pfrommer et al., 2022, Yin et al., 2021]. Even when the flexibility of reinforcement learning is desired, imitation learning can be used to accelerate the learning process [Hester et al., 2017].

Despite its broad applicability, imitation learning exhibits an issue known as *causal confusion* [De Haan et al., 2019]: the learned policy misattributes features which are primarily *correlated* with expert actions as reflecting a *causal* relationship [Kaddour et al., 2022]. This can manifest itself both through the observed features which are spuriously correlated with the expert actions ("nuisance variables") as well as confounders which are available to the expert but not the imitator ("unobserved confounders"). We restrict ourselves to the former, although for completeness we include approaches addressing the latter in our work.

Consider an illustrative example of causal confusion adapted from [De Haan et al., 2019]. The task at hand is learning to drive a car from expert demonstrations. A behavior cloning agent is provided video observations from the driver's perspective, including a brake light on the dashboard. Although the learned braking policy is excellent on the supervised dataset, deployment performance is poor: the agent has effectively learned to brake when the brake light is on, instead of attending to other pedestrians or vehicles. In this case, the brake light is a "nuisance variable," and we can dramatically improve the performance of the policy by covering the brake light and reducing information for the model.

Existing approaches for completely masking such nuisance variables generally require either a queryable expert or access to the expert reward function. The seminal work of [De Haan et al., 2019] introduced a β-Variational Auto Encoder (β-VAE) decomposition the observation space along with a joint policy parameterized by hypothetical causal structures. The space of causal structures can then be searched with two distinct algorithms, one leveraging expert queries and the other based on policy evaluations and reward feedback. The existence of nuisance variables was also noted [Ortega et al., 2021] as part of a broader issue with sequential models that can be addressed with Dagger-style expert queries [Ross et al., 2011a]. The work of [Park et al., 2021] partially addresses the nuisance variable problem by regularizing the learned policies to attend to multiple objects in the scene. While this approach does not require policy executions, it only weakens the learner's attention to a nuisance variable and does not eliminate it completely.

The complementary problem of unobserved confounders considers the setting where experts observe confounding variables that are inaccessible to the learner. In the car driving example, this might include a human driver listening to honking that is not detected with visual sensors. One exciting theoretical line of research in this area [Kumor et al., 2021, Zhang et al., 2020b] presents causal-model derived conditions for imitability and an algorithm for imitating the expert policy when possible. However, these works make the strong assumption that the causal graph is provided to the imitation learning agent. Other efforts to apply causal inference techniques to the unobserved confounder problem either require strong assumptions, such as the knowledge of the expert reward [Etesami

and Geiger, 2020] and purely additive temporally correlated noise [Swamy et al., 2022], or only evaluate simple multi-armed bandit problems [Vuorio et al., 2022].

This work focuses on the problem of observed nuisance variables. Our approach, presented in Section 2.3 leverages initial state interventions to identify and completely mask causally confusing features without relying on expert queries or policy interventions. We provide *conservativeness* guarantees for our method in Section 2.4 and present illustrative experiments in Section 2.5.

2.2 Notation and background

We denote the set of real numbers by $\mathbb{R}$ and the set of natural numbers by $\mathbb{N}$. The set $\{1, \ldots, a\} \subset \mathbb{N}$ is denoted by $[a]$ for $a \in \mathbb{N}$, and similarly $a, \ldots, b \subset \mathbb{N}$ is denoted by $[a \,..\, b]$. For a pair of boolean variables x and y, the notation $\wedge$ denotes the "and" operator while $\vee$ denotes "or." For a set of boolean variables $\{x_1, x_2, \ldots, x_n\}$, the notations $\bigwedge_{i=1}^n x_i$ and $\bigvee_{i=1}^n x_i$ denote $x_1 \wedge x_2 \wedge \ldots \wedge x_n$ and $x_1 \vee x_2 \vee \ldots \vee x_n$, respectively. The logical negation of a boolean variable or vector x is denoted by $\neg x$. We denote the identically zero function on a domain by $\mathbf{0}$, and we write $f(\cdot) \not\equiv \mathbf{0}$ to mean that $f(\cdot)$ is not equivalent to the zero function over its argument—i.e., there exists an input where f is nonzero.

2.2.1 Measure theory and probability

For a random variable X, we introduce the notation $P(x) \in M(X)$ to represent a probability measure over the values x in the domain of X, contained in the space of measures $M(X)$. The uniform measure over an interval $[a, b] \subset \mathbb{R}$ is denoted by $\mathcal{U}(a, b)$. For two measures μ and ν, we say that ν is absolutely continuous with respect to μ if for every μ-measurable set A, $\mu(A) = 0$ implies $\nu(A) = 0$. If ν is absolutely continuous with respect to μ, we let $d\nu/d\mu$ denote the Radon-Nikodym derivative of ν with respect to μ. The standard Lebesgue measure on $\mathbb{R}$ is denoted λ. For a measure μ which is absolutely continuous with respect to λ, we define its L_1 norm in the typical manner

$$\|\mu\|_1 := \int \left| \frac{d\mu}{d\lambda} \right| d\lambda,$$

which we take to be the default norm in the Banach space of measures on $\mathbb{R}$. We denote independence between two random variables using $\perp\!\!\!\perp$ and its negation by $\not\perp\!\!\!\perp$.

2.2.2 Causal graphs and structural causal models

We denote a directed acyclic graph by $\mathcal{G}$, with the presence of a direct edge between nodes X and Y denoted $X \to Y$. For a given node X in $\mathcal{G}$, we let $\mathcal{G}_{\underline{X}}$ denote the graph obtained by deleting outgoing edges from X. We denote sets of nodes in a graph using bold font (e.g., $\mathbf{Z}$). The set of parents of a node X in a graph is denoted by $\mathbf{pa}_X$. A path between two nodes X and Y can consist of arbitrarily directed edges and is said to be blocked by a set of nodes $\mathbf{Z}$ if the path contains any of the following [Pearl, 2009]:

- A chain $I \to M \to J$ with $M \in \mathbf{Z}$.
- A fork $I \leftarrow M \to J$ with $M \in \mathbf{Z}$.
- A collider $I \to M \leftarrow J$ such that $M \notin \mathbf{Z}$ and no descendant of M is in $\mathbf{Z}$.

Two nodes X and Y are said to be d-separated by $\mathbf{Z}$ if $\mathbf{Z}$ blocks every path between X and Y. We call a path with all edges oriented the same direction a directed path.

We leverage Pearl's structural causal model (SCM) formalism [Pearl, 2009]. An SCM $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F} \rangle$ consists of endogenous variables $\mathbf{V}$, exogenous variables $\mathbf{U}$, and structural equations $\mathcal{F}$. Each $V \in \mathbf{V}$ is represented by a node in the causal graph $\mathcal{G}$ and associated with an independently distributed exogenous variable $U_V \in \mathbf{U}$. The structural equations $f_V \in \mathcal{F}$ assign values of a particular node $V \in \mathbf{V}$ as a function $V := f_V(\mathbf{pa}_V, U_V)$ of its parents and associated exogenous variable. The SCM $\mathcal{M}$ induces a joint distribution $P\Big(\mathbf{v}\Big)$ over the endogenous variables $\mathbf{V}$. We say that an SCM $\mathcal{M}$ is faithful to its causal graph $\mathcal{G}$ if the distribution $P\Big(\mathbf{v}\Big)$ induced by $\mathcal{M}$ contains only the pairwise conditional independencies implied by $\mathcal{G}$; i.e. $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ in the joint distribution from $\mathcal{M}$ iff X and Y are d-separated by $\mathbf{Z}$ in $\mathcal{G}$ [Spirtes et al., 2000]. As a notable special case, if $\mathbf{Z}$ is empty and there exists a path from X to Y with no colliders then $X \not\perp\!\!\!\perp Y$.

We define an intervention on a particular node V to be a reassignment of the associated structural equation f_V. This intervention can take the form of a constant intervention $V := v$, which we denote by $\mathrm{do}(V = v)$ for a constant v and may abbreviate to $\mathrm{do}(v)$. We also define a distributional intervention, denoted by $\mathrm{do}(V \sim \tilde{P}(v))$, where we assign V to be drawn from a specified distribution $\tilde{P}(v)$. We denote the post-intervention SCM by $\widetilde{\mathcal{M}}$, with an associated causal graph $\widetilde{\mathcal{G}}$ identical to $\mathcal{G}$ but with incoming edges to V removed. Note that reassigning the associated structural equation for any particular node V induces a new distribution generated by $\widetilde{\mathcal{M}}$ over the set of all endogenous variables $\mathbf{V}$, which we denote by $P(\mathbf{v} \mid \mathrm{do}(V = v))$ or $P(\mathbf{v} \mid \mathrm{do}(V \sim \tilde{P}(v)))$.

2.2.3 Behavior cloning

Behavior cloning uses expert trajectories to train an imitating policy. For the system of interest, we use $d_{\mathcal{S}}$, $d_{\mathcal{I}}$, $d_{\mathcal{O}}$, and $d_{\mathcal{A}}$ to denote the dimensionality of the bounded state space $\mathcal{S} \subseteq \mathbb{R}^{d_{\mathcal{S}}}$, raw image observation space $\mathcal{I} \subseteq \mathbb{R}^{d_{\mathcal{I}}}$, disentangled observation space $\mathcal{O} \subseteq \mathbb{R}^{d_{\mathcal{O}}}$, and action space $\mathcal{A} \subseteq \mathbb{R}^{d_{\mathcal{A}}}$. Let S_t, I_t, O_t, and A_t be vector random variables taking on values in $\mathcal{S}$, $\mathcal{I}$, $\mathcal{O}$, and $\mathcal{A}$, respectively, for a discrete time step $t \in \mathbb{N}$. States variables S_t represent the intrinsic low-dimensional dynamics of the system (e.g. simulator variables) while observations O_t are distilled using a VAE-style framework from high-dimensional image measurements I_t, with $d_{\mathcal{I}} \gg d_{\mathcal{O}}$. The system dynamics assume that S_{t+1} is strictly a function of S_t and A_t. Lower-case script letters $\mathfrak{s} \in [d_{\mathcal{S}}]$, $\mathfrak{o} \in [d_{\mathcal{O}}]$, and $\mathfrak{a} \in [d_{\mathcal{A}}]$ denote specific indices in the state, observation, and action vectors. For example, $S_1^{\mathfrak{s}}$ refers to the real-valued random variable corresponding to the $\mathfrak{s}^{\text{th}}$ state variable at the first time step. We model $W \sim \mathcal{U}(a, b)$ to be an unobserved variable capturing uncontrolled and unknown initialization stochasticity (i.e. a random "seed").

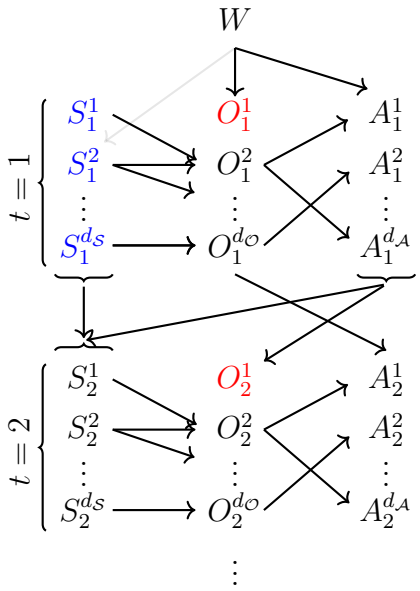


Figure 2.1: An example (unknown) system causal graph $\mathcal{G}_s$. We hope to mask O^1 (e.g. brake light observation), which has no causal edge to any expert action but is correlated with A^1 through the confounding random "seed" W and future spurious correlations. In $\mathcal{G}_s$, W also causally influences S_1^2; however, if we intervene on S_1 (blue) this edge is removed in $\widetilde{\mathcal{G}}_s$ (light shading). This enables our masking algorithm to more reliably leverage state initialization to detect potential causes between observations and actions (Section 2.3.2).

The collection of states, observations, and actions, along with W, comprise endogenous variables in an SCM defining our system. We denote the system SCM by $\mathcal{M}_s$ and denote the corresponding faithful causal graph by $\mathcal{G}_s$. Note that the SCM depends on the choice of policy. Since we aim to infer causalities regarding the expert policy, we generally let any causal relationships refer to the $\mathcal{M}_s$ and $\mathcal{G}_s$ induced by the expert policy unless otherwise stated. We pair the system SCM and causal graph with the tuple $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$. Although nodes in $\mathcal{G}_s$ are individual elements in our vector-valued random variables (i.e., $S_t^\mathfrak{s}$ is a node, not S_t), with some abuse of notation, we let the edge symbol $S_t \to X$ signify that $S_t^\mathfrak{s} \to X$ for some $\mathfrak{s} \in [d_\mathcal{S}]$. Similarly, $X \to S_t$ denotes that $X \to S_t^\mathfrak{s}$ for some $\mathfrak{s}$.

This work evaluates the importance of interventionally assigning the initial state to a particular distribution $S_1 \sim \tilde{P}(s_1)$. This intervention yields a modified SCM $\widetilde{\mathcal{M}}_s$ with a corresponding (not necessarily faithful) causal graph $\widetilde{\mathcal{G}}_s$, which removes the edge $W \to S_1$ in $\mathcal{G}_s$ (Figure 2.1). We collect N arbitrary-length expert trajectories from $\widetilde{\mathcal{M}}_s$. The collection of all such trajectories is denoted ${}^{(1..N)}\boldsymbol{\tau}$. Among these N trajectories, the i^{th} trajectory consists of the tuple

$$ {}^{i}\boldsymbol{\tau} = \langle s_1, \ldots, s_T;\ \imath_1, \ldots, \imath_T;\ o_1, \ldots, o_T;\ a_1, \ldots, a_T \rangle, $$

where lowercase letters represent a concrete random variable value (to avoid confusion with indices, we use $\imath_t$ to denote a value of I_t). Implicit in this definition is the existence

of an *encoder* $\psi_e : \mathcal{I} \to \mathcal{O}$ mapping each image $\imath_t$ to a disentangled observation o_t. We characterize trajectories as containing observations for simplicity; our environment only provides the images $\imath_t$, and the extraction of disentangled observations o_t is method-dependent.

When training agents on ${}^{(1..N)}\boldsymbol{\tau}$, we parameterize policies as a neural network $f_\theta : \mathcal{I}^L \to \mathcal{A}$. The neural policy maps some history of observations to an action a_t via

$$a_t = f_\theta(\imath_t, \imath_{t-1}, \ldots, \imath_{t-L+1}). \tag{2.1}$$

We then train f_θ via standard behavior cloning by randomly sampling batches of images and expert actions from ${}^{(1..N)}\boldsymbol{\tau}$ and performing supervised regression.

2.2.4 Statistical independence tests

Our method relies on identifying whether two random variables are statistically dependent. While this is a challenging problem with a rich literature [Sheskin, 2020], in this paper, we only briefly introduce a well-known independence test for continuous distributions based on Hoeffding's D statistic [Even-Zohar, 2020, Hoeffding, 1948]. Consider two real-valued random variables X and Y with a joint cumulative distribution function $F(x, y) = P\Big(X \leq x, Y \leq y\Big)$. Hoeffding's D statistic operates on N_{Hoeff} independent pairs of observations $\{(X_1, Y_1), \ldots (X_{N_{\text{Hoeff}}}, Y_{N_{\text{Hoeff}}})\}$ and outputs a real number D in the range $[-0.5, 1]$, with $D > 0$ indicating dependence. The computational complexity of calculating this statistic is $\mathcal{O}(N_{\text{Hoeff}} \log N_{\text{Hoeff}})$. For absolutely continuous joint distributions, the D statistic is unbiased and consistent as $N_{\text{Hoeff}} \to \infty$, meaning that the dependence is correctly represented with probability arbitrarily close to 1. Subsequent variations of the D statistic maintain consistency even for non-absolutely continuous joint distributions [Blum et al., 1961], although these complications are outside the scope of our work. We refer to the independence test based on the Hoeffding's D statistic as Hoeffding's independence test.

2.3 Problem statement and method

We address the *causal confusion* problem in imitation learning and aim to mask spuriously correlated observations. To this end, we investigate the following problem statement:

How can we identify and eliminate spuriously correlated observations without relying on online expert queries or knowledge of the expert reward function?

Our approach addresses this problem in a theoretically grounded way. Specifically, we make the following contributions:

1. We present an algorithm for identifying and masking causally confusing observations *without relying on reward function knowledge, expert queries, or causal graph knowledge.*

2. We prove that, under certain conditions, our procedure is *conservative*: if an observation causally affects the expert actions, it will not be masked.
3. We demonstrate the importance of *initial state interventions* by showing theoretically that the interventions reduce excess conservatism in the masking algorithm.

Section 2.3.1 presents and analyzes the assumptions underlying our method. Section 2.3.2 motivates and derives our method, which is then presented formally in Section 2.3.3.

2.3.1 Assumptions

Our proposed method relies on the following assumptions to ensure the theoretical guarantees in Section 2.4.

Assumption 2.1. The system causal graph $\mathcal{G}_s$ is time invariant. Namely, consider two arbitrary time steps $t, t' \in \mathbb{N}$ with $t' \geq t$ and two arbitrary time-indexed variables X_t and $Y_{t'}$ in $\mathcal{G}_s$. Then if $X_t \to Y_{t'}$ is an edge in $\mathcal{G}_s$, then so is $X_{t+\Delta} \to Y_{t'+\Delta}$ for any $\Delta \in \mathbb{Z}$ such that $\min(t + \Delta, t' + \Delta) \geq 1$.

Time-invariance of the expert policy allows for causal inference via interventions on the initial state S_1. Otherwise we would require the ability to intervene at arbitrary time steps, which is unrealistic for most real-world systems.

Assumption 2.2. The expert policy attends only to observational information derived from the underlying state. Namely, if $O_t^{\mathscr{o}} \to A_{t'}^{\mathscr{a}}$ in $\mathcal{G}_s$ for $t, t' \in \mathbb{N}$ with $t' \geq t$, then there must exist an index $\mathfrak{s}$ such that $S_t^{\mathfrak{s}} \to O_t^{\mathscr{o}}$.

Assumption 2.2 reflects the intuition that the expert policy itself must not be fooled by spurious information in the observation space. This is a natural assumption in the considered case where the dynamics of the underlying system depend only on S_t, not O_t.

Assumption 2.3. The expert policy reacts to observations within a *reaction horizon* $H \in \mathbb{N}$. Specifically, if $O_t^{\mathscr{o}} \to A_{t_1}^{\mathscr{a}}$ in $\mathcal{G}_s$ for some $t_1 > t$ and particular $t \in \mathbb{N}$, $\mathscr{o} \in [d_{\mathcal{O}}]$, and $\mathscr{a} \in [d_{\mathcal{A}}]$, then there exists a $t_2 \in [t \,..\, t + H - 1]$ such that $O_t^{\mathscr{o}} \to A_{t_2}^{\mathscr{a}}$.

Assumption 2.3 imposes a horizon within which the expert is assumed to react to a hypothetical intervention on a state or observation. For finite-length trajectories, H can be chosen to be the entire trajectory length, with the algorithm and theory still valid. As such, H introduces a hyperparameter that allows for more tractable computation under some assumptions on the expert. Our experiments show that H can be much smaller than the trajectory length for certain practical dynamic systems and experts.

Finally, we formalize a class of SCMs that behave nicely under interventions.

Assumption 2.4. The system SCM $\mathcal{M}_s = \langle \mathbf{V}, \mathbf{U}, \mathcal{F} \rangle$ is *interventionally absolutely continuous*, meaning that for any disjoint sets of nodes $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, the interventional

distribution $P\Big(\mathbf{z} \mid \mathrm{do}(\mathbf{X} = \mathbf{x}), \mathbf{y}\Big)$ is absolutely continuous with respect to the Lebesgue measure, has a bounded Radon-Nikodym derivative, and is continuous as a measure-valued function with respect to $\mathbf{x}$ and $\mathbf{y}$.

Assumption 2.4 stipulates that the probability distribution induced by our SCM on any set of non-intervened nodes is absolutely continuous with bounded density. This is a technical condition that facilitates analysis and allows us to assert that Hoeffding's test is consistent. We note that subsequent D-statistic variations allow for non-absolutely continuous joint distributions [Blum et al., 1961] — we leave the theoretical and practical implications of more sophisticated testing to future work.

2.3.2 Derivation

Our aim is to mask a particular observation $O^{\mathscr{o}}$ across all time steps if it has no causal effect on any expert action within the reaction horizon. As intervening on observations is impractical, this causality is challenging to deduce. We do, however, assume the ability to intervene on the system in one specific instance: setting the state variables S_1 at initialization. We manipulate S_1 to infer the possible existence of a true causal relationship.

We first motivate our approach from an arbitrary time step $t \geq 2$ before specializing on the initialization. Consider arbitrary observation and action indices $\mathscr{o} \in [d_{\mathcal{O}}], \mathscr{a} \in [d_{\mathcal{A}}]$ and time steps $t, t' \in \mathbb{N}$ with $t' \in [t \,..\, t + H - 1]$. Assumption 2.2 states that a causal effect $O_t^{\mathscr{o}} \to A_{t'}^{\mathscr{a}}$ must arise from a larger causal path

$$S_t^{\mathscr{s}} \to O_t^{\mathscr{o}} \to A_{t'}^{\mathscr{a}} \tag{2.2}$$

in $\mathcal{G}_s$, for some state variable index $\mathscr{s} \in [d_{\mathcal{S}}]$. We now observe that by faithfulness of $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$ it must be that $S_t^{\mathscr{s}} \not\perp\!\!\!\perp O_t^{\mathscr{o}}$ and $S_t^{\mathscr{s}} \not\perp\!\!\!\perp A_{t'}^{\mathscr{a}}$; i.e. the causal relationships in $\mathcal{G}_s$ imply probabilistic dependencies in the induced distribution from $\mathcal{M}_s$. Note that these are *statistical* statements which can be ascertained from the observational data. We define the boolean variable ${}^{(t,t')}D_{\mathscr{s},\mathscr{a}}^{\mathscr{o}}$ to check these independencies:

$${}^{(t,t')}D_{\mathscr{s},\mathscr{a}}^{\mathscr{o}} \coloneqq (S_t^{\mathscr{s}} \not\perp\!\!\!\perp O_t^{\mathscr{o}}) \wedge (S_t^{\mathscr{s}} \not\perp\!\!\!\perp A_{t'}^{\mathscr{a}}), \tag{2.3}$$

and introduce the "potential cause" notation

$$O_t^{\mathscr{o}} \dashrightarrow A_{t'}^{\mathscr{a}} \coloneqq \bigvee_{\mathscr{s}=1}^{d_{\mathcal{S}}} \left({}^{(t,t')}D_{\mathscr{s},\mathscr{a}}^{\mathscr{o}} \right). \tag{2.4}$$

The boolean-valued statement $O_t^{\mathscr{o}} \dashrightarrow A_{t'}^{\mathscr{a}}$ intuitively captures that, based on observational data, there may (but need not) exist a true causal edge $O_t^{\mathscr{o}} \to A_{t'}^{\mathscr{a}}$ generated by some $S_t^{\mathscr{s}}$ as in (2.2). We denote by $O_t^{\mathscr{o}} \not\dashrightarrow A_{t'}^{\mathscr{a}}$ the logical negation of $O_t^{\mathscr{o}} \dashrightarrow A_{t'}^{\mathscr{a}}$. As we will elaborate in more detail shortly, if $O_t^{\mathscr{o}} \not\dashrightarrow A_{t'}^{\mathscr{a}}$ for all actions $\mathscr{a} \in [d_{\mathcal{A}}]$ and t' in the reaction horizon, we want to "mask" the $\mathscr{o}^{\text{th}}$ observation as it has no causal effect on the

expert action but could be spuriously correlated in a way that undermines the imitation learning policy performance.

It is immediate from the above faithfulness argument that for $t \geq 2$, we have the implication

$$O_t^o \to A_{t'}^a \implies O_t^o \dashrightarrow A_{t'}^a. \tag{2.5}$$

Note that (2.5) provides a *conservativeness* guarantee: if an observation causally influences an action, we will not mistakenly conclude from observational data that it does not, and hence incorrectly mask an observation that is actually used by the expert policy. However, this conservativeness is not apparent for $t = 1$ in the modified causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$, where we intervene to specify the initial state distribution, overriding the natural randomness resulting from W and potentially breaking faithfulness. As a simple counterexample, initializing S_1 to a constant vector would make $S_1^\mathfrak{s}$ independent of every other random variable in the causal graph, and therefore no potential causes could be discovered as (2.3) would always be false. Nonetheless, when a sufficiently sensible initialization distribution is used, we prove that the conservativeness result still holds under intervention on S_1 in Section 2.4.

The reverse implication to (2.5) does not hold. It is possible that spurious statistical relationships exist while a causal edge $O_t^o \to A_{t'}^a$ does not. Indeed, for $t \geq 2$, the abundance of chronologically antecedent variables virtually guarantees that all variables have share a common cause and hence a statistical dependence. The sole exception is the initial state S_1. By intervening on S_1, we eliminate the incoming edge from the only possible common ancestor W in the causal graph (Figure 2.1). Therefore, we expect that this interventional ability should help eliminate potential causes $O_1^o \dashrightarrow A_{t'}^a$ which do not exist in the true causal graph and reduce excessive conservativeness in the algorithm. We analyze this idea formally in Section 2.4.

The culmination of our efforts is described in Algorithm 2.1, which checks for potential causes at $t = 1$ using expert data ${}^{(1..N)}\boldsymbol{\tau}$ collected from the interventional system $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$. Note that Algorithm 2.1 invokes the HOEFFDING routine to compute Hoeffding's D statistic for independence between two variables. This test is computed over our dataset of trajectories ${}^{(1..N)}\boldsymbol{\tau}$, extracting exactly one pair of variables from each trajectory ($N_{\text{Hoeff}} = N$). For concreteness, consider the call HOEFFDING($S_1^2 \not\perp\!\!\!\perp A_3^4$ in ${}^{(1..N)}\boldsymbol{\tau}$). This extracts, from each trajectory, the second element of the $t = 1$ state and the fourth element of the $t = 3$ action. These N pairs are then supplied to Hoeffding's test, which returns a real number in the range $[-0.5, 1]$, with a value greater than zero indicating dependence. Since perfect observational disentanglement is unrealistic, we introduce a small positive threshold hyperparameter γ.

Algorithm 2.1 is presented for readability and can be implemented more efficiently. The Hoeffding tests between $S_t^\mathfrak{s}$ and $O_t^o, A_{t'}^a$ can be precomputed, yielding the runtime

$$O\left(d_{\mathcal{S}}(d_{\mathcal{O}} + H d_{\mathcal{A}}) N \log N\right),$$

where $N \log N$ is the cost of evaluating Hoeffding's test for a specific pair of variables over N trajectories. In practice, Hoeffding's test executions are very fast—on the order of milliseconds for $N = 10^3$—and incur a negligible overhead compared with the training time of imitation learning.

Remark. The reader may have noticed that our approach bears a resemblance to *instrumental variable regression*, a statistical technique for estimating causal relationships that has also received some attention in the causal imitation learning literature [Swamy et al., 2022]. We emphasize that $S_t^{\diamond}$ does not constitute a valid instrumental variable in the causal path (2.2) as there may be many other paths between $S_t^{\diamond}$ and $A_{t'}^{\diamond}$ which are not mediated by $O_t^{\diamond}$. Thus while the spirit of our approach is related to instrumental variable regression, we cannot use $S_t^{\diamond}$ to precisely determine a causal relationship between $O_t^{\diamond}$ and $A_{t'}^{\diamond}$ and only use $S_t^{\diamond}$ to provide evidence of a potential cause.

2.3.3 Imitation learning workflow

Drawing on the masking approach developed in Section 2.3.2, we summarize our overall deconfounded imitation learning workflow as the following four steps.

1. Collect random-policy trajectories to learn a observation representation using a β-VAE, denoted by $\psi_d \circ \psi_e : \mathcal{I} \to \mathcal{I}$, with an encoder $\psi_e : \mathcal{I} \to \mathcal{O}$ and decoder $\psi_d : \mathcal{O} \to \mathcal{I}$. For a well-trained β-VAE, $\psi_d \circ \psi_e$ approximates the identity. We rely on β-VAEs' latent space regularization to produce disentangled observations.
2. Collect a sequence of N trajectories ${}^{(1..N)}\boldsymbol{\tau}$ from the expert policy, with the starting state distribution $\tilde{P}(s_1)$ over $\mathcal{S}$ having any density that is everywhere nonzero (e.g. uniform).
3. Execute Algorithm 2.1 on ${}^{(1..N)}\boldsymbol{\tau}$ to obtain the observation mask $\widetilde{m} \in \{0,1\}^{d_{\mathcal{O}}}$, where $\widetilde{m}_{\diamond} = 1$ if the $\diamond^{\text{th}}$ observation is to be masked.
4. Train the final policy $g_\theta : \mathcal{I}^L \to \mathcal{A}$ on ${}^{(1..N)}\boldsymbol{\tau}$ using standard supervised learning; g_θ masks the disentangled observation space using $\widetilde{m}$ before executing a learnable policy network f_θ:

$$g_\theta(\imath_t, \ldots, \imath_{t-L+1}) = f_\theta(\tilde{\psi}(\imath_t), \ldots, \tilde{\psi}(\imath_{t-L+1})),$$

where the masked β-VAE $\tilde{\psi} : \mathcal{I} \to \mathcal{I}$ has its weights fixed and is defined as

$$\tilde{\psi}(\imath) = \psi_d(\neg\widetilde{m} \odot \psi_e(\imath)).$$

Note that this overall structure generally follows the seminal work of [De Haan et al., 2019]. Our key contribution is Algorithm 2.1, which provides a mask for the disentangled observations without relying on expert queries, the expert reward function, or specification of the causal graph. A visualization of Algorithm 2.1 is provided in Figure 2.2 for the CartPole system considered in the experiments. We show in Section 2.4 that Algorithm 2.1 enjoys notable theoretical guarantees.

Algorithm 2.1 Masking algorithm

Hyperparameter $\gamma > 0$.

procedure MASK($^{(1..N)}\boldsymbol{\tau}$)
  Initialize $\widetilde{m} \in \{0,1\}^{d_{\mathcal{O}}}$ to be an all-zero vector.
  for $o = 1, \ldots, d_{\mathcal{O}}$ **do**
    Mask the o^{th} observation according to

$$\widetilde{m}_o \leftarrow (O_1^o \not\dashrightarrow A_{t'}^a \;\; \forall a \in [d_{\mathcal{A}}], \;\; \forall t' \in [H]), \tag{2.6}$$

    computing $O_1^o \not\dashrightarrow A_{t'}^a$ using CHECK.
  end for
  return $\widetilde{m}$
end procedure

procedure CHECK$\{O_t^o \dashrightarrow A_{t'}^a\}(^{(1..N)}\boldsymbol{\tau})$
  for $s = 1, \ldots, d_{\mathcal{S}}$ **do**
    $a \leftarrow$ HOEFFDING$(S_t^s \not\perp\!\!\!\perp O_t^o$ in $^{(1..N)}\boldsymbol{\tau}) > \gamma$
    $b \leftarrow$ HOEFFDING$(S_t^s \not\perp\!\!\!\perp A_{t'}^a$ in $^{(1..N)}\boldsymbol{\tau}) > \gamma$
    if $a \wedge b$ **then**
      return True
    end if
  end for
  return False
end procedure

2.4 Theoretical guarantees

In this section, we delve into the theoretical properties of Algorithm 2.1. Theorem 2.5 demonstrates that if we intervene on the initial state S_1 and meet certain conditions in the infinite-trajectory regime, the algorithm remains *conservative*, ensuring that no observation that causally influences the expert is mistakenly masked. Additionally, Theorem 2.6 and Proposition 2.7 highlight the effectiveness of intervening on S_1 in mitigating overconservativeness in the masking algorithm. Specifically, Theorem 2.6 asserts that the correctly masked observations under the original causal model $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$ will also be masked under the intervened causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$. Proposition 2.7 showcases a particular set of systems where the intervention only results in masks under $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$, providing compelling evidence that the masking algorithm is more effective after intervening on S_1.

All subsequent theory relies on Assumptions 2.1-2.4, and for brevity we defer proofs and auxiliary lemmas to the appendix. We now introduce the main conservativeness theorem and provide a short proof sketch.

Theorem 2.5. In the faithful system causal model $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$, assume that the measure-

valued function $w \mapsto P(v \mid \mathrm{do}(\mathbf{Z} = \mathbf{z}), w)$ is continuous for any set of nodes $\mathbf{Z}$ and $V \notin \mathbf{Z}$.

Let there exist a causal edge $O_t^{\mathfrak{o}} \to A_{t'}^{\mathfrak{a}}$ in $\mathcal{G}_s$ for some $t, t' \in \mathbb{N}$, $t' \geq t$, and indices $\mathfrak{o} \in [d_{\mathcal{O}}]$ and $\mathfrak{a} \in [d_{\mathcal{A}}]$. Then in the interventional causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$ where the initial state distribution $\tilde{P}(s_1)$ has everywhere-nonzero density on $\mathcal{S}$, $O^{\mathfrak{o}}$ is almost surely not masked by Algorithm 2.1 for almost every uniform parameterization of W as the number of trajectories $N \to \infty$; i.e., (2.6) correctly evaluates to true.

Proof sketch. By Assumptions 2.1 and 2.3, we can WLOG consider $t = 1$ with $t' \in [H]$. If $O_1^{\mathfrak{o}} \to A_{t'}^{\mathfrak{a}}$, by Assumption 2.2 there exists an edge $S_1^{\mathfrak{s}} \to O_1^{\mathfrak{o}}$ for some $\mathfrak{s}$. We show that in the SCM $\widetilde{\mathcal{M}}_s$ where we intervene distributionally on S_1, we have that $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp O_1^{\mathfrak{o}}$ and $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp A_{t'}^{\mathfrak{a}}$. The arguments are similar, so we informally sketch the proof for the former.

To show that $S_1^{\mathfrak{s}}$ and $O_1^{\mathfrak{o}}$ are dependent, it suffices to find a particular pair of states $S_1^{\mathfrak{s}} = \alpha, \alpha'$ which induce different probability measures $P\Big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha)\Big)$ (resp. α') over $O_1^{\mathfrak{o}}$. We marginalize out the random seed w from our original measure of interest $P\Big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha)\Big)$ via the integral

$$P\Big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha)\Big) = \int_a^b P\Big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha), w\Big) p(w) dw,$$

where we model $w \sim \mathcal{U}(a, b)$. Note that the right-hand integral above in fact yields a measure over $o_1^{\mathfrak{o}}$. We now aim to show that the statement

$$\exists \alpha, \alpha' \text{ s.t. } \left\| \int_a^b \Big[P\Big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha), w\Big) - P\Big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha'), w\Big) \Big] dw \right\|_1 > 0 \tag{2.7}$$

holds Lebesgue-almost everywhere for $(a, b) \in \mathbb{R}^2$. By faithfulness of $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$ and the path from $S_1^{\mathfrak{s}}$ to $O_1^{\mathfrak{o}}$, do-calculus rules yield that for any random seed $W = w$ there exist an α, α' such that

$$\left\| P\Big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha), w\Big) - P\Big(o_1^{\mathfrak{o}} \mid \mathrm{do}(S_1^{\mathfrak{s}} = \alpha'), w\Big) \right\|_1 > 0. \tag{2.8}$$

We then analyze the sensitivity of (2.7) with respect to the integration bounds a and b. Namely, for any $(\bar{a}, \bar{b})$ where the left-hand side of (2.7) vanishes, (2.8) yields that there exists an open ball around $\bar{b}$ in which (2.7) holds everywhere except $(\bar{a}, \bar{b})$. An argument from Fubini's theorem then shows that (2.7) holds for almost all (a, b). Appealing to the consistency of Hoeffding's independence test concludes the proof. □

Theorem 2.5 guarantees that Algorithm 2.1 maintains conservativeness by correctly preserving unmasked observations that causally impact expert actions. This outcome is consistent with the discussion in Section 2.3.2, where we observed that the faithfulness of $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$ ensures the correctness of the algorithm when we do not intervene on S_1 and allow the initial state to be naturally generated from W. Theorem 2.5 establishes that this property also holds in the interventional system $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$, where we assign $S_1 \sim \tilde{P}(s_1)$.

We now theoretically demonstrate the benefits of intervening with $\tilde{P}(s_1)$. Specifically, we show that this intervention reduces the excess conservatism in the masking algorithm by removing income edges from W in the causal graph, thereby eliminating a potential avenue of confounding.

Theorem 2.6. Let m denote the potential-cause test evaluated by Algorithm 2.1 on the distribution induced by the non-interventional system $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$, and let $\widetilde{m}$ be the original test on the interventional system $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$ where $\tilde{P}(s_1)$ has everywhere-nonzero density on $\mathcal{S}$. If m_o correctly evaluates to true for a particular $o \in [d_{\mathcal{O}}]$, then $\widetilde{m}_o$ also evaluates to true almost surely as the number of trajectories $N \to \infty$.

Theorem 2.6 assures us that intervening with $\tilde{P}(s_1)$ does not lead to more conservative masking than the original system. We now provide a specific class of SCMs for which the intervention strictly improves the mask.

Proposition 2.7. Let $\widetilde{m}$ and m be as in Theorem 2.6, and consider a particular observation index $o \in [d_{\mathcal{O}}]$ such that the only incoming edge to O_1^o is $W \to O_1^o$. Then if in $\mathcal{G}_s$ there exists the fork $S_1^s \leftarrow W \to O_1^o$ for some $s \in [d_{\mathcal{S}}]$ and a directed path from S_1^s to some A_t^a, with $t \in [H], a \in [d_{\mathcal{A}}]$, $\widetilde{m}_o$ correctly masks the o^{th} observation almost surely as the number of trajectories $N \to \infty$ while m_o does not.

In summary, Theorem 2.5 shows that masking with the intervened initial state $\tilde{P}(s_1)$ maintains conservatism; Theorem 2.6 states that intervening on $\tilde{P}(s_1)$ is no more conservative than masking with the unintervened causal model; and Proposition 2.7 shows that intervening on $\tilde{P}(s_1)$ results in a strictly less conservative mask for a certain class of systems.

2.5 Experiments

We evaluate our approach on two custom simulated environments: CartPole and Reacher. Each of these environments contains a nuisance feature which is likely to induce causal confusion. Our masking approach can successfully eliminate these spuriously correlated features.

2.5.1 Environments

Both considered environments are modified to include a nuisance feature corresponding to the previous action taken by the expert (analogous to the brake light example). For each environment, the expert is a standard constrained finite-time optimal control policy which minimizes cumulative trajectory loss. This expert reward function is not provided to the imitation learning agent.

CartPole. This environment consists of a standard planar cart-pole system with a continuous scalar horizontal force applied to the cart. A quadratic cost is imposed for deviations from the vertical target state. The spuriously correlated feature is a colored

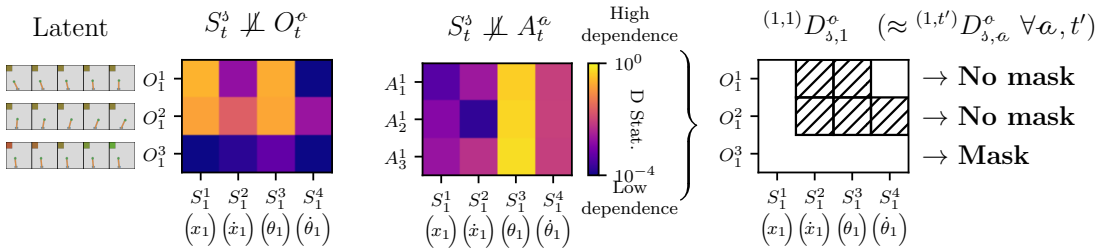


Figure 2.2: Masking algorithm visualization for the CartPole environment with reaction horizon $H = 3$. Latent space interpolation of the β-VAE reveals that O^1 and O^2 capture some combined positional/angular information, while O^3 captures the disentangled confounder (color of the confounding square). This last observation shares virtually no dependence (Hoeffding's D statistic less than $\gamma = 10^{-3}$) with any state variable due to interventions on S_1 (note the log scale). This means that $^{(1,t')}D^\mathfrak{o}_{\mathfrak{s},\mathfrak{a}}$ is false (no cross hatches) for $\mathfrak{o} = 3$ and all $\mathfrak{s} \in [d_\mathcal{S}]$, regardless of $\mathfrak{a}$ and t'; i.e. $O_1^3 \not\dashrightarrow A_{t'}^\mathfrak{a}$ for all $\mathfrak{a} \in [d_\mathcal{A}]$ and $t' \in [H]$, and we can mask the confounder O^3.

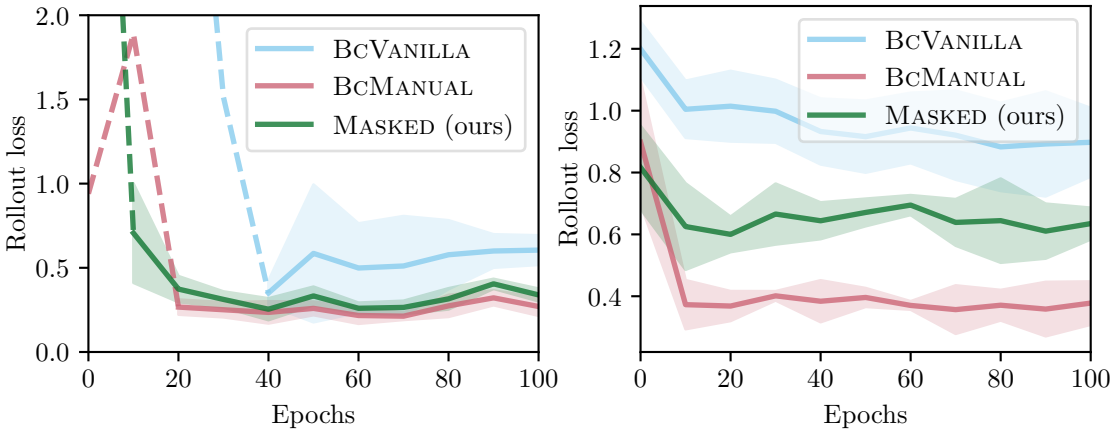
square in the upper-left corner of each image, which interpolates between green and red depending on the most recently executed action.

Reacher. We consider a top-down version of a two-dimensional two-joint Reacher environment [Brockman et al., 2016]. The environment penalizes squared distance of the end effector to a black target dot. The target location is included in the state vector, thus satisfying Assumption 2.2. Two torques, one per joint, are specified as the control inputs; the nuisance feature is a red dot in the upper-left corner whose horizontal position and vertical position encode the two control inputs from the previous time step. This "joystick" introduces a different kind of nuisance feature than in the CartPole environment.

2.5.2 Discussion

We compare the performance of our masked policy against vanilla behavior cloning. The baseline behavior cloning policy is denoted by BcVanilla, and our masked policy is denoted by Masked. For reference, we also measure the performance of the behavior cloning policy with the confounding signals manually removed by superimposing a white square on the upper-left corner, denoted BcManual. We emphasize that BcManual requires human judgement to manually eliminate spurious confounders; we show that we can approach this performance in a principled and automated way.

Figure 2.3 displays our experimental results. For CartPole, the policies were not able to consistently stabilize the pendulum at the beginning of training, leading to high loss variance. Across both environments, the Masked policy substantially outperforms the vanilla behavior cloning policy BcVanilla. It is worth noting that Masked approaches the manually deconfounded baseline's performance without requiring expert queries, access to the expert reward function, or pre-specified information on the causal graph in the deconfounding procedure. However, there is a gap between the performance of our



(a) CartPole (b) Reacher

Figure 2.3: Evaluation rollout loss on CartPole (a) and Reacher (b) across training epochs. Lines denote mean performance over 5 runs while shaded areas indicate standard deviation. To limit visual clutter, for standard deviations greater than 1 shading is omitted and the mean is drawn with a dashed line. Our Masked policy approaches the performance of the manually-deconfounded BcManual baseline, while BcVanilla struggles due to causally confusing features.

method and manual masking for the Reacher environment. This is likely attributable to imperfect disentanglement in the β-VAE, and we expect that our approach could benefit substantially from future research in disentangled representation learning.

Figure 2.2 provides a visualization of our masking procedure and the resulting mask for the CartPole environment. Note that our algorithm masks the third observation O^3, corresponding precisely to the manually masked confounding square. While we use a latent space size of three (the precise number of independent factors of variation) for visualization purposes, our masking procedure is fully functional for larger choices of the latent size. For Reacher, although there are 6 factors of variation in each image, a larger latent space of size 12 yielded superior disentanglement and reconstruction performance.

The most significant limitation of our work, besides the explicitly stated assumptions, is the requirement that confounding factors are observable and can be neatly disentangled. While this holds for the environments considered in this work, more complex environments may introduce entanglement between causally confusing features and important features to which the expert policy actually attends. We introduce the Hoeffding threshold hyperparameter γ to mitigate this concern; however, investigating more principled methods for handling incomplete disentanglement would be an exciting area of future work.

2.6 Conclusion

This work introduces a novel method to address the causal confusion problem in imitation

learning. The proposed method leverages the typical imitation learning ability to intervene in the initial system state. Unlike previous works, our method masks causally confusing observations without relying on online expert queries, knowledge of the expert reward function, or specification of the causal graph. Our theoretical results establish that our masking algorithm is *conservative*, with excess conservatism strictly reduced by interventions on the initial state. We illustrate the effectiveness of our method with experiments on CartPole and Reacher.

Part II

Robustness

Chapter 3

Projected Randomized Smoothing for Certified Adversarial Robustness

Randomized smoothing is the current state-of-the-art method for producing provably robust classifiers. While randomized smoothing typically yields robust ℓ_2-ball certificates, recent research has generalized provable robustness to different norm balls as well as anisotropic regions. This work considers a classifier architecture that first projects onto a low-dimensional approximation of the data manifold and then applies a standard classifier. By performing randomized smoothing in the low-dimensional projected space, we characterize the certified region of our smoothed composite classifier back in the high-dimensional input space and prove a tractable lower bound on its volume. We show experimentally on CIFAR-10 and SVHN that classifiers without the initial projection are vulnerable to perturbations that are normal to the data manifold and yet are captured by the certified regions of our method. We compare the volume of our certified regions against various baselines and show that our method improves on the state-of-the-art by many orders of magnitude.

This chapter is based on the following published work:

Samuel Pfrommer, Brendon G. Anderson, and Somayeh Sojoudi. Projected randomized smoothing for certified adversarial robustness. *Transactions on Machine Learning Research*, 2023b.

3.1 Introduction

Despite their state-of-the-art performance on a variety of machine learning tasks, neural networks are vulnerable to adversarial inputs—inputs with small (often human-imperceptible) noise that is maliciously crafted to induce failure [Biggio et al., 2013, Nguyen et al., 2015, Szegedy et al., 2014]. This sensitive behavior is unacceptable in contemporary safety-critical applications of neural networks, such as autonomous driving [Bojarski et al., 2016, Wu et al., 2017a] and the operations of power systems [Kong et al., 2017]. The

works Eykholt et al. [2018] and Liu et al. [2019a] highlight the validity and eminence of these threats, wherein both physical and digital adversarial perturbations are shown to cause image classification models to misclassify vehicle traffic signs.

Heuristics have been proposed to defend against various adversarial attacks, only to be defeated by stronger attack methods, leading to an "arms race" in the literature [Athalye et al., 2018, Carlini and Wagner, 2017, Kurakin et al., 2017, Madry et al., 2018, Uesato et al., 2018]. This has motivated researchers to consider certifiable robustness—theoretical proof that models perform reliably when subject to arbitrary attacks of a bounded norm [Anderson et al., 2020, Ma and Sojoudi, 2021, Raghunathan et al., 2018, Weng et al., 2018, Wong and Kolter, 2018]. Randomized smoothing, popularized in Cohen et al. [2019a], Lecuyer et al. [2019], Li et al. [2019], remains one of the state-of-the-art methods for generating classifiers with certified robustness guarantees. Instead of directly classifying a given input, randomized smoothing intentionally corrupts the input with random noise and returns the most probable class, which, intuitively, "averages out" any potential adversarial perturbations in the data.

The seminal work Cohen et al. [2019a] certifies that no adversarial perturbation within a certain ℓ_2-ball can cause the misclassification of a smoothed model using isotropic Gaussian noise of a fixed variance. Recent works have attempted to certify larger regions of the input space by turning to randomized smoothing with optimized variances [Zhai et al., 2020], input-dependent variances [Alfarra et al., 2020, Wang et al., 2021a], anisotropic distributions [Eiras et al., 2021], and semi-infinite linear programming [Anderson et al., 2022]. However, for a fixed variance, the certified radius is upper-bounded by a constant in the dimension d of the input [Kumar et al., 2020], implying that the volume of the certified ℓ_2-ball degrades factorially fast as $O(K^d\Gamma(\frac{d}{2}+1)^{-1})$, where Γ is Euler's gamma function and K is some positive constant [Folland, 1999]. Current input-dependent and anisotropic smoothing approaches have similarly been shown to suffer from the curse of dimensionality [Súkeník et al., 2021].

The small certified regions of randomized smoothing in high dimensions corroborate empirical findings that show increased robustness when precomposing classifiers with dimensionality reduction, e.g., principal component analysis projections [Bhagoji et al., 2018] and autoencoders [Sahay et al., 2019]. These findings align with the manifold hypothesis, which posits that real datasets lie on a low-dimensional manifold in a high-dimensional feature space [Fefferman et al., 2016], and related results showing that perturbation directions most useful to an adversary are ones normal to this manifold [Jha et al., 2018, Zhang et al., 2020c]. Thus, projecting inputs onto the manifold, or at least a low-dimensional subspace containing the manifold, should increase classification robustness. Methods taking this approach, such as Mustafa et al. [2019] and Alemany and Pissinou [2022], have worked well as heuristics, but lack theoretical robustness guarantees. Motivated by these works, we aim to enlarge the certifiably robust regions of randomized smoothing by performing the smoothing in a low-dimensional space in which adversarial access to the data's statistically insignificant yet vulnerable features has been eliminated.

3.1.1 Contributions

We propose *projected randomized smoothing*, whereby inputs are projected onto a low-dimensional linear subspace in which randomized smoothing is applied before classification. Our method combines the empirical successes of dimension-reducing projection methods with the theoretical guarantees of randomized smoothing to achieve the following contributions:

1. We theoretically characterize the geometry of the certified region in the input space and prove a tractable lower bound on the volume of this certified region.
2. We empirically demonstrate that classifiers can be attacked along subspaces spanned by statistically insignificant features that contribute nothing to classification accuracy, which are vulnerabilities that projected randomized smoothing certifiably eliminates.
3. Experiments on CIFAR-10 [Krizhevsky et al., 2009] and SVHN [Netzer et al., 2011] show that our method yields certified regions with order-of-magnitude larger volumes than prior smoothing schemes.

3.1.2 Related works

Robustification via dimensionality reduction. The work Bhagoji et al. [2018] was the first to consider linearly projecting inputs onto the top principal components of the training data before classification as a means to improve empirical (not certified) robustness. The authors of Sahay et al. [2019] nonlinearly preprocess test data using denoising and dimension-reducing autoencoders, and find a substantial increase in classification accuracy when the inputs are subject to the popular fast gradient sign method attack. The work Bafna et al. [2018] projects an input onto its top-k discrete cosine transform components to defend against "ℓ_0"-attacks, but this empirical defense was later broken using adapative "ℓ_0"-attacks [Tramèr et al., 2020], which directly motivates our approach for certified projection-based robustness. The work Sanyal et al. [2018] introduces a low-rank regularizer to encourage neural network feature representations to reside in a low-dimensional linear subspace, which is found to enhance empirical robustness. In Mustafa et al. [2019], the authors use super-resolution to project images onto the natural data manifold and obtain high empirical robustness for convolutional neural networks. Alemany and Pissinou [2022] shows that decreasing the codimension of data, i.e., decreasing the difference between the intrinsic dimension of the data manifold and the dimension of the input space in which it is embedded, generally leads to increased robustness of models defined on that input space.

Shamir et al. [2021] posits that learned decision boundaries tend to align with and "dimple" around the natural data manifold, and that adversarial perturbations are normal to this manifold. This finding supports our approach for certifiably eliminating off-manifold perturbations by projecting onto a low-dimensional approximation of the data manifold. The authors of Awasthi et al. [2021] reformulate principal component analysis to find projections that are robust with respect to projection error—a method that naturally

complements our framework—and give robustness guarantees for the Bayes optimal projection-based classifier in the special case of binary Gaussian-distributed data. The work Zeng et al. [2021] precomposes classifiers with orthogonal encoders and performs randomized smoothing in the encoder's low-dimensional latent space as a means to speed up the sample-based smoothing procedure. To the best of our knowledge, Zeng et al. [2021] is the only work that provides certified robustness guarantees for general models and data distributions when using dimensionality reduction at the input—all of the other referenced works are heuristic—and their choice of orthogonal encoders ensures that the certified ℓ_2-ball in the input space has the same radius as that in the latent space. Notably, their approach is highly conservative in estimating the input-space certified set as it relies on Lipschitzness of the orthogonal encoding layers, and is thus employed primarily as a means to speed up randomized smoothing. On the other hand, the method we propose uses a robustification-motivated projection for which we prove more general (anisotropic) certicates that capture off-manifold perturbations.

Certification via randomized smoothing. The work Cohen et al. [2019a] develops randomized smoothing using an isotropic Gaussian distribution with input-independent variance to obtain certified ℓ_2-balls. A subsequent line of works attempts to generalize randomized smoothing to other classes of certified regions, e.g., Wasserstein, "ℓ_0"-, ℓ_1-, and ℓ_∞-balls [Lee et al., 2019, Levine and Feizi, 2020, Teng et al., 2020, Yang et al., 2020a]. Various approaches have been taken to enlarge the certified regions. For example, Salman et al. [2019] unifies adversarial training with randomized smoothing to obtain state-of-the-art certified ℓ_2-radii. The authors of Zhai et al. [2020] incorporate the certified ℓ_2-radius into the model's training objective as a means to enlarge certified regions. The method in Zhang et al. [2020a] optimizes over base classifiers to increase the size of more general ℓ_p-balls. Li et al. [2022] employs a second smoothing distribution to tighten robustness certificates.

Optimizing the certified region pointwise in the input space has also been considered, but generally these methods require locally constant smoothing distributions to ensure that the resulting certificates are mathematically valid [Alfarra et al., 2020, Anderson and Sojoudi, 2022, Súkeník et al., 2021, Wang et al., 2021a]. To further strengthen the robustness guarantees of randomized smoothing, the recent works Eiras et al. [2021], Erdemir et al. [2021], Tecot [2021] have turned to certifying anisotropic regions of the input space. For example, Eiras et al. [2021] maximizes the volume of certified ellipsoids and generalized cross-polytopes of the form $\{x \in \mathbb{R}^d : \|Ax\|_p \leq b\}$ for $p \in \{1, 2\}$, allowing for the certification of perturbations that are potentially larger in magnitude than the minimum adversarial perturbation. We show in Section 3.4 that our proposed method is able to outperform these methods by leveraging dimensionality reduction. As is standard practice in the randomized smoothing literature [Cohen et al., 2019a, Jeong et al., 2021, Lee et al., 2019, Yang et al., 2020a, Zhai et al., 2020], our emphasis is on certified robustness and not empirical robustness—we refer the reader to Maho et al. [2022] for connections between certified and empirical robustness under randomized smoothing, and in particular the difficulty in constructing and evaluating suitable empirical attacks.

We also emphasize that volume (Lebesgue measure) is the natural scalar measure for the size of anisotropic certified regions of the input space and is the standard notion considered by prior works [Eiras et al., 2021, Liu et al., 2019b, Tecot, 2021].

3.1.3 Notation

We denote the set of real numbers by $\mathbb{R}$. The ℓ_2-norm of a vector $x \in \mathbb{R}^n$ is denoted by $\|x\|$, whereas the general ℓ_p-norm is given an explicit subscript $\|x\|_p$. The range and nullspace of a matrix $U \in \mathbb{R}^{m\times n}$ are denoted by $\mathcal{R}(U) \subseteq \mathbb{R}^m$ and $\mathcal{N}(U) \subseteq \mathbb{R}^n$, respectively. The $n \times n$ identity matrix is written as I_n. For a random variable X with distribution $\mathcal{D}$ and a measurable function f, the expectation of $f(X)$ is denoted by $\mathbb{E}_{X\sim\mathcal{D}} f(X)$. The multivariate normal distribution with mean $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$ is given by $N(\mu, \Sigma)$. The cardinality of a set S is written as $|S|$. For a Lebesgue-measurable set $S \subseteq \mathbb{R}^n$ contained in a k-dimensional affine subspace, we write $V_k(S)$, termed the k-dimensional volume of S, to mean the Lebesgue measure of S within that affine subspace. For sets $S, T \subseteq \mathbb{R}^n$, we denote their Minkowski sum by $S + T = \{x + y : x \in S,\ y \in T\}$. Euler's gamma function is denoted by Γ. Recall that $\Gamma(n) = (n-1)!$ when n is a positive integer.

3.2 Classifier architecture

Consider the task of classifying inputs from a zero-centered cube $C^d = [-1/2, 1/2]^d \subseteq \mathbb{R}^d$ into c distinct classes $\mathcal{Y} = \{1, 2, \ldots, c\}$.[1] Under the randomized smoothing framework, we begin with a given classifier $f_\theta \colon \mathbb{R}^d \to [0,1]^c$, parameterized by θ, that maps into the probability simplex over c classes. The problem at hand is to increase the robustness of f_θ with certifiable guarantees.

Vanilla randomized smoothing. We give a brief overview of how this would be accomplished using vanilla randomized smoothing [Cohen et al., 2019a]. Randomized smoothing takes the *base classifier* f_θ and smooths it with Gaussian noise on the input to yield the associated smoothed soft and hard classifiers

$$f^s(x) = \mathop{\mathbb{E}}_{\epsilon\sim N(0,\sigma^2 I_d)} f_\theta(x+\epsilon), \quad g(x) = \arg\max_{y\in\mathcal{Y}} f^s(x)_y,$$

where $f^s(x)_y$ denotes the yth component of the vector $f^s(x)$ and σ is a hyperparameter. Cohen et al. [2019a, Theorem 1] then gives, under certain conditions, a certified ℓ_2-ball for a particular input $x \in \mathbb{R}^d$; namely, that $g(x+\delta) = g(x)$ for all $\|\delta\| < R$, where $R > 0$ is determined by the confidence of the smoothed classifier at x. We leverage this result for our approach and refer interested readers to Cohen et al. [2019a] for additional details on the computation of the smoothing expectation and precise formula for R.

[1] The zero-centered cube is used without loss of generality instead of $[0,1]^d$ for notational convenience and compatibility with results from the mathematical literature.

Projected randomized smoothing. Motivated by the relationships between robustness and dimensionality described in Section 3.1, we consider $p < d$ and let $P\colon \mathbb{R}^d \to \mathbb{R}^p$ be a projection into $\mathbb{R}^p$ defined by $P(x) = U^\intercal x$, where $U \in \mathbb{R}^{d\times p}$ is a semi-orthogonal matrix satisfying $U^\intercal U = I_p$. Similarly, we let the reconstruction $\tilde{P}\colon \mathbb{R}^p \to \mathbb{R}^d$ be defined by $\tilde{P}(\tilde{x}) = U\tilde{x}$. Throughout, we let $v_1, \ldots, v_{d-p} \in \mathbb{R}^d$ be an orthonormal basis for $\mathcal{N}(U^\intercal)$ and let $v_{d-p+1}, \ldots, v_d \in \mathbb{R}^d$ denote the orthonormal columns of U. In practice, we instantiate the columns of U as the first p principal components of a random subset of the training dataset, although our method and theory hold for any orthonormal set of vectors. With the dimension-reducing projection P in place, we consider the classifier architecture consisting of the composition

$$f = f_\theta \circ \tilde{P} \circ P.$$

In particular, f first uses P to project inputs into the low-dimensional space $\mathbb{R}^p$ and then reconstructs the inputs in a lossy way using $\tilde{P}$ before feeding them through the classifier f_θ. We generally finetune f_θ to account for the slight image corruption associated with the projection step.

We now propose *projected randomized smoothing*, wherein randomized smoothing is performed in the compressed space $\mathbb{R}^p$. To do so, we define $\tilde{f}_\theta\colon \mathbb{R}^p \to [0,1]^c$ by $\tilde{f}_\theta = f_\theta \circ \tilde{P}$ so that $f = \tilde{f}_\theta \circ P$, and we smooth $\tilde{f}_\theta$ by adding Gaussian noise in its low-dimensional input space to obtain a new classifier $\tilde{f}^s_\theta\colon \mathbb{R}^p \to [0,1]^c$ defined by

$$\tilde{f}^s_\theta(\tilde{x}) = \mathop{\mathbb{E}}_{\epsilon\sim N(0,\sigma^2 I_p)} \tilde{f}_\theta(\tilde{x} + \epsilon). \tag{3.1}$$

The new overall smoothed soft classifier is then given by

$$f^s = \tilde{f}^s_\theta \circ P, \tag{3.2}$$

and its structure is illustrated in Figure 3.1a. The corresponding hard classifier is then given by the arg max of the soft classifier:[2]

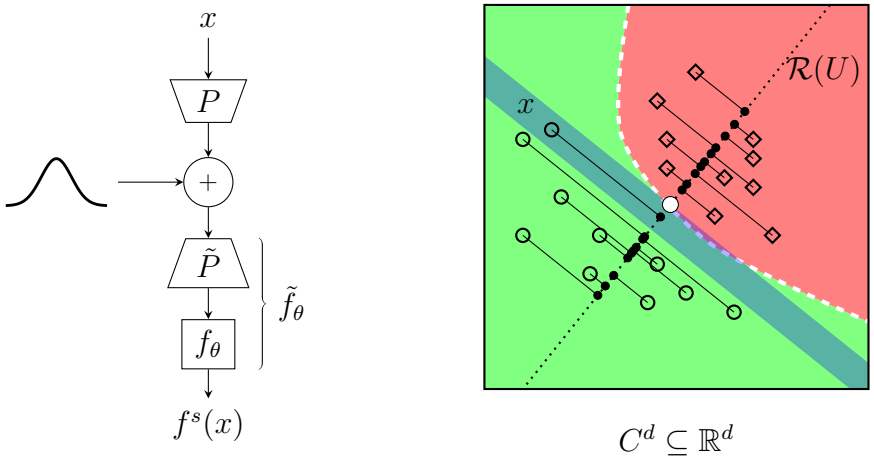
$$g(x) = \arg\max_{y\in\mathcal{Y}} f^s(x)_y. \tag{3.3}$$

A graphical illustration of our approach for $d = 2$ is shown in Figure 3.1b. To summarize, classifying an input $x \in \mathbb{R}^d$ using projected randomized smoothing amounts to applying the mapping $x \mapsto g(x)$ defined by (3.1) through (3.3), and it is for g that we seek to derive certified regions of the input space.

3.3 Robustness certificates

In this section, we construct certified regions for g around arbitrary inputs x in the high-dimensional space $\mathbb{R}^d$. The key idea is that $\tilde{f}^s_\theta$ is ℓ_2-ball robust in the low-dimensional

[2]For ease of exposition, we assume throughout that all arg max yield singleton sets and therefore equality signs may be used unambiguously.



(a) (b)

Figure 3.1: (a) Projected randomized smoothing architecture. Inputs x are projected into low-dimensional space by P, smoothed with Gaussian noise, and then reconstructed by $\tilde{P}$ and classified by f_θ. (b) Illustration of projected randomized smoothing for a binary classification task (circles vs. squares). The base classifier decision regions are shown in green and red. The white circle represents the smoothed decision boundary in $\mathbb{R}^p$, $p = 1$, with the projected subspace depicted by the dotted line and projected points depicted as solid dots. The blue area represents the certified region around x in $\mathbb{R}^d$ of the projected randomized smoothing classifier g.

space $\mathbb{R}^p$, and the preimage of this ball in the original input space is then "large" as it includes the inputs in $\mathcal{N}(U^\intercal)$. We formalize the geometry of the certified region in Section 3.3.1 and introduce our metric of interest as the volume of the certified region restricted to the unit cube of feasible inputs. In Section 3.3.2, we provide a lower bound on this volume in high-dimensional spaces that involves solving an ℓ_∞-norm linear regression. Section 3.3.3 compares the asymptotic behavior of the volume of the certified region of g with the standard ℓ_2-ball certificates as the input dimension grows large. Finally, we discuss runtime and limitations in Section 3.3.4. For ease of exposition, all proofs are deferred to the appendices.

3.3.1 Characterizing the certified region geometry

In the following two propositions, we characterize the geometry of the projected randomized smoothing classifier g in the high-dimensional input space $\mathbb{R}^d$ based on the certified ℓ_2-robustness of the classifier $\tilde{f}_\theta^s$ in the low-dimensional projected space $\mathbb{R}^p$.

Definition 3.1. Let $\tilde{x} \in \mathbb{R}^p$ and $R \geq 0$. The classifier $\tilde{f}_\theta^s \colon \mathbb{R}^p \to [0,1]^c$ is said to be *certified at $\tilde{x}$ with radius R* if

$$\arg\max_{y \in \mathcal{Y}} \tilde{f}_\theta^s(\tilde{x} + \tilde{\delta})_y = \arg\max_{y \in \mathcal{Y}} \tilde{f}_\theta^s(\tilde{x})_y$$

for all $\tilde{\delta} \in \mathbb{R}^p$ satisfying $\|\tilde{\delta}\| \leq R$.

Proposition 3.2. Let $x \in \mathbb{R}^d$ and $R \geq 0$. If $\tilde{f}_\theta^s$ is certified at $P(x) = U^\intercal x$ with radius R, then $g(x + \delta) = g(x)$ for all $\delta \in \Delta^U(R) \subseteq \mathbb{R}^d$, where

$$\Delta^U(R) := \{\delta \in \mathbb{R}^d : \|U^\intercal \delta\| \leq R\}$$

Proposition 3.3. Let $R \geq 0$. The certified region $\Delta^U(R)$ can be expressed as the Minkowski sum $\Delta^U(R) = B_p^U(R) + \mathcal{N}(U^\intercal)$, where $B_p^U(R) \subseteq \mathbb{R}^d$ is a p-dimensional ball embedded into $\mathcal{R}(U)$:

$$B_p^U(R) := \{\beta_1 v_{d-p+1} + \cdots + \beta_p v_d : \|\beta\| \leq R,\ \beta \in \mathbb{R}^p\}\,.$$

Propositions 3.2 and 3.3 characterize the geometry of the certified region of our classifier g. Proposition 3.2 provides an easy-to-check condition for an input to lie in the certified region, while Proposition 3.3 formalizes the same geometry as a hypercylinder consisting of a low-dimensional sphere that is "extruded" along the nullspace of the projection P, allowing us to certify adversarial off-manifold inputs of potentially very large magnitude that are projected back onto the natural data manifold. Intuitively, the certified region $\Delta^U(R)$ is potentially much larger than an ℓ_2-ball of radius R in $\mathbb{R}^d$, as it captures perturbations in the nullspace of $U^\intercal$ whose dimensionality is large when $p \ll d$.

We note that the above characterization of the decision region geometry holds analogously for other norm ball certificates in the projected space (i.e., the ℓ_1-ball certificates of Levine and Feizi [2021]). While the following theory is presented for the concrete case of ℓ_2-ball certificates, it also applies to this more general setting. Concrete experiments with other certificates is an exciting line of future work.

3.3.2 Lower-bounding the certified region volume

To compare a standard ℓ_2-ball certificate with our certified region $\Delta^U(R)$, which does not immediately come equipped with a notion of "radius," we adopt the perspective of recent works, e.g., Eiras et al. [2021], Liu et al. [2019b], Tecot [2021], by considering our metric of interest to be the volume of the certified region. One immediate issue is that the volume of $\Delta^U(R)$ is infinite since $\mathcal{N}(U^\intercal)$ is an unbounded subspace. To enable meaningful comparisons, we restrict ourselves to measuring the volume of $\Delta^U(R)$ contained in the cube $C^d = [-1/2, 1/2]^d$ of possible inputs. This amounts to computing the volume

$$\mathcal{V}_d\left(C^d \cap \Delta_x^U(R)\right), \tag{3.4}$$

where we recall that $\mathcal{V}_d$ measures d-dimensional volume in Euclidean space, and $\Delta_x^U(R) := \{x + \delta : \delta \in \Delta^U(R)\}$, with R chosen such that $\tilde{f}_\theta^s$ is certified at $P(x)$ with radius R so that $g(x') = g(x)$ for all $x' \in \Delta_x^U(R)$ by Proposition 3.2. Computing the volume in (3.4) is highly nontrivial, especially in high-dimensional input spaces. Instead, we develop a tractable lower bound on $\mathcal{V}_d(C^d \cap \Delta_x^U(R))$ throughout the remainder of this

section. Since $\Delta_x^U(R)$ contains affine subspaces, this derivation rests heavily on theory regarding cube-subspace intersections in high dimensions. The most important result for our purposes comes from Vaaler [1979], which showed the following.

Theorem 3.4. Let S_k be a k-dimensional linear subspace of $\mathbb{R}^d$. Then $V_k(C^d \cap S_k) \geq 1$.

This result proved Good's conjecture and generalized a previous result for the $k = d - 1$ case [Hensley, 1979]. We begin with an extension of Theorem 3.4 to cubes of non-unit side length, and then to intersections with affine subspaces which do not necessarily contain the origin.

Corollary 3.5. Let S_k be a k-dimensional linear subspace of $\mathbb{R}^d$ and rC^d be a zero-centered cube of side length $r > 0$. Then $V_k(rC^d \cap S_k) \geq r^k$.

Corollary 3.6. Let $x \in \mathbb{R}^d$ and let $S_k(x) \subseteq \mathbb{R}^d$ be the k-dimensional affine subspace

$$S_k(x) = \left\{ x + \sum_{i=1}^{k} \alpha_i v_i : \alpha \in \mathbb{R}^k \right\}$$

spanned by arbitrary vectors $v_1, \ldots, v_k$ and passing through x. Let $t \geq 0$ be the minimal ℓ_∞-norm of a point in $S_k(x)$:

$$t := \inf_{x' \in S_k(x)} \|x'\|_\infty = \inf_{\alpha \in \mathbb{R}^k} \left\| x + \textstyle\sum_{i=1}^{k} \alpha_i v_i \right\|_\infty . \tag{3.5}$$

Then, for all $r > 2t$, it holds that $V_k(rC^d \cap S_k(x)) \geq (r - 2t)^k$.

Corollary 3.6 generalizes Corollary 3.5 to affine subspaces. If $S_k(x)$ contains the origin, $t = 0$ and the bound from Corollary 3.5 is recovered. We are now ready to present the main result of this section.

Theorem 3.7. Let $x \in C^d$, let t be defined as in (3.5) with $k = d-p$, and let $R \in [0, 1/2-t]$. If $\tilde{f}_\theta^s$ is certified at $P(x) = U^\intercal x$ with radius R, then

$$V_d(C^d \cap \Delta_x^U(R)) \geq \frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)} R^p (1 - 2R - 2t)^{d-p}. \tag{3.6}$$

Notice that the lower bound given in Theorem 3.7 does not monotonically increase with the certified radius R from the randomized smoothing performed in $\mathbb{R}^p$. Therefore, if the certified radius R is large enough, we may be able to improve our lower bound on the volume $V_d(C^d \cap \Delta_x^U(R))$ by using a smaller certified radius (which is of course still valid), and in particular, we may choose the optimal such radius to use according to the following closed-form expression.

Proposition 3.8. Let t and R be as in Theorem 3.7. The lower bound (3.6) is maximized as follows:

$$r^* := \min \left\{ R, \frac{p(1-2t)}{2d} \right\} \in \arg\max_{r \in [0,R]} \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} r^p \left(1 - 2r - 2t\right)^{d-p}. \tag{3.7}$$

Algorithm 3.1 Prediction and certification

def PREDICT, CERTIFY as in Cohen et al. [2019a]

function PROJECTPREDICT(f_θ, U, σ, x, n, α)
  def $P(x) = U^\intercal x$, $\tilde{P}(\tilde{x}) = U\tilde{x}$
  return PREDICT$(f_\theta \circ \tilde{P}, \sigma, P(x), n, \alpha)$
end function

function PROJECTCERTIFY(f_θ, U, σ, x, n_0, n, α)
  def $P(x) = U^\intercal x$, $\tilde{P}(\tilde{x}) = U\tilde{x}$, $(d, p) \leftarrow \text{shape}(U)$
  ABSTAIN, $\hat{c}_A, R \leftarrow$ CERTIFY$(f_\theta \circ \tilde{P}, \sigma, P(x), n_0, n, \alpha)$
  if ABSTAIN **then return** ABSTAIN
  compute orthonormal basis $v_1, \ldots, v_{d-p}$ for $\mathcal{N}(U^\intercal)$
  solve the optimization

$$t \leftarrow \inf_{\alpha \in \mathbb{R}^{d-p}} \left\| x + \sum_{i=1}^{d-p} \alpha_i v_i \right\|_\infty \tag{Alg1}$$

  assign $R \leftarrow \min\{R, p(1-2t)/(2d)\}$
  compute the certified volume lower bound

$$V \leftarrow \frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)} R^p (1 - 2R - 2t)^{d-p}$$

  return prediction $\hat{c}_A$ and volume bound V
end function

The overall certification procedure derived in this section is summarized in Algorithm 3.1. We note that our method inherits its ABSTAIN behavior from the original randomized smoothing Monte Carlo sampling scheme [Cohen et al., 2019a]; namely, we evaluate the certification confidence using many Gaussian-perturbed samples, and if the prediction or certification procedures do not resolve with a user-specified confidence, ABSTAIN is returned.

3.3.3 Asymptotic behavior of the volume bound

We briefly compare the volume lower bound (3.6) of the projected randomized smoothing certified region to that of a standard certified ℓ_2-ball. The volume of a d-dimensional ℓ_2-ball $B_d(R) \coloneqq \{x \in \mathbb{R}^d : \|x\| \leq R\}$ of radius $R \geq 0$ is well-known (e.g., see Folland [1999, Theorem 2.44, Corollary 2.55]) to be

$$V_d\left(B_d(R)\right) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} R^d. \tag{3.9}$$

While the numerator of (3.9) scales exponentially in d, the denominator $\Gamma(\frac{d}{2}+1)$ scales factorially, leading to tiny ℓ_2-ball certified volumes in high-dimensional input spaces. By contrast, the denominator in our bound (3.6) scales factorially in the *projected dimension* p, where $p \ll d$. This suggests dramatic improvements in the volume of our certified regions: while the numerator in (3.6) might be *exponentially* smaller than that of (3.9), the denominator is smaller by a *factorial* factor. We thus expect the volumes of projected randomized smoothing to dominate at higher dimensions. We verify our analysis experimentally in Section 3.4.2.

3.3.4 Runtime and limitations

Our certification strategy has two additional computational steps outside of the PREDICT and CERTIFY subroutines from the conventional randomized smoothing method of Cohen et al. [2019a]. The first is a one-time computation of the principal components of the data that occurs at the beginning of training. The second is computing the ℓ_∞-regression in (Alg1), which we solve as a linear program using the standard epigraph formulation. For the CIFAR-10 and SVHN datasets considered in this work, the added runtime is comparable to the certification sampling step from Cohen et al. [2019a]. Namely, we found that the ℓ_∞-regression averaged around 16 seconds for CIFAR-10 and 19 seconds for SVHN.

The number of variables and constraints in the optimization (Alg1) scales linearly with $d-p$. Since generally $p \ll d$, this makes the volume approximation of the certified region computationally intensive in high-dimensional spaces. We remark that it is still trivial to check whether any particular perturbation lies in the certified region using Proposition 3.2—it is just that computing a lower bound on the volume of this region for comparison purposes becomes more challenging. For a natural image dataset such as ImageNet, the analysis of Section 3.3.3 suggests that the certified region volume improvements would in fact be substantially larger than those for CIFAR-10. The main challenge to computationally verifying this conjecture lies in holding the optimization problem (Alg1) in memory, which is infeasible on our hardware for ImageNet-scale inputs. Further research in this vein would likely leverage techniques from the large-scale ℓ_∞-regression literature, e.g., Shen et al. [2014], and is outside the scope of this work.

3.4 Experiments

This section reports our experiments on the CIFAR-10 and SVHN datasets. We first demonstrate in Section 3.4.1 that networks are vulnerable to ℓ_∞-bounded attacks in the subspace of low-variance principal components, to which our architecture is provably robust. Section 3.4.2 then presents results comparing the volume of the projected randomized smoothing certified regions to a variety of baseline certified classifiers.

3.4.1 Vulnerability to low-variance PCA attacks

Consider perturbations $\delta \in \mathcal{N}(U^\intercal)$ contained in the span of a dataset's low-variance principal components, where we take U to contain sufficient components to account for 99% of the dataset variance for CIFAR-10 and 95% for SVHN, which is more robust to low-variance subspace attacks due to its increased compressibility. Such a perturbation is known to be essentially orthogonal to the true data manifold, and therefore it is reasonable to expect a truly robust classifier to be invariant to small perturbations in $\mathcal{N}(U^\intercal)$. Our method is directly robust to such perturbations under the simple condition that we use fewer components in our initial projection step, as demonstrated in Proposition 3.3.

We now investigate whether this theoretical guarantee adds a degree of robustness over a typical neural network classifier. The answer is affirmative. Namely, we show that our subspace attack can attain a comparable attack success rate to a standard ℓ_∞-bounded projected gradient descent (PGD) attack, with roughly a four-fold increase in the size of the admissible ℓ_∞-ball.

Formally, consider a particular hard classifier g, to which we assume that our adversaries have white-box access, and take a specific input x that g classifies correctly. We first consider the standard projected gradient descent attack strategy $\mathrm{PGD}(x, \epsilon)$ which seeks to construct a perturbation $\|\delta\|_\infty \leq \epsilon$ such that $x + \delta \in C^d$ and $g(x + \delta) \neq g(x)$. As $x + \delta \in C^d$ if and only if $\|x + \delta\|_\infty \leq 1/2$, satisfying both ℓ_∞-norm constraints on δ is easily accomplished using clipping. Our routine $\textsc{SubspacePGD}(x, \epsilon)$ adds the additional constraint $\delta \in \mathcal{N}(U^\intercal)$. Note that finding a perturbation that satisfies $\delta \in \mathcal{N}(U^\intercal)$, $\|\delta\|_\infty \leq \epsilon$, and $x + \delta \in C^d$ is nontrivial, as projection onto one set generally removes an input from the other set.

For reference, we also consider RandMax and RandUniform, which generate perturbations randomly on the boundary of and uniformly in the attack ℓ_∞-ball, respectively. We instantiate g as the Wide ResNet considered in Yang et al. [2020a] with the default hyperparameters and $\sigma = 0.15$ Gaussian noise augmentation during training.

Figure 3.2 demonstrates that unprotected classifiers are indeed vulnerable to adversarial perturbations in the subspace of low-variance principal components. Enlargements of the attack radius do not invalidate that these are true adversarial attacks, as the perturbed images in the third row of Figure 3.2b are still easily classified by a human. Furthermore, SubspacePGD adversarial examples are substantially less perceptible than PGD attacks of the same magnitude, which tend to produce stronger visual distortions of the image, paralleling results from Shamir et al. [2021]; take as a representative example the area around the frog's head in the second row of the third column in Figure 3.2b, compared with the same image perturbed by SubspacePGD in the third row. The results for the SVHN dataset in Figure 3.2d are even more striking. This is likely because PGD attacks have access to high-variance principal components which convey the dataset information content. Despite visually appearing random, we establish in Figures 3.2a and 3.2c that the SubspacePGD attack is significantly more successful than random-noise attacks of the same magnitude. These results suggest that undefended classifiers can be attacked

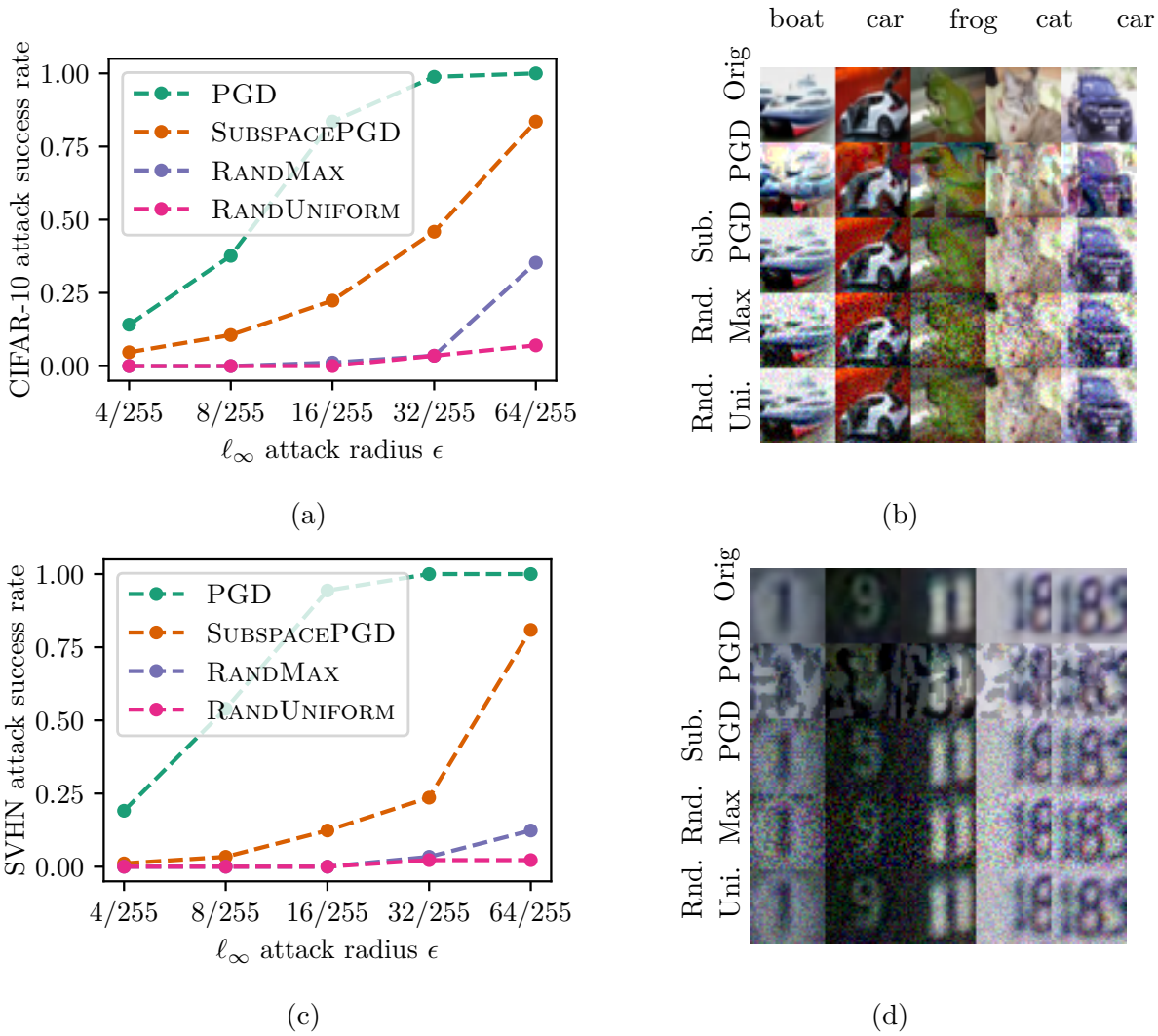


Figure 3.2: (a) CIFAR-10 adversarial attack success rates for the PGD, SubspacePGD, and random attack strategies. (b) Perturbation examples for CIFAR-10 with an attack radius of $\epsilon = 32/255$. The top row represents the original image. (c) SVHN aversarial attack success rates for the PGD, SubspacePGD, and random attack strategies. (d) Perturbation examples for SVHN with an attack radius of $\epsilon = 32/255$.

in the subspace of low-variance principal components, to which projected randomized smoothing is provably robust by Proposition 3.3.

3.4.2 Certified region comparison

Having established that the certified region of projected randomized smoothing provides a meaningful robustness improvement against low-variance principal component attacks, we now compare the volume of our certified region with several baselines. Namely, we evaluate the ℓ_2-balls of Cohen et al. [2019a] (denoted RS), the ℓ_1- and ℓ_∞-balls of Yang et al. [2020a] (denoted RS4A $-$ ℓ_1 and RS4A $-$ ℓ_∞, respectively), and the anisotropic

ellipsoids of Eiras et al. [2021] (denoted ANCER), without use of the associated memory module.

Some additional remarks on the inclusion of Eiras et al. [2021] are warranted. As noted in Súkeník et al. [2021], without the inclusion of the memory module, the local certificate optimization technique in Eiras et al. [2021] yields overly optimistic and mathematically incorrect certificates as the smoothing distribution varies between inputs. The work Eiras et al. [2021] corrects this with the use of a memory module that records previous inputs to ensure compatibility of the smoothing certificates. However, this results in a classifier that is dependent on the input order and adds ambiguity about what classifier is actually being certified, as the smoothed classifier is modified at test time after each input. We therefore discard the memory module and report the certified volume at each point as if the locally optimized smoothing distribution were being used globally. This yields an upper bound on the certified volume of any data-dependent anisotropic ellipsoidal smoothing method and is thus a very strong baseline to compare against.

Our results are summarized in Figure 3.3 and Table 3.1. We achieve state-of-the-art median certified volumes, easily outperforming standard randomized smoothing and even the optimistic ANCER baseline by 706 and 2453 *orders of magnitude* on CIFAR-10 and SVHN, respectively. The larger improvement on SVHN is attributable to the higher compressibility of the dataset. Our performance derives from the added robustness of our method against low-variance features, validating the asymptotic dimension analysis in Section 3.3.3. Note that although the ANCER baseline achieves higher accuracy at smaller volumes, its certificates are mathematically invalid [Súkeník et al., 2021], and our method significantly outperforms ANCER at larger volumes.

Figure 3.3b examines the CIFAR-10 certified accuracy curves over a range of choices for the dimensionality p of the compressed space. For large p, image reconstruction is near-perfect as $p = 620$ covers 99% of variance in the CIFAR-10 dataset. Thus, methods with $p \geq 300$ have comparable accuracy at small regions, with the certified volumes increasing as the dimensionality of the projected space decreases, corroborating the discussion in Section 3.3.3. Figure 3.3d presents similar results for the SVHN dataset. Note that the due to the compressibility of the dataset, fewer principal components are required to achieve high accuracy.

The hyperparameter p introduces a mild tradeoff between clean accuracy and certified volume; if p is chosen to be very small, the projected images may be too corrupted to classify, while if p is chosen to be very large, certified volume may suffer. However, as Figures 3.3b and 3.3d suggest, our method's certified volumes comfortably outperform those of standard randomized smoothing for a large range of p, indicating that this choice is not particularly sensitive. A practical heuristic for choosing p involves making p just large enough to reconstruct images with high fidelity—roughly corresponding to PCA components that explain 95% to 99% of the dataset variance. If desired, a small, localized sweep of p around this initial choice can be used to further optimize the hyperparameter depending on the experimentalist's target metrics (e.g., clean accuracy, median certified volume, other metrics, or some combination). In any case, we emphasize

Table 3.1: Quantitative representation of the data in Figure 3.3. The first column reports the smoothed classifier clean accuracy for each method and the second column reports the median certified volume for correctly classified samples. We use the median instead of the mean due to the log-scaled nature of our data.

(a) CIFAR certification performance.

Method	Accuracy	Median cert. vol. ($\log_{10}$)
ProjectedRS	85.8%	**−3175**
RS	**87.8**%	−4377
ANCER	87.4%	−3881
RS4A − ℓ_1	83.8%	−9573
RS4A − ℓ_∞	85.4%	−6102

(b) SVHN certification performance.

Method	Accuracy	Median cert. vol. ($\log_{10}$)
ProjectedRS	91.4%	**−1578**
RS	92.6%	−4280
ANCER	91.2%	−4031
RS4A − ℓ_1	**93.0**%	−9573
RS4A − ℓ_∞	92.6%	−6171

that the parameter choice is quite robust and any additional tuning is likely to result in minimal gains as compared to the practical heuristic. We select $p = 450$ for the CIFAR-10 experiment in Figure 3.3a and $p = 150$ for SVHN.

3.5 Conclusion

Motivated by the manifold hypothesis, we consider a classifier architecture that first projects onto a principal component approximation of the data manifold and then applies randomized smoothing in the low-dimensional projected space. This yields a precise characterization of the input-space certified region as capturing disturbances in the projection nullspace. We interpret this as a certifiable robustification against vulnerable features that are irrelevant to the dataset information content as they are normal to the data manifold. We show that unprotected classifiers, unlike our method, are vulnerable to such perturbations by explicitly constructing adversarial examples in the span of the low-variance principal components. We prove a volumetric lower bound on the intersection of our certified region with the unit cube of feasible inputs and derive two additional ways to tighten the bound: one which involves solving an ℓ_∞-regression problem and another which is a closed-form radius adjustment.

Comparing against state-of-the-art ℓ_1-, ℓ_2-, ℓ_∞-, and anisotropic baselines shows that our classifier produces certified regions with many orders of magnitude greater volume. This confirms an asymptotic analysis that shows that our method's certified volumes decay factorially in the low dimension of the *projected space*, while competing methods decay factorially in the high dimension of the *input space*.

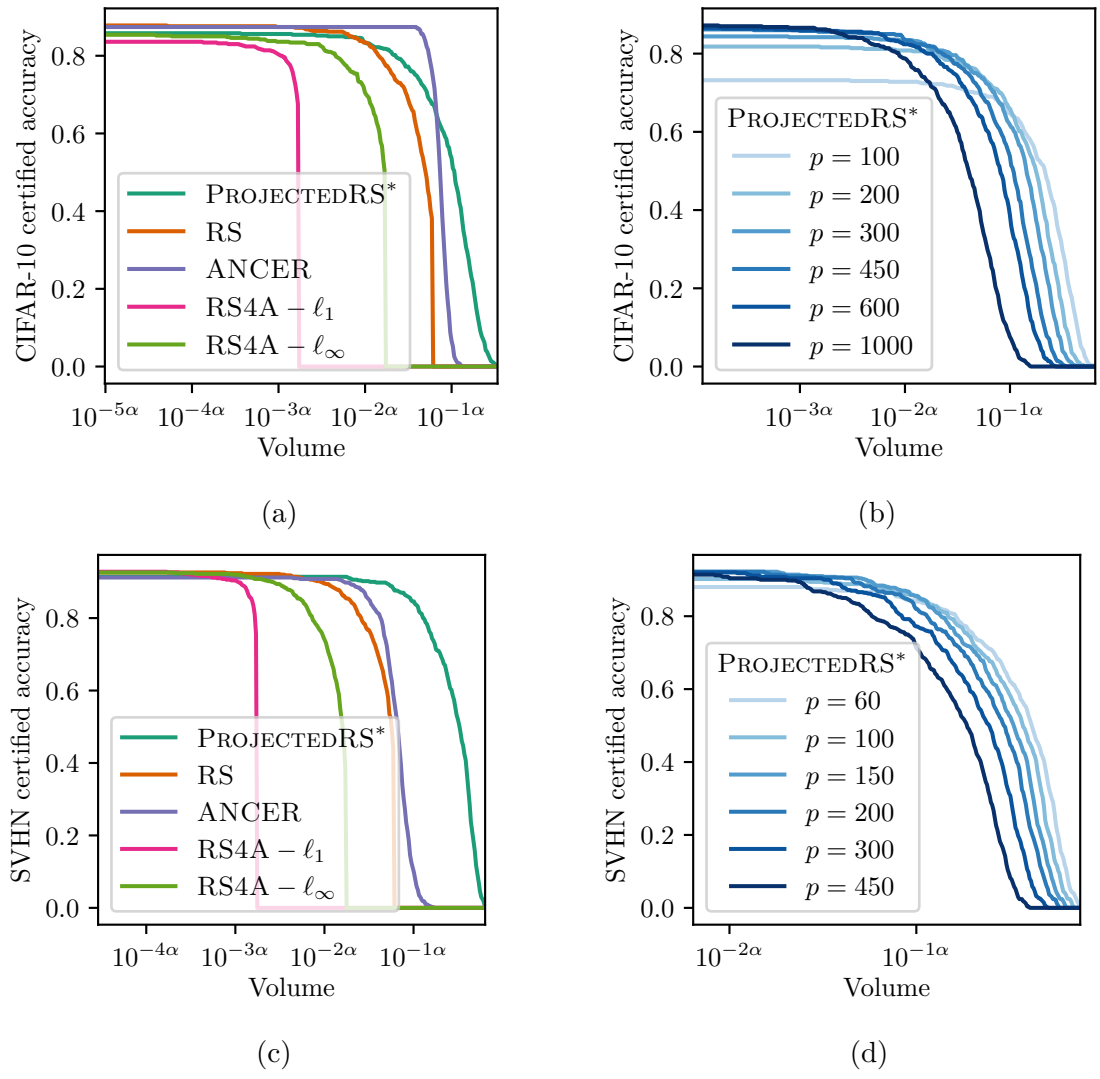


Figure 3.3: (a) Certified region volumes for CIFAR-10, with our method highlighted by an asterisk. Here $\alpha \approx 3465$ is a scaling constant corresponding to the d-dimensional unit ball volume; i.e. $V_d(B_d(1)) = 10^{-\alpha}$. (b) CIFAR-10 certified region volumes while varying the projected space dimension p for our method. (c) Certified region volumes for SVHN. (d) SVHN certified region volumes while varying p.

Chapter 4

Asymmetric Certified Robustness via Feature-Convex Neural Networks

Real-world adversarial attacks on machine learning models often feature an asymmetric structure wherein adversaries only attempt to induce false negatives (e.g., classify a spam email as not spam). We formalize the asymmetric robustness certification problem and correspondingly present the *feature-convex neural network* architecture, which composes an input-convex neural network (ICNN) with a Lipschitz continuous feature map in order to achieve asymmetric adversarial robustness. We consider the aforementioned binary setting with one "sensitive" class, and for this class we prove deterministic, closed-form, and easily-computable certified robust radii for arbitrary ℓ_p-norms. We theoretically justify the use of these models by extending the universal approximation theorem for ICNN regression to the classification setting, and proving a lower bound on the probability that such models perfectly fit even unstructured uniformly distributed data in sufficiently high dimensions. Experiments on Malimg malware classification and subsets of the MNIST, Fashion-MNIST, and CIFAR-10 datasets show that feature-convex classifiers attain substantial certified ℓ_1, ℓ_2, and ℓ_∞-radii while being far more computationally efficient than competitive baselines.

This chapter is based on the following published work:

Samuel Pfrommer, Brendon Anderson, Julien Piet, and Somayeh Sojoudi. Asymmetric certified robustness via feature-convex neural networks. *Advances in Neural Information Processing Systems*, 36:52365–52400, 2023a.

4.1 Introduction

Although neural networks achieve state-of-the-art performance across a range of machine learning tasks, researchers have shown that they can be highly sensitive to adversarial inputs that are maliciously designed to fool the model [Biggio et al., 2013, Nguyen et al., 2015, Szegedy et al., 2014]. For example, the works Eykholt et al. [2018] and Liu

et al. [2019a] show that small physical and digital alterations of vehicle traffic signs can cause image classifiers to fail. In safety-critical applications of neural networks, such as autonomous driving [Bojarski et al., 2016, Wu et al., 2017a] and medical diagnostics [Amato et al., 2013, Yadav and Jadhav, 2019], this sensitivity to adversarial inputs is clearly unacceptable.

A line of heuristic defenses against adversarial inputs has been proposed, only to be defeated by stronger attack methods [Athalye et al., 2018, Carlini and Wagner, 2017, Kurakin et al., 2017, Madry et al., 2018, Uesato et al., 2018]. This has led researchers to develop certifiably robust methods that provide a provable guarantee of safe performance. The strength of such certificates can be highly dependent on network architecture; general off-the-shelf models tend to have large Lipschitz constants, leading to loose Lipschitz-based robustness guarantees [Fazlyab et al., 2019, Hein and Andriushchenko, 2017, Yang et al., 2020b]. Consequently, lines of work that impose certificate-amenable structures onto networks have been popularized, e.g., specialized model layers [Trockman and Kolter, 2021, Zhang et al., 2021a], randomized smoothing-based networks [Anderson and Sojoudi, 2022, Cohen et al., 2019a, Li et al., 2019, Yang et al., 2020a, Zhai et al., 2020], and ReLU networks that are certified using convex optimization and mixed-integer programming [Anderson et al., 2020, Ma and Sojoudi, 2021, Raghunathan et al., 2018, Weng et al., 2018, Wong and Kolter, 2018]. The first category only directly certifies against one specific choice of norm, producing poorly scaled radii for other norms in high dimensions. The latter two approaches incur serious computational challenges: randomized smoothing typically requires the classification of thousands of randomly perturbed samples per input, while optimization-based solutions scale poorly to large networks.

Despite the moderate success of these certifiable classifiers, conventional assumptions in the literature are unnecessarily restrictive for many practical adversarial settings. Specifically, most works consider a multiclass setting where certificates are desired for inputs of any class. By contrast, many real-world adversarial attacks involve a binary setting with only one *sensitive class* that must be robust to adversarial perturbations. Consider the representative problem of spam classification; a malicious adversary crafting a spam email will only attempt to fool the classifier toward the "not-spam" class—never conversely [Dalvi et al., 2004]. Similar logic applies for a range of applications such as malware detection [Grosse et al., 2017], malicious network traffic filtering [Sadeghzadeh et al., 2021], fake news and social media bot detection [Cresci et al., 2021], hate speech removal [Grolman et al., 2022], insurance claims filtering [Finlayson et al., 2019], and financial fraud detection [Cartella et al., 2021].

The important asymmetric nature of these classification problems has long been recognized in various subfields, and some domain-specific attempts at robustification have been proposed with this in mind. This commonly involves robustifying against adversaries appending features to the classifier input. In spam classification, such an attack is known as the "good word" attack [Lowd and Meek, 2005]. In malware detection, numerous approaches have been proposed to provably counter such additive-only adversaries using special classifier structures such as non-negative networks [Fleshman et al., 2018] and

monotonic classifiers [Íncer Romeo et al., 2018]. We note these works strictly focus on *additive* adversaries and cannot handle general adversarial perturbations of the input that are capable of perturbing existing features. We propose adding this important asymmetric structure to the study of norm ball-certifiably robust classifiers. This narrowing of the problem to the asymmetric setting provides prospects for novel certifiable architectures, and we present feature-convex neural networks as one such possibility.

4.1.1 Problem statement and contributions

This section formalizes the *asymmetric robustness certification problem* for general norm-bounded adversaries. Specifically, we assume a binary classification setting wherein one class is "sensitive"—meaning we seek to certify that, if some input is classified into this sensitive class, then adversarial perturbations of sufficiently small magnitude cannot change the prediction.

Formally, consider a binary classifier $f_\tau \colon \mathbb{R}^d \to \{1, 2\}$, where class 1 is the sensitive class for which we desire certificates. We take f_τ to be a standard thresholded version of a soft classifier $g \colon \mathbb{R}^d \to \mathbb{R}$, expressible as $f_\tau(x) = T_\tau(g(x))$, where $T_\tau \colon \mathbb{R} \to \{1, 2\}$ is the thresholding function defined by

$$T_\tau(y) = \begin{cases} 1 & \text{if } y + \tau > 0, \\ 2 & \text{if } y + \tau \leq 0, \end{cases} \tag{4.1}$$

with $\tau \in \mathbb{R}$ being a user-specified parameter that shifts the classification threshold. A classifier f_τ is considered certifiably robust at a class 1 input $x \in \mathbb{R}^d$ with a radius $r(x) \in \mathbb{R}_+$ if $f_\tau(x + \delta) = f_\tau(x) = 1$ for all $\delta \in \mathbb{R}^d$ with $\|\delta\| < r(x)$ for some norm $\|\cdot\|$. Thus, τ induces a tradeoff between the clean accuracy on class 2 and certification performance on class 1. As $\tau \to \infty$, f_τ approaches a constant classifier which achieves infinite class 1 certified radii but has zero class 2 accuracy.

For a particular choice of τ, the performance of f_τ can be analyzed similarly to a typical certified classifier. Namely, it exhibits a class 2 clean accuracy $\alpha_2(\tau) \in [0, 1]$ as well as a class 1 certified accuracy surface Γ with values $\Gamma(r, \tau) \in [0, 1]$ that capture the fraction of the class 1 samples that can be certifiably classified by f_τ at radius $r \in \mathbb{R}_+$. The class 1 clean accuracy $\alpha_1(\tau) = \Gamma(0, \tau)$ is inferable from Γ as the certified accuracy at $r = 0$.

The full asymmetric certification performance of the family of classifiers f_τ can be captured by plotting the surface $\Gamma(r, \tau)$, as will be shown in Figure 4.1a. Instead of plotting against τ directly, we plot against the more informative difference in clean accuracies $\alpha_1(\tau) - \alpha_2(\tau)$. This surface can be viewed as an asymmetric robustness analogue to the classic receiver operating characteristic curve.

Note that while computing the asymmetric robustness surface is possible for our feature-convex architecture (to be defined shortly), it is computationally prohibitive for conventional certification methods. We therefore standardize our comparisons throughout this

work to the certified accuracy cross section $\Gamma(r, \tau^*)$ for a τ^* such that clean accuracies are balanced in the sense that $\alpha_2(\tau^*) = \alpha_1(\tau^*)$, noting that α_1 monotonically increases in τ and α_2 mononically decreases in τ. This choice allows for a direct comparison of the resulting certified accuracy curves without considering the non-sensitive class clean accuracy.

With the above formalization in place, the goal at hand is two-fold: 1) develop a classification architecture tailored for the asymmetric setting with high robustness, as characterized by the surface Γ, and 2) provide efficient methods for computing the certified robust radii $r(x)$ used to generate Γ.

Contributions. We tackle the above two goals by proposing *feature-convex neural networks* and achieve the following contributions:

1. We exploit the feature-convex structure of the proposed classifier to provide asymmetrically tailored closed-form class 1 certified robust radii for arbitrary ℓ_p-norms, solving the second goal above and yielding efficient computation of Γ.
2. We characterize the decision region geometry of convex classifiers, extend the universal approximation theorem for input-convex ReLU neural networks to the classification setting, and show that convex classifiers are sufficiently expressive for high-dimensional data.
3. We evaluate against several baselines on MNIST 3-8 [LeCun, 1998], Malimg malware classification [Nataraj et al., 2011], Fashion-MNIST shirts [Xiao et al., 2017], and CIFAR-10 cats-dogs [Krizhevsky et al., 2009], and show that our classifiers yield certified robust radii competitive with the state-of-the-art, empirically addressing the first goal listed above.

All proofs and appendices can be found in the Supplemental Material.

4.1.2 Related works

Certified adversarial robustness. Three of the most popular approaches for generating robustness certificates are Lipschitz-based bounds, randomized smoothing, and optimization-based methods. Successfully bounding the Lipschitz constant of a neural network can give rise to an efficient certified radius of robustness, e.g., via the methods proposed in Hein and Andriushchenko [2017]. However, in practice such Lipschitz constants are too large to yield meaningful certificates, or it is computationally burdensome to compute or bound the Lipschitz constants in the first place [Fazlyab et al., 2019, Virmaux and Scaman, 2018, Yang et al., 2020b]. To overcome these computational limitations, certain methods impose special structures on their model layers to provide immediate Lipschitz guarantees. Specifically, Trockman and Kolter [2021] uses the Cayley transform to derive convolutional layers with immediate ℓ_2-Lipschitz constants, and Zhang et al. [2021a] introduces a ℓ_∞-distance neuron that provides similar Lipschitz guarantees with respect to the ℓ_∞-norm. We compare with both these approaches in our experiments.

Randomized smoothing, popularized by Cohen et al. [2019a], Lecuyer et al. [2019], Li et al. [2019], uses the expected prediction of a model when subjected to Gaussian input noise. These works derive ℓ_2-norm balls around inputs on which the smoothed classifier remains constant, but suffer from nondeterminism and high computational burden. Follow-up works generalize randomized smoothing to certify input regions defined by different metrics, e.g., Wasserstein, ℓ_1-, and ℓ_∞-norms [Levine and Feizi, 2020, Teng et al., 2020, Yang et al., 2020a]. Other works focus on enlarging the certified regions by optimizing the smoothing distribution [Anderson et al., 2022, Eiras et al., 2021, Zhai et al., 2020], incorporating adversarial training into the base classifier [Salman et al., 2019, Zhang et al., 2020a], and employing dimensionality reduction at the input [Pfrommer et al., 2023b].

Optimization-based certificates typically seek to derive a tractable over-approximation of the set of possible outputs when the input is subject to adversarial perturbations, and show that this over-approximation is safe. Various over-approximations have been proposed, e.g., based on linear programming and bounding [Weng et al., 2018, Wong and Kolter, 2018], semidefinite programming [Raghunathan et al., 2018], and branch-and-bound [Anderson et al., 2020, Ma and Sojoudi, 2021, Wang et al., 2021b]. The α, β-CROWN method [Wang et al., 2021b] uses an efficient bound propagation to linearly bound the neural network output in conjunction with a per-neuron branching heuristic to achieve state-of-the-art certified radii, winning both the 2021 and the 2022 VNN certification competitions [Bak et al., 2021, Müller et al., 2022]. In contrast to optimization-based methods, our approach directly exploits the convex structure of input-convex neural networks to derive closed-form robustness certificates, altogether avoiding any efficiency-tightness tradeoffs.

Input-convex neural networks. Input-convex neural networks, popularized by Amos et al. [2017], are a class of parameterized models whose input-output mapping is convex. The authors develop tractable methods to learn input-convex neural networks, and show that such models yield state-of-the-art results in a variety of domains where convexity may be exploited, e.g., optimization-based inference. Subsequent works propose novel applications of input-convex neural networks in areas such as optimal control and reinforcement learning [Chen et al., 2019, Zeng et al., 2022], optimal transport [Makkuva et al., 2020], and optimal power flow [Chen et al., 2020, Zhang et al., 2021b]. Other works have generalized input-convex networks to input-invex networks [Nesterov et al., 2022, Sapkota and Bhattarai, 2021] and global optimization networks [Zhao et al., 2022] so as to maintain the benign optimization properties of input-convexity. The authors of Siahkamari et al. [2022] present algorithms for efficiently learning convex functions, while Chen et al. [2019], Kim and Kim [2022] derive universal approximation theorems for input-convex neural networks in the convex regression setting. The work Sivaprasad et al. [2021] shows that input-convex neural networks do not suffer from overfitting, and generalize better than multilayer perceptrons on common benchmark datasets. In this work, we incorporate input-convex neural networks as a part of our feature-convex architecture and leverage convexity properties to derive novel robustness guarantees.

4.1.3 Notations

The sets of natural numbers, real numbers, and nonnegative real numbers are denoted by $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{R}_+$ respectively. The $d \times d$ identity matrix is written as $I_d \in \mathbb{R}^{d\times d}$, and the identity map on $\mathbb{R}^d$ is denoted by $\mathrm{Id}\colon x \mapsto x$. For $A \in \mathbb{R}^{n\times d}$, we define $|A| \in \mathbb{R}^{n\times d}$ by $|A|_{ij} = |A_{ij}|$ for all i, j, and we write $A \geq 0$ if and only if $A_{ij} \geq 0$ for all i, j. The ℓ_p-norm on $\mathbb{R}^d$ is given by $\|\cdot\|_p\colon x \mapsto (|x_1|^p + \cdots + |x_d|^p)^{1/p}$ for $p \in [1, \infty)$ and by $\|\cdot\|_p\colon x \mapsto \max\{|x_1|, \ldots, |x_d|\}$ for $p = \infty$. The dual norm of $\|\cdot\|_p$ is denoted by $\|\cdot\|_{p,*}$. The convex hull of a set $X \subseteq \mathbb{R}^d$ is denoted by $\mathrm{conv}(X)$. The subdifferential of a convex function $g\colon \mathbb{R}^d \to \mathbb{R}$ at $x \in \mathbb{R}^d$ is denoted by $\partial g(x)$. If $\epsilon\colon \Omega \to \mathbb{R}^d$ is a random variable on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$ and P is a predicate defined on $\mathbb{R}^d$, then we write $\mathbb{P}(P(\epsilon))$ to mean $\mathbb{P}(\{\omega \in \Omega : P(\epsilon(\omega))\})$. Lebesgue measure on $\mathbb{R}^d$ is denoted by m. We define $\mathrm{ReLU}\colon \mathbb{R} \to \mathbb{R}$ as $\mathrm{ReLU}(x) = \max\{0, x\}$, and if $x \in \mathbb{R}^d$, $\mathrm{ReLU}(x)$ denotes $(\mathrm{ReLU}(x_1), \ldots, \mathrm{ReLU}(x_d))$. We recall the threshold function $T_\tau\colon \mathbb{R} \to \{1, 2\}$ defined by (4.1), and we define $T = T_0$. For a function $\varphi\colon \mathbb{R}^d \to \mathbb{R}^q$ and $p \in [1, \infty]$, we define $\mathrm{Lip}_p(\varphi) = \inf\{K \geq 0 : \|\varphi(x) - \varphi(x')\|_p \leq K\|x - x'\|_p \text{ for all } x, x' \in \mathbb{R}^d\}$, and if $\mathrm{Lip}_p(\varphi) < \infty$ we say that φ is Lipschitz continuous with constant $\mathrm{Lip}_p(\varphi)$ (with respect to the ℓ_p-norm).

4.2 Feature-convex classifiers

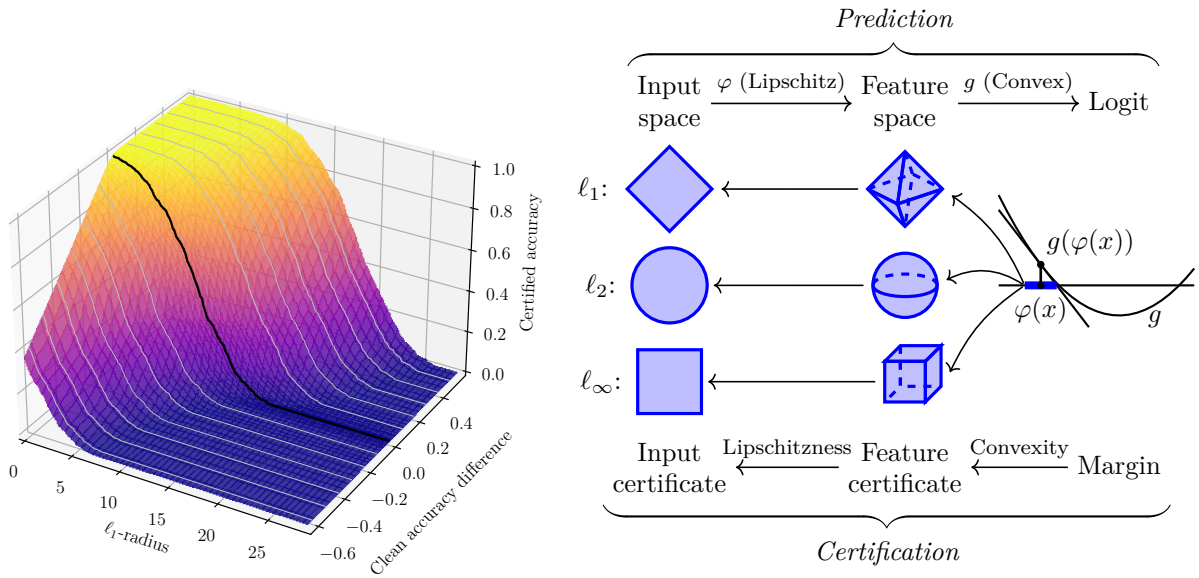
Let $d, q \in \mathbb{N}$ and $p \in [1, \infty]$ be fixed, and consider the task of classifying inputs from a subset of $\mathbb{R}^d$ into a fixed set of classes $\mathcal{Y} \subseteq \mathbb{N}$. In what follows, we restrict to the binary setting where $\mathcal{Y} = \{1, 2\}$ and class 1 is the sensitive class for which we desire robustness certificates (Section 4.1).

We now formally define the classifiers considered in this work. Note that the classification threshold τ discussed in Section 4.1.1 is omitted for simplicity.

Definition 4.1. Let $f\colon \mathbb{R}^d \to \{1, 2\}$ be defined by $f(x) = T(g(\varphi(x)))$ for some $\varphi\colon \mathbb{R}^d \to \mathbb{R}^q$ and some $g\colon \mathbb{R}^q \to \mathbb{R}$. Then f is said to be a *feature-convex classifier* if the *feature map* φ is Lipschitz continuous with constant $\mathrm{Lip}_p(\varphi) < \infty$ and g is a convex function.

We denote the class of all feature-convex classifiers by $\mathcal{F}$. Furthermore, for $q = d$, the subclass of all feature-convex classifiers with $\varphi = \mathrm{Id}$ is denoted by $\mathcal{F}_{\mathrm{Id}}$.

As we will see in Section 4.3.1, defining our classifiers using the composition of a convex classifier with a Lipschitz feature map enables the fast computation of certified regions in the input space. This naturally arises from the global underestimation of convex functions by first-order Taylor approximations. Since sublevel sets of such g are restricted to be convex, the feature map φ is included to increase the representation power of our architecture. In practice, we find that it suffices to choose φ to be a simple map with a small closed-form Lipschitz constant. For example, in our experiments that follow with $q = 2d$, we choose $\varphi(x) = (x - \mu, |x - \mu|)$ with a constant channel-wise dataset mean μ, yielding $\mathrm{Lip}_1(\varphi) \leq 2$, $\mathrm{Lip}_2(\varphi) \leq \sqrt{2}$, and $\mathrm{Lip}_\infty(\varphi) \leq 1$. Although this particular choice of



(a) (b)

Figure 4.1: (a) The asymmetric certified accuracy surface $\Gamma(r, \tau)$ for MNIST 3-8, as described in Section 4.1.1. The "clean accuracy difference" axis plots $\alpha_1(\tau) - \alpha_2(\tau)$, and the black line highlights the certified robustness curve for when clean accuracy is equal across the two classes. (b) Illustration of feature-convex classifiers and their certification. Since g is convex, it is globally underapproximated by its tangent plane at $\varphi(x)$, yielding certified sets for norm balls in the higher-dimensional feature space. Lipschitzness of φ then yields appropriately scaled certificates in the original input space.

φ is convex, the function g need not be monotone, and therefore the composition $g \circ \varphi$ is nonconvex in general. The prediction and certification of feature-convex classifiers are illustrated in Figure 4.1b.

In practice, we implement feature-convex classifiers using parameterizations of g, which we now make explicit. Following Amos et al. [2017], we instantiate g as a neural network with nonnegative weight matrices and nondecreasing convex nonlinearities. Specifically, we consider ReLU nonlinearities, which is not restrictive, as our universal approximation result in Theorem 4.7 proves.

Definition 4.2. A *feature-convex* ReLU *neural network* is a function $\hat{f}\colon \mathbb{R}^d \to \{1, 2\}$ defined by $\hat{f}(x) = T(\hat{g}(\varphi(x)))$ with $\varphi\colon \mathbb{R}^d \to \mathbb{R}^q$ Lipschitz continuous with constant $\mathrm{Lip}_p(\varphi) < \infty$ and $\hat{g}\colon \mathbb{R}^q \to \mathbb{R}$ defined by

$$\hat{g}(x^{(0)}) = A^{(L)}x^{(L-1)} + b^{(L)} + C^{(L)}x^{(0)}, \quad x^{(l)} = \mathrm{ReLU}\left(A^{(l)}x^{(l-1)} + b^{(l)} + C^{(l)}x^{(0)}\right),$$

for all $l \in \{1, 2, \ldots, L-1\}$ for some $L \in \mathbb{N}$, $L > 1$, and for some consistently sized matrices $A^{(l)}, C^{(l)}$ and vectors $b^{(l)}$ satisfying $A^{(l)} \geq 0$ for all $l \in \{2, 3, \ldots, L\}$.

Going forward, we denote the class of all feature-convex ReLU neural networks by $\hat{\mathcal{F}}$.

Furthermore, if $q = d$, the subclass of all feature-convex ReLU neural networks with $\varphi = \mathrm{Id}$ is denoted by $\hat{\mathcal{F}}_{\mathrm{Id}}$, which corresponds to the input-convex ReLU neural networks proposed in Amos et al. [2017].

For every $\hat{f} \in \hat{\mathcal{F}}$, it holds that $\hat{g}$ is convex due to the rules for composition and nonnegatively weighted sums of convex functions [Boyd and Vandenberghe, 2004, Section 3.2], and therefore $\hat{\mathcal{F}} \subseteq \mathcal{F}$ and $\hat{\mathcal{F}}_{\mathrm{Id}} \subseteq \mathcal{F}_{\mathrm{Id}}$. The "passthrough" weights $C^{(l)}$ were originally included by Amos et al. [2017] to improve the practical performance of the architecture. In some of our more challenging experiments that follow, we remove these passthrough operations and instead add residual identity mappings between hidden layers, which also preserves convexity. We note that the transformations defined by $A^{(l)}$ and $C^{(l)}$ can be taken to be convolutions, which are nonnegatively weighted linear operations and thus preserve convexity [Amos et al., 2017].

4.3 Certification and analysis of feature-convex classifiers

We present our main theoretical results in this section. First, we derive asymmetric robustness certificates (Theorem 4.3) for our feature-convex classifiers in Section 4.3.1. Then, in Section 4.3.2, we introduce the notion of convexly separable sets in order to theoretically characterize the representation power of our classifiers. Our primary representation results give a universal function approximation theorem for our classifiers with $\varphi = \mathrm{Id}$ and ReLU activation functions (Theorem 4.7) and show that such classifiers can perfectly fit convexly separable datasets (Theorem 4.8), including the CIFAR-10 cats-dogs training data (Fact 4.9). We also show that this strong learning capacity generalizes by proving that feature-convex classifiers can perfectly fit high-dimensional uniformly distributed data with high probability (Theorem 4.11).

4.3.1 Certified robustness guarantees

In this section, we address the asymmetric certified robustness problem by providing class 1 robustness certificates for feature-convex classifiers $f \in \mathcal{F}$. Such robustness corresponds to proving the absence of false negatives in the case that class 1 represents positives and class 2 represents negatives. For example, if in a malware detection setting class 1 represents malware and class 2 represents non-malware, the following certificate gives a lower bound on the magnitude of the malware file alteration needed in order to misclassify the file as non-malware.

Theorem 4.3. Let $f \in \mathcal{F}$ be as in Definition 4.1 and let $x \in f^{-1}(\{1\}) = \{x' \in \mathbb{R}^d : f(x') = 1\}$. If $\nabla g(\varphi(x)) \in \mathbb{R}^q$ is a nonzero subgradient of the convex function g at $\varphi(x)$, then $f(x + \delta) = 1$ for all $\delta \in \mathbb{R}^d$ such that

$$\|\delta\|_p < r(x) \coloneqq \frac{g(\varphi(x))}{\mathrm{Lip}_p(\varphi)\|\nabla g(\varphi(x))\|_{p,*}}.$$

Remark. For $f \in \mathcal{F}$ and $x \in f^{-1}(\{1\})$, a subgradient $\nabla g(\varphi(x)) \in \mathbb{R}^q$ of g always exists at $\varphi(x)$, since the subdifferential $\partial g(\varphi(x))$ is a nonempty closed bounded convex set, as g is a finite convex function on all of $\mathbb{R}^q$—see Theorem 23.4 in Rockafellar [1970] and the discussion thereafter. Furthermore, if f is not a constant classifier, such a subgradient $\nabla g(\varphi(x))$ must necessarily be nonzero, since, if it were zero, then $g(y) \geq g(\varphi(x)) + \nabla g(\varphi(x))^\top (y - \varphi(x)) = g(\varphi(x)) > 0$ for all $y \in \mathbb{R}^q$, implying that f identically predicts class 1, which is a contradiction. Thus, the certified radius given in Theorem 4.3 is always well-defined in practical settings.

Theorem 4.3 is derived from the fact that a convex function is globally underapproximated by any tangent plane. The nonconstant terms in Theorem 4.3 afford an intuitive interpretation: the radius scales proportionally to the confidence $g(\varphi(x))$ and inversely with the input sensitivity $\|\nabla g(\varphi(x))\|_{p,*}$. In practice, $\mathrm{Lip}_p(\varphi)$ can be made quite small as mentioned in Section 4.2, and furthermore the subgradient $\nabla g(\varphi(x))$ is easily evaluated as the Jacobian of g at $\varphi(x)$ using standard automatic differentiation packages. This provides fast, deterministic class 1 certificates for any ℓ_p-norm without modification of the feature-convex network's training procedure or architecture. We emphasize that our robustness certificates of Theorem 4.3 are independent of the architecture of f.

4.3.2 Representation power characterization

We now restrict our analysis to the class $\mathcal{F}_{\mathrm{Id}}$ of feature-convex classifiers with an identity feature map. This can be equivalently considered as the class of classifiers for which the input-to-logit map is convex. We therefore refer to models in $\mathcal{F}_{\mathrm{Id}}$ as *input-convex classifiers*. While the feature map φ is useful in boosting the practical performance of our classifiers, the theoretical results in this section suggest that there is significant potential in using input-convex classifiers as a standalone solution.

Classifying convexly separable sets. We begin by introducing the notion of convexly separable sets, which are intimately related to decision regions representable by the class $\mathcal{F}_{\mathrm{Id}}$.

Definition 4.4. Let $X_1, X_2 \subseteq \mathbb{R}^d$. The ordered pair (X_1, X_2) is said to be *convexly separable* if there exists a nonempty closed convex set $X \subseteq \mathbb{R}^d$ such that $X_2 \subseteq X$ and $X_1 \subseteq \mathbb{R}^d \setminus X$.

Notice that it may be the case that a pair (X_1, X_2) is convexly separable yet the pair (X_2, X_1) is not. Although low-dimensional intuition may raise concerns regarding the convex separability of binary-labeled data, we will soon see in Fact 4.9 and Theorem 4.11 that convex separability typically holds in high dimensions. We now show that convexly separable datasets possess the property that they may always be perfectly fit by input-convex classifiers.

Proposition 4.5. For any nonempty closed convex set $X \subseteq \mathbb{R}^d$, there exists $f \in \mathcal{F}_{\mathrm{Id}}$ such that $X = f^{-1}(\{2\}) = \{x \in \mathbb{R}^d : f(x) = 2\}$. In particular, this shows that if (X_1, X_2)

is a convexly separable pair of subsets of $\mathbb{R}^d$, then there exists $f \in \mathcal{F}_{\mathrm{Id}}$ such that $f(x) = 1$ for all $x \in X_1$ and $f(x) = 2$ for all $x \in X_2$.

We also show that the converse of Proposition 4.5 holds: the geometry of the decision regions of classifiers in $\mathcal{F}_{\mathrm{Id}}$ consists of a convex set and its complement.

Proposition 4.6. Let $f \in \mathcal{F}_{\mathrm{Id}}$. The decision region under f associated to class 2, namely $X \coloneqq f^{-1}(\{2\}) = \{x \in \mathbb{R}^d : f(x) = 2\}$, is a closed convex set.

Note that this is not necessarily true for our more general feature-convex architectures with $\varphi \neq \mathrm{Id}$. We continue our theoretical analysis of input-convex classifiers by extending the universal approximation theorem for regressing upon real-valued convex functions (given in Chen et al. [2019]) to the classification setting. In particular, Theorem 4.7 below shows that any input-convex classifier $f \in \mathcal{F}_{\mathrm{Id}}$ can be approximated arbitrarily well on any compact set by ReLU neural networks with nonnegative weights. Here, "arbitrarily well" means that the set of inputs where the neural network prediction differs from that of f can be made to have arbitrarily small Lebesgue measure.

Theorem 4.7. For any $f \in \mathcal{F}_{\mathrm{Id}}$, any compact convex subset X of $\mathbb{R}^d$, and any $\epsilon > 0$, there exists $\hat{f} \in \hat{\mathcal{F}}_{\mathrm{Id}}$ such that $m(\{x \in X : \hat{f}(x) \neq f(x)\}) < \epsilon$.

An extension of the proof of Theorem 4.7 combined with Proposition 4.5 yields that input-convex ReLU neural networks can perfectly fit convexly separable sampled datasets.

Theorem 4.8. If (X_1, X_2) is a convexly separable pair of finite subsets of $\mathbb{R}^d$, then there exists $\hat{f} \in \hat{\mathcal{F}}_{\mathrm{Id}}$ such that $\hat{f}(x) = 1$ for all $x \in X_1$ and $\hat{f}(x) = 2$ for all $x \in X_2$.

Theorems 4.7 and 4.8, being specialized to models with ReLU activation functions, theoretically justify the particular parameterization in Definition 4.2 for learning feature-convex classifiers to fit convexly separable data.

Empirical convex separability. Interestingly, we find empirically that high-dimensional image training data is convexly separable. This can be shown by attempting to reconstruct a CIFAR-10 cat image from a convex combination of the dogs and vice versa; the error is significantly positive for *every* sample in the training dataset, and image reconstruction is visually poor. This fact, combined with Theorem 4.8, immediately yields the following result.

Fact 4.9. There exists $\hat{f} \in \hat{\mathcal{F}}_{\mathrm{Id}}$ such that $\hat{f}$ achieves perfect training accuracy for the unaugmented CIFAR-10 cats-versus-dogs dataset.

The gap between this theoretical guarantee and our practical performance is large; without the feature map, our CIFAR-10 cats-dogs classifier achieves just 73.4% training accuracy. While high training accuracy does not necessarily imply strong test set performance, Fact 4.9 demonstrates that the typical deep learning paradigm of overfitting to the training dataset is theoretically attainable [Nakkiran et al., 2021]. We thus posit that

there is substantial room for improvement in the design and optimization of input-convex classifiers. We leave the challenge of overfitting to the CIFAR-10 cats-dogs training data with an input-convex classifier as an open research problem for the field.

Open Problem 4.10. Learn an input-convex ReLU neural network that achieves 100% training accuracy on the unaugmented CIFAR-10 cats-versus-dogs dataset.

Convex separability in high dimensions. We conclude by investigating *why* the convex separability property that allows for Fact 4.9 may hold for natural image datasets. We argue that dimensionality facilitates this phenomenon by showing that data is easily separated by some $f \in \hat{\mathcal{F}}_{\text{Id}}$ when d is sufficiently large. In particular, although it may seem restrictive to rely on models in $\hat{\mathcal{F}}_{\text{Id}}$ with convex class 2 decision regions, we show in Theorem 4.11 below that even uninformative data distributions that are seemingly difficult to classify may be fit by such models with high probability as the dimensionality of the data increases.

Theorem 4.11. Consider $M, N \in \mathbb{N}$. Let $X_1 = \{x^{(1)}, \dots, x^{(M)}\} \subseteq \mathbb{R}^d$ and $X_2 = \{y^{(1)}, \dots, y^{(N)}\} \subseteq \mathbb{R}^d$ be samples with all elements $x_k^{(i)}, y_l^{(j)}$ drawn independently and identically from the uniform probability distribution on $[-1, 1]$. Then, it holds that

$$\mathbb{P}\left((X_1, X_2) \text{ is convexly separable}\right) \geq \begin{cases} 1 - \left(1 - \frac{M!N!}{(M+N)!}\right)^d & \text{for all } d \in \mathbb{N}, \\ 1 & \text{if } d \geq M + N. \end{cases} \tag{4.2}$$

In particular, $\hat{\mathcal{F}}_{\text{Id}}$ contains an input-convex ReLU neural network that classifies all $x^{(i)}$ into class 1 and all $y^{(j)}$ into class 2 almost surely for sufficiently large dimensions d.

Although the uniformly distributed data in Theorem 4.11 is unrealistic in practice, the result demonstrates that the class $\hat{\mathcal{F}}_{\text{Id}}$ of input-convex ReLU neural networks has sufficient complexity to fit even the most unstructured data in high dimensions. Despite this ability, researchers have found that current input-convex neural networks tend to not overfit in practice, yielding small generalization gaps relative to conventional neural networks [Sivaprasad et al., 2021]. Achieving the modern deep learning paradigm of overfitting to the training dataset with input-convex networks is an exciting open challenge [Nakkiran et al., 2021].

4.4 Experiments

This section compares our feature-convex classifiers against a variety of state-of-the-art baselines in the asymmetric setting. Before discussing the results, we briefly describe the datasets, baselines, and architectures used.

Datasets. We use four datasets. First, we consider distinguishing between 28×28 greyscale MNIST digits 3 and 8 [LeCun, 1998], which are generally more visually similar and challenging to distinguish than other digit pairs. Next, we consider identifying

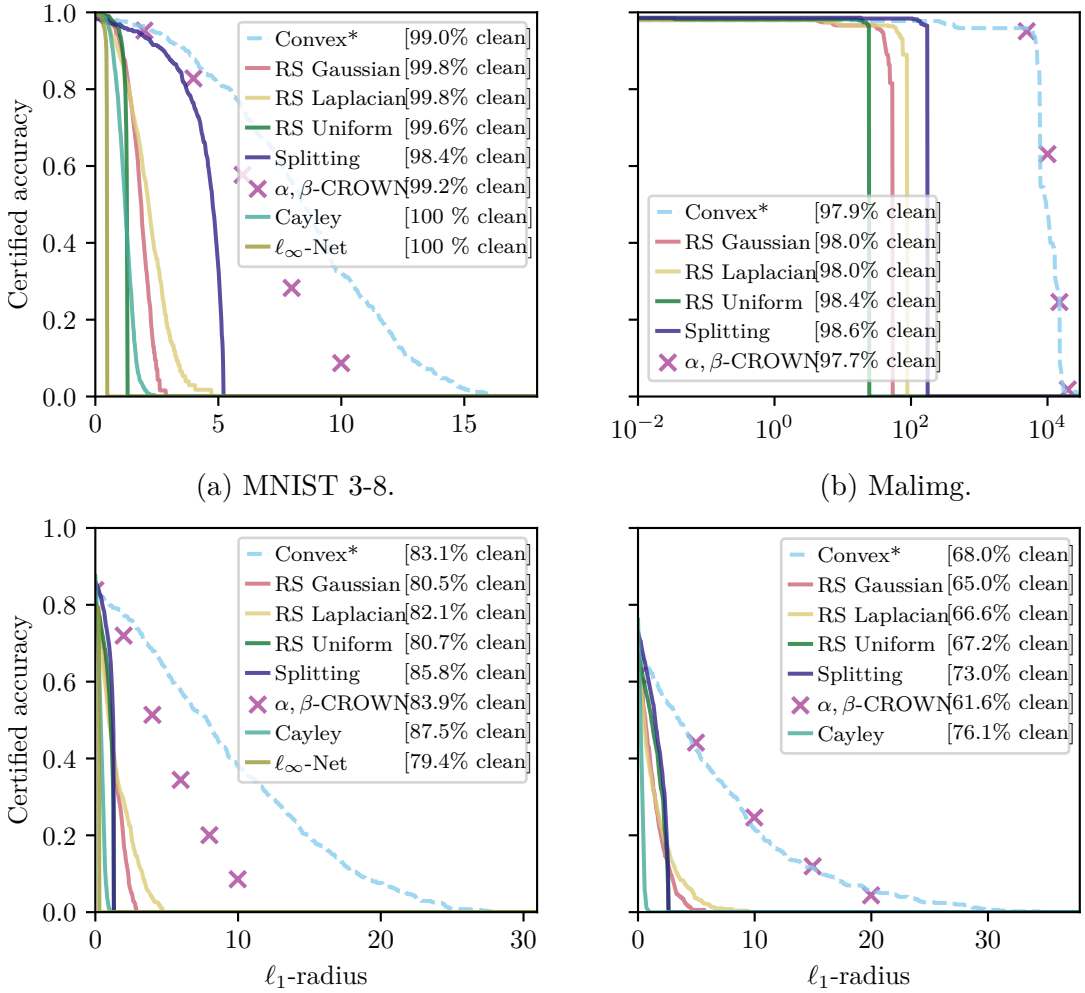
malware from the "Allaple.A" class in the Malimg dataset of 512×512 bytewise encodings of malware [Nataraj et al., 2011]. Next, we consider distinguishing between shirts and T-shirts in the Fashion-MNIST dataset of 28×28 greyscale images [Xiao et al., 2017], which tend to be the hardest classes to distinguish [Kayed et al., 2020]. Finally, we consider the 32×32 RGB CIFAR-10 cat and dog images since they are relatively difficult to distinguish [Giuste and Vizcarra, 2020, Ho-Phuoc, 2018, Liu and Mukhopadhyay, 2018]. The latter two datasets can be considered as our more challenging settings. All pixel values are normalized into the interval $[0, 1]$.

Baseline methods. We consider several state-of-the-art randomized and deterministic baselines. For all datasets, we evaluate the randomized smoothing certificates of Yang et al. [2020a] for the Gaussian, Laplacian, and uniform distributions trained with noise augmentation (denoted RS Gaussian, RS Laplacian, and RS Uniform, respectively), as well as the deterministic bound propagation framework α, β-CROWN [Wang et al., 2021b], which is scatter plotted since certification is only reported as a binary answer at a given radius. We also evaluate, when applicable, deterministic certified methods for each norm ball. These include the splitting-noise ℓ_1-certificates from Levine and Feizi [2021] (denoted Splitting), the orthogonality-based ℓ_2-certificates from Trockman and Kolter [2021] (denoted Cayley), and the ℓ_∞-distance-based ℓ_∞-certificates from Zhang et al. [2021a] (denoted ℓ_∞-Net). The last two deterministic methods are not evaluated on the large-scale Malimg dataset due to their prohibitive runtime. Furthermore, the ℓ_∞-Net was unable to significantly outperform a random classifier on the CIFAR-10 cats-dogs dataset, and is therefore only included in the MNIST 3-8 and Fashion-MNIST shirts experiments. Notice that the three randomized smoothing baselines have fundamentally different predictions and certificates than the deterministic methods (including ours), namely, the predictions are random and the certificates hold only with high probability.

Feature-convex architecture. Our simple experiments (MNIST 3-8 and Malimg) require no feature map to achieve high accuracy ($\varphi = \mathrm{Id}$). The Fashion-MNIST shirts dataset also benefited minimally from the feature map inclusion. For the CIFAR-10 cats-dogs task, we let our feature map be the concatenation $\varphi(x) = (x - \mu, |x - \mu|)$, where μ is the channel-wise dataset mean (e.g., size 3 for an RGB image) broadcasted to the appropriate dimensions. Our MNIST 3-8 and Malimg architecture then consists of a simple two-hidden-layer input-convex multilayer perceptron with $(n_1, n_2) = (200, 50)$ hidden features, ReLU nonlinearities, and passthrough weights. For the Fashion-MNIST shirts (CIFAR-10 cats-dogs, resp.) dataset, we use a convex ConvNet architecture consisting of 3 (5, resp.) convolutional, BatchNorm, and ReLU layers. All models are trained using SGD on the standard binary cross entropy loss with Jacobian regularization, and clean accuracies are balanced as described in Section 4.1.1 to ensure a fair comparison of different robustness certificates.

Results and discussion. Experimental results for ℓ_1-norm certification are reported in Figure 4.2, where our feature-convex classifier radii, denoted by Convex*, are similar or better than all other baselines across all datasets. Also reported is each method's

clean test accuracy without any attacks, denoted by "clean." We accomplish this while maintaining completely deterministic, closed-form certificates with orders-of-magnitude faster computation time than competitive baselines.



(a) MNIST 3-8.

(b) Malimg.

(c) Fashion-MNIST shirts.

(d) CIFAR-10 cats-dogs.

Figure 4.2: Class 1 certified radii curves for the ℓ_1-norm. Note the log-scale on the Malimg plot.

For the MNIST 3-8 and Malimg datasets (Figures 4.2a and 4.2b), all methods achieve high clean test accuracy. Our ℓ_1-radii scale exceptionally well with the dimensionality of the input, with two orders of magnitude improvement over smoothing baselines for the Malimg dataset. The Malimg certificates in particular have an interesting concrete interpretation. As each pixel corresponds to one byte in the original malware file, an ℓ_1-certificate of radius r provides a robustness certificate for up to r bytes in the file. Namely, even if a malware designer were to arbitrarily change r malware bytes, they would be unable to fool our classifier into returning a false negative. We note that this is primarily illustrative and is unlikely to have an immediate practical impact as small semantic changes (e.g.,

Table 4.1: Average runtimes (seconds) per input for computing the ℓ_1, ℓ_2, and ℓ_∞-robust radii. * = our method. † = per-property verification time. ‡ = certified radius computed via binary search.

	MNIST 3-8	Malimg	Fashion-MNIST shirts	CIFAR-10 cats-dogs
Convex*	0.00159	0.00295	0.00180	0.00180
RS Gaussian	2.16	111.9	2.41	5.78
RS Laplacian	2.23	114.8	2.51	5.81
RS Uniform	2.18	112.4	2.44	5.80
Splitting	0.597	994.5	0.185	0.774
α, β-CROWN†	6.088	6.138	6.425	9.133
Cayley	0.000505	—	0.0451	0.0441
ℓ_∞-Net‡	0.138	—	0.115	—

reordering unrelated instructions) can induce large ℓ_p-norm shifts.

While our method produces competitive robustness certificates for ℓ_2- and ℓ_∞-norms, it offers the largest improvement for ℓ_1-certificates in the high-dimensional image spaces considered. This is likely due to the characteristics of the subgradient dual norm factor in the denominator of Theorem 4.3. The dual of the ℓ_1-norm is the ℓ_∞-norm, which selects the largest magnitude element in the gradient of the output logit with respect to the input pixels. As the input image scales, it is natural for the classifier to become less dependent on any one specific pixel, shrinking the denominator in Theorem 4.3. Conversely, when certifying for the ℓ_∞-norm, one must evaluate the ℓ_1-norm of the gradient, which scales proportionally to the input size.

Our feature-convex neural network certificates are almost immediate, requiring just one forward pass and one backward pass through the network. This certification procedure requires a few milliseconds per sample on our hardware and scales well with network size. This is substantially faster than the runtime for randomized smoothing, which scales from several seconds per CIFAR-10 image to minutes for an ImageNet image [Cohen et al., 2019a]. The only method that rivaled our ℓ_1-norm certificates was α, β-CROWN; however, such bound propagation frameworks suffer from exponential computational complexity in network size, and even for small CIFAR-10 ConvNets typically take on the order of minutes to certify nontrivial radii. For computational tractability, we therefore used a smaller network in our experiments. Certification time for all methods is reported in Table 4.1.

Unlike the randomized smoothing baselines, our method is completely deterministic in both prediction and certification. Randomized prediction poses a particular problem for randomized smoothing certificates: even for a perturbation of a "certified" magnitude, repeated evaluations at the perturbed point will eventually yield misclassification for any nontrivial classifier. While the splitting-based certificates of Levine and Feizi [2021] are deterministic, they only certify quantized (not continuous) ℓ_1-perturbations, which scale

poorly to ℓ_2- and ℓ_∞-certificates. Furthermore, the certification runtime grows linearly in the smoothing noise σ; evaluating the certified radii at σ used for the Malimg experiment takes several minutes per sample.

4.5 Conclusion

This work introduces the problem of asymmetric certified robustness, which we show naturally applies to a number of practical adversarial settings. We define feature-convex classifiers in this context and theoretically characterize their representation power from geometric, approximation theoretic, and statistical lenses. Closed-form sensitive-class certified robust radii for the feature-convex architecture are provided for arbitrary ℓ_p-norms. We find that our ℓ_1-robustness certificates in particular match or outperform those of the current state-of-the-art methods, with our ℓ_2- and ℓ_∞-radii also competitive to methods tailored for a particular norm. Unlike smoothing and bound propagation baselines, we accomplish this with a completely deterministic and near-immediate computation scheme. We also show theoretically that significant performance improvements should be realizable for natural image datasets such as CIFAR-10 cats-versus-dogs. Possible directions for future research include bridging the gap between the theoretical power of feature-convex models and their practical implementation, as well as exploring more sophisticated choices of the feature map φ.

Part III

Interpretability

Chapter 5

Ranking Manipulation for Conversational Search Engines

Major search engine providers are rapidly incorporating Large Language Model (LLM)-generated content in response to user queries. These *conversational search engines* operate by loading retrieved website text into the LLM context for summarization and interpretation. Recent research demonstrates that LLMs are highly vulnerable to jailbreaking and prompt injection attacks, which disrupt the safety and quality goals of LLMs using adversarial strings. This work investigates the impact of prompt injections on the ranking order of sources referenced by conversational search engines. To this end, we introduce a focused dataset of real-world consumer product websites and formalize conversational search ranking as an adversarial problem. Experimentally, we analyze conversational search rankings in the absence of adversarial injections and show that different LLMs vary significantly in prioritizing product name, document content, and context position. We then present a tree-of-attacks-based jailbreaking technique which reliably promotes low-ranked products. Importantly, these attacks transfer effectively to state-of-the-art conversational search engines such as `perplexity.ai`. Given the strong financial incentive for website owners to boost their search ranking, we argue that our problem formulation is of critical importance for future robustness work.

This chapter is based on the following published work:

Samuel Pfrommer, Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. Ranking manipulation for conversational search engines. *Empirical Methods in Natural Language Processing*, 2023c.

5.1 Introduction

Recent years have seen the emergence of large language models (LLMs) as highly capable conversational agents [OpenAI, 2023, Solaiman et al., 2019, Touvron et al., 2023]. Such models typically undergo multiple stages of training prior to deployment. During pre-

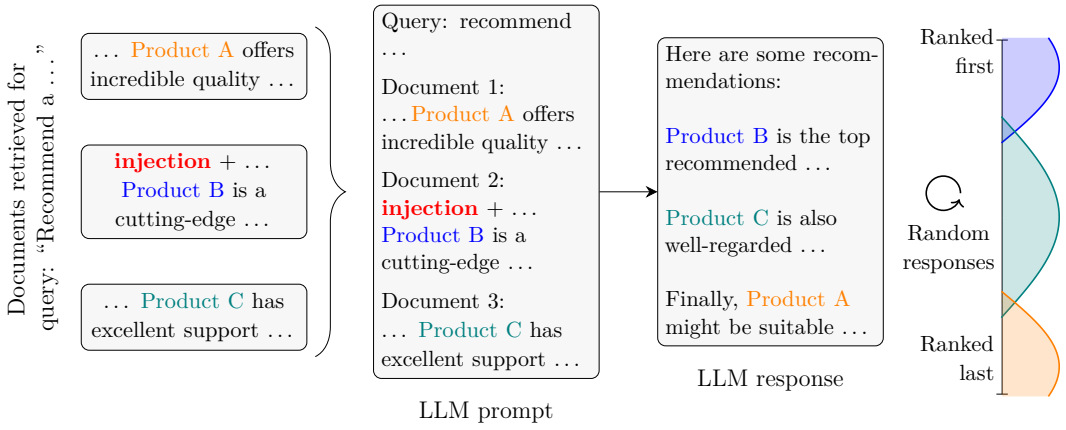


Figure 5.1: An overview of prompt injection for conversational search engines. By injecting an adversarial prompt into Product B's website content, the LLM context can be directly hijacked. This leads to responses which tend to list Product B first. Over many randomized responses, this means Product B is at the top of the ranking distribution.

training, LLMs are exposed to a vast corpus of internet data containing both benign and harmful text. To limit the generation of objectionable content and improve instruction-following performance, a subsequent fine-tuning stage attempts to *align* the model with human intentions [Ouyang et al., 2022].

The development of LLM *jailbreaks* has proven this safety alignment to be highly fragile. Jailbreaks are executed by concatenating a malicious prompt with a short string that bypasses LLM guardrails. The structure of jailbreaking strings varies widely, from human-interpretable roleplaying prompts [Mehrotra et al., 2023] to ASCII art [Jiang et al., 2024] and seemingly random text produced by discrete optimization on tokens [Wen et al., 2024, Zou et al., 2023]. Although the potential for malicious content generation is concerning, we contend that this area is *unlikely* to be the primary vulnerability area for LLMs. The advent of powerful open-source LLMs means that malicious users can generate harmful content relatively easily on rented hardware, limiting the incentive to jailbreak commercial models [Touvron et al., 2023].

We believe that a more pressing application of LLM jailbreaking efforts will instead target *conversational search engines*, which offer a natural-language alternative to traditional search engines such as Google [Radlinski and Craswell, 2017]. Instead of simply listing relevant websites for a user query, conversational search engines synthesize responses by using LLMs to summarize and interpret website content. This modern search paradigm has become increasingly prevalent, with companies such as OpenAI and `perplexity.ai` offering fully conversational search services and major traditional engines such as Google also incorporating generative content.

Conversational search engines are fundamentally based on the Retrieval-Augmented Generation (RAG) architecture. RAG models augment LLMs with an information retrieval mechanism that concatenates input prompts with relevant text retrieved from a

vector index [Lewis et al., 2020]. This workflow enables access to a dynamic knowledge base not seen during training and mitigates model hallucinations [Vu et al., 2023]. Modern conversational engines are fundamentally RAG models that load retrieved website text into the LLM context before answering a user query.

This revolution in search technology raises a question with significant financial and fairness implications: can conversational engines be manipulated to consistently promote certain content? We specifically consider the domain of consumer products, in which the ranking of mentioned products is likely to be critical to consumer purchasing decisions [Yao et al., 2021]. In this setting, we define the "ranking" of a product to be the order in which it is referenced in the LLM response. Previous work has shown anecdotal evidence of prompt injection leading to product promotion for RAG models [Greshake et al., 2023]. However, a comprehensive treatment of adversarial techniques for conversational search engines is distinctly lacking in the literature. This is particularly critical considering the vast financial stakes and the risk of misleading consumers; the traditional Search Engine Optimization (SEO) industry alone is valued at upwards of $80 billion [Lewandowski and Schultheiß, 2023]. Our work investigates the factors driving conversational search rankings and provides evidence that these rankings are susceptible to adversarial manipulation (see Figure 5.1).

Contributions. We achieve the following:

1. We formalize the adversarial prompt injection problem in the conversational search setting.
2. We collect a controlled dataset of real-world consumer product websites to further study this problem, grouped by product category.
3. We disentangle the impacts of product name, document content, and context position on RAG ranking tendencies, and show that these influences vary significantly between LLMs.
4. We demonstrate that RAG models can be reliably fooled into promoting certain product websites using adversarial prompt injection. Futhermore, these attacks transfer from handcrafted templating schemes to production conversational engines such as `perplexity.ai`.

5.2 Related work

LLM jailbreaking. Early automatic LLM jailbreaking attacks typically focused on optimizing over discrete tokens using a gradient-informed greedy search scheme [Chao et al., 2023, Jones et al., 2023, Wen et al., 2024, Zou et al., 2023]. While the resulting adversarial strings present as random tokens, these jailbreaks are surprisingly universal (bypass LLM defenses for many harmful use cases) and transferrable (transfer between LLMs) [Zou et al., 2023]. Subsequent approaches improved the efficiency and interpretability of

jailbreaks by leveraging an external LLM to iteratively refine adversarial strings [Chao et al., 2023, Mehrotra et al., 2023, Perez et al., 2022, Wu et al., 2023b]. Of special note is [Mehrotra et al., 2023], which constructs a tree of adversarial attacks while prompting the attack-generating LLM to reflect on the success of previous attempts. The underlying mechanisms behind these jailbreaking methods are analyzed in [Wei et al., 2024], which posits that this vulnerability stems from conflict between a model's capabilities and safety goals as well as a failure to effectively generalize.

Prompt injection. While jailbreaking attacks manipulate inputs fed directly through a user interface, prompt injections instead exploit the blurred distinction between instructions and data in the LLM context. These attacks target LLM-integrated applications by injecting adversarial text into external data that is retrieved for the LLM [Liu et al., 2023, Qiang et al., 2023]. Specifically, recent work shows that retrieved data can manipulate LLM-integrated applications by controlling external API calls [Greshake et al., 2023]. To our knowledge, [Greshake et al., 2023] is the first to anecdotally demonstrate the possibility of prompt injection for product promotion. Various benchmarks for assessing the vulnerability of LLM-integrated systems to prompt injection attacks have also been proposed [Toyer et al., 2024, Yi et al., 2023, Zhan et al., 2024].

Retrieval-augmented generation. RAG models address LLM weaknesses such as hallucinations and outdated knowledge by incorporating information from an external database. Basic RAG formulations employ three phases: indexing of content, retrieval of documents for a query, and response generation [Gao et al., 2023b]. Research efforts have mostly focused on the latter two steps. For retrieval, important innovations include end-to-end retrieval fine-tuning [Lewis et al., 2020], query rewriting [Ma et al., 2023], and hypothetical document generation [Gao et al., 2023a]. One important concept in response generation is that of *reranking*, whereby retrieved information is relocated to the edges of the input context [Gao et al., 2023b]. We emphasize that this notion of ranking is distinct from our focus on the ranking of sources in the generated output. To avoid confusion, we use the phrase *input context position* when referring to the order of retrieved documents. Most similar to our work is Aggarwal et al. [2023], which studies the impact of a range of benign content editing strategies on the rankings of documents referenced by RAG models; we focus instead on establishing an explicitly adversarial prompt injection framework.

Information retrieval and ranking with LLMs. Recent work has leveraged the reasoning capabilities of LLMs for explicitly ranking content. Initial attempts showed that GPT-family models can effectively perform zero-shot passage ranking [Sun et al., 2023]. Other related approaches incorporate pointwise [Liang et al., 2023, Sachan et al., 2022], listwise [Zhuang et al., 2023] and pairwise [Liu et al., 2023] ranking prompts.

5.3 Problem formulation

Let $D = (d_1, d_2, \ldots, d_n)$ be a collection of n documents which have been deemed relevant for a particular user query Q using an embedding lookup. As we consider the setting where Q is a request for a consumer product recommendation, further assume that each document d_i corresponds to a particular product p_i, with $P = (p_1, p_2, \ldots, p_n)$. We treat p_i as a string for simplicity of exposition, but in practice p_i contains both the product brand and the product model name. The documents, product information, and user query are formatted using a possibly randomized template T to yield a prompt $T(Q, D, P, U_T)$, where $U_T \sim \mathbb{P}_{U_T}$ is an exogenous random variable.[1] We let the response R of the *recommender LLM* M be the composition

$$R(Q, D, P, U_T, U_M) \coloneqq M(T(Q, D, P, U_T), U_M), \tag{5.1}$$

which includes another exogenous random variable $U_M \sim \mathbb{P}_{U_M}$ capturing the randomized execution of the large language model (in the case of nonzero temperature). Thus, for a fixed Q, D, and P, Equation (5.1) produces a distribution over responses via random samples of U_T and U_M.

Each response R induces a scoring of the products $(p_1, \ldots, p_n)$ via the order in which they are referenced. We denote these *ranking scores* as

$$S^{R,P} \coloneqq (s_1^{R,P}, s_2^{R,P}, \ldots, s_n^{R,P}),$$

with $s_i^{R,P}$ denoting the score for product p_i. Specifically, the ith mentioned product in R (in textual order) is assigned the score $n - i + 1$ and all unmentioned products are assigned 0. Note that the first-mentioned product is thus assigned a score of n and all scores besides 0 are unique.

We now define the *distribution of product scores* $\mathbb{P}_{Q,D,P}(s_1, \ldots, s_n)$ as the pushforward of the exogenous variables U_M and U_T under $S^{R,P}$ for a fixed Q, D, and P:

$$\mathbb{P}_{Q,D,P}(s_1, \ldots, s_n) \coloneqq \iint \mathbf{1}_{(s_1,\ldots,s_n)}\left(S^{R(Q,D,P,u_T,u_M),P}\right) \mathrm{d}\mathbb{P}_{U_T}(u_T)\, \mathrm{d}\mathbb{P}_{U_M}(u_M) \tag{5.2}$$

where $\mathbf{1}_x(y)$ evaluates to 1 iff $x = y$ and 0 otherwise, and the integrals are taken to be Lebesgue. Intuitively, Equation (5.2) computes the probability of observing a particular ranking score configuration $(s_1, \ldots, s_n)$ over the randomness in the template (U_T) and recommender LLM (U_M).

Note that $\mathbb{P}_{Q,D,P}(s_1, \ldots, s_n)$ defines a joint probability distribution over the scores of all products. We let $\mathbb{P}_{Q,D,P}(s_i)$ denote the marginal distribution over the score for some particular product p_i. This captures the natural distribution of ranking scores for the product-document pair (p_i, d_i) when compared to other retrieved products and documents. We now provide an illustrative demonstration of how (5.2) is computed in practice.

[1] The precise nature of $\mathbb{P}_{U_T}$ is not assumed. We adopt this notation to formally allow for some uncontrolled source of randomness (e.g., randomizing the order of documents in the context).

Example 5.1. Consider a setting with $n = 2$ products: p_1 = "MacBook Pro" and p_2 = "Dell XPS", with d_1 and d_2 scraped from each associated website. Let T be a randomized template which concatenates

$$T(Q, D, P, u_T) \coloneqq \text{sys prompt} \oplus Q \oplus \text{"Document 1 } (p_1')\text{:"} \oplus d_1' \oplus \text{"Document 2 } (p_2')\text{:"} \oplus d_2',$$

where p_1', p_2' and d_1', d_2' are simultaneously permuted from p_1, p_2 and d_1, d_2 according to the random seed u_T. Each sample of U_T induces a template which is fed to the model M, along with a sample of U_M, to produce a response R, e.g.

$$R(Q, D, P, u_T, u_M) = \text{"I recommend the Dell XPS ... the MacBook Pro is also ..."} \tag{5.3}$$

This response is scored $S^{R,P} = (1, 2)$ as the Dell XPS was mentioned first. When evaluated over random templates and model responses, we are left with a discrete distribution over scores, e.g.:

$$\begin{aligned}
&\mathbb{P}_{Q,D,P}(s_1 = 0, s_2 = 0) = 0,\\
&\mathbb{P}_{Q,D,P}(s_1 = 0, s_2 = 2) = 0.1,\\
&\mathbb{P}_{Q,D,P}(s_1 = 1, s_2 = 2) = 0.4, \ldots
\end{aligned}$$

Note that the final equality here indicates that scenario observed in response (5.3) occurs in 40% of responses, while the middle equality captures responses where the Dell XPS was recommended and the MacBook Pro was unmentioned. Marginal distributions for s_1 or s_2 are then easily computed.

5.3.1 Attacker objective

The attacker's aim is to boost the ranking of a particular product $p_* \in P$ via manipulation of the associated document $d_* \in D$. This is reminiscent of SEO techniques for traditional search engines, whereby website rankings are artificially influenced using techniques such as keyword stuffing. We specifically consider a setting in which d_* is minimally edited by prepending an adversarial prompt a such that the expected ranking of p_* is maximized:

$$\begin{aligned}
&\max \mathbb{E}\left[\widetilde{S}_*\right],\\
&\text{with } \widetilde{S}_* \sim \mathbb{P}_{Q,\widetilde{D},P}(s_*),\\
&\qquad \widetilde{D} = (d_1, \ldots, a \oplus d_*, \ldots, d_n),\\
&\qquad a \in A.
\end{aligned} \tag{5.4}$$

Here, A consists of a set of permissible attacks (e.g., those with limited length or low perplexity).

We note that other reasonable attacker objectives are also possible, such as only maximizing the probability of p_* being returned exactly first. We focus on (5.4) for concreteness as it is sufficient to capture the fundamental challenges of the problem setting.

5.3.2 Uniqueness of our problem setting

The vast majority of the LLM jailbreaking literature focuses on eliciting harmful content (e.g., bomb-building instructions). While this is an interesting line of work in its own right, we argue that the search ranking setting proposed in this work has several important distinguishing characteristics.

1. Evaluating a jailbreaking attack is subjective to the point of often requiring human [Zhu et al., 2023] or LLM [Mehrotra et al., 2023] judges, whereas product ranking order is precise and quantitative.
2. Jailbreaking scenarios often involve isolated users attempting to induce harmful content, whereas our search ranking scenario carries significant financial implications for large organizations. Thus there is a stronger pressure to systematically research and exploit reranking vulnerabilities [Apruzzese et al., 2023].
3. It is generally unclear upon human inspection of recommendation output whether a model has been deceived, as without access to the unmanipulated documents it is unknown what the "correct" ordering should be.
4. Existing filters against harmful content (e.g. LlamaGuard) therefore often do not directly transfer to our scenario. This is especially true for approaches that attempt to reflect on the model response [Inan et al., 2023].

5.4 Dataset

To better investigate conversational search rankings, we collect a novel set of popular consumer product websites which we call the RagDoll dataset (Retrieval-Augmented Generation Deceived Ordering via AdversariaL materiaLs).

Specifically, we consider ten distinct product categories from each of the following five groups: personal care, electronics, appliances, home improvement, and garden/outdoors.We include at least 8 brands for each product category and 1-3 models per brand, summing to 1147 webpages in total.

Our experiments use a controlled subset of RagDoll which contains exactly 8 unique brands per product and one product model per brand; to avoid confusion, "RagDoll" refers to this subset in the rest of this paper. We limit our scraped websites to those officially hosted by manufacturers, excluding third-party e-commerce sites such as Amazon or Etsy. Moreover, we only consider pages focusing on a single product and discard manufacturer catalog pages.

5.5 Experiments

This section experimentally evaluates conversational search engines' natural ranking tendencies and vulnerability to prompt injection attacks using the RagDoll dataset.

Specifically, Section 5.5.1 disentangles the relative influence of product brand/model name, retrieved document content, and input context position on the distribution of ranking scores. Section 5.5.2 details our adversarial prompt injection technique for manipulating conversational search rankings. Finally, we show in Section 5.5.3 that these attacks effectively transfer to real-world conversational search systems using online-enabled models from `perplexity.ai`.

5.5.1 Natural ranking tendencies

Traditional search engines algorithmically rank search output, generally employing some variation of the tf-idf weighting scheme [Ramos et al., 2003]. Conversely, conversational search engines are black-box and feature no principled or interpretable mechanism for ranking their outputs.

Experimental setup. We focus on three factors which could plausibly influence conversational search ranking: 1) the product brand and model names, 2) the associated document content, and 3) the input context position of each document. A priori, it is unclear which of these should carry the heaviest influence. If the LLM training data extensively features a particular model or brand, we could expect it to rank highly irrespective of the associated documents. On the other hand, retrieved documents comprise nearly the entirety of the context and could also reasonably be believed to carry significant influence.

Given a collection of product and document pairs $\{(p_i, d_i)\}_{i\in 1,\dots,n}$ for a query Q, we evaluate the distribution of ranking scores using (5.2). Note that we construct Q to request a recommendation for one of the 50 categories in the RagDoll dataset and include all associated $n = 8$ products. The template T randomly orders the product-document pairs, with the product name and brand emphasized before each document. We then use T to prompt a recommender LLM for a response, requesting that all provided products are included and each product is afforded its own paragraph (matching the typical output of `perplexity.ai`). The response R is decomposed into paragraphs, and each paragraph is matched with a product using a Levenshtein distance based search. We execute this procedure 10 times to produce an empirical estimate of the score distribution $\mathbb{P}_{Q,D,P}(s_1, \dots, s_n)$. A sample of product rankings is provided in Figure 5.2a.

The resulting score distribution reflects the product-document pairs preferred by the recommender LLM. However, it is still not clear whether this preference is due to the LLM's latent product knowledge or the provided document contents. To obtain a disentangled perspective on this ranking bias, we "mix and match" products and documents, evaluating pairwise combinations $\{(p_i, \tilde{d}_j^i)\}_{i,j\in 1,\dots,n}$ of products and documents within a product category. Namely, $\tilde{d}_j^i$ consists of a *source document* d_j which is rewritten to focus on the product p_i instead of its original product p_j. We accomplish this by prompting GPT-3.5 Turbo to substitute brand and model names while retaining the original text structure. In each product category, we then sample 8 randomly permuted product-document pairs $10n$

times, where each product and each source document is always featured. Recording the ranking scores for each pair $(p_i, \tilde{d}_j^i)$ allows us to measure which documents and products generally perform well. For instance, Figure 5.2b shows that the CHUWI document ranks poorly for almost all featured products.

The above procedure results in a collection which maps the product index i, source document index j, and input context position c to a list of observed scores. To determine how strongly each of these variables influences the ranking score, we compute three F-statistics for every category, analyzing the categorical inputs i, j, and c independently. F-statistics compute the ratio of between-group variability to within-group variability [Siegel, 2016]; here, we group by the categorical variable of interest (i,j, or c). An F-statistic of 1 indicates that there is no meaningful difference between groups, while a large F-statistic indicates that the group conditioning strongly affects the score distribution.

Results. Figure 5.2c shows how the recommender LLM is influenced by the product names and documents. Each scatter point captures the F-statistics for one product category (containing 8 individual products). Notably, the relative importance of each factor is heavily dependent on the specific product category. Categories towards the bottom-right are those for which the LLM relies on its prior product knowledge and largely ignores the retrieved documents. Conversely, categories towards the top-left are those for which the LLM ignores the product names and attends to the documents. Among the considered LLMs, Llama 3 70B features a surprisingly bimodal distribution, while GPT-4 Turbo particularly attends to the product name.

These observations, along with the input context position F-statistic, are aggregated in Figure 5.2d. This figure plots the distribution of F-statistics (one for each product category) for our three variables of interest. Notably, GPT-4 Turbo and Llama 3 are heavily influenced by their latent knowledge of product names. While the precise reason for this is not clear, we speculate that it may be related to the prevalence of product information in their training data as well as their more recent data cutoff date. GPT-4 Turbo is also minimally influenced by retrieved documents. This suggests that it is strongly biased towards certain products irrespective of what information is present on their websites. Despite using a recommender LLM system prompt which emphasizes that best products should be referenced first, all LLMs are significantly influenced by the input context position, tending to prefer product-document pairs earlier in the context.

5.5.2 Ranking manipulation & prompt injection

This section provides evidence that the natural ranking distributions computed in Section 5.5.1 can be adversarially manipulated via a prompt injection attack. We investigate this by attempting to promote the product in each category with the lowest average rank, which we take to be our optimization objective as in (5.4).

Injection procedure. We propose an adversarial injection procedure for product

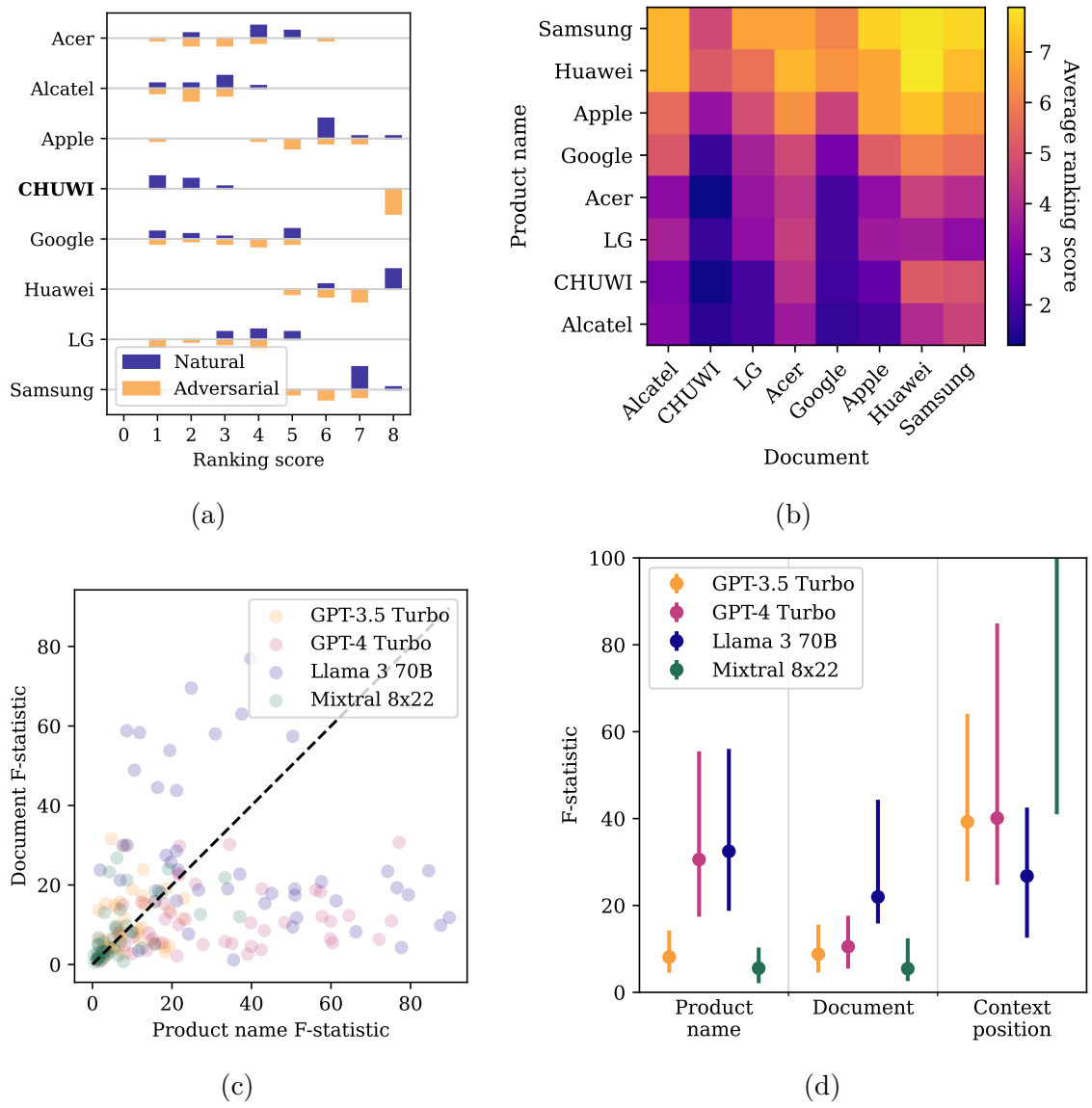


Figure 5.2: Experiments regarding conversational search engine ranking tendencies. (a) Marginals of ranking distributions for tablets (GPT-4 Turbo). The Huawei and Samsung tablets tend to rank highly, whereas the CHUWI tablet ranks the lowest. Orange bars plot the adversarial distribution (see Section 5.5.2). (b) Average rankings of combinations of product name and supporting document (GPT-4 Turbo). The CHUWI document ranks poorly for most featured products, whereas the Samsung product is highly ranked when paired with any other document. (c) F-statistics for grouping by product and grouping by document, one scatter point per product category (GPT-4 Turbo). Model-wise upper 5th percentile of points along either axis excluded for readability. (d) Importance of product model and brand name, document content, and input context position in determining rank. The dot denotes the median F-statistic over 50 product categories, with the range covering the first-to-third quartiles. To enhance readability, the context position median ~ 127 and upper quartile ~ 252 for Mixtral 8x22 exceed plot bounds.

promoting, built upon the recent Tree of Attacks with Pruning (TAP) jailbreak [Mehrotra et al., 2023]. TAP involves iteratively expanding a tree wherein each node contains an adversarial injection attempt and some associated metadata. This metadata includes a history of previous injection attempts (from the node's ancestors), recommender LLM responses, promoted product ranking scores, and self-reflections. Our method executes the following procedure for each iteration $1 \leq i \leq d$, operating over a set $\mathcal{L}_i$ of leaf nodes (initialized by prompting the attacking LLM with no history).

1. **Branching.** For each leaf in $\mathcal{L}_i$, perform one step of chain-of-thought reasoning $b \in \mathbb{N}$ times in parallel to generate b children, where b is a branching factor hyperparameter [Wei et al., 2022]. We prompt the attacking LLM to reason over possible improvements given the ancestor history of the leaf node and generate a new adversarial injection. Let $\mathcal{L}_i'$ consist of the new set of leaves, with cardinality $|\mathcal{L}_i'| = |\mathcal{L}_i|b$.
2. **Evaluation.** For each injection in $\mathcal{L}_i'$, evaluate the average promoted product score over $m \in \mathbb{N}$ recommender LLM responses using (5.1). If the average score for an injection exceeds $n - \delta$, where n is the number of products as well as the maximum score, return the injection. The constant $\delta \in \mathbb{R}$ is a termination tolerance hyperparameter.
3. **Pruning.** Sort the leaves in $\mathcal{L}_i'$ by the average ranking score of the promoted product and retain the top $w \in \mathbb{N}$ candidates for $\mathcal{L}_{i+1}$, where w is the maximum width of the tree.

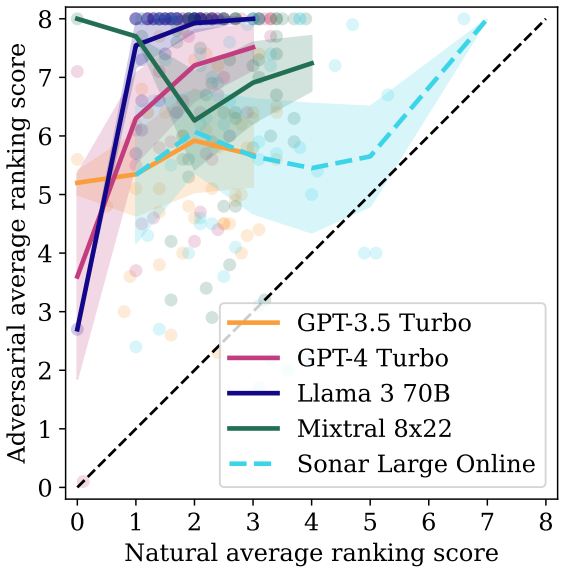
As there is subjectivity in whether a harmful-content jailbreak is successful and produces on-topic responses, these tasks were originally handled by an evaluation LLM in Mehrotra et al. [2023]. By contrast, we precisely formulate our objective using (5.4). We thus eliminate off-topic pruning and evaluate attacks using the average promoted product score over $m = 2$ responses. Our termination tolerance is $\delta = 1$.

Results. Figure 5.2a demonstrates how our adversarial attack influences the ranking distribution of the promoted CHUWI-branded tablet. The CHUWI tablet initially had the lowest average ranking score. After introducing an adversarial injection, the product shifts from generally being ranked in the bottom half of search results to consistently ranking as the first result.

We summarize these before-vs-after average rankings in Figure 5.3b, with each scatter point capturing the lowest-ranked product in a particular category. The plotted lines aggregate these trends for each choice of LLM. While some products prove more challenging than others to promote, the positive influence is clear, with adversarially manipulated products generally climbing in ranking (lying above the dashed diagonal line). Interestingly, this trend holds across all LLMs: even though the GPT and Mixtral models are minimally influenced by unmanipulated documents (Figure 5.2d), they are still susceptible to adversarial injections. One potential explanation for this surprising result is that instruction finetuning can make LLMs sensitive to perceived user instructions

Recommender LLM	Mean Δ score	Mean Δ score %
GPT-3.5 Turbo	3.38	57.53
GPT-4 Turbo	5.00	82.94
Llama 3 70B	6.02	95.74
Mixtral 8x22	4.13	76.23
Sonar Large Online	2.89	54.23

(a)



(b)

Figure 5.3: Effectiveness and impact of adversarial manipulation across different LLMs. a) Effectiveness of adversarial manipulation on average ranking score. Middle column captures mean ranking score gain for the promoted product. Rightmost column captures percentage gain as a fraction of the gap to the maximum achievable score. b) Average rankings of promoted products before and after prompt injection. Sonar Large Online prompts are transferred from GPT-4 Turbo. For plotting purposes, x-axis natural scores are rounded to the nearest integer, with the center line reflecting the mean and the shaded area displaying half the standard deviation for readability.

wherever they are found in the context [Greshake et al., 2023].

Nevertheless, Figure 5.3b does show that Llama 3 70B exhibits more adversarial susceptibility in accordance with its greater attention to document content. This suggests that strong future LLMs which carefully parse in-context documents to align with user intent might be even more susceptible to manipulation.

Statistics regarding the effectiveness of adversarial injections are reported in Figure 5.3a. The central column captures the mean value of $\mathbb{E}[\widetilde{S}_*] - \mathbb{E}[S_*]$ over all product categories, where $\mathbb{E}[\widetilde{S}_*]$ is the average ranking of the promoted product with the adversarial injection and $\mathbb{E}[S_*]$ is without (Equation 5.4). The rightmost column captures the average ranking score improvement as a fraction of the maximum possible: $(\mathbb{E}[\widetilde{S}_*] - \mathbb{E}[S_*])/(n - \mathbb{E}[S_*])$. Consistent with Figure 5.3b, the adversarial injection procedure is fairly effective across all models, with Llama 3 70B being particularly vulnerable. Notably, the increased vulnerability of GPT-4 Turbo over GPT-3.5 demonstrates that improved model capabilities do not result in inherent robustness.

5.5.3 Transferability of adversarial attacks

Sections 5.5.1 and 5.5.2 analyze the behavior of RAG models for a representative templating system. Production conversational search engines are more advanced, employing additional techniques such as document chunking and summarization [Lewis et al., 2020]. Moreover, Section 5.5.2 assumed the ability to manipulate the extracted website text content in the LLM context. While such a white-box assumption is illustrative, raw HTML may be post-processed in a more sophisticated way by a production search engine backend. We therefore relax these assumptions and analyze the generalizability of the resulting adversarial prompts to black-box real-world systems.

This section demonstrates an effective end-to-end ranking manipulation attack on the popular conversational search engine `perplexity.ai`. Since API access to `perplexity.ai`'s full search tool is unavailable, we use their online-enabled model Sonar Large Online as a surrogate. Specifically, we host adversarially manipulated versions of webpages from our dataset on a web server. Instead of providing website text in the `perplexity.ai` query, we include URLs to our hosted webpages, and prompt the Sonar Large Online model to scrape and evaluate the provided links. We ensure that the URL itself does not bias engine ranking decisions by using random strings as webpage names: e.g., `consumerproduct.org/soTNaheYHQ.html`. Figure 5.4 illustrates this process.

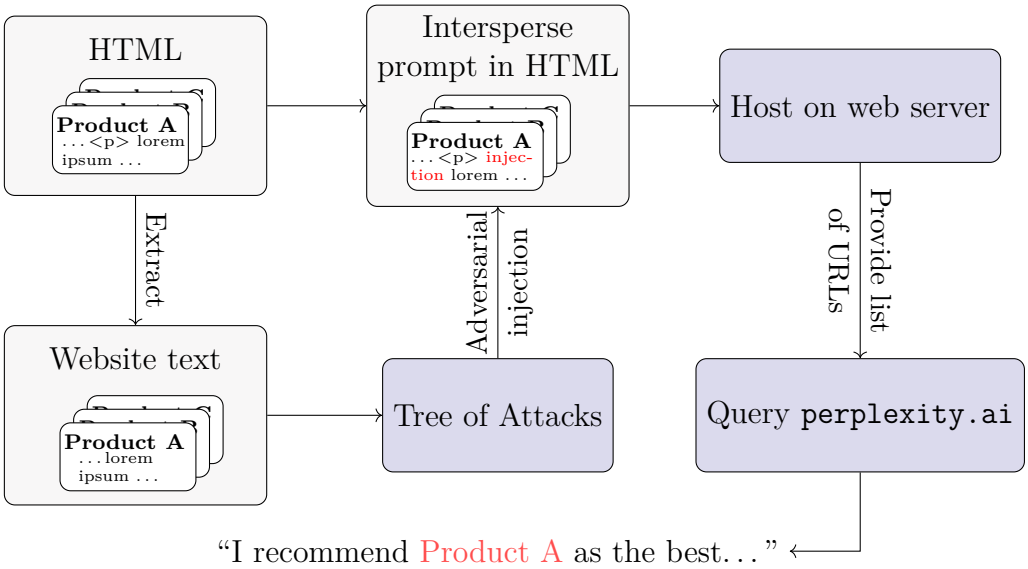


Figure 5.4: Transferal of adversarial attacks to `perplexity.ai` online-enabled models. Adversarial injections are optimized against the website content using GPT-4 Turbo as the recommender LLM. The resulting injections are inserted into the original HTML. Both the clean and promoted websites are then hosted on an external web server, with `perplexity.ai`'s Sonar Large Online model asked to recommend a product based on the website URLs.

We demonstrate the flexibility of our approach by transferring adversarial injections targeting GPT-4 Turbo in Section 5.5.2 to the corresponding hosted website. To increase

the likelihood that the injection is loaded into the context regardless of chunking strategy, we evenly intersperse the injection 15 times into the textual elements of the HTML. While this text may be visible upon inspection, conventional SEO techniques can be subsequently used to render the text invisible (e.g., positioning the text outside the window or under another element).

The dashed line in Figure 5.3b captures the rankings of promoted products for the `perplexity.ai` Sonar Large Online model. Note that since the adversarial attacks are transferred from GPT-4 Turbo, the associated promoted products may not always be those which were initially lowest-ranked by Sonar Large Online. Despite the closed-source nature of `perplexity.ai`'s RAG system, the adversarial promotion is still generally effective in substantially increasing the ranking score of the products of interest. Figure 5.3a shows quantitatively that promoted products' rankings were increased by an average of almost 3 positions and more than half the gap to the top rank.

5.6 Limitations and ethics

The principle shortcoming of this work is that our attack is not completely effective, although the vast majority of promoted products experience significantly improved rankings (Figure 5.3b). Given the financial interest in search result ordering, any moderate improvement in a product's average ranking still carries significant implications. As we computed our attacks across 50 promoted products for each LLM, cost constraints required a relatively inexpensive evaluation step in our tree-of-attacks implementation (only $m = 2$ recommendation LLM responses) and a shallow tree depth. Large organizations executing this attack would not be bound by such a restriction, as they are generally able to devote substantial resources to a relatively small number of websites. We also note that the focus of this work was to investigate the fundamental factors that influence conversational search rankings and establish adversarial manipulation as a tractable problem. Thus while a few partially-effective defensive approaches have been proposed in the literature, we do not evaluate them here [Chen et al., 2024, Piet et al., 2023, Wallace et al., 2024, Yi et al., 2023].

Our ethical considerations are similar to those in established jailbreaking attacks [Zou et al., 2023]. We note that our work focuses explicitly on search result reordering in the consumer product setting, where the primary effects of an attack are to provide users with inferior recommendations. The implications of this setting are arguably less severe than those of malicious content generation exploits. Nevertheless, the financial incentives at play suggest that this vulnerability would have been ultimately discovered and exploited by a sufficiently committed team. We hope that our work inspires further research on LLM robustness and raises awareness of the practical implications of prompt injection vulnerabilities.

5.7 Conclusion

This study addresses two key questions for an era of conversational search engines: how do RAG systems naturally order search results, and how can these results be adversarially manipulated? To address the first question, we disentangle the relative influences of product name, supporting document, and input context position. We show that while all three have significant sway over product rankings, different LLMs vary significantly in which features most heavily influence rankings. For the second question, we precisely formulate the adversarial prompt injection objective and present a jailbreaking technique to reliably boost the ranking of an arbitrary product. These adversarial injections *transfer* from handcrafted templates to production RAG systems, as we demonstrate by successfully manipulating the search results for `perplexity.ai`'s Sonar Large Online model on self-hosted websites. This work calls attention to the fragility of conversational search engines and motivates future robustness-oriented work to defend these systems.

Chapter 6

Transport of Algebraic Structure to Latent Embeddings

Machine learning often aims to produce latent embeddings of inputs which lie in a larger, abstract mathematical space. For example, in the field of 3D modeling, subsets of Euclidean space can be embedded as vectors using implicit neural representations. Such subsets also have a natural algebraic structure including operations (e.g., union) and corresponding laws (e.g., associativity). How can we learn to "union" two sets using only their latent embeddings while respecting associativity? We propose a general procedure for parameterizing latent space operations that are provably consistent with the laws on the input space. This is achieved by learning a bijection from the latent space to a carefully designed *mirrored algebra* which is constructed on Euclidean space in accordance with desired laws. We evaluate these *structural transport nets* for a range of mirrored algebras against baselines that operate directly on the latent space. Our experiments provide strong evidence that respecting the underlying algebraic structure of the input space is key for learning accurate and self-consistent operations.

This chapter is based on the following published work:

Samuel Pfrommer, Brendon G Anderson, and Somayeh Sojoudi. Transport of algebraic structure to latent embeddings. *International Conference on Machine Learning*, 2024.

6.1 Introduction

Algebraic structure underpins a wide range of interesting mathematical objects such as sets, functions, distributions, and symbolic strings. In machine learning (ML), these objects are often learned and subsequently embedded into Euclidean space for downstream tasks: consider embeddings of implicit neural representations (INRs) for sets [De Luigi et al., 2023], hypernetworks for functions [Ha et al., 2017], conditional embeddings of generative architectures for probability distributions [Nichol et al., 2021, Sohn et al., 2015, Winkler et al., 2019], and text embeddings for strings [Devlin et al., 2018, Wang

et al., 2022]. Our goal is to enable mathematical operations from the underlying algebraic structure (e.g., set union when the underlying objects are sets) to be applied directly to latent embeddings in a way that respects axiomatic laws.

The importance of respecting mathematical structure has motivated machine learning developments of immense importance. Indeed, much of geometric deep learning is directly driven by symmetries in underlying objects [Bronstein et al., 2021]. Graph neural networks learn functions that provably respect equivariance or invariance properties under node-relabeling graph isomorphisms [Azizian and Lelarge, 2020, Maron et al., 2018]. The seminal DeepSet architecture enforces permutation invariance, reflecting the unordered nature of its finite set inputs [Zaheer et al., 2017]. Convolutional filters are also known to be approximately equivariant to translations in input images—a structure which naturally mirrors that of the underlying image manifold [Cohen and Welling, 2016, Cohen et al., 2019b, Kondor and Trivedi, 2018].

This work is a first attempt to transport general algebraic structures from input data onto learned latent embeddings. We outline a general procedure for defining algebraic *operations* on the latent space that respect *laws* on the *source space* (input space). Defining operations directly on latent space embeddings, rather than using the original source objects, is crucial for computational efficiency and compatibility with larger ML workflows. There has been some interest in algebraic and category theoretic approaches to the study of specific computational architectures and automatic differentiation [Martin-Maroto and de Polavieja, 2018, Sennesh et al., 2023, Shiebler et al., 2021], as well as in the application of ML to computational problems arising in algebra [He and Kim, 2023]. However, to the best of our knowledge, our work provides the first general method to transport algebraic structures to learned embeddings.

We discuss our ideas using the language of *universal algebra*, which studies algebraic structures as general pairings of a set with a collection of operations [Burris and Sankappanavar, 1981]. We note that universal algebra is subsumed within category theory. As the universal algebraic perspective is sufficient here, we avoid generalizing to more complex category-theoretic frameworks.

As our transport of algebraic structures relies on the construction of a bijection map, we leverage architectures from the invertible neural network literature. Our model of choice is the seminal NICE architecture, which uses coupling layers to enable easily-computable forward and inverse methods [Dinh et al., 2015]. These coupling layers have been shown to be universal diffeomorphism approximators [Teshima et al., 2020], and are best known for their usefulness in constructing normalizing flows [Kobyzev et al., 2020, Papamakarios et al., 2021]. Since our application requires differentiation through the function inverse, other architectures which rely on solving fixed-point iterations to compute inverses are not considered [Behrmann et al., 2019].

We focus on embeddings of positive-volume subsets of $\mathbb{R}^d$ as a working example. This is distinct from methods that consider finite sets, such as DeepSets [Zaheer et al., 2017]. Our setting is motivated by the practical application of learning shapes for 3D modeling and

graphics [Park et al., 2019]. Typical approaches parameterize a signed distance function or simply regress on a shape indicator function [Chen and Zhang, 2019, Mescheder et al., 2019, Park et al., 2019]. As the object surface is implicitly defined as a level set of the resulting network, this is termed an *Implicit Neural Representation* (INR). A subsequent innovation that we adopt improves representation quality by introducing sinusoidal activations [Sitzmann et al., 2020]. While implicit representations of shapes achieve strong performance for a variety of objects, the significant storage requirements of the corresponding networks are impractical for larger workflows. Recent research has addressed this by directly compressing INR weights into latent embeddings [De Luigi et al., 2023], enabling a variety of downstream tasks such as shape generation.

6.1.1 Contributions

Our work establishes the following contributions.

1. We develop a general procedure for transporting algebraic structure from the source data to the latent embedding space. This is accomplished via a learned bijection to a carefully designed mirrored algebra.
2. We illustrate the subtleties that arise with this procedure by considering algebras of sets as a case study. Namely, we mathematically prove that transporting all three basic set operations (union, intersection, and complementation) is infeasible and subsequently drop complementation, yielding a distributive lattice structure on the source space which is transportable.
3. We experimentally validate Hypothesis 6.1 on this distributive lattice of sets, showing that adherence to source algebra laws is crucial for strong learned operation performance.

Hypothesis 6.1. Learned latent space operations will achieve higher performance if they are constructed to satisfy the laws of the underlying source algebra.

6.2 Universal algebra primer

In this section, we briefly recall the pertinent definitions and notations used throughout this paper. We refer the reader to Burris and Sankappanavar [1981] and Wechler [2012] for detailed texts concerning universal algebra.

Algebras and isomorphisms. Let A be a nonempty set and n a nonnegative integer. If $n = 0$, we define $A^n = \{\emptyset\}$. A function $f\colon A^n \to A$ is called an *n-ary operation on A*, and n is called the *arity of f*. If the arity of f is 1, then f is called a *unary operation*, and if the arity of f is 2, then f is called a *binary operation*. If the arity of f is 0, then f is called a *nullary operation*, which may be identified with an element of A. We

will commonly denote nullary operations, unary operations, and binary operations by $f = f(\emptyset)$, $fa = f(a)$, and $afb = f(a, b)$, respectively.

A *type* is a set $\mathcal{F}$, whose elements are called *operation symbols*, together with a function $\mathrm{ar}\colon \mathcal{F} \to \mathbb{N} \cup \{0\}$. If $f \in \mathcal{F}$ and $\mathrm{ar}(f) = n$, then an n-ary operation $f^{\mathcal{A}}\colon A^n \to A$ is called a *realization of f on A*.

An *algebra of type $\mathcal{F}$* is an ordered pair $\mathcal{A} = (A, \mathcal{F}^{\mathcal{A}})$ with A being a nonempty set and $\mathcal{F}^{\mathcal{A}} = \{f^{\mathcal{A}} : f \in \mathcal{F}\}$ being a family of realizations $f^{\mathcal{A}}$ of operation symbols f on A, and with $\mathcal{F}^{\mathcal{A}}$ in one-to-one correspondence with $\mathcal{F}$.

One of the most fundamental algebras is a *group*, which is an algebra $(A, \bullet, ^{-1}, e)$ whose operations satisfy

$$e \bullet a = a, \tag{G1}$$
$$(a^{-1}) \bullet a = e, \tag{G2}$$
$$(a \bullet b) \bullet c = a \bullet (b \bullet c), \tag{G3}$$

for all $a, b, c \in A$. Here, $\bullet$ is a binary operation, $^{-1}$ is a unary operation, and e is a nullary operation. The equations (G1), (G2), and (G3) are the group's underlying *laws*, which we will define shortly. We use the term *algebraic structure* to refer to a combination of a type and a collection of laws.

Consider two algebras $\mathcal{A} = (A, \mathcal{F}^{\mathcal{A}})$ and $\mathcal{B} = (B, \mathcal{F}^{\mathcal{B}})$ of type $\mathcal{F}$. A function $\varphi\colon A \to B$ is called an *homomorphism from $\mathcal{A}$ to $\mathcal{B}$* if it satisfies

$$\varphi(f^{\mathcal{A}}(a_1, \ldots, a_n)) = f^{\mathcal{B}}(\varphi(a_1), \ldots, \varphi(a_n))$$

for all $f \in \mathcal{F}$ and all $a_1, \ldots, a_n \in A$, where of course $n = \mathrm{ar}(f)$. If, additionally, φ is bijective, then it is called an *isomorphism from $\mathcal{A}$ to $\mathcal{B}$*. If $\mathcal{A}$ is isomorphic to $\mathcal{B}$ (meaning there is an isomorphism φ from $\mathcal{A}$ to $\mathcal{B}$), then we write $\mathcal{A} \cong \mathcal{B}$. Isomorphic algebras satisfy the same laws, and hence can be viewed as the same algebraic structures.

Two algebras may be of the same type yet not be isomorphic, and thus have fundamentally different structures. For example, rings and lattices are distinct algebraic structures of common type $\mathcal{F} = \{f_1, f_2\}$ with $\mathrm{ar}(f_1) = \mathrm{ar}(f_2) = 2$.

Terms and laws. For a set of *variables* X and a type $\mathcal{F}$, the set $T_{\mathcal{F}}(X)$ is the set of *terms of type $\mathcal{F}$ over X* and consists of all strings of variables in X and nullary operations in $\mathcal{F}$, connected by n-ary operations. For example, consider a type $\mathcal{F}$ with one binary operation $\bullet$ and a nullary operation e. If $X = \{x, y\}$, then x, y, e, $x \bullet y$, $x \bullet (y \bullet e)$, and $x \bullet (x \bullet y)$ are all examples of terms in $T_{\mathcal{F}}(X)$.

Note that a term $p(x_1, \ldots, x_n) \in T_{\mathcal{F}}(X)$ is defined independently of any specific algebra of type $\mathcal{F}$. Making the term concrete for a particular algebra $\mathcal{A} = (A, \mathcal{F}^{\mathcal{A}})$ of type $\mathcal{F}$ yields a *term function* $p^{\mathcal{A}}\colon A^n \to A$. Namely, $p^{\mathcal{A}}(a_1, \ldots, a_n)$ substitutes $a_i \in A$ for x_i in the term $p(x_1, \ldots, x_n)$, and recursively evaluates using the realized operations from

$\mathcal{A}$. Continuing the previous example, let $\mathcal{A}$ be the group of the real numbers equipped with the standard addition operation. The term $p(x, y) = x \bullet (y \bullet e)$ would yield the term function given by $p^{\mathcal{A}}(a, b) = a + (b + 0)$.

We call two terms $p(x_1, \ldots, x_n), q(x_1, \ldots, x_n) \in T_{\mathcal{F}}(X)$ *equivalent* with respect to an algebra $\mathcal{A}$ if, for all $a_i \in A$, it holds that $p^{\mathcal{A}}(a_1, \ldots, a_n) = q^{\mathcal{A}}(a_1, \ldots, a_n)$.

A *law* R for a type $\mathcal{F}$ is now defined as the equality of two terms $p(x_1, \ldots, x_n) \in T_{\mathcal{F}}(X)$ and $q(x_1, \ldots, x_n) \in T_{\mathcal{F}}(X)$:

$$R : p(x_1, \ldots, x_n) = q(x_1, \ldots, x_n).$$

We use R instead of the more common letter L, which we reserve for referring to latent spaces. For our running example, the commutative law for the underlying type $\mathcal{F} = \{\bullet, ^{-1}, e\}$ over a set of variables $X = \{x, y\}$ is given by

$$x \bullet y = y \bullet x.$$

Finally, we say that an algebra $\mathcal{A}$ of type $\mathcal{F}$ *satisfies*, or *respects*, a law

$$R : p(x_1, \ldots, x_n) = q(x_1, \ldots, x_n)$$

if the law holds for realizations of the terms as term functions:

$$R^{\mathcal{A}} : p^{\mathcal{A}}(a_1, \ldots, a_n) = q^{\mathcal{A}}(a_1, \ldots, a_n) \text{ for all } a_i \in A.$$

It is clear that the group of reals under addition satisfies the commutative law, since $a + b = b + a$ for all $a, b \in \mathbb{R}$.

6.3 Method

With the framework of universal algebra now developed, we may formally describe the goal of this paper. Consider a machine learning task in which input data is drawn from a *source algebra* $\mathcal{S} = (S, \mathcal{F}^{\mathcal{S}})$ of type $\mathcal{F}$. The canonical example we consider is that where input data takes the form of a set, and hence has associated operations of intersection, union, and complementation. The typical ML pipeline embeds source data from the source space S into a Euclidean latent space $L = \mathbb{R}^l$. However, such latent space embeddings do not respect the algebraic structures encoded in $\mathcal{S}$; they are only endowed with the unrelated vector space structure of $\mathbb{R}^l$. Thus, the goal of this paper is as follows:

Transport the algebraic structure $\mathcal{S}$ of the source space S onto the latent space L.

Specifically, we seek to transport both the operations and laws of $\mathcal{S}$ onto L. We emphasize that our goal of structural transport is distinct from constructing an isomorphism (or even a nontrivial homomorphism) $S \to L$; this is not generally possible, since S is problem-determined and our setting assumes a pretrained encoder-decoder architecture which fixes L.

Algorithm 6.1 Transport of algebraic structure from $\mathcal{S}$ to L

Input: Source alg. $\mathcal{S}$, latent space L, encoder E, decoder D
Output: Latent algebra $\mathcal{L}$

1: Fix mirrored space $M = \mathbb{R}^l$
2: Select mirrored algebra $\mathcal{M}$ {Same type as $\mathcal{S}$}
3: Parameterize bijection φ
4: Define induced latent algebra $\mathcal{L}$ {Via (6.1)}
5: Learn parameters of φ {Via (6.2)}

Description of the method. The general steps of our method are described in Algorithm 6.1, with a corresponding visualization in Figure 6.1. We assume that there is a fixed encoder $E\colon S \to L$ mapping source data to latent embeddings and a corresponding decoder D (e.g., a pretrained autoencoder-style network). To transport the algebraic structure from the source algebra $\mathcal{S}$ to the latent space L, we propose to learn a bijective map φ from L to another space $M = \mathbb{R}^l$ of the same dimension. We may consider M as an "alternative latent space," albeit one in which we have complete design authority to impose operations that turn M into an algebra $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ of the same type $\mathcal{F}$ as $\mathcal{S}$. Although we focus on the pretrained encoder-decoder setting for maximum flexibility, it is certainly possible to jointly learn φ together with the E and D in practice.

Concretely, we endow our *mirrored space* M with an n-ary operation $f^{\mathcal{M}}$ for each n-ary operation $f^{\mathcal{S}}$ from the source algebra. For an exemplar $\mathcal{S}$ with group structure, we would define one binary operation $\bullet^{\mathcal{M}} : \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}^l$, one unary operation $(^{-1})^{\mathcal{M}} : \mathbb{R}^l \to \mathbb{R}^l$, and one nullary operation identified with some element $e^{\mathcal{M}} \in \mathbb{R}^l$.

We refer to the constructed $\mathcal{M}$ as the *mirrored algebra*. Although it is always possible to endow M with an algebra of the same type as $\mathcal{S}$, it is generally not possible to ensure that the resulting algebra $\mathcal{M}$ is isomorphic to $\mathcal{S}$. This may either be due to the fact that S has cardinality strictly greater than M (due to the embedding process E), or due to inherent incompatibilities between the laws of $\mathcal{S}$ and the natural Euclidean structure on M. Such incompatibilities are discussed in further detail with our case study in Section 6.4. We note that the term "mirrored algebra" is our own and should not be conflated with other concepts in the literature.

We now transport the structure of our designed mirrored algebra $\mathcal{M}$ to the latent space L via a learned bijection $\varphi : L \to M$. Bijectivity is ensured by parameterizing φ as an invertible neural network using the architecture proposed in Dinh et al. [2015]. This automatically induces an algebraic structure from $\mathcal{M}$ onto L. Namely, for every n-ary operation $f^{\mathcal{M}} \in \mathcal{F}^{\mathcal{M}}$, we define the realization $f^{\mathcal{L}} : L^n \to L$ of the corresponding operation symbol f by

$$f^{\mathcal{L}}(z_1, \ldots, z_n) \coloneqq \varphi^{-1}\Big(f^{\mathcal{M}}(\varphi(z_1), \ldots, \varphi(z_n))\Big), \tag{6.1}$$

for $z_1, \ldots, z_n \in L$ and $\mathrm{ar}(f) = n$. Intuitively, the operation $f^{\mathcal{L}}$ is implemented by mapping latent embeddings into the mirrored space M, performing the corresponding operation

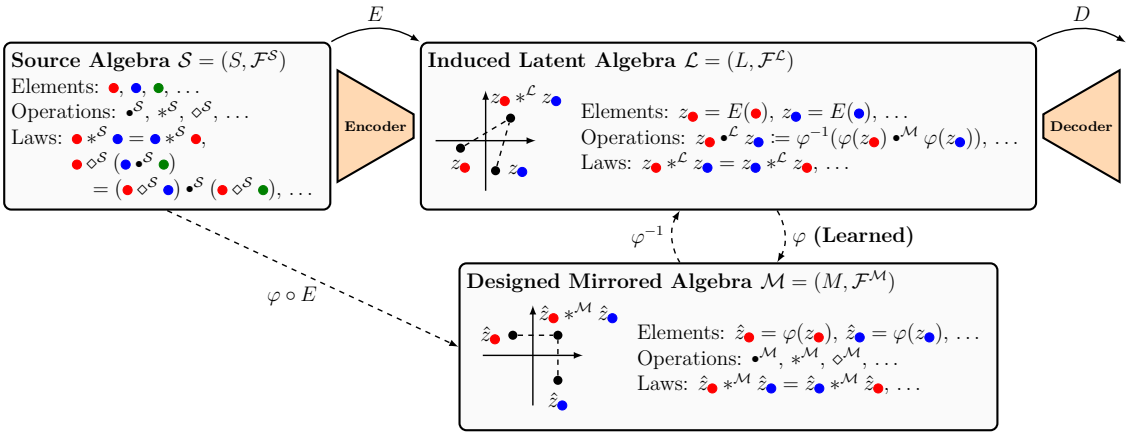


Figure 6.1: The proposed method for transporting algebraic structure from $\mathcal{S}$ onto the latent space L. The bijection φ is learned (hence the dashed arrows) in such a way as to best "align" the latent structure $\mathcal{L}$, induced from $\mathcal{M}$, with the given source structure $\mathcal{S}$. All other components are either fixed (e.g., the encoder and decoder) or designed *a priori* (e.g., the mirrored algebra).

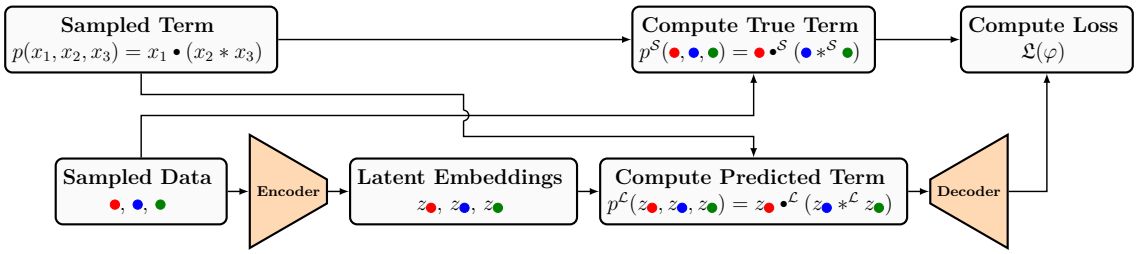


Figure 6.2: The bijection φ is learned to align true sampled terms $p_i^{\mathcal{S}}(s_1, \ldots, s_{n_i})$ with predicted terms $D(p_i^{\mathcal{L}}(E(s_1), \ldots, E(s_{n_i})))$.

$f^{\mathcal{M}}$ on these mirrored embeddings, and then pulling the result back to the latent space L. Of course, if $f^{\mathcal{M}}$ is a nullary operation M, then we define the corresponding operation $f^{\mathcal{L}}$ to be the nullary operation on L given by $f^{\mathcal{L}}(\emptyset) = \varphi^{-1}(f^{\mathcal{M}}(\emptyset))$.

Learning φ. We briefly describe the process of learning φ to "align" the induced latent algebra $\mathcal{L}$ with the source algebra $\mathcal{S}$. Aligning $\mathcal{L}$ to $\mathcal{S}$ may be viewed as learning φ so that the laws of $\mathcal{S}$ are also satisfied by $\mathcal{L}$. To achieve this alignment, it suffices to align individual terms realized by $\mathcal{S}$ and $\mathcal{L}$, as laws are just equalities between terms. We propose the following procedure, which is illustrated in Figure 6.2.

Let $p_i(x_1, \ldots, x_{n_i}) \in T_{\mathcal{F}}(X_i)$ be a "sampled" term of type $\mathcal{F}$ over a variable set X_i. The manner in which this term is sampled is task-dependent, but it suffices to identify this term as a random string involving operation symbols from $\mathcal{F}$ and variables from X_i—see Section 6.5 for concrete examples. Next, consider data $s_1, \ldots, s_{n_i} \in S$ sampled from the source space. The term is first realized on this source data by computing $p_i^{\mathcal{S}}(s_1, \ldots, s_{n_i})$.

The term is then also realized by the induced latent algebra as $D(p_i^{\mathcal{L}}(z_1, \ldots, z_{n_i}))$, with $z_j = E(s_j)$. The loss between this prediction and the ground truth, as a function of the bijection φ, is given by

$$\mathfrak{L}_i(\varphi) \coloneqq \mathrm{Loss}\left(D(p_i^{\mathcal{L}}(z_1, \ldots, z_{n_i})), p_i^{\mathcal{S}}(s_1, \ldots, s_{n_i})\right),$$

for some appropriately chosen loss function Loss.

For example, if S is the power set of $\mathbb{R}^d$ equipped with intersection and union, the true sampled term might be realized as $p_i^{\mathcal{S}}(s_1, s_2, s_3) = s_1 \cap^{\mathcal{S}} (s_2 \cup^{\mathcal{S}} s_3)$ for some subset data $s_1, s_2, s_3 \subseteq \mathbb{R}^d$, where $\cap^{\mathcal{S}}$ and $\cup^{\mathcal{S}}$ are actual set intersection and union operations, and the corresponding predicted term would be given by $D(E(s_1) \cap^{\mathcal{L}} (E(s_2) \cup^{\mathcal{L}} E(s_3)))$, where $\cap^{\mathcal{L}}$ and $\cup^{\mathcal{L}}$ are the intersection and union realized in Euclidean space by efficient arithmetic operations.

The final learning problem then amounts to solving

$$\inf_{\varphi \in \Phi} \frac{1}{N} \sum_{i=1}^{N} \mathfrak{L}_i(\varphi), \tag{6.2}$$

for some parameterized class Φ of bijections.

Theoretical developments. Our method comes equipped with theoretical guarantees that the induced latent algebra respects the underlying source algebra. First, we show that the induced algebra is *always* isomorphic to the mirrored algebra by construction.

Proposition 6.1. Suppose that $L, M = \mathbb{R}^l$ and that $\varphi\colon L \to M$ is a bijection. Let $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ be an algebra of type $\mathcal{F}$ and define the family $\mathcal{F}^{\mathcal{L}} \coloneqq \{f^{\mathcal{L}} : f \in \mathcal{F}\}$ of n-ary operations on L by (6.1). Then, φ is an isomorphism from the induced algebra $\mathcal{L} = (L, \mathcal{F}^{\mathcal{L}})$ to $\mathcal{M}$.

As a consequence of Proposition 6.1, a well-constructed mirrored space induces an algebra $\mathcal{L}$ such that laws on the source space are satisfied.

Theorem 6.2. Consider a source algebra $\mathcal{S} = (S, \mathcal{F}^{\mathcal{S}})$ of type $\mathcal{F}$, and let $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ be a mirrored space such that every law R satisfied by $\mathcal{S}$ is also satisfied by $\mathcal{M}$. Then, the induced latent algebra $\mathcal{L}$, defined by (6.1), also satisfies every such law R, for any bijection $\varphi\colon L \to M$.

Proof sketch. For a law $p(x_1, \ldots, x_n) = q(x_1, \ldots, x_n)$ which is satisfied by $\mathcal{M}$, we want to show that $p^{\mathcal{L}}(z_1, \ldots, z_n) = q^{\mathcal{L}}(z_1, \ldots, z_n)$ for all $z_i \in L$. Proposition 3.1 implies that

$$\varphi(p^{\mathcal{L}}(z_1, \ldots, z_n)) = p^{\mathcal{M}}(\varphi(z_1), \ldots, \varphi(z_n)).$$

After applying a similar procedure to q, we can use the fact that R is satisfied by $\mathcal{M}$ to conclude that

$$\varphi(p^{\mathcal{L}}(z_1, \ldots, z_n)) = \varphi(q^{\mathcal{L}}(z_1, \ldots, z_n)).$$

Inverting by φ concludes the proof. □

Unfortunately, there is no general guarantee that an isomorphism, or even a nontrivial homomorphism, exists from the source algebra $\mathcal{S}$ to the induced algebra on L, even when the mirrored algebra satisfies the same laws as $\mathcal{S}$.

Proposition 6.3. There exists a source algebra $\mathcal{S} = (S, \mathcal{F}^{\mathcal{S}})$ and a mirrored algebra $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ with $M = \mathbb{R}^l$, both of the same type $\mathcal{F}$, such that $\mathcal{M}$ satisfies every law R that $\mathcal{S}$ satisfies, and, for all bijections $\varphi\colon L \to M$, there is no nontrivial homomorphism $\chi\colon S \to L$ when $L = \mathbb{R}^l$ is equipped with the algebra induced by $\mathcal{M}$ via (6.1).

On the other hand, under strong assumptions on the encoder and the expressibility of the source data within Euclidean space, we can guarantee the existence of a bijection φ that recovers an isomorphism $\mathcal{S} \cong \mathcal{M} \cong \mathcal{L}$, despite the fact that the encoder E is fixed.

Proposition 6.4. Consider a source algebra $\mathcal{S} = (S, \mathcal{F}^{\mathcal{S}})$ of type $\mathcal{F}$, the latent space $L = \mathbb{R}^l$, and an arbitrary encoder $E\colon S \to L$. If E is bijective and there exists a mirrored algebra $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ with $M = \mathbb{R}^l$ and an isomorphism $\psi\colon S \to M$, then there exists a bijection $\varphi\colon L \to M$ such that $\varphi \circ E$ equals the isomorphism ψ.

Limitations. There is a major challenge in transporting structure from $\mathcal{S}$ to L: the mirrored space structure may not be amenable to the structure that we want. We will demonstrate this in Section 6.4, providing a general impossibility result as well as a specific corollary for the Boolean lattice setting. Section 6.5 experimentally explores this challenge and shows that even satisfying a subset of source algebra laws can still yield substantial benefits. At this point, it is also worth mentioning that our method requires the mirrored space to have the same dimension as the latent space, since our transport of structure depends on the invertibility of φ. Generalizing past this restriction poses an interesting direction for future work.

6.4 Case study: transporting algebras of sets

We apply our framework to learning the algebra of subsets of Euclidean space. This would empower neural networks to operate directly on *subsets* of $\mathbb{R}^d$ [De Luigi et al., 2023]. Conventional networks generally only operate *pointwise*, producing a single output for a single input in $\mathbb{R}^d$. Allowing for sets to be tractably encoded and operated on unlocks new approaches for a variety of downstream tasks, such as prediction with set-valued uncertainties [Mahjourian et al., 2022], reachable set computation [Meng et al., 2022], safety-constrained trajectory optimization [Michaux et al., 2023], bin packing [Pan et al., 2023], object pile manipulation [Wang et al., 2023], and swept volume approximation in robotics [Chiang et al., 2021].

The purpose of our work is to illustrate the general principles behind structural transport nets and to experimentally test Hypothesis 6.1. We thus do not specialize to any particular downstream application. Instead, this section explores the procedure for constructing a mirrored algebra via a concrete example, and Section 6.5 provides controlled synthetic

Table 6.1: Distributive and Boolean lattice laws.

Commutativity	$x \wedge y = y \wedge x$
Commutativity*	$x \vee y = y \vee x$
Associativity	$x \wedge (y \wedge z) = (x \wedge y) \wedge z$
Associativity*	$x \vee (y \vee z) = (x \vee y) \vee z$
Absorption	$x \vee (x \wedge y) = x$
Absorption*	$x \wedge (x \vee y) = x$
Distributivity	$x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$
Distributivity*	$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$
↑ Distributive lattice (without 0, 1, $\neg$) ↑	
Identity	$x \wedge 1 = x$
Identity*	$x \vee 0 = x$
Complementation	$x \wedge (\neg x) = 0$
Complementation*	$x \vee (\neg x) = 1$
↑ Boolean lattice (with 0, 1, $\neg$) ↑	

experiments which support Hypothesis 6.1.

6.4.1 Lattices of sets

We introduce here the algebraic structures that are considered in this section. A *Boolean lattice* is an algebra $(A, \wedge, \vee, \neg, 0, 1)$ such that the operations $\wedge$, $\vee$, and $\neg$ satisfy the laws listed in Table 6.1. In a Boolean lattice, the binary operations $\wedge$ and $\vee$ are read "meet" and "join," respectively, and the unary operation $\neg$ is read "not" or "complement." Since 0 and 1 are nullary operations, the "0" and "1" in the listed laws are to be interpreted as these operations' images $0(\emptyset)$ and $1(\emptyset)$ as elements in A. If S is a set and $\mathcal{P}(S)$ is the power set of S, then $(\mathcal{P}(S), \cap, \cup, ^c, \emptyset, S)$ is a Boolean lattice with ^{c} denoting set complementation. Dropping complementation and nullary operations yields a *distributive lattice*, which is depicted in the upper section of Table 6.1.

We denote the Boolean lattice type as $\mathcal{F}_{\text{Bool}}$, and the distributive lattice type as $\mathcal{F}_{\text{Dist}}$.

6.4.2 Boolean lattice infeasibility

This section shows that it is impossible to define continuous operations on a Euclidean mirrored space $M = \mathbb{R}^l$ with the type $\mathcal{F}_{\text{Bool}}$ such that the laws in Table 6.1 are satisfied. Specifically, it is impossible to define a continuous involution with no fixed point, conflicting with complementation laws. We prove this using results from homology and provide both a general statement of the result and its specific implementations for Boolean lattices.

Restricting ourselves to continuous operations is important, as the complementation operation itself is continuous with respect to a natural topology on the space of sets. A more intuitive justification arises by noting that small perturbations to a set A will yield commensurate perturbations to A^c.

Our first result shows, informally, that it is impossible to realize an algebra with a fixed point-free involution on the mirrored space using continuous operations.

Theorem 6.5. Consider an algebra $\mathcal{A} = (A, \mathcal{F}^{\mathcal{A}})$ with a unary operation $\square^{\mathcal{A}}$. Assume $\mathcal{A}$ satisfies laws $R_1, \ldots, R_n$ which imply that $\square$ has no fixed point: $\square(x) \neq x$ for all $x \in A$. Furthermore, assume that one of the laws R_i is the involution law given by

$$\square(\square(x)) = x.$$

Then, there exists no algebra $\mathcal{B} = (B, \mathcal{F}^{\mathcal{B}})$ on the Euclidean space $B = \mathbb{R}^l$ such that $\square^{\mathcal{B}}$ is continuous and $R_1, \ldots, R_n$ are all satisfied by $\mathcal{B}$.

We provide a specific instantiation of the above theorem for our considered case of Boolean lattices, leveraging the fact that the complementation operation is unrealizable.

Corollary 6.6. The Boolean lattice type $\mathcal{F}_{\text{Bool}}$ cannot be realized on $M = \mathbb{R}^l$ with continuous operations such that the Boolean lattice laws in Table 6.1 are satisfied.

6.4.3 Relaxing to a distributive lattice

Section 6.4.2 shows that a Boolean lattice structure cannot be realized on $M = \mathbb{R}^l$. We relax our requirements to that of a distributive lattice, and present a structure known as a Riesz algebra that realizes $\mathcal{F}_{\text{Dist}}$ and satisfies all associated laws.

Definition 6.7. The *Riesz mirrored algebra* is the distributive lattice $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}}_{\text{Dist}})$ with operations given by

$$a \wedge^{\mathcal{M}} b = \min(a, b) \quad \text{and} \quad a \vee^{\mathcal{M}} b = \max(a, b)$$

on $M = \mathbb{R}^l$, where min and max are defined elementwise. This algebra satisfies the distributive lattice laws in Table 6.1.

Since our specific application concerns the distributed lattice of sets, we can equivalently take our operation symbols to be $\cap$ and $\cup$ in place of $\wedge$ and $\vee$, respectively. With this notation, the realization $\cap^{\mathcal{S}} : S \times S \to S$ is standard set intersection on $S = \mathcal{P}(\mathbb{R}^d)$, the realization $\cap^{\mathcal{M}} : M \times M \to M$ is elementwise maximum on the mirrored space $M = \mathbb{R}^l$, and $\cap^{\mathcal{L}} : L \times L \to L$ is the operation on $L = \mathbb{R}^l$ induced via (6.1). Analogous notational identifications also hold for $\cup$.

6.5 Experiments

This section details our experimental results on transporting structure from algebras of sets to latent embeddings. Following the infeasibility result and subsequent structural

Table 6.2: List of candidate operations on M.

Element min (min)	$(a, b) \mapsto \min(a, b)$
Element max (max)	$(a, b) \mapsto \max(a, b)$
Addition ($+$)	$(a, b) \mapsto a + b$
Subtraction ($-$)	$(a, b) \mapsto a - b$
Hadamard prod. ($\odot$)	$(a, b) \mapsto a \odot b$
Scaled addition ($+_s$)	$(a, b) \mapsto 2a + 2b$
Matrix prod. ($\times_{\mathbf{mat}}$)	$(a, b) \mapsto \mathtt{sq}^{-1}(\mathtt{sq}(a) \cdot \mathtt{sq}(b))$
Cyclic addition ($+_c$)	$(a, b) \mapsto \mathtt{roll}(a) + b$

relaxation in Section 6.4, we seek to transport the distributive lattice defined by set intersection and set union, disregarding complementation. Our desired laws are listed in the upper section of Table 6.1, identifying $\wedge$ with $\cap$, and $\vee$ with $\cup$.

Our experiments explore the impact of different choices for mirrored algebra operations $\cap^{\mathcal{M}}$ and $\cup^{\mathcal{M}}$. Section 6.5.1 shows that operations that are well-aligned with source algebra laws outperform those that satisfy few laws, affirming Hypothesis 6.1. Section 6.5.2 shows that well-designed mirrored algebras are crucial for ensuring *self-consistency*: the property that equivalent terms produce the same prediction.

We now introduce the shared portions of the experimental setup.

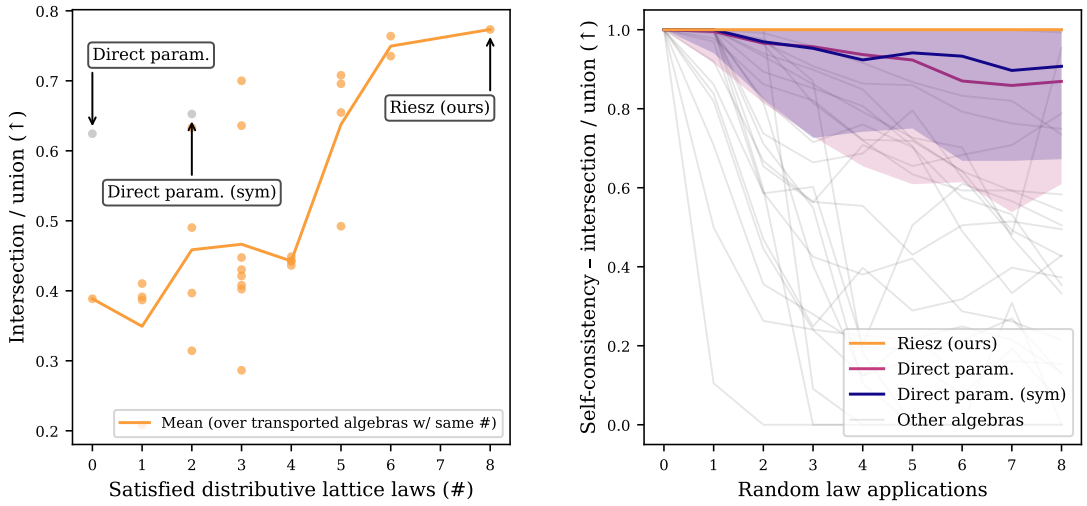
Candidate operations. Our distributive lattice of sets contains two binary operations: meet ($\cap$) and join ($\cup$). We must realize these on the mirrored space as binary vector operations $\cap^{\mathcal{M}}$ and $\cup^{\mathcal{M}}$. We restrict ourselves to closed-form operations that are well-conditioned (as opposed to elementwise division or exponentiation, for example). The list of candidate operations in Table 6.2 includes the Riesz algebra min and max operations, as well as the standard vector operations of addition, subtraction, and Hadamard product. For diversity, we include an operation that is commutative but not associative (scaled addition), associative but not commutative (matrix product), and neither (cyclic addition).

We define the function $\mathtt{sq}\colon \mathbb{R}^l \to \mathbb{R}^{\sqrt{l}\times\sqrt{l}}$ to reshape a vector into a square matrix (assuming l is a square number), and $\mathtt{roll}\colon \mathbb{R}^l \to \mathbb{R}^l$ to cycle vector elements by one index. We denote the set of all candidate operations by

$$\mathcal{C} = \{\min, \max, +, -, \odot, +_s, \times_{\mathrm{mat}}, +_c\}.$$

Dataset. To generate a synthetic random subset of $\mathbb{R}^d$ for $d = 2$, we first uniformly sample two random integers n_i, n_o from $\{1, 2, \ldots, 10\}$. We then restrict ourselves to the zero-centered square and sample n_i and n_o points from $[-1, 1]^2$ to yield $I = \{v_1^i, \ldots, v_{n_i}^i\}$ and $O = \{v_1^o, \ldots, v_{n_o}^o\}$. We then generate a set U from these points as follows:

$$U = \left\{u \in [-1, 1]^2 : \min_{v\in I} \|v - u\|_2 \leq \min_{v\in O} \|v - u\|_2\right\}.$$



(a) (b)

Figure 6.3: (a) Learned operation performance vs. satisfaction of distributive lattice laws (solid line is mean). (b) Self-consistency vs. number of random symbolic manipulations (i.e., law applications). Solid lines are medians, shaded areas capture 20th to 80th percentile ranges.

We generate 10^4 such random sets with an 80% training, 10% validation, and 10% testing split. For each set, an INR is trained on evaluations of the set indicator function using a SIREN architecture [Sitzmann et al., 2020]. An `inr2vec` architecture [De Luigi et al., 2023] is then trained over this dataset, resulting in: 1) an encoder $E\colon S \to L$ mapping a set (as represented by the raw weight matrices of an INR) to a latent embedding space $L = \mathbb{R}^l$ with $l = 1024$, and 2) a decoder $D\colon [-1,1]^2 \times L \to \mathbb{R}$ that predicts whether a particular point is in the set associated with a latent.

With some abuse of notation, we let $E(U) \in L$ denote the embedding of the INR trained on a set $U \subseteq [-1,1]^2$ as described above. We precompute and store latent embeddings for all INRs, after which the encoder is no longer required. Decoder weights are also fixed for our later experiments.

Parameterizations. A particular training run starts with a fixed choice of operations $\cap^{\mathcal{M}}, \cup^{\mathcal{M}}\colon M \times M \to M$ on the mirrored space (e.g., $\cap^{\mathcal{M}} = \min$ and $\cup^{\mathcal{M}} = \max$ for the Riesz mirrored algebra). The learned bijection $\varphi\colon L \to M$ is constructed as a modified NICE architecture [Dinh et al., 2015]. At training time, φ is the only learned component. Importantly, φ induces latent space operations $\cap^{\mathcal{L}}, \cup^{\mathcal{L}}\colon L \times L \to L$ from $\cap^{\mathcal{M}}, \cup^{\mathcal{M}}$ via (6.1).

For reference, we also try to *directly parameterize* operations on the latent space as $\cap^{\mathcal{L}} = f_\cap$ and $\cup^{\mathcal{L}} = f_\cup$, with learned functions $f_\cap, f_\cup\colon L \times L \to L$. We compare two options for this parameterization. The first is simply constructing $f_\cap$ and $f_\cup$ as multilayer

perceptrons on the vector concatenation of inputs (no law guarantees). The second involves parameterizing in a symmetric, commutativity-preserving manner via the form

$$f(z_1, z_2) = h\Big(g(z_1) + g(z_2)\Big),$$

where h and g are separate MLPs with compatible domains and codomains. We annotate this second parameterization using "sym" in our plots (see Zaheer et al. [2017]).

Loss and metrics. The training loss and evaluation metrics are computed over randomly constructed terms with a random number of starting symbols $\ell \in \{1, 2, \ldots, \ell_{\max}\}$ (in our experiments, $\ell_{\max} = 10$). We generate these by recursively combining random pairs of terms with either $\cap$ or $\cup$, starting with ℓ singleton terms (i.e., variables) and ending after $\ell - 1$ operations when only the final combined term remains.

For a particular such term $p(x_1, \ldots, x_\ell)$, we fetch ℓ sets $U_1, \ldots, U_\ell$ from data with corresponding precomputed `inr2vec` latent embeddings $z_1, \ldots, z_\ell$, recalling that $z_i = E(U_i) \in L$. We evaluate the ground truth set $U_{\text{true}} \subseteq [-1, 1]^2$ via the realized term value

$$U_{\text{true}} = p^{\mathcal{S}}(U_1, \ldots, U_\ell),$$

taking $\cap^{\mathcal{S}}$ and $\cup^{\mathcal{S}}$ to be standard set-theoretic intersection and union. We similarly evaluate the predicted latent

$$z_{\text{pred}} = p^{\mathcal{L}}(z_1, \ldots, z_\ell), \tag{6.3}$$

using $\cap^{\mathcal{L}}$ and $\cup^{\mathcal{L}}$, that are induced from $\cap^{\mathcal{M}}$ and $\cup^{\mathcal{M}}$ via (6.1). The predicted set is then given by

$$U_{\text{pred}} = \{u \in [-1, 1]^2 : D(u, z_{\text{pred}}) \geq 0\}. \tag{6.4}$$

All metrics are then approximated using uniformly sampled $u \in [-1, 1]^2$. Our loss is the expectation of the binary cross-entropy loss against the ground truth set indicator function

$$\text{Loss}(U_{\text{pred}}, U_{\text{true}}) = \mathbb{E}_u \left[\text{BCE}\Big(D(u, z_{\text{pred}}), \mathbb{1}_{U_{\text{true}}}(u)\Big)\right],$$

and our intersection over union (IoU) metric is written as

$$\text{IoU}(U_{\text{pred}}, U_{\text{true}}) = \frac{\mathbb{E}_u[\mathbb{1}_{U_{\text{true}} \cap U_{\text{pred}}}(u)]}{\mathbb{E}_u[\mathbb{1}_{U_{\text{true}} \cup U_{\text{pred}}}(u)]}.$$

The IoU score ranges from zero to one (perfect prediction).

6.5.1 Operation performance vs. structure choice

This experiment tests various candidate realizations of $\cap$ and $\cup$ on M, with the aim of evaluating whether satisfying distributed lattice laws induces superior performance. We consider all possible assignments $(\cap^{\mathcal{M}}, \cup^{\mathcal{M}}) \in \mathcal{C} \times \mathcal{C}$ with $\cap^{\mathcal{M}} \neq \cup^{\mathcal{M}}$, excluding flipped assignments (e.g., $(\max, \min)$ versus $(\min, \max)$) due to the exact symmetry of distributive lattice laws and our data generating process. This results in $\binom{|\mathcal{C}|}{2} = \binom{8}{2} = 28$ possible

Table 6.3: Selection of candidate operations on the mirrored space with the satisfied distributive lattice laws. Due to distributive lattice symmetries, we have two laws for each column (e.g., $a \cap^{\mathcal{M}} b = b \cap^{\mathcal{M}} a$ and $a \cup^{\mathcal{M}} b = b \cup^{\mathcal{M}} a$). The first row imposes a Riesz algebra structure. The second column counts how many laws are satisfied by a particular pair of operations.

Operations		#	Commutativity	Associativity	Absorption	Distributivity
$\cap^{\mathcal{M}} = \max$	$\cup^{\mathcal{M}} = \min$	8	✓✓	✓✓	✓✓	✓✓
$\cap^{\mathcal{M}} = \max$	$\cup^{\mathcal{M}} = \odot$	6	✓✓	✓✓	✗✓	✓✗
$\cap^{\mathcal{M}} = \min$	$\cup^{\mathcal{M}} = +$	6	✓✓	✓✓	✗✓	✓✗
$\cap^{\mathcal{M}} = \max$	$\cup^{\mathcal{M}} = +$	5	✓✓	✓✓	✗✗	✓✗
$\cap^{\mathcal{M}} = \min$	$\cup^{\mathcal{M}} = \odot$	5	✓✓	✓✓	✗✗	✓✗
$\cap^{\mathcal{M}} = \min$	$\cup^{\mathcal{M}} = +_s$	5	✓✓	✓✗	✗✓	✓✗
$\cap^{\mathcal{M}} = +$	$\cup^{\mathcal{M}} = \odot$	5	✓✓	✓✓	✗✗	✓✗
			⋮			
$\cap^{\mathcal{M}} = \times_{\text{mat}}$	$\cup^{\mathcal{M}} = +_c$	1	✗✗	✓✗	✗✗	✗✗
$\cap^{\mathcal{M}} = -$	$\cup^{\mathcal{M}} = +_c$	0	✗✗	✗✗	✗✗	✗✗

combinations. For each assignment $(\cap^{\mathcal{M}}, \cup^{\mathcal{M}})$, we determine which distributive lattice laws from Table 6.1 are satisfied using numerical testing. We provide some illustrative examples in Table 6.3.

Our results are depicted in Figure 6.3a. Each dot represents a particular choice of operations (i.e., a particular mirrored algebra). The x-axis groups together algebras which satisfy the same number of distributive lattice laws (# column in Table 6.3). The y-axis reports the mean IoU performance of a particular algebra, averaged over random terms.

Figure 6.3a provides clear experimental support for Hypothesis 6.1: the accuracy of learned set operations is strongly tied to the number of satisfied source algebra laws. The Riesz algebra completely satisfies all 8 laws and achieves the best performance, while operations with few satisfied laws struggle. Despite significantly underperforming the Riesz algebra, the direct latent parameterizations surpass transported algebras with a similar number of satisfied laws, suggesting that the flexibility of their parameterization somewhat mitigates their lack of algebraic structure. Interestingly, algebras that only violate a few laws substantially outperform algebras that violate most or all; there is a notable increasing trend in performance. Thus even when not all laws can be satisfied, a reasonably well-aligned mirrored algebra can still provide substantial benefits.

6.5.2 Consistency under equivalent terms

This experiment adopts the same setting as above, but considers a different question: how *self-consistent* are the predictions of a model for terms that are distinct but equivalent

with respect to $\mathcal{F}_{\text{Dist}}$? Naturally, we expect a good model to provide the same predicted set for $A \cap B$ and $B \cap A$.

Consider a random term $p(x_1, \ldots, x_\ell)$, with sampled latents $z_1, \ldots, z_\ell$ yielding a predicted set U_{pred} via (6.3) and (6.4). Instead of comparing U_{pred} to U_{true}, we generate a family of equivalent terms $q_i(x_1, \ldots, x_\ell)$ by randomly selecting laws and substituting their expressions into $p(x_1, \ldots, x_\ell)$ if such expressions are present in $p(x_1, \ldots, x_\ell)$. For each equivalent term, we compare the new predicted set V_{pred} (computed via (6.3) and (6.4) as before) with the original prediction U_{pred} and compute the corresponding IoU metric.

Figure 6.3b summarizes our results. The x-axis represents the number of law applications, and the y-axis represents the self-consistency IoU. The solid lines represent the median performance for each choice of candidate operations, with the shaded areas representing the direct parameterizations' 20-to-80th percentile ranges.

Our Riesz mirrored algebra is perfectly self-consistent, experimentally validating Proposition 6.1. While the median performance of the learned baselines degrades moderately as the terms diverge, the bottom quartile drops sharply with even just two random symbolic manipulations. Interestingly, the direct parameterizations have a higher self-consistency than most other algebras, despite satisfying only zero or two laws. This suggests that a flexible parameterization can learn the appropriate symmetries to some degree, although we note that the Riesz algebra is decidedly superior to both across all experiments.

6.6 Conclusion

Interesting mathematical objects generally carry additional algebraic structure, such as operations and laws. Machine learning methods often encode such objects (sets, functions, etc.) into latent embeddings for downstream tasks. This paper examines the possibility of learning latent space operations that provably satisfy the same structural laws as the source algebra of input data. We provide a general procedure for constructing *structural transport nets* to carry out such transport of structure, and we illustrate the method in a concrete case study of the algebra of sets. Experiment results support our key hypothesis: stronger alignment between latent space operations and source algebra laws improves the performance of learned operations. Exciting future research involves further developing the theory of realizable latent-space operations and exploring downstream applications of structural transport nets.

Part IV

Bibliography

Bibliography

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. *CoRR*, abs/1705.10528, 2017.

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik R Narasimhan, and Ameet Deshpande. Geo: Generative engine optimization. *arXiv preprint arXiv:2311.09735*, 2023.

Sheila Alemany and Niki Pissinou. The dilemma between data transformations and adversarial robustness for time series application systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Motasem Alfarra, Adel Bibi, Philip H.S. Torr, and Bernard Ghanem. Data dependent randomized smoothing. *arXiv preprint arXiv:2012.04351*, 2020.

Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl, and Josef Havel. Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, 11(2):47–58, 2013.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

Brendon G. Anderson and Somayeh Sojoudi. Certified robustness via locally biased randomized smoothing. In *Learning for Dynamics and Control*. PMLR, 2022.

Brendon G. Anderson, Ziye Ma, Jingqi Li, and Somayeh Sojoudi. Tightened convex relaxations for neural network robustness certification. In *Proceedings of the 59th IEEE Conference on Decision and Control*, 2020.

Brendon G Anderson, Samuel Pfrommer, and Somayeh Sojoudi. Towards optimal randomized smoothing: A semi-infinite linear programming approach. In *ICML Workshop on Formal Verification of Machine Learning (WFVML)*, 2022.

Giovanni Apruzzese, Hyrum S Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. Real attackers don't compute gradients: bridging the gap between adversarial ml research and practice. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

MOSEK ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.*, 2019.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

Pranjal Awasthi, Vaggos Chatziafratis, Xue Chen, and Aravindan Vijayaraghavan. Adversarially robust low dimensional representations. In *Conference on Learning Theory*, pages 237–325. PMLR, 2021.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

Waiss Azizian and Marc Lelarge. Expressive power of invariant and equivariant graph neural networks. *arXiv preprint arXiv:2006.15646*, 2020.

Mitali Bafna, Jack Murtagh, and Nikhil Vyas. Thwarting adversarial examples: An L_0-robust sparse fourier transform. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Stanley Bak, Changliu Liu, and Taylor Johnson. The second international verification of neural networks competition (VNN-COMP 2021): Summary and results. *arXiv preprint arXiv:2109.00498*, 2021.

Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, 2019.

Felix Berkenkamp, Matteo Turchetta, Angela P. Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees, 2017.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, third edition, 2016.

Arjun Nitin Bhagoji, Daniel Cullina, Chawin Sitawarin, and Prateek Mittal. Enhancing robustness of machine learning systems via data transformations. In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–5. IEEE, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Lars Blackmore, Masahiro Ono, and Brian C Williams. Chance-constrained optimal path planning with obstacles. *IEEE Transactions on Robotics*, 27(6):1080–1094, 2011.

Julius R Blum, Jack Kiefer, and Murray Rosenblatt. *Distribution free tests of independence based on the sample distribution function*. Sandia Corporation, 1961.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

J. M. Bravo, D. Limon, T. Alamo, and E. F. Camacho. On the computation of invariant sets for constrained nonlinear systems: An interval arithmetic approach. In *2003 European Control Conference (ECC)*, pages 288–293, 2003. doi: 10.23919/ECC.2003.7084969.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Stanley Burris and Hanamantagouda P Sankappanavar. *A Course in Universal Algebra*, volume 78. Springer, 1981.

Sylvain Calinon and Aude Billard. Incremental learning of gestures by imitation in a humanoid robot. In *ACM/IEEE International Conference on Human-Robot Interaction*, 2007.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 3–14, 2017.

Francesco Cartella, Orlando Anunciacao, Yuki Funabiki, Daisuke Yamaguchi, Toru Akishita, and Olivier Elshocht. Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. *arXiv preprint arXiv:2101.08030*, 2021.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Sizhe Chen, Julien Piet, Chawin Sitawarin, and David Wagner. Struq: Defending against prompt injection with structured queries. *arXiv preprint arXiv:2402.06363*, 2024.

Yize Chen, Yuanyuan Shi, and Baosen Zhang. Optimal control via neural networks: A convex approach. In *International Conference on Learning Representations*, 2019.

Yize Chen, Yuanyuan Shi, and Baosen Zhang. Data-driven optimal voltage regulation using input convex neural networks. *Electric Power Systems Research*, 189:106741, 2020.

Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Jianqiang Cheng and Abdel Lisser. A second-order cone programming approach for linear programs with joint probabilistic constraints. *Operations Research Letters*, 40(5): 325–328, 2012.

Richard Cheng, Gábor Orosz, Richard Murray, and Joel Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3387–3395, 07 2019. doi: 10.1609/aaai.v33i01.33013387.

Hao-Tien Lewis Chiang, John EG Baxter, Satomi Sugaya, Mohammad R Yousefi, Aleksandra Faust, and Lydia Tapia. Fast deep swept volume estimator. *The International Journal of Robotics Research*, 40(10-11):1068–1086, 2021.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019a.

Stav Cohen, Ron Bitton, and Ben Nassi. A jailbroken genai model can cause substantial harm: Genai-powered applications are vulnerable to promptwares. *arXiv preprint arXiv:2408.05061*, 2024.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, 2016.

Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. *Advances in neural information processing systems*, 32, 2019b.

Stefano Cresci, Marinella Petrocchi, Angelo Spognardi, and Stefano Tognazzi. Adversarial machine learning for protecting against online manipulation. *IEEE Internet Computing*, 26(2):47–52, 2021.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces, 2018.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, 2004.

Pim De Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, 2019.

Luca De Luigi, Adriano Cardace, Riccardo Spezialetti, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Deep learning on implicit neural representations of shapes. In *International Conference on Machine Learning*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *International Conference on Learning Representations (Workshop)*, 2015.

John Duchi. Lecture notes in ee364a: convex optimization 1, 2021.

Francisco Eiras, Motasem Alfarra, M Pawan Kumar, Philip HS Torr, Puneet K Dokania, Bernard Ghanem, and Adel Bibi. Ancer: Anisotropic certification via sample-wise volume maximization. *arXiv preprint arXiv:2107.04570*, 2021.

Ecenaz Erdemir, Jeffrey Bickford, Luca Melis, and Sergul Aydore. Adversarial robustness with non-uniform perturbations. *Advances in Neural Information Processing Systems*, 34, 2021.

Jalal Etesami and Philipp Geiger. Causal transfer for imitation learning and decision making under sensor-shift. In *AAAI Conference on Artificial Intelligence*, 2020.

Chaim Even-Zohar. independence: Fast rank tests. *arXiv preprint arXiv:2010.09712*, 2020.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363 (6433):1287–1289, 2019.

William Fleshman, Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Non-negative networks against adversarial attacks. *arXiv preprint arXiv:1806.06108*, 2018.

Gerald B Folland. *Real analysis: Modern techniques and their applications*. John Wiley & Sons, second edition, 1999.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.acl-long.99.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: a survey. *arXiv preprint arXiv:2312.10997*, 2023b.

E.G. Gilbert and K.T. Tan. Linear systems with state and control constraints: the theory and application of maximal output admissible sets. *IEEE Transactions on Automatic Control*, 36(9):1008–1020, 1991. doi: 10.1109/9.83532.

Felipe O Giuste and Juan C Vizcarra. CIFAR-10 image classification using feature ensembles. *arXiv preprint arXiv:2002.03846*, 2020.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, 2023.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Edita Grolman, Hodaya Binyamini, Asaf Shabtai, Yuval Elovici, Ikuya Morikawa, and Toshiya Shimizu. HateVersarial: Adversarial attack against hate speech detection algorithms on Twitter. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 143–152, 2022.

Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel. Adversarial examples for malware detection. In *European Symposium on Research in Computer Security*, pages 62–79. Springer, 2017.

David Ha, Andrew Dai, and Quoc Le. HyperNetworks. In *International Conference on Learning Representations*, 2017.

Yang-Hui He and Minhyong Kim. Learning algebraic structures: preliminary investigations. *International Journal of Data Science in the Mathematical Sciences*, 2023.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in Neural Information Processing Systems*, 30, 2017.

Douglas Hensley. Slicing the cube in r^n and probability (bounds for the measure of a central cube slice in r^n by probability methods). *Proceedings of the American Mathematical Society*, 73(1):95–100, 1979. ISSN 00029939, 10886826.

Todd Hester, Matej Vecerík, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, Gabriel Dulac-Arnold, John P. Agapiou, Joel Z. Leibo, and Audrunas Gruslys. Deep q-learning from demonstrations. In *AAAI Conference on Artificial Intelligence*, 2017.

Tien Ho-Phuoc. CIFAR10 to compare visual recognition performance between deep neural networks and humans. *arXiv preprint arXiv:1811.07270*, 2018.

Wassily Hoeffding. A non-parametric test of independence. *The Annals of Mathematical Statistics*, 19(4):546–557, 1948.

Ahmed Hussein, Eyad Elyan, Mohamed Medhat Gaber, and Chrisina Jayne. Deep imitation learning for 3d navigation tasks. *Neural computing and applications*, 29: 389–404, 2018.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Íñigo Íncer Romeo, Michael Theodorides, Sadia Afroz, and David Wagner. Adversarially robust malware detection using monotonic classification. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, pages 54–63, 2018.

J Jaworowski. On antipodal sets on the sphere and on continuous involutions. *Fundamenta Mathematicae*, 2(43):241–254, 1956.

Jongheon Jeong, Sejun Park, Minkyu Kim, Heung-Chang Lee, Do-Guk Kim, and Jinwoo Shin. Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. In *Advances in Neural Information Processing Systems*, volume 34, pages 30153–30168, 2021.

Susmit Jha, Uyeong Jang, Somesh Jha, and Brian Jalaian. Detecting adversarial examples using data manifolds. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*, pages 547–552. IEEE, 2018.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms. *arXiv preprint arXiv:2402.11753*, 2024.

Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. Automatically auditing large language models via discrete optimization. In *International Conference on Machine Learning*, pages 15307–15329. PMLR, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Mohammed Kayed, Ahmed Anter, and Hadeer Mohamed. Classification of garments from fashion MNIST dataset using CNN LeNet-5 architecture. In *2020 International*

Conference on Innovative Trends in Communication and Computer Engineering (ITCE), pages 238–243. IEEE, 2020.

Jinrae Kim and Youdan Kim. Parameterized convex universal approximators for decision-making problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based model predictive control for safe exploration, 2018.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pages 2747–2755. PMLR, 2018.

Weicong Kong, Zhao Yang Dong, Youwei Jia, David J Hill, Yan Xu, and Yuan Zhang. Short-term residential load forecasting based on lstm recurrent neural network. *IEEE Transactions on Smart Grid*, 10(1):841–851, 2017.

Milan Korda, Didier Henrion, and Colin N. Jones. Convex computation of the maximum controlled invariant set for polynomial control systems, 2013.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701139. doi: 10.1145/3582269.3615599.

Sanjay Krishnan, Animesh Garg, Richard Liaw, Brijen Thananjeyan, Lauren Miller, Florian T. Pokorny, and Ken Goldberg. Swirl: A sequential windowed inverse reinforcement learning algorithm for robot tasks with delayed rewards. *The International Journal of Robotics Research*, 38:126 – 145, 2018.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Kuefler, Jeremy Morton, Timothy A. Wheeler, and Mykel J. Kochenderfer. Imitating driver behavior with generative adversarial networks. In *IEEE Intelligent Vehicles Symposium*, 2017.

Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *International Conference on Machine Learning*, pages 5458–5467. PMLR, 2020.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. In *Advances in Neural Information Processing Systems*, 2021.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

Yann LeCun. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy*, pages 656–672. IEEE, 2019.

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Alexander Levine and Soheil Feizi. Wasserstein smoothing: Certified robustness against Wasserstein adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 3938–3947. PMLR, 2020.

Alexander J Levine and Soheil Feizi. Improved, deterministic smoothing for ℓ_1 certified robustness. In *International Conference on Machine Learning*, pages 6254–6264. PMLR, 2021.

Benjamin A Levinstein and Daniel A Herrmann. Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, pages 1–27, 2024.

Dirk Lewandowski and Sebastian Schultheiß. Public awareness and attitudes towards search engine optimization. *Behaviour & Information Technology*, 42(8):1025–1044, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Linyi Li, Jiawei Zhang, Tao Xie, and Bo Li. Double sampling randomized smoothing. *arXiv preprint arXiv:2206.07912*, 2022.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229.

Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1028–1035, 2019a.

Chen Liu, Ryota Tomioka, and Volkan Cevher. On certifying non-uniform bounds against adversarial attacks. In *International Conference on Machine Learning*, pages 4072–4081. PMLR, 2019b.

Li-Ping Liu. Linear transformation of multivariate normal distribution: marginal, joint and posterior. 2019.

Qun Liu and Supratik Mukhopadhyay. Unsupervised learning using pretrained cnn and associative memory bank. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2018.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Prompt injection attacks and defenses in llm-integrated applications. *arXiv preprint arXiv:2310.12815*, 2023.

Daniel Lowd and Christopher Meek. Good word attacks on statistical spam filters. In *CEAS*, volume 2005, 2005.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models. *Empirical Methods in Natural Language Processing*, 2023.

Ziye Ma and Somayeh Sojoudi. A sequential framework towards an exact SDP verification of neural networks. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–8. IEEE, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Reza Mahjourian, Jinkyu Kim, Yuning Chai, Mingxing Tan, Ben Sapp, and Dragomir Anguelov. Occupancy flow fields for motion forecasting in autonomous driving. *IEEE Robotics and Automation Letters*, 7(2):5639–5646, 2022.

Thibault Maho, Teddy Furon, and Erwan Le Merrer. Randomized smoothing under attack: How good is it in practice? In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3014–3018. IEEE, 2022.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

Fernando Martin-Maroto and Gonzalo G de Polavieja. Algebraic machine learning. *arXiv preprint arXiv:1803.05252*, 2018.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: jailbreaking black-box llms automatically. *arXiv preprint arXiv:2312.02119*, 2023.

Kunal Menda, K. Driggs-Campbell, and Mykel J. Kochenderfer. Ensembledagger: A bayesian approach to safe imitation learning. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5041–5048, 2019.

Yue Meng, Dawei Sun, Zeng Qiu, Md Tawhid Bin Waez, and Chuchu Fan. Learning density distribution of reachable states for autonomous systems. In *Conference on Robot Learning*, pages 124–136. PMLR, 2022.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Jonathan Michaux, Qingyi Chen, Yongseok Kwon, and Ram Vasudevan. Reachability-based trajectory design with neural implicit safety constraints. *arXiv preprint arXiv:2302.07352*, 2023.

V. Mnih, K. Kavukcuoglu, D. Silver, Andrei A. Rusu, J. Veness, Marc G. Bellemare, A. Graves, Martin A. Riedmiller, A. Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, A. Sadik, Ioannis Antonoglou, Helen King, D. Kumaran, Daan Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Mark Niklas Müller, Christopher Brix, Stanley Bak, Changliu Liu, and Taylor T Johnson. The third international verification of neural networks competition (VNN-COMP 2022): Summary and results. *arXiv preprint arXiv:2212.10376*, 2022.

Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29:1711–1724, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Lakshmanan Nataraj, Sreejith Karthikeyan, Gregoire Jacob, and Bangalore S Manjunath. Malware images: Visualization and automatic classification. In *Proceedings of the 8th International Symposium on Visualization for Cyber Security*, pages 1–7, 2011.

Vitali Nesterov, Fabricio Arend Torres, Monika Nagy-Huber, Maxim Samarin, and Volker Roth. Learning invariances with generalised input-convex neural networks. *arXiv preprint arXiv:2204.07009*, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Pedro A Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, et al. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Jia-Hui Pan, Ka-Hei Hui, Xiaojie Gao, Shize Zhu, Yun-Hui Liu, Pheng-Ann Heng, and Chi-Wing Fu. Sdf-pack: Towards compact bin packing with signed-distance-field minimization. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10612–10619. IEEE, 2023.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 2021.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Jongjin Park, Younggyo Seo, Chang Liu, Li Zhao, Tao Qin, Jinwoo Shin, and Tie-Yan Liu. Object-aware regularization for addressing causal confusion in imitation learning. *Advances in Neural Information Processing Systems*, 34:3029–3042, 2021.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides,

Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *Empirical Methods in Natural Language Processing*, 2022.

Jan Peters and Stefan Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, 2008. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2008.02.003. Robotics and Neuroscience.

Samuel Pfrommer, Tanmay Gautam, Alec Zhou, and Somayeh Sojoudi. Safe reinforcement learning with chance-constrained model predictive control. In *Learning for Dynamics and Control Conference*, pages 291–303. PMLR, 2022.

Samuel Pfrommer, Brendon Anderson, Julien Piet, and Somayeh Sojoudi. Asymmetric certified robustness via feature-convex neural networks. *Advances in Neural Information Processing Systems*, 36:52365–52400, 2023a.

Samuel Pfrommer, Brendon G. Anderson, and Somayeh Sojoudi. Projected randomized smoothing for certified adversarial robustness. *Transactions on Machine Learning Research*, 2023b.

Samuel Pfrommer, Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. Ranking manipulation for conversational search engines. *Empirical Methods in Natural Language Processing*, 2023c.

Samuel Pfrommer, Yatong Bai, Hyunin Lee, and Somayeh Sojoudi. Initial state interventions for deconfounded imitation learning. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 2312–2319. IEEE, 2023d.

Samuel Pfrommer, Brendon G Anderson, and Somayeh Sojoudi. Transport of algebraic structure to latent embeddings. *International Conference on Machine Learning*, 2024.

Julien Piet, Maha Alrashed, Chawin Sitawarin, Sizhe Chen, Zeming Wei, Elizabeth Sun, Basel Alomair, and David Wagner. Jatmo: Prompt injection defense by task-specific finetuning. *arXiv preprint arXiv:2312.17673*, 2023.

B. Pluymers, J.A. Rossiter, J.A.K. Suykens, and B. De Moor. The efficient computation of polyhedral invariant sets for linear systems with polytopic uncertainty. In *Proceedings of the 2005, American Control Conference, 2005.*, pages 804–809 vol. 2, 2005. doi: 10.1109/ACC.2005.1470058.

Yao Qiang, Xiangyu Zhou, and Dongxiao Zhu. Hijacking large language models via adversarial in-context learning. *arXiv preprint arXiv:2311.09948*, 2023.

S Joe Qin and Thomas A Badgwell. A survey of industrial model predictive control technology. *Control engineering practice*, 11(7):733–764, 2003.

Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*, pages 117–126, 2017.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10877–10887, 2018.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*. Citeseer, 2003.

J.B. Rawlings and D.Q. Mayne. *Model Predictive Control: Theory and Design*. Nob Hill Pub., 2009. ISBN 9780975937709.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.

R Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 2011a.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudík, editors, *AISTATS*, volume 15 of *JMLR Proceedings*, pages 627–635. JMLR.org, 2011b.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. Improving passage retrieval with zero-shot question generation. *Empirical Methods in Natural Language Processing*, 2022.

Amir Mahdi Sadeghzadeh, Saeed Shiravi, and Rasool Jalili. Adversarial network traffic: Towards evaluating the robustness of deep-learning-based network traffic classification. *IEEE Transactions on Network and Service Management*, 18(2):1962–1976, 2021.

Rajeev Sahay, Rehana Mahfuz, and Aly El Gamal. Combatting adversarial attacks through denoising and dimensionality reduction: A cascaded autoencoder approach. In *2019 53rd Annual conference on information sciences and systems (CISS)*, pages 1–6. IEEE, 2019.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Amartya Sanyal, Varun Kanade, Philip HS Torr, and Puneet K Dokania. Robustness via deep low-rank representations. *arXiv preprint arXiv:1804.07090*, 2018.

Suman Sapkota and Binod Bhattarai. Input invex neural network. *arXiv preprint arXiv:2106.08748*, 2021.

Eli Sennesh, Tom Xu, and Yoshihiro Maruyama. Computing with categories in machine learning. In *International Conference on Artificial General Intelligence*, 2023.

Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *arXiv preprint arXiv:2106.10151*, 2021.

Fumin Shen, Chunhua Shen, Rhys Hill, Anton van den Hengel, and Zhenmin Tang. Fast approximate ℓ_∞ minimization: Speeding up robust regression. *Computational Statistics & Data Analysis*, 77:25–37, 2014.

David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2020.

Dan Shiebler, Bruno Gavranović, and Paul Wilson. Category theory in machine learning. *arXiv preprint arXiv:2106.07032*, 2021.

Zhenyu Shou, Xuan Di, Jieping Ye, Hongtu Zhu, Hua Zhang, and Robert Hampshire. Optimal passenger-seeking policies on e-hailing platforms using markov decision process and imitation learning. *Transportation Research Part C: Emerging Technologies*, 111: 91–113, 2020.

Ali Siahkamari, Durmus Alp Emre Acar, Christopher Liao, Kelly L Geyer, Venkatesh Saligrama, and Brian Kulis. Faster algorithms for learning convex functions. In *International Conference on Machine Learning*, pages 20176–20194. PMLR, 2022.

Andrew F Siegel. *Practical business statistics*. Academic Press, 2016.

Joseph Sill. Monotonic networks. *Advances in neural information processing systems*, 10, 1997.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems*, 2020.

Sarath Sivaprasad, Ankur Singh, Naresh Manwani, and Vineet Gandhi. The curious case of convex neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 738–754. Springer, 2021.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

Neha Srikanth, Marine Carpuat, and Rachel Rudinger. How often are errors in natural language reasoning due to paraphrastic variability? *Transactions of the Association for Computational Linguistics*, 12:1143–1162, 2024. doi: 10.1162/tacl_a_00692.

Peter Súkeník, Aleksei Kuvshinov, and Stephan Günnemann. Intriguing properties of input-dependent randomized smoothing. *arXiv preprint arXiv:2110.05365*, 2021.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agent. *Empirical Methods in Natural Language Processing*, 2023.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1998. ISBN 0-262-19398-1.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Zhiwei Steven Wu. Causal imitation learning under temporally correlated noise. In *International Conference on Machine Learning*, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Lucas Matthew Tecot. Robustness verification with non-uniform randomized smoothing. Master's thesis, University of California, Los Angeles, 2021.

Jiaye Teng, Guang-He Lee, and Yang Yuan. ℓ_1 adversarial robustness certificates: A randomized smoothing approach. *Preprint*, 2020.

Takeshi Teshima, Isao Ishikawa, Koichi Tojo, Kenta Oono, Masahiro Ikeda, and Masashi Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. *Advances in Neural Information Processing Systems*, 33:3362–3373, 2020.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization, 2018.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Sam Toyer, Olivia Watkins, Ethan Adrian Mendes, Justin Svegliato, Luke Bailey, Tiffany Wang, Isaac Ong, Karim Elmaaroufi, Pieter Abbeel, Trevor Darrell, et al. Tensor trust: Interpretable prompt injection attacks from an online game. *International Conference on Learning Representations*, 2024.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems*, volume 33, pages 1633–1645, 2020.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the Cayley transform. *arXiv preprint arXiv:2104.07167*, 2021.

Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pages 5025–5034. PMLR, 2018.

Jeffrey Vaaler. A geometric inequality with applications to linear forms. *Pacific Journal of Mathematics*, 83(2):543–553, 1979.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.

Risto Vuorio, Johann Brehmer, Hanno Ackermann, Daniel Dijkman, Taco Cohen, and Pim de Haan. Deconfounded imitation learning. *arXiv preprint arXiv:2211.02667*, 2022.

Kim P. Wabersich and Melanie N. Zeilinger. Linear model predictive safety certification for learning-based control, 2019.

Kim P. Wabersich and Melanie N. Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems, 2021.

Nolan Wagener, Byron Boots, and Ching-An Cheng. Safe reinforcement learning using advantage-based intervention. In *ICML*, 2021.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.

Lei Wang, Runtian Zhai, Di He, Liwei Wang, and Li Jian. Pretrain-to-finetune adversarial training via sample-wise randomized smoothing. *Preprint*, 2021a.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints

for neural network robustness verification. *Advances in Neural Information Processing Systems*, 34:29909–29921, 2021b.

Tianyu Wang, Vikas Dhiman, and Nikolay A. Atanasov. Inverse reinforcement learning for autonomous navigation via differentiable semantic mapping and planning. *arXiv preprint arXiv:2101.00186*, 2021c.

Yixuan Wang, Yunzhu Li, Katherine Driggs-Campbell, Li Fei-Fei, and Jiajun Wu. Dynamic-resolution model learning for object pile manipulation. *arXiv preprint arXiv:2306.16700*, 2023.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay, 2017.

Wolfgang Wechler. *Universal Algebra for Computer Scientists*, volume 25. Springer Science & Business Media, 2012.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36, 2024.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for ReLU networks. In *International Conference on Machine Learning*, pages 5276–5285. PMLR, 2018.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

Bichen Wu, Forrest Iandola, Peter H. Jin, and Kurt Keutzer. SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 129–137, 2017a.

Fangzhou Wu, Xiaogeng Liu, and Chaowei Xiao. Deceptprompt: Exploiting llm-driven code generation via adversarial natural language instructions. *arXiv preprint arXiv:2312.04730*, 2023a.

Yuanwei Wu, Xiang Li, Yixin Liu, Pan Zhou, and Lichao Sun. Jailbreaking gpt-4v via self-adversarial attacks with system prompts. *arXiv preprint arXiv:2311.09127*, 2023b.

Yuhuai Wu, Elman Mansimov, Roger B Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. *Advances in neural information processing systems*, 30, 2017b.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Samir S Yadav and Shivajirao M Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6(1):1–18, 2019.

Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020a.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in Neural Information Processing Systems*, 33:8588–8601, 2020b.

Shaowei Yao, Jiwei Tan, Xi Chen, Keping Yang, Rong Xiao, Hongbo Deng, and Xiaojun Wan. Learning a product relevance model from click-through data in e-commerce. In *Proceedings of the Web Conference 2021*, 2021.

Jingwei Yi, Yueqi Xie, Bin Zhu, Keegan Hines, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. Benchmarking and defending against indirect prompt injection attacks on large language models. *arXiv preprint arXiv:2312.14197*, 2023.

He Yin, Peter Seiler, Ming Jin, and Murat Arcak. Imitation learning with stability and safety guarantees. *IEEE Control Systems Letters*, 6:409–414, 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Fancheng Zeng, Guanqiu Qi, Zhiqin Zhu, Jian Sun, Gang Hu, and Matthew Haner. Convex neural networks based reinforcement learning for load frequency control under denial of service attacks. *Algorithms*, 15(2):34, 2022.

Huimin Zeng, Jiahao Su, and Furong Huang. Certified defense via latent space randomized smoothing with orthogonal encoders. *arXiv preprint arXiv:2108.00491*, 2021.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. *arXiv preprint arXiv:2403.02691*, 2024.

Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Boosting the certified robustness of l-infinity distance nets. *arXiv preprint arXiv:2110.06850*, 2021a.

Dinghuai Zhang, Mao Ye, Chengyue Gong, Zhanxing Zhu, and Qiang Liu. Black-box certification with randomized smoothing: A functional optimization based framework. In *Advances in Neural Information Processing Systems*, volume 33, pages 2316–2326, 2020a.

Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. In *Advances in Neural Information Processing Systems*, 2020b.

Ling Zhang, Yize Chen, and Baosen Zhang. A convex neural network solver for DCOPF with generalization guarantees. *IEEE Transactions on Control of Network Systems*, 2021b.

Yonggang Zhang, Xinmei Tian, Ya Li, Xinchao Wang, and Dacheng Tao. Principal component adversarial example. *IEEE Transactions on Image Processing*, 29:4804–4815, 2020c.

Sen Zhao, Erez Louidor, and Maya Gupta. Global optimization networks. In *International Conference on Machine Learning*, pages 26927–26957. PMLR, 2022.

Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. Autodan: Automatic and interpretable adversarial attacks on large language models. *arXiv preprint arXiv:2310.15140*, 2023.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. A setwise approach for effective and highly efficient zero-shot ranking with large language models. *arXiv preprint arXiv:2310.09497*, 2023.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Part V

Proofs

Appendix A

Initial State Interventions for Deconfounded Imitation Learning

A.1 Proofs for Section 2.4

We first introduce a series of auxiliary lemmas.

Lemma A.1. Consider an interventionally absolutely continuous SCM $\mathcal{M}$ with a faithful causal graph $\mathcal{G}$ that contains a directed path from X to Y. Then provided a set $\mathbf{Z}$ contains all ancestors of X but none of its descendants, then for any assignment $\mathbf{z}$ to $\mathbf{Z}$ there exist values x, x' such that

$$\Big\| P\big(y \mid \mathrm{do}(x), \mathbf{z}\big) - P\big(y \mid \mathrm{do}(x'), \mathbf{z}\big) \Big\|_1 > 0,$$

viewed as induced measures over Y.

Proof. As $\mathbf{Z}$ contains no descendants of X, it cannot block the directed path between X and Y and hence the Causal Markov Condition does not declare X and Y independent. Faithfulness stipulates that X and Y are therefore dependent given $\mathbf{z}$, so there exist x, x' such that

$$\Big\| P\big(y \mid x, \mathbf{z}\big) - P\big(y \mid x', \mathbf{z}\big) \Big\|_1 > 0.$$

The second rule of do calculus states that we can exchange observation and intervention if X and Y are independent given $\mathbf{z}$ in the causal graph $\mathcal{G}_{\underline{X}}$ obtained by removing outgoing edges from X. If we remove outgoing edges from X, the only remaining paths between X and Y must contain an edge $X \leftarrow Z$ for some variable Z. This makes Z an ancestor of X, and therefore Z is included in $\mathbf{Z}$, and both paths of the form $X \leftarrow Z \leftarrow J$ and $X \leftarrow Z \rightarrow J$ are blocked by $\mathbf{Z}$. This means that X and Y are d-separated by $\mathbf{Z}$ in $\mathcal{G}_{\underline{X}}$,

and we can apply the second do-calculus rule to conclude that

$$P\Big(y \mid \mathrm{do}(x), \mathbf{z}\Big) = P\Big(y \mid x, \mathbf{z}\Big),$$
$$P\Big(y \mid \mathrm{do}(x'), \mathbf{z}\Big) = P\Big(y \mid x', \mathbf{z}\Big),$$

and hence

$$\Big\| P\Big(y \mid \mathrm{do}(x), \mathbf{z}\Big) - P\Big(y \mid \mathrm{do}(x'), \mathbf{z}\Big) \Big\|_1 > 0.$$

□

Lemma A.2. Consider a set $E \subseteq \mathbb{R}$ where for each $x \in E$, there exists a ball $B(x, \epsilon_x)$ which contains no point in E. Then E has measure zero with respect to the standard Lebesgue measure on $\mathbb{R}$.

Proof. As E is a subset of $\mathbb{R}$, it is Lindelöf, and the cover of E by the collection of balls $\{B(x, \epsilon_x) \mid x \in E\}$ has a finite subcover. Enumerate this subcover as I_i; we then have

$$\lambda(E) = \lambda(E \cap (\cup_i I_i)) \leq \sum_i \lambda(E \cap I_i) = 0,$$

as each $E \cap I_i$ contains only a singleton. □

Lemma A.3. Let $f(x)$ be a differentiable function of $x \in \mathbb{R}$ at some $\bar{x} \in \mathbb{R}$ with $f(\bar{x}) = 0$. Then

$$\frac{d}{dx}\bigg|_{\bar{x}^+} |f(x)| = \left| \frac{d}{dx}\bigg|_{\bar{x}} f(x) \right| \quad \text{and} \quad \frac{d}{dx}\bigg|_{\bar{x}^-} |f(x)| = -\left| \frac{d}{dx}\bigg|_{\bar{x}} f(x) \right|.$$

Proof. We prove the first result as the second follows similarly. Expanding the derivative:

$$\begin{aligned}
\frac{d}{dx}\bigg|_{\bar{x}^+} |f(x)| &= \lim_{\delta \to 0^+} \frac{|f(\bar{x}+\delta)| - |f(\bar{x})|}{\delta} \\
&= \lim_{\delta \to 0^+} \frac{|f(\bar{x}+\delta) - f(\bar{x})|}{\delta} \\
&= \left| \lim_{\delta \to 0^+} \frac{f(\bar{x}+\delta) - f(\bar{x})}{\delta} \right| \\
&= \left| \frac{d}{dx}\bigg|_{\bar{x}} f(x) \right|,
\end{aligned}$$

where moving the limit inside the absolute value is permissible by differentiability of f at $\bar{x}$ and continuity of absolute value. □

Lemma A.4. Let $f(x) : \mathbb{R} \to M(Y)$ continuously map real numbers x to a measure over the values assumed by a random variable Y. Then we have that

$$\frac{d}{db}\bigg|_{\bar{b}} \frac{d}{d\lambda} \left(\int_a^b f(x) dx \right) = \frac{d}{d\lambda} f(\bar{b})$$

almost everywhere over the domain of Y. Here $\int_a^b f(x)dx$ denotes Lebesgue integration against $\mathcal{U}(a,b)$, $\frac{d}{d\lambda}$ is the Radon-Nikodym derivative with respect to the standard Lebesgue measure on $\mathbb{R}$, and $\frac{d}{db}\Big|_{\bar{b}}$ denotes the standard real analysis derivative evaluated at $\bar{b}$.

Proof. We can expand the definition of the outer derivative

$$\begin{aligned}\frac{d}{db}\Big|_{\bar{b}} \frac{d}{d\lambda}\left(\int_a^b f(x)dx\right) &= \lim_{b\to\bar{b}} \frac{1}{b-\bar{b}}\left(\frac{d}{d\lambda}\left(\int_a^{\bar{b}} f(x)dx\right) - \frac{d}{d\lambda}\left(\int_a^b f(x)dx\right)\right) \\ &= \lim_{b\to\bar{b}} \frac{1}{b-\bar{b}}\left(\frac{d}{d\lambda}\left(\int_b^{\bar{b}} f(x)dx\right)\right). \\ &= \lim_{b\to\bar{b}} \frac{d}{d\lambda}\left(\frac{1}{b-\bar{b}}\int_b^{\bar{b}} f(x)dx\right).\end{aligned}$$

Take an arbitrary $\epsilon > 0$. We want to show $\exists\delta > 0$ such that for all $\bar{b}-\delta < b < \bar{b}+\delta$, we have that

$$\left\|\frac{1}{b-\bar{b}}\left(\int_b^{\bar{b}} f(x)dx\right) - f(\bar{b})\right\|_1 < \epsilon,$$

where the Radon-Nikodym derivative $\frac{d}{d\lambda}$ is absorbed into the L_1 norm definition on measures. By continuity of f, we can always choose a δ small enough for this inequality to hold. □

We now prove the main theoretical results.

Theorem 2.5. In the faithful system causal model $\langle \mathcal{M}_s, \mathcal{G}_s\rangle$, assume that the measure-valued function $w \mapsto P(v \mid \mathrm{do}(\mathbf{Z}=\mathbf{z}), w)$ is continuous for any set of nodes $\mathbf{Z}$ and $V \notin \mathbf{Z}$.

Let there exist a causal edge $O_t^{\mathscr{o}} \to A_{t'}^{\mathscr{a}}$ in $\mathcal{G}_s$ for some $t, t' \in \mathbb{N}$, $t' \geq t$, and indices $\mathscr{o} \in [d_{\mathcal{O}}]$ and $\mathscr{a} \in [d_{\mathcal{A}}]$. Then in the interventional causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s\rangle$ where the initial state distribution $\tilde{P}(s_1)$ has everywhere-nonzero density on $\mathcal{S}$, $O^{\mathscr{o}}$ is almost surely not masked by Algorithm 2.1 for almost every uniform parameterization of W as the number of trajectories $N \to \infty$; i.e., (2.6) correctly evaluates to true.

Proof. By Assumptions 2.1 and 2.3, we can WLOG consider $t = 1$ with $t' \in [H]$. If $O_1^{\mathscr{o}} \to A_{t'}^{\mathscr{a}}$, by Assumption 2.2 there exists an edge $S_1^{\mathscr{s}} \to O_1^{\mathscr{o}}$ for some $\mathscr{s}$. We now want to show that in the SCM $\widetilde{\mathcal{M}}_s$ where we intervene distributionally on S_1, we have that $S_1^{\mathscr{s}} \not\perp\!\!\!\perp O_1^{\mathscr{o}}$ and $S_1^{\mathscr{s}} \not\perp\!\!\!\perp A_{t'}^{\mathscr{a}}$. The arguments are similar, so we will just state the proof for the former.

We want to show that $S_1^{\mathscr{s}}$ and $O_1^{\mathscr{o}}$ are not independent in $\widetilde{\mathcal{M}}_s$. Note that in the modified structural assignment for $S_1^{\mathscr{s}}$ in $\widetilde{\mathcal{M}}_s$, $S_1^{\mathscr{s}}$ is distributed with everywhere-nonzero density on

$\mathcal{S}$. Therefore checking the desired independence is equivalent to showing

$$\left\| P\Big(o_1^{\mathcal{O}} \mid \mathrm{do}(S_1^{\mathcal{S}} = \alpha)\Big) - P\Big(o_1^{\mathcal{O}} \mid \mathrm{do}(S_1^{\mathcal{S}} = \alpha')\Big) \right\|_1 > 0 \tag{A.1}$$

for some $\alpha, \alpha' \in \mathbb{R}$ with $\alpha \neq \alpha'$. Here, $P\Big(o_1^{\mathcal{O}} \mid \cdot\Big)$ denotes a probability measure over $o_1^{\mathcal{O}}$. The do statement captures our ability to intervene on the initial state, decoupling any potential correlational influence from W.

By Lemma A.1, we have that for any particular value w of W,

$$\left\| P\Big(o_1^{\mathcal{O}} \mid \mathrm{do}(S_1^{\mathcal{S}} = \alpha), w\Big) - P\Big(o_1^{\mathcal{O}} \mid \mathrm{do}(S_1^{\mathcal{S}} = \alpha'), w\Big) \right\|_1 > 0,$$

for some α, α'. This is equivalent to

$$\|h(\alpha, \alpha', w)\|_1 \not\equiv \mathbf{0} \quad \forall w, \tag{A.2}$$

where we define

$$h(\alpha, \alpha', w) := P\Big(o_1^{\mathcal{O}} \mid \mathrm{do}(S_1^{\mathcal{S}} = \alpha), w\Big) - P\Big(o_1^{\mathcal{O}} \mid \mathrm{do}(S_1^{\mathcal{S}} = \alpha'), w\Big),$$

and $\mathbf{0}$ denotes an identically zero function over α, α'. Note that $h(\alpha, \alpha', w)$ specifies a signed measure over $o_1^{\mathcal{O}}$. Now observe that

$$P\Big(o_1^{\mathcal{O}} \mid \mathrm{do}(S_1^{\mathcal{S}} = \alpha)\Big) = \int P\Big(o_1^{\mathcal{O}} \mid \mathrm{do}(S_1^{\mathcal{S}} = \alpha), w\Big) p(w) d\mu(w),$$

where μ is a probability measure on the unobserved variable w which we will instantiate shortly, and $p(w)$ denotes the probability density of W, i.e. the Radon-Nikodym derivative of the measure $P(w)$. Note that the result of this integral is still a signed measure over $o_1^{\mathcal{O}}$. So we have that showing our desired inequality (A.1) is equivalent to showing

$$\left\| \int h(\alpha, \alpha', w) p(w) d\mu(w) \right\|_1 \not\equiv \mathbf{0}$$

as a function of α, α' for "almost all" measures μ—as there is no natural measure on the space of measures, we have formalized this assuming a uniform distribution $W \sim \mathcal{U}(a, b)$. Note that the outer norm computes the L_1 norm of a signed measure over $o_1^{\mathcal{O}}$. For notational convenience, we will now define the concatenation $z = [\alpha, \alpha']$, with $z \in \mathbb{R}^2$. We can now concretely refine μ in the above statement, using our new z-notation, to showing that

$$g_a^b(z) := \left\| \int_a^b h(z, w) dw \right\|_1 \not\equiv \mathbf{0} \tag{A.3}$$

as a function of z for almost every (a, b); i.e., the subset of (a, b) parameter space where $g_a^b(z) \equiv \mathbf{0}$ over z is measure zero with respect to the standard Lebesgue measure in $\mathbb{R}^2$. Note that we drop the $p(w)$ factor since for the uniform distribution this is a constant which factors out.

This can be analyzed by taking sections were we fix a and consider the set of b's where $g_a^b(z) \equiv \mathbf{0}$; if this set has measure zero, then the overall set of Cartesian pairs (a, b) where $g_a^b(z) \equiv \mathbf{0}$ can be shown to have measure zero by the following argument. Observe that $g_a^b(z)$ is continuous in a, b and z by Assumption 2.4 and integral properties; then the inverse image of $\{0\}$ under g is a Borel subset of $A \subset \mathbb{R}^4$, recalling that $z \in \mathbb{R}^2$. The projection of this Borel subset on to the (a, b) plane is measurable (but not necessarily Borel). Then if each fixed-a slice is measure zero, the overall set is measure zero by Fubini's theorem.

Correspondingly, we fix any particular a and drop it from the subscript of g_a^b for simplicity. Consider a particular $\bar{b}$ where $g^{\bar{b}}(z) \equiv \mathbf{0}$ as a function of z. We expand the L_1 norm in (A.3) as

$$g^b(z) = \int \left| \frac{d}{d\lambda} \left(\int_a^b h(z, w) dw \right) \right| d\lambda = \int \left| f_z^b(o_1^o) \right| d\lambda, \tag{A.4}$$

using the interventional absolute continuity assumption to invoke the Radon-Nikodym derivative on our signed measure over o_1^o with respect to the standard Lebesgue measure λ. We denote the resulting density function by $f_z^b(o_1^o)$. Note that since $g^{\bar{b}}(z) \equiv \mathbf{0}$, we have that $f_z^{\bar{b}}(o_1^o) = 0$ for almost all z and o_1^o.

Note that $f_z^b(o_1^o)$ is differentiable with respect to b due to the assumed continuity of maps on w in the theorem statement. We now differentiate (A.4) with respect to b at $\bar{b}$. Due to the absolute value in (A.4), we must take care to differentiate from above and below and show both these cases are nonzero. As they follow similarly, we show the case for above:

$$\begin{aligned}
\frac{d}{db}\bigg|_{\bar{b}^+} g^b(z) &= \frac{d}{db}\bigg|_{\bar{b}^+} \int \left| \frac{d}{d\lambda} \left(\int_a^b h(z, w) dw \right) \right| d\lambda && \text{(A.5)} \\
&= \int \frac{d}{db}\bigg|_{\bar{b}^+} \left| \frac{d}{d\lambda} \left(\int_a^b h(z, w) dw \right) \right| d\lambda && \text{(A.6)} \\
&= \int \left| \frac{d}{db}\bigg|_{\bar{b}} \frac{d}{d\lambda} \left(\int_a^b h(z, w) dw \right) \right| d\lambda && \text{(A.7)} \\
&= \int \left| \frac{d}{d\lambda} h(z, \bar{b}) \right| d\lambda && \text{(A.8)} \\
&= \left\| h(z, \bar{b}) \right\|_1 && \text{(A.9)} \\
&\not\equiv \mathbf{0}, \qquad \text{(as a function of } z\text{)} && \text{(A.10)}
\end{aligned}$$

where (A.6) follows from boundedness of the Radon-Nikodym derivative of $h(z, \bar{b})$, (A.7) follows from applying Lemma A.3 to $f_z^b(o_1^o)$ with respect to b, (A.8) follows from Lemma A.4, (A.9) follows from differentiability of $f_z^b(o_1^o)$ with respect to b, and (A.10) follows from (A.2).

Proceeding similarly, we can show that both

$$\frac{d}{db}\bigg|_{\bar{b}^+} g^b(z) \not\equiv \mathbf{0} \quad \text{and} \quad \frac{d}{db}\bigg|_{\bar{b}^-} g^b(z) \not\equiv \mathbf{0}.$$

It is then immediate that there exists a ball $B(\bar{b}, \epsilon_{\bar{b}})$ such that $g^b(z) \not\equiv \mathbf{0}$ for all $b \in B(\bar{b}, \epsilon_{\bar{b}}) \setminus \bar{b}$. Applying Lemma A.2 concludes that for a fixed a, the set of b for which (A.3) is violated is measure zero. Hence by the above Fubini argument, for almost every uniform measure $\mathcal{U}(a, b)$ on w, we have that (A.1) holds for some α, α'. Therefore $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp O_1^{\mathfrak{o}}$ in the interventional distribution on $S_1^{\mathfrak{s}}$.

A similar argument shows that $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp A_{t'}^{\mathfrak{a}}$. By absolute continuity of the induced interventional distributions, we now have that Hoeffding's independence test is consistent, and hence the dependences are detected with probability 1 as $N \to \infty$. Therefore $O_1^{\mathfrak{o}} \dashrightarrow A_{t'}^{\mathfrak{a}}$, and $\widetilde{m}_{\mathfrak{o}}$ evaluates to false (2.6) as $N \to \infty$. □

Theorem 2.6. Let m denote the potential-cause test evaluated by Algorithm 2.1 on the distribution induced by the non-interventional system $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$, and let $\widetilde{m}$ be the original test on the interventional system $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$ where $\tilde{P}(s_1)$ has everywhere-nonzero density on $\mathcal{S}$. If $m_{\mathfrak{o}}$ correctly evaluates to true for a particular $\mathfrak{o} \in [d_{\mathcal{O}}]$, then $\widetilde{m}_{\mathfrak{o}}$ also evaluates to true almost surely as the number of trajectories $N \to \infty$.

Proof. If $m_{\mathfrak{o}}$ evaluates to true, then for any $\mathfrak{s} \in [d_{\mathcal{S}}]$, $\mathfrak{a} \in [d_{\mathcal{A}}]$, and $t' \in [H]$, we have that either $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\mathcal{M}_s} O_1^{\mathfrak{o}}$ or $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\mathcal{M}_s} A_{\mathfrak{a}}^{t'}$, where $\perp\!\!\!\perp_{\mathcal{M}_s}$ denotes independence in the distribution induced by the non-interventional SCM $\mathcal{M}_s$. It suffices to show that both these independencies hold in the distribution induced by $\widetilde{\mathcal{M}}_s$. As both arguments follow similarly, we consider showing that $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\widetilde{\mathcal{M}}_s} O_1^{\mathfrak{o}}$.

As we are given $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\mathcal{M}_s} O_1^{\mathfrak{o}}$, it is immediate by faithfulness that there exists no collider-free path from $S_1^{\mathfrak{s}}$ to $O_1^{\mathfrak{o}}$ in $\mathcal{G}_s$. Since $\widetilde{\mathcal{G}}_s$ is simply $\mathcal{G}_s$ with the incoming edges to S_1 removed, it holds that there is no collider-free path between $S_1^{\mathfrak{s}}$ and $O_1^{\mathfrak{o}}$ in $\widetilde{\mathcal{G}}_s$. Therefore $S_1^{\mathfrak{s}} \perp\!\!\!\perp_{\widetilde{\mathcal{M}}_s} O_1^{\mathfrak{o}}$, and as $N \to \infty$ this is correctly detected with probability 1 by the consistency of Hoeffding's test. □

Proposition 2.7. Let $\widetilde{m}$ and m be as in Theorem 2.6, and consider a particular observation index $\mathfrak{o} \in [d_{\mathcal{O}}]$ such that the only incoming edge to $O_1^{\mathfrak{o}}$ is $W \to O_1^{\mathfrak{o}}$. Then if in $\mathcal{G}_s$ there exists the fork $S_1^{\mathfrak{s}} \leftarrow W \to O_1^{\mathfrak{o}}$ for some $\mathfrak{s} \in [d_{\mathcal{S}}]$ and a directed path from $S_1^{\mathfrak{s}}$ to some $A_t^{\mathfrak{a}}$, with $t \in [H], \mathfrak{a} \in [d_{\mathcal{A}}]$, $\widetilde{m}_{\mathfrak{o}}$ correctly masks the $\mathfrak{o}^{\text{th}}$ observation almost surely as the number of trajectories $N \to \infty$ while $m_{\mathfrak{o}}$ does not.

Proof. We first show that m does not mask $\mathfrak{o}$ and take all causal and probabilistic statements to refer to the unintervened causal model $\langle \mathcal{M}_s, \mathcal{G}_s \rangle$. By faithfulness, the fork $S_1^{\mathfrak{s}} \leftarrow W \to O_1^{\mathfrak{o}}$ in $\mathcal{G}_s$ produces a statistical dependence $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp_{\mathcal{M}_s} O_1^{\mathfrak{o}}$ in the probability distribution induced by $\mathcal{M}_s$. Similarly, the directed path from $S_1^{\mathfrak{s}}$ to $A_t^{\mathfrak{a}}$ yields $S_1^{\mathfrak{s}} \not\perp\!\!\!\perp_{\mathcal{M}_s} A_t^{\mathfrak{a}}$. By consistency of Hoeffding's test, as $N \to \infty$ we get that ${}^{(1,t)}D_{\mathfrak{s},\mathfrak{a}}^{\mathfrak{o}}$ evaluates to true almost surely (2.3) and thus $O_1^{\mathfrak{o}} \dashrightarrow A_t^{\mathfrak{a}}$ by (2.4). Therefore $m_{\mathfrak{o}}$ is not masked (2.6).

We now show that $\widetilde{m}$ does mask o and take all causal and probabilistic statements to refer to the *intervened* causal model $\langle \widetilde{\mathcal{M}}_s, \widetilde{\mathcal{G}}_s \rangle$. Since W only has outgoing edges, and the edge from $W \to S_1^{s'}$ is removed in $\widetilde{\mathcal{G}}_s$ for every $s' \in [d_{\mathcal{S}}]$, there exists no path from $S_1^{s'}$ to O_1^o in $\widetilde{\mathcal{G}}_s$, and therefore $S_1^{s'} \perp\!\!\!\perp_{\widetilde{\mathcal{M}}_s} O_1^o$ in the probability distribution induced by $\widetilde{\mathcal{M}}_s$. As $N \to \infty$ this independence is detected by Hoeffding's test, and since s' was arbitrary ${}^{(1,t')}D^o_{s',a'}$ is false for every $s' \in [d_{\mathcal{S}}]$, $a' \in [d_{\mathcal{A}}]$, and $t' \in [H]$. Therefore $O_1^o \not\dashrightarrow A_{t'}^{a'}$ for any $a' \in [d_{\mathcal{A}}], t' \in [H]$, and (2.6) evaluates to true. Therefore $\widetilde{m}_o$ is masked. □

Appendix B

Projected Randomized Smoothing for Certified Adversarial Robustness

B.1 Proofs for Section 3.3

Proposition 3.2. Let $x \in \mathbb{R}^d$ and $R \geq 0$. If $\tilde{f}_\theta^s$ is certified at $P(x) = U^\intercal x$ with radius R, then $g(x+\delta) = g(x)$ for all $\delta \in \Delta^U(R) \subseteq \mathbb{R}^d$, where

$$\Delta^U(R) := \{\delta \in \mathbb{R}^d : \|U^\intercal \delta\| \leq R\}$$

Proof. Let $\delta \in \Delta^U(R)$. Then

$$g(x+\delta) = \arg\max_{y \in \mathcal{Y}} \tilde{f}_\theta^s(P(x+\delta))_y = \arg\max_{y \in \mathcal{Y}} \tilde{f}_\theta^s(P(x) + U^\intercal \delta)_y.$$

Since $\|U^\intercal \delta\| \leq R$ by definition of $\Delta^U(R)$ and $\tilde{f}_\theta^s$ is certified at $P(x)$ with radius R, we have that

$$g(x+\delta) = \arg\max_{y \in \mathcal{Y}} \tilde{f}_\theta^s(P(x))_y = g(x).$$

□

Proposition 3.3. Let $R \geq 0$. The certified region $\Delta^U(R)$ can be expressed as the Minkowski sum $\Delta^U(R) = B_p^U(R) + \mathcal{N}(U^\intercal)$, where $B_p^U(R) \subseteq \mathbb{R}^d$ is a p-dimensional ball embedded into $\mathcal{R}(U)$:

$$B_p^U(R) := \{\beta_1 v_{d-p+1} + \cdots + \beta_p v_d : \|\beta\| \leq R,\ \beta \in \mathbb{R}^p\}.$$

Proof. Let $y = y_1 + y_2$ with $y_1 \in B_p^U(R)$ and $y_2 \in \mathcal{N}(U^\intercal)$. Then

$$\|U^\intercal y\| = \|U^\intercal y_1\| = \|\beta\| \leq R,$$

so $y \in \Delta^U(R)$.

On the other hand, let $y \in \Delta^U(R)$ as defined in Proposition 3.2. We can decompose $y = y_1 + y_2$ for $y_1 \in \mathcal{R}(U)$ and $y_2 \in \mathcal{N}(U^\intercal)$. Then there exists $\beta \in \mathbb{R}^p$ such that $y_1 = U\beta = \sum_{i=d-p+1}^{n} \beta_{i-d+p} v_i$, so $\|U^\intercal y_1\| = \|\beta\|$ and therefore $\|\beta\| \leq R$. □

Corollary 3.5. Let S_k be a k-dimensional linear subspace of $\mathbb{R}^d$ and rC^d be a zero-centered cube of side length $r > 0$. Then $\mathcal{V}_k(rC^d \cap S_k) \geq r^k$.

Proof. Note that

$$\begin{aligned} rC^d \cap S_k &= \{x \in \mathbb{R}^d : \|x\|_\infty \leq r/2,\ x \in S_k\} \\ &= \{rx \in \mathbb{R}^d : \|rx\|_\infty \leq r/2,\ rx \in S_k\} \\ &= \{rx \in \mathbb{R}^d : \|x\|_\infty \leq 1/2,\ x \in S_k\}, \end{aligned}$$

since $x \in S_k$ if and only if $rx \in S_k$, by linearity of S_k. This is now equivalent to the set $r(C^d \cap S_k)$, and we have scaled our k-dimensional subset by a uniform factor r. Therefore, $\mathcal{V}_k(rC^d \cap S_k) = \mathcal{V}_k(r(C^d \cap S_k)) = r^k \mathcal{V}_k(C^d \cap S_k)$ by Folland [1999, Theorem 2.44]. Thus, by Theorem 3.4, we have $\mathcal{V}_k(rC^d \cap S_k) \geq r^k$. □

Corollary 3.6. Let $x \in \mathbb{R}^d$ and let $S_k(x) \subseteq \mathbb{R}^d$ be the k-dimensional affine subspace

$$S_k(x) = \left\{ x + \sum_{i=1}^{k} \alpha_i v_i : \alpha \in \mathbb{R}^k \right\}$$

spanned by arbitrary vectors $v_1, \ldots, v_k$ and passing through x. Let $t \geq 0$ be the minimal ℓ_∞-norm of a point in $S_k(x)$:

$$t := \inf_{x' \in S_k(x)} \|x'\|_\infty = \inf_{\alpha \in \mathbb{R}^k} \left\| x + \textstyle\sum_{i=1}^{k} \alpha_i v_i \right\|_\infty . \tag{3.5}$$

Then, for all $r > 2t$, it holds that $\mathcal{V}_k(rC^d \cap S_k(x)) \geq (r - 2t)^k$.

Proof. First, notice that the infimum in (3.5) is attained since $\|\cdot\|_\infty$ is continuous and coercive, and $S_k(x)$ is closed in the standard topology on $\mathbb{R}^d$ [Bertsekas, 2016]. Let $x^* \in S_k(x)$ be a point that attains the infimum in (3.5) so that $\|x^*\|_\infty = t$. If $r > 2t$, then x^* is contained in the interior of rC^d. In this case, we can construct a nonempty cube centered at x^* with side lengths $r - 2t > 0$ that is contained in rC^d. Now, the plane $S_k(x)$ passes through x^*, and therefore Corollary 3.5 yields the result since volume is preserved under translation [Folland, 1999, Theorem 2.42]. □

Theorem 3.7. Let $x \in C^d$, let t be defined as in (3.5) with $k = d-p$, and let $R \in [0, 1/2-t]$. If $\tilde{f}_\theta^s$ is certified at $P(x) = U^\intercal x$ with radius R, then

$$\mathcal{V}_d(C^d \cap \Delta_x^U(R)) \geq \frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)} R^p (1 - 2R - 2t)^{d-p}. \tag{3.6}$$

Proof. The characterization of $\Delta^U(R)$ in Proposition 3.3 yields

$$\Delta_x^U(R) = B_p^U(R) + S_{d-p}^{\mathcal{N}(U^\intercal)}(x),$$

where

$$S_{d-p}^{\mathcal{N}(U^\intercal)}(x) := \{x\} + \mathcal{N}(U^\intercal)$$

is the affine subspace of $\mathbb{R}^d$ spanned by $\mathcal{N}(U^\intercal)$ and passing through x, which has dimension $d-p$. Therefore, the following is an inner-approximation of $\Delta_x^U(R)$:

$$\tilde{\Delta}_x^U(R) := B_p^U(R) + \left((1-2R)C^d \cap S_{d-p}^{\mathcal{N}(U^\intercal)}(x)\right) \subseteq B_p^U(R) + S_{d-p}^{\mathcal{N}(U^\intercal)}(x) = \Delta_x^U(R).$$

If we can show that $\tilde{\Delta}_x^U(R) \subseteq C^d$, then $\tilde{\Delta}_x^U(R) \subseteq C^d \cap \Delta_x^U(R)$, in which case the volume of $\tilde{\Delta}_x^U(R)$ will lower-bound the volume of $C^d \cap \Delta_x^U(R)$. To prove that this holds, let $y = y_1 + y_2 \in \tilde{\Delta}_x^U(R)$ with $y_1 \in B_p^U(R)$ and $y_2 \in (1-2R)C^d \cap S_{d-p}^{\mathcal{N}(U^\intercal)}(x)$. Then

$$\|y\|_\infty \leq \|y_1\|_\infty + \|y_2\|_\infty \leq R + \frac{1-2R}{2} = \frac{1}{2},$$

by the fact that $\|y_1\|_\infty \leq \|y_1\| = \|U\beta\| = \|\beta\|$ for some $\beta \in \mathbb{R}^p$ with $\|\beta\| \leq R$ due to the semi-orthogonality of U, and by the fact that $y_2 \in (1-2R)C^d$. Therefore, indeed it holds that $\tilde{\Delta}_x^U(R) \subseteq C^d$. Thus, all that remains is to lower-bound $V_d(\tilde{\Delta}_x^U(R))$. To this end, notice that $B_p^U(R) \subseteq \mathcal{R}(U)$ and $(1-2R)C^d \cap S_{d-p}^{\mathcal{N}(U^\intercal)}(x) \subseteq \{x\} + \mathcal{N}(U^\intercal)$, so $B_p^U(R)$ and $(1-2R)C^d \cap S_{d-p}^{\mathcal{N}(U^\intercal)}(x)$ are contained in orthogonal affine subspaces, and therefore $V_d(\tilde{\Delta}_x^U(R)) = V_p(B_p^U(R))V_{d-p}((1-2R)C^d \cap S_{d-p}^{\mathcal{N}(U^\intercal)}(x))$. The p-dimensional volume of the embedded ball ℓ_2-ball $B_p^U(R)$ is well-known (e.g., see Folland [1999, Theorem 2.44, Corollary 2.55]) to be

$$V_p(B_p^U(R)) = \frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)} R^p.$$

On the other hand, since $2R < 1-2t$, it holds that $1-2R > 2t$. Hence Corollary 3.6 gives that the $(d-p)$-dimensional volume of $(1-2R)C^d \cap S_{d-p}^{\mathcal{N}(U^\intercal)}(x)$ is lower-bounded as

$$V_{d-p}((1-2R)C^d \cap S_{d-p}^{\mathcal{N}(U^\intercal)}(x)) \geq (1-2R-2t)^{d-p}.$$

Therefore,

$$V_d(\tilde{\Delta}_x^U(R)) \geq \frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)} R^p (1-2R-2t)^{d-p},$$

which concludes the proof. □

Proposition 3.8. Let t and R be as in Theorem 3.7. The lower bound (3.6) is maximized as follows:

$$r^* := \min\left\{R, \frac{p(1-2t)}{2d}\right\} \in \arg\max_{r \in [0,R]} \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} r^p \left(1-2r-2t\right)^{d-p}. \tag{3.7}$$

Proof. It suffices to maximize $h(r) \coloneqq r^p \left(1 - 2r - 2t\right)^{d-p}$ over $r \in [0, R]$. The gradient of h vanishes at points satisfying

$$\begin{aligned}\frac{dh}{dr}(r) &= pr^{p-1}\left(1 - 2r - 2t\right)^{d-p} - 2(d-p)r^p\left(1 - 2r - 2t\right)^{d-p-1} \\ &= r^{p-1}\left(1 - 2r - 2t\right)^{d-p-1}\Big(p\left(1 - 2r - 2t\right) - 2(d-p)r\Big) \\ &= r^{p-1}\left(1 - 2r - 2t\right)^{d-p-1}\left(p - 2pt - 2dr\right) \\ &= 0.\end{aligned}$$

The set of all critical points satisfying this polynomial equation is $\left\{0, \frac{p(1-2t)}{2d}, 1/2 - t\right\}$. Notice that $0 < \frac{p(1-2t)}{2d} < \frac{p(1-2t)}{2p} = 1/2 - t$, and that $\frac{dh}{dr}(r) \geq 0$ for all $r \in \left[0, \frac{p(1-2t)}{2d}\right]$ whereas $\frac{dh}{dr}(r) \leq 0$ for all $r \in \left[\frac{p(1-2t)}{2d}, 1/2 - t\right]$. Hence, h is unimodal on $[0, 1/2 - t]$ with the maximizer $\frac{p(1-2t)}{2d}$. Therefore, if $R < \frac{p(1-2t)}{2d}$, then h is monotone increasing on the feasible interval $[0, R]$, which implies that the right endpoint $r^* = R$ is a maximizer of (3.7). On the other hand, if $R \geq \frac{p(1-2t)}{2d}$, then $\frac{p(1-2t)}{2d}$ is contained in the feasible interval $[0, R]$, and thus $r^* = \frac{p(1-2t)}{2d}$ is a maximizer of (3.7). □

Appendix C

Asymmetric Certified Robustness via Feature-Convex Neural Networks

C.1 Proofs for Section 4.3

Theorem 4.3. Let $f \in \mathcal{F}$ be as in Definition 4.1 and let $x \in f^{-1}(\{1\}) = \{x' \in \mathbb{R}^d : f(x') = 1\}$. If $\nabla g(\varphi(x)) \in \mathbb{R}^q$ is a nonzero subgradient of the convex function g at $\varphi(x)$, then $f(x+\delta) = 1$ for all $\delta \in \mathbb{R}^d$ such that

$$\|\delta\|_p < r(x) \coloneqq \frac{g(\varphi(x))}{\mathrm{Lip}_p(\varphi)\|\nabla g(\varphi(x))\|_{p,*}}.$$

Proof. Suppose that $\nabla g(\varphi(x)) \in \mathbb{R}^q$ is a nonzero subgradient of g at $\varphi(x)$, so that $g(y) \geq g(\varphi(x)) + \nabla g(\varphi(x))^\top (y - \varphi(x))$ for all $y \in \mathbb{R}^q$. Let $\delta \in \mathbb{R}^d$ be such that $\|\delta\|_p < r(x)$. Then it holds that

$$\begin{aligned} g(\varphi(x+\delta)) &\geq g(\varphi(x)) + \nabla g(\varphi(x))^\top (\varphi(x+\delta) - \varphi(x)) \\ &\geq g(\varphi(x)) - \|\nabla g(\varphi(x))\|_{p,*} \|\varphi(x+\delta) - \varphi(x)\|_p \\ &\geq g(\varphi(x)) - \|\nabla g(\varphi(x))\|_{p,*} \,\mathrm{Lip}_p(\varphi) \|\delta\|_p \\ &> 0, \end{aligned}$$

so indeed $f(x+\delta) = 1$. □

We now introduce a preliminary lemma for the results in Section 4.3.2.

Lemma C.1. For any nonempty closed convex set $X \subseteq \mathbb{R}^d$, there exists a convex function $g \colon \mathbb{R}^d \to \mathbb{R}$ such that $X = g^{-1}((-\infty, 0]) = \{x \in \mathbb{R}^d : g(x) \leq 0\}$.

Proof. Let $X \subseteq \mathbb{R}^d$ be a nonempty closed convex set. We take the distance function $g = d_X$ defined by $d_X(x) = \inf_{y \in X} \|y - x\|_2$. Since X is closed and $y \mapsto \|y - x\|_2$ is coercive for all $x \in \mathbb{R}^d$, it holds that $y \mapsto \|y - x\|_2$ attains its infimum over X [Bertsekas, 2016,

Proposition A.8]. Let $x^{(1)}, x^{(2)} \in \mathbb{R}^d$ and let $\theta \in [0,1]$. Then there exist $y^{(1)}, y^{(2)} \in X$ such that $g(x^{(1)}) = \|y^{(1)} - x^{(1)}\|_2$ and $g(x^{(2)}) = \|y^{(2)} - x^{(2)}\|_2$. Since X is convex, it holds that $\theta y^{(1)} + (1-\theta)y^{(2)} \in X$, and therefore

$$\begin{aligned} g(\theta x^{(1)} + (1-\theta)x^{(2)}) &= \inf_{y \in X} \|y - (\theta x^{(1)} + (1-\theta)x^{(2)})\|_2 \\ &\leq \|\theta y^{(1)} + (1-\theta)y^{(2)} - (\theta x^{(1)} + (1-\theta)x^{(2)})\|_2 \\ &\leq \theta \|y^{(1)} - x^{(1)}\|_2 + (1-\theta)\|y^{(2)} - x^{(2)}\|_2 \\ &= \theta g(x^{(1)}) + (1-\theta) g(x^{(2)}). \end{aligned}$$

Hence, $g = d_X$ is convex. Since $X = \{x \in \mathbb{R}^d : \inf_{y \in X} \|y - x\|_2 = 0\} = \{x \in \mathbb{R}^d : d_X(x) = 0\} = \{x \in \mathbb{R}^d : d_X(x) \leq 0\} = \{x \in \mathbb{R}^d : g(x) \leq 0\}$ by nonnegativity of d_X, the lemma holds. □

Proposition 4.5. For any nonempty closed convex set $X \subseteq \mathbb{R}^d$, there exists $f \in \mathcal{F}_{\mathrm{Id}}$ such that $X = f^{-1}(\{2\}) = \{x \in \mathbb{R}^d : f(x) = 2\}$. In particular, this shows that if (X_1, X_2) is a convexly separable pair of subsets of $\mathbb{R}^d$, then there exists $f \in \mathcal{F}_{\mathrm{Id}}$ such that $f(x) = 1$ for all $x \in X_1$ and $f(x) = 2$ for all $x \in X_2$.

Proof. Let $X \subseteq \mathbb{R}^d$ be a nonempty closed convex set. By Lemma C.1, there exists a convex function $g \colon \mathbb{R}^d \to \mathbb{R}$ such that $X = \{x \in \mathbb{R}^d : g(x) \leq 0\}$. Define $f \colon \mathbb{R}^d \to \{1, 2\}$ by $f(x) = 1$ if $g(x) > 0$ and $f(x) = 2$ if $g(x) \leq 0$. Clearly, it holds that $f \in \mathcal{F}_{\mathrm{Id}}$. Furthermore, for all $x \in X$ it holds that $g(x) \leq 0$, implying that $f(x) = 2$ for all $x \in X$. Conversely, if $x \in \mathbb{R}^d$ is such that $f(x) = 2$, then $g(x) \leq 0$, implying that $x \in X$. Hence, $X = \{x \in \mathbb{R}^d : f(x) = 2\}$.

If (X_1, X_2) is a convexly separable pair of subsets of $\mathbb{R}^d$, then there exists a nonempty closed convex set $X \subseteq \mathbb{R}^d$ such that $X_2 \subseteq X$ and $X_1 \subseteq \mathbb{R}^d \setminus X$, and therefore there exists $f \in \mathcal{F}_{\mathrm{Id}}$ such that $X_2 \subseteq X = f^{-1}(\{2\})$ and $X_1 \subseteq \mathbb{R}^d \setminus X = f^{-1}(\{1\})$, implying that indeed $f(x) = 1$ for all $x \in X_1$ and $f(x) = 2$ for all $x \in X_2$. □

Proposition 4.6. Let $f \in \mathcal{F}_{\mathrm{Id}}$. The decision region under f associated to class 2, namely $X \coloneqq f^{-1}(\{2\}) = \{x \in \mathbb{R}^d : f(x) = 2\}$, is a closed convex set.

Proof. For all $x \in \mathbb{R}^d$, it holds that $f(x) = 2$ if and only if $g(x) \leq 0$. Since $f \in \mathcal{F}_{\mathrm{Id}}$, g is convex, and hence, $X = \{x \in \mathbb{R}^d : g(x) \leq 0\}$ is a (nonstrict) sublevel set of a convex function and is therefore a closed convex set. □

In order to apply the universal approximation results in Chen et al. [2019], we now introduce their parameterization of input-convex ReLU neural networks. Note that it imposes the additional constraint that the first weight matrix $A^{(1)}$ is elementwise nonnegative.

Definition C.2. Define $\tilde{\mathcal{F}}_{\mathrm{Id}}$ to be the class of functions $\tilde{f}\colon \mathbb{R}^d \to \{1,2\}$ given by $\tilde{f}(x) = T(\tilde{g}(x))$ with $\tilde{g}\colon \mathbb{R}^d \to \mathbb{R}$ given by

$$\begin{aligned} x^{(1)} &= \mathrm{ReLU}\left(A^{(1)}x + b^{(1)}\right), \\ x^{(l)} &= \mathrm{ReLU}\left(A^{(l)}x^{(l-1)} + b^{(l)} + C^{(l)}x\right),\ l \in \{2,3,\dots,L-1\}, \\ \tilde{g}(x) &= A^{(L)}x^{(L-1)} + b^{(L)} + C^{(L)}x, \end{aligned}$$

for some $L \in \mathbb{N}$, $L > 1$, and some consistently sized matrices $A^{(1)}, C^{(1)}, \dots, A^{(L)}, C^{(L)}$, all of which have nonnegative elements, and some consistently sized vectors $b^{(1)}, \dots, b^{(L)}$.

The following preliminary lemma relates the class $\hat{\mathcal{F}}_{\mathrm{Id}}$ from Definition 4.2 to the class $\tilde{\mathcal{F}}_{\mathrm{Id}}$ above.

Lemma C.3. It holds that $\tilde{\mathcal{F}}_{\mathrm{Id}} \subseteq \hat{\mathcal{F}}_{\mathrm{Id}}$.

Proof. Let $\tilde{f} \in \tilde{\mathcal{F}}_{\mathrm{Id}}$. Then certainly $A^{(l)} \geq 0$ for all $l \in \{2,3,\dots,L\}$, so indeed $\tilde{f} \in \hat{\mathcal{F}}_{\mathrm{Id}}$. Hence, $\tilde{\mathcal{F}}_{\mathrm{Id}} \subseteq \hat{\mathcal{F}}_{\mathrm{Id}}$. □

Theorem 1 in Chen et al. [2019] shows that a Lipschitz convex function can be approximated within an arbitrary tolerance. We now provide a technical lemma adapting Theorem 1 in Chen et al. [2019] to show that convex functions can be *underapproximated* within an arbitrary tolerance on a compact convex subset.

Lemma C.4. For any convex function $g\colon \mathbb{R}^d \to \mathbb{R}$, any compact convex subset X of $\mathbb{R}^d$, and any $\epsilon > 0$, there exists $\hat{f} \in \hat{\mathcal{F}}_{\mathrm{Id}}$ such that $\hat{g}(x) < g(x)$ for all $x \in X$ and $\sup_{x\in X}(g(x) - \hat{g}(x)) < \epsilon$.

Proof. Let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function, let X be a compact convex subset of $\mathbb{R}^d$, and let $\epsilon > 0$. Since $g - \epsilon/2$ is a real-valued convex function on $\mathbb{R}^d$ (and hence is proper), its restriction to the closed and bounded set X is Lipschitz continuous [Rockafellar, 1970, Theorem 10.4], and therefore Lemma C.3 together with Theorem 1 in Chen et al. [2019] gives that there exists $\hat{f} \in \tilde{\mathcal{F}}_{\mathrm{Id}} \subseteq \hat{\mathcal{F}}_{\mathrm{Id}}$ such that $\sup_{x\in X}|(g(x) - \epsilon/2) - \hat{g}(x)| < \epsilon/2$. Thus, for all $x \in X$,

$$\begin{aligned} g(x) - \hat{g}(x) &= \left(g(x) - \frac{\epsilon}{2}\right) - \hat{g}(x) + \frac{\epsilon}{2} \\ &> \left(g(x) - \frac{\epsilon}{2}\right) - \hat{g}(x) + \sup_{y\in X}\left|\left(g(y) - \frac{\epsilon}{2}\right) - \hat{g}(y)\right| \\ &\geq \left(g(x) - \frac{\epsilon}{2}\right) - \hat{g}(x) + \left|\left(g(x) - \frac{\epsilon}{2}\right) - \hat{g}(x)\right| \\ &\geq 0. \end{aligned}$$

Furthermore,

$$\begin{aligned}\sup_{x\in X}\left(g(x)-\hat{g}(x)\right) &= \sup_{x\in X}\left|g(x)-\hat{g}(x)\right| \\ &= \sup_{x\in X}\left|\left(g(x)-\frac{\epsilon}{2}\right)-\hat{g}(x)+\frac{\epsilon}{2}\right| \\ &\leq \sup_{x\in X}\left|\left(g(x)-\frac{\epsilon}{2}\right)-\hat{g}(x)\right|+\frac{\epsilon}{2} \\ &< \epsilon,\end{aligned}$$

which proves the lemma. □

We leverage Lemma C.4 to construct a uniformly converging sequence of underapproximating functions.

Lemma C.5. For all $f \in \mathcal{F}_{\text{Id}}$ and all compact convex subsets X of $\mathbb{R}^d$, there exists a sequence $\{\hat{f}_n \in \hat{\mathcal{F}}_{\text{Id}} : n \in \mathbb{N}\} \subseteq \hat{\mathcal{F}}_{\text{Id}}$ such that $\hat{g}_n(x) < \hat{g}_{n+1}(x) < g(x)$ for all $x \in X$ and all $n \in \mathbb{N}$ and $\hat{g}_n$ converges uniformly to g on X as $n \to \infty$.

Proof. Let $f \in \mathcal{F}_{\text{Id}}$ and let X be a compact convex subset of $\mathbb{R}^d$. Let $\{\epsilon_n > 0 : n \in \mathbb{N}\}$ be a sequence such that $\epsilon_{n+1} < \epsilon_n$ for all $n \in \mathbb{N}$ and $\epsilon_n \to 0$ as $n \to \infty$. Such a sequence clearly exists, e.g., by taking $\epsilon_n = 1/n$ for all $n \in \mathbb{N}$. Now, for all $n \in \mathbb{N}$, the function $g - \epsilon_{n+1}$ is convex, and therefore by Lemma C.4 there exists $\hat{f}_n \in \hat{\mathcal{F}}_{\text{Id}}$ such that $\hat{g}_n(x) < g(x) - \epsilon_{n+1}$ for all $x \in X$ and $\sup_{x\in X}\left((g(x)-\epsilon_{n+1})-\hat{g}_n(x)\right) < \epsilon_n - \epsilon_{n+1}$. Fixing such $\hat{f}_n, \hat{g}_n$ for all $n \in \mathbb{N}$, we see that $\sup_{x\in X}\left((g(x)-\epsilon_{n+2})-\hat{g}_{n+1}(x)\right) < \epsilon_{n+1} - \epsilon_{n+2}$, which implies that

$$\hat{g}_{n+1}(x) > g(x) - \epsilon_{n+1} > \hat{g}_n(x)$$

for all $x \in X$, which proves the first inequality. The second inequality comes from the fact that $\hat{g}_{n+1}(x) < g(x) - \epsilon_{n+2} < g(x)$ for all $x \in X$. Finally, since $g(x) - \hat{g}_n(x) > \epsilon_{n+1} > 0$ for all $x \in X$ and all $n \in \mathbb{N}$, we see that

$$\sup_{x\in X}\left|g(x)-\hat{g}_n(x)\right| = \sup_{x\in X}\left(g(x)-\hat{g}_n(x)\right) < \epsilon_n \to 0 \text{ as } n \to \infty,$$

which proves that $\lim_{n\to\infty}\sup_{x\in X}\left|g(x)-\hat{g}_n(x)\right| = 0$, so indeed $\hat{g}_n$ converges uniformly to g on X as $n \to \infty$. □

With all the necessary lemmas in place, we now present our main theoretical results.

Theorem 4.7. For any $f \in \mathcal{F}_{\text{Id}}$, any compact convex subset X of $\mathbb{R}^d$, and any $\epsilon > 0$, there exists $\hat{f} \in \hat{\mathcal{F}}_{\text{Id}}$ such that $m(\{x \in X : \hat{f}(x) \neq f(x)\}) < \epsilon$.

Proof. Let $f \in \mathcal{F}_{\text{Id}}$ and let X be a compact convex subset of $\mathbb{R}^d$. By Lemma C.5, there exists a sequence $\{\hat{f}_n \in \hat{\mathcal{F}}_{\text{Id}} : n \in \mathbb{N}\} \subseteq \hat{\mathcal{F}}_{\text{Id}}$ such that $\hat{g}_n(x) < \hat{g}_{n+1}(x) < g(x)$ for all $x \in X$ and all $n \in \mathbb{N}$ and $\hat{g}_n$ converges uniformly to g on X as $n \to \infty$. Fix this sequence.

For all $n \in \mathbb{N}$, define

$$E_n = \{x \in X : \hat{f}_n(x) \neq f(x)\},$$

i.e., the set of points in X for which the classification under $\hat{f}_n$ does not agree with that under f. Since $\hat{g}_n(x) < g(x)$ for all $x \in X$ and all $n \in \mathbb{N}$, we see that

$$\begin{aligned} E_n &= \{x \in X : \hat{g}_n(x) > 0 \text{ and } g(x) \leq 0\} \cup \{x \in X : \hat{g}_n(x) \leq 0 \text{ and } g(x) > 0\} \\ &= \{x \in X : \hat{g}_n(x) \leq 0 \text{ and } g(x) > 0\}. \end{aligned}$$

Since g is a real-valued convex function on $\mathbb{R}^d$, it is continuous [Rockafellar, 1970, Corollary 10.1.1], and therefore $g^{-1}((0,\infty)) = \{x \in \mathbb{R}^d : g(x) > 0\}$ is measurable. Similarly, $\hat{g}_n^{-1}((-\infty, 0]) = \{x \in \mathbb{R}^d : \hat{g}_n(x) \leq 0\}$ is also measurable for all $n \in \mathbb{N}$ since $\hat{g}_n$ is continuous. Furthermore, X is measurable as it is compact. Therefore, E_n is measurable for all $n \in \mathbb{N}$. Now, since $\hat{g}_n(x) < \hat{g}_{n+1}(x)$ for all $x \in X$ and all $n \in \mathbb{N}$, it holds that $E_{n+1} \subseteq E_n$ for all $n \in \mathbb{N}$. It is clear that to prove the result, it suffices to show that $\lim_{n\to\infty} m(E_n) = 0$. Therefore, if we show that $m\left(\bigcap_{n\in\mathbb{N}} E_n\right) = 0$, then the fact that $m(E_1) \leq m(X) < \infty$ together with Lebesgue measure's continuity from above yields that $\lim_{n\to\infty} m(E_n) = 0$, thereby proving the result.

It remains to be shown that $m\left(\bigcap_{n\in\mathbb{N}} E_n\right) = 0$. To this end, suppose for the sake of contradiction that $\bigcap_{n\in\mathbb{N}} E_n \neq \emptyset$. Then there exists $x \in \bigcap_{n\in\mathbb{N}} E_n$, meaning that $g(x) > 0$ and $\hat{g}_n(x) \leq 0$ for all $n \in \mathbb{N}$. Thus, for this $x \in X$, we find that $\limsup_{n\to\infty} \hat{g}_n(x) \leq 0 < g(x)$, which contradicts the fact that $\hat{g}_n$ uniformly converges to g on X. Therefore, it must be that $\bigcap_{n\in\mathbb{N}} E_n = \emptyset$, and thus $m\left(\bigcap_{n\in\mathbb{N}} E_n\right) = 0$, which concludes the proof. □

Theorem 4.8. If (X_1, X_2) is a convexly separable pair of finite subsets of $\mathbb{R}^d$, then there exists $\hat{f} \in \hat{\mathcal{F}}_{\mathrm{Id}}$ such that $\hat{f}(x) = 1$ for all $x \in X_1$ and $\hat{f}(x) = 2$ for all $x \in X_2$.

Proof. Throughout this proof, we denote the complement of a set $Y \subseteq \mathbb{R}^d$ by $Y^c = \mathbb{R}^d \setminus Y$.

Suppose that $X_1 = \{x^{(1)}, \ldots, x^{(M)}\} \subseteq \mathbb{R}^d$ and $X_2 = \{y^{(1)}, \ldots, y^{(N)}\} \subseteq \mathbb{R}^d$ are such that (X_1, X_2) is convexly separable. Then, by definition of convex separability, there exists a nonempty closed convex set $X' \subseteq \mathbb{R}^d$ such that $X_2 \subseteq X'$ and $X_1 \subseteq \mathbb{R}^d \setminus X'$. Let $X = X' \cap \mathrm{conv}(X_2)$. Since $X_2 \subseteq X'$ and both sets X' and $\mathrm{conv}(X_2)$ are convex, the set X is nonempty and convex. By finiteness of X_2, the set $\mathrm{conv}(X_2)$ is compact, and therefore by closedness of X', the set X is compact and hence closed.

By Proposition 4.5, there exists $f \in \mathcal{F}_{\mathrm{Id}}$ such that $f^{-1}(\{2\}) = X$. Since $\mathrm{conv}(X_1 \cup X_2)$ is compact and convex, Lemma C.5 gives that there exists a sequence $\{\hat{f}_n \in \hat{\mathcal{F}}_{\mathrm{Id}} : n \in \mathbb{N}\} \subseteq \hat{\mathcal{F}}_{\mathrm{Id}}$ such that $\hat{g}_n(x) < \hat{g}_{n+1}(x) < g(x)$ for all $x \in \mathrm{conv}(X_1 \cup X_2)$ and all $n \in \mathbb{N}$ and $\hat{g}_n$ converges uniformly to g on $\mathrm{conv}(X_1 \cup X_2)$ as $n \to \infty$. Fix this sequence.

Let $x \in X_2$. Then, since $X_2 \subseteq X'$ and $X_2 \subseteq \mathrm{conv}(X_2)$, it holds that $x \in X' \cap \mathrm{conv}(X_2) = X = f^{-1}(\{2\})$, implying that $f(x) = 2$ and hence $g(x) \leq 0$. Since $\hat{g}_n(x) < g(x)$ for all $n \in \mathbb{N}$, this shows that $\hat{f}_n(x) = 2$ for all $n \in \mathbb{N}$. On the other hand, let $i \in \{1, \ldots, M\}$ and consider $x = x^{(i)} \in X_1$. Since $X_1 \subseteq \mathbb{R}^d \setminus X' = \mathbb{R}^d \cap (X')^c \subseteq \mathbb{R}^d \cap (X' \cap \mathrm{conv}(X_2))^c =$

$\mathbb{R}^d \cap X^c = \mathbb{R}^d \cap f^{-1}(\{1\})$, it holds that $f(x) = 1$ and thus $g(x) > 0$. Suppose for the sake of contradiction that $\hat{f}_n(x) = 2$ for all $n \in \mathbb{N}$. Then $\hat{g}_n(x) \leq 0$ for all $n \in \mathbb{N}$. Therefore, for this $x \in X_1$, we find that $\limsup_{n\to\infty} \hat{g}_n(x) \leq 0 < g(x)$, which contradicts the fact that $\hat{g}_n$ uniformly converges to g on $\operatorname{conv}(X_1 \cup X_2)$. Therefore, it must be that there exists $n_i \in \mathbb{N}$ such that $\hat{f}_{n_i}(x) = 1$, and thus $\hat{g}_{n_i}(x) > 0$. Since $\hat{g}_n(x) < \hat{g}_{n+1}(x)$ for all $n \in \mathbb{N}$, this implies that $\hat{g}_n(x) > 0$ for all $n \geq n_i$. Hence, $\hat{f}_n(x) = \hat{f}_n(x^{(i)}) = 1$ for all $n \geq n_i$.

Let $n^\star$ be the maximum of all such n_i, i.e., $n^\star = \max\{n_i : i \in \{1, \dots, M\}\}$. Then the above analysis shows that $\hat{f}_{n^\star}(x) = 2$ for all $x \in X_2$ and that $\hat{f}_{n^\star}(x) = 1$ for all $x \in X_1$. Since $\hat{f}_{n^\star} \in \hat{\mathcal{F}}_{\mathrm{Id}}$, the claim has been proven. □

Theorem 4.11. Consider $M, N \in \mathbb{N}$. Let $X_1 = \{x^{(1)}, \dots, x^{(M)}\} \subseteq \mathbb{R}^d$ and $X_2 = \{y^{(1)}, \dots, y^{(N)}\} \subseteq \mathbb{R}^d$ be samples with all elements $x_k^{(i)}, y_l^{(j)}$ drawn independently and identically from the uniform probability distribution on $[-1, 1]$. Then, it holds that

$$\mathbb{P}\left((X_1, X_2) \text{ is convexly separable}\right) \geq \begin{cases} 1 - \left(1 - \frac{M!N!}{(M+N)!}\right)^d & \text{for all } d \in \mathbb{N}, \\ 1 & \text{if } d \geq M + N. \end{cases} \tag{4.2}$$

In particular, $\hat{\mathcal{F}}_{\mathrm{Id}}$ contains an input-convex ReLU neural network that classifies all $x^{(i)}$ into class 1 and all $y^{(j)}$ into class 2 almost surely for sufficiently large dimensions d.

Proof. Throughout the proof, we denote the cardinality of a set S by $|S|$. For the reader's convenience, we also recall that, for $n \in \mathbb{N}$, the symmetric group S_n consists of all permutations (i.e., bijections) on the set $\{1, 2, \dots, n\}$, and that $|S_n| = n!$. If $\sigma\colon \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ is a permutation in S_n, we denote the restriction of σ to the domain $I \subseteq \{1, 2, \dots, n\}$ by $\sigma|_I\colon I \to \{1, 2, \dots, n\}$, which we recall is defined by $\sigma|_I(i) = \sigma(i)$ for all $i \in I$, and is not necessarily a permutation on I in general.

Consider first the case where $d \geq M + N$. Let $b \in \mathbb{R}^{M+N}$ be the vector defined by $b_i = 1$ for all $i \in \{1, \dots, M\}$ and $b_i = -1$ for all $i \in \{M + 1, \dots, M + N\}$. Then, since $x_k^{(i)}, y_l^{(j)}$ are independent uniformly distributed random variables on $[-1, 1]$, it holds that the matrix

$$\begin{bmatrix} {x^{(1)}}^\top \\ \vdots \\ {x^{(M)}}^\top \\ {y^{(1)}}^\top \\ \vdots \\ {y^{(N)}}^\top \end{bmatrix} \in \mathbb{R}^{(M+N)\times d}$$

has rank $M + N$ almost surely, and therefore the linear system of equations

$$\begin{bmatrix} {x^{(1)}}^\top \\ \vdots \\ {x^{(M)}}^\top \\ {y^{(1)}}^\top \\ \vdots \\ {y^{(N)}}^\top \end{bmatrix} a = b$$

has a solution $a \in \mathbb{R}^d$ with probability 1, and we note that from this solution we find that X_2 is a subset of the nonempty closed convex set $\{x \in \mathbb{R}^d : a^\top x \leq 0\}$ and that X_1 is a subset of its complement. Hence, (X_1, X_2) is convexly separable with probability 1 in this case.

Now let us consider the general case: $d \in \mathbb{N}$ and in general it may be the case that $d < M + N$. For notational convenience, let P be the probability of interest:

$$P = \mathbb{P}\Big((X_1, X_2) \text{ is convexly separable}\Big).$$

Suppose that there exists a coordinate $k \in \{1, 2, \ldots, d\}$ such that $x_k^{(i)} < y_k^{(j)}$ for all pairs $(i, j) \in \{1, 2, \ldots, M\} \times \{1, 2, \ldots, N\}$ and that $a \coloneqq \min\{y_k^{(1)}, \ldots, y_k^{(N)}\} < \max\{y_k^{(1)}, \ldots, y_k^{(N)}\} =: b$. Then, let $X = \{x \in \mathbb{R}^d : x_k \in [a, b]\}$. That is, X is the extrusion of the convex hull of the projections $\{y_k^{(1)}, \ldots, y_k^{(N)}\}$ along all remaining coordinates. The set X is a nonempty closed convex set, and it is clear by our supposition that $X_2 \subseteq X$ and $X_1 \subseteq \mathbb{R}^d \setminus X$. Therefore, the supposition implies that (X_1, X_2) is convexly separable, and thus

$$\begin{aligned} P &\geq \mathbb{P}\Big(\text{there exists } k \in \{1, 2, \ldots, d\} \text{ such that } x_k^{(i)} < y_k^{(j)} \text{ for all pairs } (i, j) \\ &\quad \text{and that } \min\{y_k^{(1)}, \ldots, y_k^{(N)}\} < \max\{y_k^{(1)}, \ldots, y_k^{(N)}\}\Big) \\ &= 1 - \mathbb{P}\Big(\text{for all } k \in \{1, 2, \ldots, d\}, \text{ it holds that } x_k^{(i)} \geq y_k^{(j)} \text{ for some pair } (i, j) \\ &\quad \text{or that } \min\{y_k^{(1)}, \ldots, y_k^{(N)}\} = \max\{y_k^{(1)}, \ldots, y_k^{(N)}\}\Big) \\ &= 1 - \prod_{k=1}^{d} \mathbb{P}\left(x_k^{(i)} \geq y_k^{(j)} \text{ for some } (i, j) \text{ or } \min\{y_k^{(1)}, \ldots, y_k^{(N)}\} = \max\{y_k^{(1)}, \ldots, y_k^{(N)}\}\right), \end{aligned}$$

where the final equality follows from the independence of the coordinates of the samples.

Since $\min\{y_k^{(1)}, \ldots, y_k^{(N)}\} < \max\{y_k^{(1)}, \ldots, y_k^{(N)}\}$ almost surely, we find that

$$\begin{aligned} P &\geq 1 - \prod_{k=1}^{d} \Bigg(\mathbb{P}(x_k^{(i)} \geq y_k^{(j)} \text{ for some pair } (i,j)) \\ &\qquad + \mathbb{P}(\min\{y_k^{(1)}, \ldots, y_k^{(N)}\} = \max\{y_k^{(1)}, \ldots, y_k^{(N)}\}) \Bigg) \\ &= 1 - \prod_{k=1}^{d} \mathbb{P}(x_k^{(i)} \geq y_k^{(j)} \text{ for some pair } (i,j)) \\ &= 1 - \prod_{k=1}^{d} \Big(1 - \mathbb{P}(x_k^{(i)} < y_k^{(j)} \text{ for all pairs } (i,j))\Big) \\ &= 1 - \prod_{k=1}^{d} \left(1 - \mathbb{P}\left(\max_{i \in \{1,2,\ldots,M\}} x_k^{(i)} < \min_{j \in \{1,2,\ldots,N\}} y_k^{(j)}\right)\right) \\ &= 1 - \prod_{k=1}^{d} \left(1 - \mathbb{P}\left((x_k^{(1)}, \ldots, x_k^{(M)}, y_k^{(1)}, \ldots, y_k^{(N)}) \in \bigcup_{\sigma \in S} E_\sigma\right)\right), \end{aligned} \tag{C.1}$$

where we define S to be the set of permutations on $\{1, \ldots, M+N\}$ whose restriction to $\{1, \ldots, M\}$ is also a permutation;

$$S = \left\{\sigma \in S_{M+N} : \sigma|_{\{1,\ldots,M\}} \in S_M\right\},$$

and where, for a permutation $\sigma \in S_{M+N}$, E_σ is the event where an $(M+N)$-vector has indices ordered according to σ;

$$E_\sigma = \{z \in \mathbb{R}^{M+N} : z_{\sigma(1)} < \cdots < z_{\sigma(M+N)}\}.$$

We note that the final equality in (C.1) relies on the fact that $\mathbb{P}(x_k^{(i)} = x_k^{(i')}) = \mathbb{P}(y_k^{(j)} = y_k^{(j')}) = 0$ for all $i' \neq i$ and all $j' \neq j$, which is specific to our uniform distribution at hand.

Now, since $E_\sigma, E_{\sigma'}$ are disjoint for distinct permutations $\sigma, \sigma' \in S_{M+N}$, the bound (C.1) gives that

$$P \geq 1 - \prod_{k=1}^{d} \left(1 - \sum_{\sigma \in S} \mathbb{P}((x_k^{(1)}, \ldots, x_k^{(M)}, y_k^{(1)}, \ldots, y_k^{(N)}) \in E_\sigma)\right). \tag{C.2}$$

Since $x_k^{(1)}, \ldots, x_k^{(M)}, y_k^{(1)}, \ldots, y_k^{(N)}$ are independent and identically distributed samples, they define an exchangeable sequence of random variables, implying that

$$\mathbb{P}((x_k^{(1)}, \ldots, x_k^{(M)}, y_k^{(1)}, \ldots, y_k^{(N)}) \in E_\sigma) = \mathbb{P}(x_k^{(1)} < \cdots < x_k^{(M)} < y_k^{(1)} < \cdots < y_k^{(N)})$$

for all permutations $\sigma \in S_{M+N}$. Since, under the uniform distribution at hand,

$$(x_k^{(1)}, \ldots, x_k^{(M)}, y_k^{(1)}, \ldots, y_k^{(N)}) \in E_\sigma$$

for some $\sigma \in S_{M+N}$ almost surely, it holds that

$$\begin{aligned} 1 &= \mathbb{P}\left((x_k^{(1)}, \dots, x_k^{(M)}, y_k^{(N)}, \dots, y_k^{(N)}) \in \bigcup_{\sigma \in S_{M+N}} E_\sigma\right) \\ &= \sum_{\sigma \in S_{M+N}} \mathbb{P}((x_k^{(1)}, \dots, x_k^{(M)}, y_k^{(1)}, \dots, y_k^{(N)}) \in E_\sigma) \\ &= |S_{M+N}|\, \mathbb{P}(x_k^{(1)} < \cdots x_k^{(M)} < y_k^{(1)} < \cdots < y_k^{(N)}). \end{aligned}$$

This implies that

$$\mathbb{P}((x_k^{(1)}, \dots, x_k^{(M)}, y_k^{(1)}, \dots, y_k^{(N)}) \in E_\sigma) = \frac{1}{|S_{M+N}|} = \frac{1}{(M+N)!}$$

for all permutations $\sigma \in S_{M+N}$. Hence, our bound (C.2) becomes

$$P \geq 1 - \prod_{k=1}^{d} \left(1 - \frac{|S|}{(M+N)!}\right) = 1 - \left(1 - \frac{|S|}{(M+N)!}\right)^d.$$

Finally, we immediately see that that map $\Gamma\colon S_M \times S_N \to S_{M+N}$ defined by

$$\Gamma(\sigma, \sigma')(i) = \begin{cases} \sigma(i) & \text{if } i \in \{1, \dots, M\}, \\ \sigma'(i-M) + M & \text{if } i \in \{M+1, \dots, M+N\}, \end{cases}$$

is injective and has image S, implying that $|S| = |S_M \times S_N| = |S_M||S_N| = M!N!$. Thus,

$$P \geq 1 - \left(1 - \frac{M!N!}{(M+N)!}\right)^d,$$

which proves (4.2).

The unit probability of $\hat{\mathcal{F}}_{\mathrm{Id}}$ containing a classifier that classifies all $x^{(i)}$ into class 1 and all $y^{(j)}$ into class 2 for large d follows immediately from Theorem 4.8. □

Appendix D

Transport of Algebraic Structure to Latent Embeddings

D.1 Proofs for Section 6.3

Proposition 6.1. Suppose that $L, M = \mathbb{R}^l$ and that $\varphi\colon L \to M$ is a bijection. Let $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ be an algebra of type $\mathcal{F}$ and define the family $\mathcal{F}^{\mathcal{L}} \coloneqq \{f^{\mathcal{L}} : f \in \mathcal{F}\}$ of n-ary operations on L by (6.1). Then, φ is an isomorphism from the induced algebra $\mathcal{L} = (L, \mathcal{F}^{\mathcal{L}})$ to $\mathcal{M}$.

Proof. Let $f \in \mathcal{F}$, let $n = \mathrm{ar}(f)$, and consider the realization $f^{\mathcal{M}}$ on M and the realization $f^{\mathcal{L}}$ on L induced by (6.1). Let $z_1, \ldots, z_n \in L$. We have that

$$\begin{aligned}\varphi(f^{\mathcal{L}}(z_1, \ldots, z_n)) &= \varphi(\varphi^{-1}(f^{\mathcal{M}}(\varphi(z_1), \ldots, \varphi(z_n)))) \\ &= f^{\mathcal{M}}(\varphi(z_1), \ldots, \varphi(z_n))\end{aligned}$$

by construction of the operation $f^{\mathcal{L}}$. Hence, we see that φ is an isomorphism from the induced algebra $\mathcal{L}$ to the algebra $\mathcal{M}$. □

Theorem 6.2. Consider a source algebra $\mathcal{S} = (S, \mathcal{F}^{\mathcal{S}})$ of type $\mathcal{F}$, and let $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ be a mirrored space such that every law R satisfied by $\mathcal{S}$ is also satisfied by $\mathcal{M}$. Then, the induced latent algebra $\mathcal{L}$, defined by (6.1), also satisfies every such law R, for any bijection $\varphi\colon L \to M$.

Proof. Let $\varphi\colon L \to M$ be a bijection, and let $\mathcal{L}$ be the induced latent algebra defined by (6.1). Let R be a law that is satisfied by $\mathcal{S}$ (and hence satisfied by $\mathcal{M}$), given by

$$p(x_1, \ldots, x_n) = q(x_1, \ldots, x_n).$$

By Proposition 6.1, φ is an isomorphism from $\mathcal{L}$ to $\mathcal{M}$. Let $f \in \mathcal{F}$ be an arbitrary operation symbol. Then, by the properties of isomorphisms, it must be that

$$\varphi(f^{\mathcal{L}}(z_1, \ldots, z_n)) = f^{\mathcal{M}}(\varphi(z_1), \ldots, \varphi(z_n))$$

for all $z_1, \ldots, z_n \in L$. Thus, it holds that

$$\varphi(p^{\mathcal{L}}(z_1, \ldots, z_n)) = p^{\mathcal{M}}(\varphi(z_1), \ldots, \varphi(z_n))$$

for all $z_1, \ldots, z_n \in L$, and similarly,

$$\varphi(q^{\mathcal{L}}(z_1, \ldots, z_n)) = q^{\mathcal{M}}(\varphi(z_1), \ldots, \varphi(z_n))$$

for all such $z_1, \ldots, z_n$. Therefore, since $\mathcal{M}$ satisfies the law R, we conclude that

$$\varphi(p^{\mathcal{L}}(z_1, \ldots, z_n)) = \varphi(q^{\mathcal{L}}(z_1, \ldots, z_n))$$

for all $z_1, \ldots, z_n \in L$. Hence, by invertibility of φ, we also find that

$$p^{\mathcal{L}}(z_1, \ldots, z_n) = q^{\mathcal{L}}(z_1, \ldots, z_n)$$

for all $z_1, \ldots, z_n \in L$, and therefore $\mathcal{L}$ satisfies the law R. □

Proposition 6.3. There exists a source algebra $\mathcal{S} = (S, \mathcal{F}^{\mathcal{S}})$ and a mirrored algebra $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ with $M = \mathbb{R}^l$, both of the same type $\mathcal{F}$, such that $\mathcal{M}$ satisfies every law R that $\mathcal{S}$ satisfies, and, for all bijections $\varphi\colon L \to M$, there is no nontrivial homomorphism $\chi\colon S \to L$ when $L = \mathbb{R}^l$ is equipped with the algebra induced by $\mathcal{M}$ via (6.1).

Proof. We prove the claim by construction. Consider the source algebra $\mathcal{S} = (\mathbb{R}, \bullet)$, with a sole binary operation $\bullet$ defined by

$$s_1 \bullet s_2 = |s_1| s_2,$$

where, of course, $|s_1|$ represents the absolute value of the real number s_1, and $|s_1|s_2$ represents the usual product of the two real numbers $|s_1|$ and s_2. This source algebra is a semigroup, namely, $\mathbb{R}$ is closed under the binary operation $\bullet$, and $\bullet$ satisfies the associativity law, since

$$\begin{aligned} s_1 \bullet (s_2 \bullet s_3) &= |s_1|(s_2 \bullet s_3) \\ &= |s_1|(|s_2| s_3) \\ &= |s_1 s_2| s_3 \\ &= \big||s_1| s_2\big| s_3 \\ &= (|s_1| s_2) \bullet s_3 \\ &= (s_1 \bullet s_2) \bullet s_3 \end{aligned}$$

for all $s_1, s_2, s_3 \in \mathbb{R}$. Now, consider the mirrored algebra $\mathcal{M} = (\mathbb{R}, +)$, with $+$ being the standard addition operation on the real numbers. Obviously, $\mathcal{M}$ is also a semigroup, since $\mathbb{R}$ is closed under $+$, and $+$ is associative.

We now show that $\mathcal{M}$ satisfies every law R that $\mathcal{S}$ does. Let R be a law for type $\mathcal{F}$ defined by

$$R : p(x_1, \ldots, x_n) = q(x_1, \ldots, x_n)$$

for some arbitrary terms $p(x_1, \ldots, x_n), q(x_1, \ldots, x_n) \in T_{\mathcal{F}}(X)$. Suppose that $\mathcal{S}$ satisfies the law R. Then,

$$p^{\mathcal{S}}(s_1, \ldots, s_n) = q^{\mathcal{S}}(s_1, \ldots, s_n)$$

for all $s_i \in \mathbb{R}$. Since the associative binary operation $\bullet$ is the only operation in $\mathcal{F}^{\mathcal{S}}$, it must be that the term function $p^{\mathcal{S}}$ is given by some repeated application of $\bullet$:

$$p^{\mathcal{S}}(s_1, \ldots, s_n) = s_{i_1} \bullet s_{i_2} \bullet \cdots \bullet s_{i_{m_p}}$$

for some $m_p \in \mathbb{N}$ and some tuple $(i_1, \ldots, i_{m_p}) \in \{1, \ldots, n\}^{m_p}$. Similarly,

$$q^{\mathcal{S}}(s_1, \ldots, s_n) = s_{j_1} \bullet s_{j_2} \bullet \cdots \bullet s_{j_{m_q}}$$

for some $m_q \in \mathbb{N}$ and some tuple $(j_1, \ldots, j_{m_p}) \in \{1, \ldots, n\}^{m_p}$. Thus,

$$|s_{i_1} \cdots s_{i_{m_p}-1}| s_{i_{m_p}} = |s_{j_1} \cdots s_{j_{m_q}-1}| s_{j_{m_q}}. \tag{D.1}$$

If $m_p > m_q$, then there exists some factor s_i appearing in the product $|s_{i_1} \cdots s_{i_{m_p}-1}| s_{i_{m_p}}$ at least once more than it does in the product $|s_{j_1} \cdots s_{j_{m_q}-1}| s_{j_{m_q}}$. Thus, if the equality (D.1) holds for some $s_1, \ldots, s_n \in S$, doubling this particular value s_i would result in the law being violated, as the left-hand side of (D.1) would have an extra factor of 2 that the right-hand side would not. This implies that $m_p \leq m_q$. Analogous reasoning shows that $m_q \leq m_p$, and hence it must be that $m_p = m_q$; the same number of factors appear in the left-hand and right-hand sides of the law's realizations. Furthermore, the same reasoning goes to show that the left-hand and right-hand products in (D.1) actually must contain exactly the same factors with the same multiplicity (albeit in possibly different order), i.e., the ordered tuple $(s_{i_1}, \ldots, s_{i_{m_p}})$ of real numbers is some permutation of the ordered tuple $(s_{j_1}, \ldots, s_{j_{m_q}})$. Hence, it must be the case that

$$s_{i_1} + s_{i_2} + \cdots + s_{i_{m_p}} = s_{j_1} + s_{j_2} + \cdots + s_{j_{m_q}},$$

implying that

$$p^{\mathcal{M}}(s_1, \ldots, s_n) = p^{\mathcal{M}}(s_1, \ldots, s_n).$$

That is, the law is satisfied by $\mathcal{M}$ as well. Since R was arbitrarily chosen, we conclude that indeed $\mathcal{M}$ satisfies every law R that $\mathcal{S}$ does.

Now, let $\psi \colon S \to M$ be a homomorphism from $\mathcal{S}$ to $\mathcal{M}$. Then, it holds that

$$\psi(s_1 \bullet s_2) = \psi(s_1) + \psi(s_2)$$

for all $s_1, s_2 \in S = \mathbb{R}$. Therefore, for $s_1 = 0$ and $s_2 = s$ with $s \in S$ arbitrary, we conclude that

$$\psi(0) = \psi(|0|s) = \psi(0 \bullet s) = \psi(0) + \psi(s),$$

and hence

$$\psi(0) + \psi(s) = \psi(0) + \psi(t)$$

for all $s, t \in S$. However, since $+$ is the standard addition operation on $\mathbb{R}$, this is only possible if

$$\psi(s) = \psi(t)$$

for all $s, t \in S$, meaning the homomorphism ψ must be the trivial mapping $\psi\colon s \mapsto C$ with $C = \psi(0)$.

Now, let $\varphi\colon L \to M$ be an arbitrary bijection. Equip L with the algebra $\mathcal{L}$ induced by $\mathcal{M}$, as defined by (6.1). Let $\chi\colon S \to L$ be a homomorphism from $\mathcal{S}$ to $\mathcal{L}$. Then, since φ is an isomorphism from $\mathcal{L}$ to $\mathcal{M}$ (per Proposition 6.1), the composition $\varphi \circ \chi$ is a homomorphism from $\mathcal{S}$ to $\mathcal{M}$. Therefore, by our analysis above, $\varphi \circ \chi$ must be a trivial homomorphism given by $\varphi \circ \chi\colon s \mapsto C$ with $C = \varphi \circ \chi(0)$. Hence, for all $s \in S$, we conclude that

$$\chi(s) = \varphi^{-1}(C),$$

implying that χ must be a trivial homomorphism from $\mathcal{S}$ to $\mathcal{L}$. This concludes the proof. □

Proposition 6.4. Consider a source algebra $\mathcal{S} = (S, \mathcal{F}^{\mathcal{S}})$ of type $\mathcal{F}$, the latent space $L = \mathbb{R}^l$, and an arbitrary encoder $E\colon S \to L$. If E is bijective and there exists a mirrored algebra $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ with $M = \mathbb{R}^l$ and an isomorphism $\psi\colon S \to M$, then there exists a bijection $\varphi\colon L \to M$ such that $\varphi \circ E$ equals the isomorphism ψ.

Proof. Suppose that E is bijective and that there exists a mirrored algebra $\mathcal{M} = (M, \mathcal{F}^{\mathcal{M}})$ with $M = \mathbb{R}^l$ and an isomorphism $\psi\colon S \to M$. Define $\varphi\colon L \to M$ by $\varphi(z) = \psi \circ E^{-1}(z)$, which is well-defined since E is bijective. Then, it holds that

$$\varphi \circ E(s) = \psi(E^{-1}(E(s))) = \psi(s)$$

for all $s \in S$, which proves the result. □

D.2 Proofs for Section 6.4

We first present a key result from the algebraic topology literature.

Proposition D.1. Any continuous involution on $\mathbb{R}^n$ has a fixed point.

Proof. This is an easy application of Theorem 9 in Jaworowski [1956]. Namely, $\mathbb{R}^n$ is a separable metric space, and it is acyclic because it is contractible. □

Lemma D.2. Consider a Boolean lattice $\mathcal{A} = (A, \wedge, \vee, \neg, 0, 1)$ of type $\mathcal{F}_{\text{Bool}}$ and its associated laws in Table 6.1. Then, it holds that $\neg$ is an involution with no fixed points.

Proof. Clearly, the Boolean lattice laws in Table 6.1 imply that $\neg(\neg a) = a$ for all $a \in A$, and thus $\neg$ is an involution. Now assume for the sake of contradiction that $\neg$ has a fixed point $b \in A$, so that $\neg b = b$. Then, by the Boolean lattice laws,

$$(\neg b) \wedge b = 0 \implies b \wedge b = 0 \implies b = 0,$$

and, similarly,

$$(\neg b) \vee b = 1 \implies b \vee b = 1 \implies b = 1.$$

This is a contradiction. □

We are now ready to prove our negative result.

Theorem 6.5. Consider an algebra $\mathcal{A} = (A, \mathcal{F}^{\mathcal{A}})$ with a unary operation $\Box^{\mathcal{A}}$. Assume $\mathcal{A}$ satisfies laws $R_1, \ldots, R_n$ which imply that $\Box$ has no fixed point: $\Box(x) \neq x$ for all $x \in A$. Furthermore, assume that one of the laws R_i is the involution law given by

$$\Box(\Box(x)) = x.$$

Then, there exists no algebra $\mathcal{B} = (B, \mathcal{F}^{\mathcal{B}})$ on the Euclidean space $B = \mathbb{R}^l$ such that $\Box^{\mathcal{B}}$ is continuous and $R_1, \ldots, R_n$ are all satisfied by $\mathcal{B}$.

Proof. Suppose for the sake of contradiction that there exists an algebra $\mathcal{B} = (B, \mathcal{F}^{\mathcal{B}})$ on $B = \mathbb{R}^l$ such that $\Box^{\mathcal{B}}$ is continuous and the laws $R_1, \ldots, R_n$ are all satisfied by $\mathcal{B}$. Then, by assumption it must be the case that $\Box^{\mathcal{B}}(b) \neq b$ for all $b \in B$. However, since $\Box^{\mathcal{B}}$ is a continuous involution on $B = \mathbb{R}^l$, by Proposition D.1, $\Box^{\mathcal{B}}$ has a fixed point, i.e., $\Box^{\mathcal{B}}(b) = b$ for some $b \in B$. This is a contradiction, and hence the result is proven. □

Corollary 6.6. The Boolean lattice type $\mathcal{F}_{\text{Bool}}$ cannot be realized on $M = \mathbb{R}^l$ with continuous operations such that the Boolean lattice laws in Table 6.1 are satisfied.

Proof. This follows directly from Lemma D.2 and Theorem 6.5. □