

Data-Driven Performance Guarantees for Classical and Learned Optimizers

Rajiv Sambharya

SAMBHAR9@SEAS.UPENN.EDU

Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA

Bartolomeo Stellato

BSTELLATO@PRINCETON.EDU

Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA

Editor: Silvia Villa

Abstract

We introduce a data-driven approach to analyze the performance of continuous optimization algorithms using generalization guarantees from statistical learning theory. We study classical and learned optimizers to solve families of parametric optimization problems. We build generalization guarantees for classical optimizers, using a sample convergence bound, and for learned optimizers, using the Probably Approximately Correct (PAC)-Bayes framework. To train learned optimizers, we use a gradient-based algorithm to directly minimize the PAC-Bayes upper bound. Numerical experiments in signal processing, control, and meta-learning showcase the ability of our framework to provide strong generalization guarantees for both classical and learned optimizers given a fixed budget of iterations. For classical optimizers, our bounds which hold with high probability are much tighter than those that worst-case guarantees provide. For learned optimizers, our bounds outperform the empirical outcomes observed in their non-learned counterparts.

1. Introduction

This paper studies continuous parametric optimization problems of the form

$$\text{minimize } f(z, x), \tag{1}$$

where $z \in \mathbf{R}^n$ is the decision variable, $x \in \mathbf{R}^d$ is the parameter or context drawn from some distribution \mathcal{X} , and $f : \mathbf{R}^n \times \mathbf{R}^d \rightarrow \mathbf{R} \cup \{+\infty\}$ is the objective. Problem (1) implicitly defines a (potentially non-unique) solution $z^*(x) \in \mathbf{R}^n$. Many applications require repeatedly solving problem (1) with varying x . For instance, in robotics and control, we repeatedly solve optimization problems to update the inputs (*e.g.*, torques and thrusts) while the state (*e.g.*, position and velocity) and the goals (*e.g.*, reference trajectory) change (Borrelli et al., 2017). This problem structure is also observed in other domains, such as sparse coding, where sparse signals are recovered from noisy measurements (Gregor and LeCun, 2010), and image restoration, where images are recovered from their corrupted versions (Elad and Aharon, 2006). These optimization problems usually do not admit closed-form solutions, so instead, iterative algorithms are needed to search for an optimal solution. First-order methods, which only rely on first-order derivatives (Beck, 2017), are a popular approach to solve problem (1) due to their cheap per-iteration cost. Typically, first-order methods

repeatedly apply a mapping $T : \mathbf{R}^n \times \mathbf{R}^d \rightarrow \mathbf{R}^n$, obtaining iterations of the form

$$z^{k+1}(x) = T(z^k(x), x). \quad (2)$$

Due to the limited time available to compute the solutions between instances of problem (1), in several applications we can only afford a fixed number of iterations of algorithm (2). In such settings, obtaining strong performance guarantees on the quality of the solution within this iteration budget is essential, particularly for safety-critical applications.

Analyzing the worst-case performance of the first-order method (2) has been intensely studied in optimization literature by constructing asymptotic convergence rates of the algorithms (see (Beck, 2017, Section 5) and (Ryu and Yin, 2022, Section 2)). In contrast, the performance estimation problem (PEP) (Drori and Teboulle, 2014; Taylor et al., 2015) approach recently emerged as a powerful tool for the numerical computation of exact worst-case guarantees for first-order methods after only a finite number of iterations. There are two main drawbacks of both the theoretical and computer-assisted worst-case analyses. First, worst-case guarantees are pessimistic by definition; they provide guarantees for the most adverse problem instance among a class of problems, even if such instance occur very infrequently. Second, these analyses typically consider a general class of functions (*e.g.*, strongly convex and smooth functions) without leveraging the specific parametric nature inherent in problem (1). In contrast to worst-case analysis, a probabilistic approach may provide less pessimistic results for applications where high-probability bounds are acceptable instead of strict worst-case guarantees (Pedregosa and Scieur, 2020).

While guarantees for first-order methods are important for ensuring reliability, these algorithms often suffer from slow convergence in practice (Zhang et al., 2020). To mitigate this limitation, the *learning to optimize* paradigm (Amos, 2023; Chen et al., 2022; Balcan, 2020) takes advantage of the parametric setting of our interest, and uses machine learning to predict the solutions to problem (1), thereby significantly reducing the solve time compared with classical solvers (*i.e.*, those without learned components). A common strategy is to learn algorithm steps (Gregor and LeCun, 2010) or initializations (Sambharya et al., 2024). Learned optimizers have shown promise in a range of domains, *e.g.*, in inverse problems (Gregor and LeCun, 2010), convex optimization (Sambharya et al., 2024; Ichnowski et al., 2021), meta-learning (Finn et al., 2017), and non-convex optimization (Kotary et al., 2021; Bertsimas and Stellato, 2022). However, guaranteeing convergence of learned optimizers is a challenge since the algorithm steps have been replaced with learned variants (Chen et al., 2022; Amos, 2023). While asymptotic convergence can sometimes be guaranteed by construction (Sambharya et al., 2024) or by safeguarding (Heaton et al., 2023; Prémont-Schwarz et al., 2022), these approaches do not provide performance bounds within a fixed number of iterations. To address this shortcoming, several methodologies have been developed to construct generalization bounds for learned optimizers, for example, using Rademacher complexity (Chen et al., 2020; Sambharya et al., 2023) and the PAC-Bayes framework (Bartlett et al., 2022; Gupta and Roughgarden, 2017; Sambharya et al., 2024). However, such bounds tend to be loose (in many cases not reported), or sometimes, the generalization bound itself can be difficult to compute.

Our contributions. In this paper, we present a data-driven approach based on statistical learning theory to obtain performance guarantees for both classical and learned optimizers

based on fixed-point iterations. For classical optimizers, our approach differs from existing worst-case analysis frameworks. Instead of worst-case guarantees, we construct data-driven guarantees that hold with high probability over the parametric family of optimization problems. Meanwhile, for learned optimizers, we rely on PAC-Bayes theory (McAllester, 1998; Alquier, 2023) to provide generalization bounds, and moreover, use gradient-based methods to optimize the bounds themselves. Our method is not limited to standard metrics to analyze optimization algorithms (*e.g.*, distance to the optimal solution or fixed-point residual); rather, we provide guarantees on any metric as long as it can be evaluated. We summarize our contributions as follows.

- We provide probabilistic guarantees for classical optimizers in two steps: first, we run the optimizer for a given number of iterations on each problem instance in a given dataset; then, we apply a sample convergence bound by solving a one-dimensional convex optimization problem.
- We construct generalization bounds for learned optimizers using PAC-Bayes theory. In addition, we develop a framework to learn optimizers by directly minimizing the PAC-Bayes bounds using gradient-based methods. After training, we calibrate the PAC-Bayes bounds by sampling the weights of the learned optimizer, and subsequently running the optimizer for a fixed number of steps for each problem instance in a given dataset. Then, we compute the PAC-Bayes bounds via solving two one-dimensional convex optimization problems.
- We apply our method to compute guarantees for classical optimizers on several examples including image deblurring and robust Kalman filtering, illustrating that our bounds that hold with high probability outperform bounds from worst-case analyses. We also showcase our generalization guarantees for several learned optimizers: LISTA (Gregor and LeCun, 2010) and its variants (Liu et al., 2019; Wu et al., 2020), learned warm starts (Sambharya et al., 2024), and model-agnostic meta-learning (Finn et al., 2017). Our generalization guarantees accurately represent the benefits of learning by outperforming the empirical performance observed in their non-learned counterparts.

Notation. We denote the set of non-negative vectors of length n as \mathbf{R}_+^n and the set of vectors with positive entries of length n as \mathbf{R}_{++}^n . We let the set of vectors consisting of natural numbers of length n be \mathbf{N}^n . The set of $n \times n$ symmetric matrices is denoted as \mathbf{S}^n , and the set of $n \times n$ positive semidefinite matrices is denoted as \mathbf{S}_+^n . We denote the trace of a square matrix A with $\text{tr}(A)$ and its determinant as $\det A$. For a matrix A , we denote its spectral norm with $\|A\|_2$ and its Frobenius norm with $\|A\|_F$. For two vectors $u \in \mathbf{R}^n$ and $v \in \mathbf{R}^n$, we denote its element-wise multiplication with $u \odot v$. We round a vector v element-wise to the nearest integer with $\text{round}(v)$. For a vector $v \in \mathbf{R}^n$, the diagonal matrix $V \in \mathbf{S}^n$ with entries $V_{ii} = v_i$ for $i = 1, \dots, n$ is given by $\text{diag}(v)$. The all ones vector of length d is denoted as $\mathbf{1}_d$. For a vector v , the operation $\text{sign}(v)$ returns, for each element, a value of $+1$ if the corresponding element in v is non-negative and -1 otherwise. For any closed and convex set C , we let $\text{dist}_C : \mathbf{R}^n \rightarrow \mathbf{R}$ be the distance function: $\text{dist}_C(x) = \min_{s \in C} \|s - x\|_2$. We denote expectation and probability with \mathbf{E} and \mathbf{P} respectively. Finally, for a boolean condition c , we let $\mathbf{1}(x) = 1$ if c is true, and 0 otherwise.

Outline. We structure the rest of the paper as follows. Section 2 reviews the literature on i) guarantees on classical optimizers and ii) learned optimizers, focusing on existing methods and generalization guarantees associated with them. In Section 3, we introduce the mechanics of both classical and learned optimizers. In Section 4 we introduce our method for obtaining data-driven guarantees for classical optimizers. We then focus on learned optimizers in the next two sections. In Section 5, we provide our generalization guarantees for learned optimizers derived from the PAC-Bayes framework. Then in Section 6, we present a gradient-based algorithm designed to optimize the PAC-Bayes bound itself. After that, in Section 7, we present numerous numerical experiments with data-driven guarantees for both classical and learned optimizers. Finally, in Section 8 we conclude.

2. Related work

Theoretical and computer-assisted worst-case analysis. Theoretical convergence analysis techniques for first-order methods typically focus on general classes of problems (Ryu and Yin, 2022; Beck, 2017). Many analyses provide upper bounds on the asymptotic rate of convergence for an algorithm (Giselsson and Boyd, 2014; Hong and Luo, 2012) that are tight in certain cases (Nesterov, 1983). However, there are cases where upper bounds are not tight, because they either lack corresponding lower bounds (Taylor et al., 2015) or they are only known up to a constant (Ryu et al., 2020). Even if the asymptotic rate is tight (*i.e.*, there exists at least one iteration where the worst-case rate is exactly met), it may be pessimistic: the algorithm may still perform significantly better during most iterations (*e.g.*, the local convergence rate may be better than the global one (Boley, 2013)). Most importantly, these analyses are fundamentally pessimistic and do not exploit the parametric structure. A less-explored area is average-case analysis which analyzes an algorithm’s performance in expectation over a class of problems. This approach, while not pessimistic, is designed to analyze the asymptotic convergence rate rather than provide numerical guarantees, and is further limited by its focus on unconstrained problems, as highlighted in existing works (Pedregosa and Scieur, 2020; Paquette et al., 2022).

Computer-assisted approaches like PEP (Drori and Teboulle, 2014; Taylor et al., 2015; Ryu et al., 2020) and integral quadratic constraints (Lessard et al., 2016; Fazlyab et al., 2017; Taylor et al., 2018) have emerged as methods to obtain numerical worst-case guarantees, but they do not take advantage of the parametric nature of problem (1). To bridge this gap, Ranjan and Stellato (2024) introduced a technique inspired by neural network verification (Fazlyab et al., 2022) to compute worst-case guarantees of fixed-point algorithms for parametric quadratic programs (QPs). However, this approach deals with relaxations that become looser and more computationally expensive as the number of steps increases. Our probabilistic guarantees for classical optimizers complement worst-case analysis by demonstrating that, for those willing to accept high-probability bounds instead of stricter worst-case guarantees, our approach provides significantly stronger performance bounds over the parametric family.

Learning initializations and algorithm steps. A common strategy in learning to optimize is to learn high-quality initializations. Sambharya et al. (2023) and Sambharya et al. (2024) unroll, *i.e.*, differentiate through (Monga et al., 2021; Chen et al., 2020), algorithm steps to learn warm starts, thereby reducing solve times for convex problems.

Some works learn initializations in a decoupled fashion (Baker, 2019; Mak et al., 2023; Briden et al., 2023; Misra et al., 2022), while others directly learn the optimal solution, and rather than warm-starting an algorithm, ensure feasibility and optimality with a correction step (Donti et al., 2021; Karg and Lucia, 2020; Chen et al., 2018a).

An alternate approach is to learn the algorithm steps. In convex optimization, learned algorithm steps have been shown to decrease solve times through learned hyperparameters (Jung et al., 2022; Ichnowski et al., 2021; King et al., 2024), and learned acceleration schemes (Venkataraman and Amos, 2021). While a lack of convergence guarantees was seen as a potential downside of learning algorithm steps (Amos, 2023), some works have addressed this by safeguarding (Heaton et al., 2023; Prémont-Schwarz et al., 2022), providing convergence rate bounds (Tan et al., 2023), and constraining the updates (Banert et al., 2021). The idea of learning algorithm steps has also been used to solve non-convex problems (Bai et al., 2022; Sjölund and Bånestad, 2022; Balcan et al., 2017, 2018) and inverse problems (Gregor and LeCun, 2010; Liu et al., 2019; Chen et al., 2018b, 2021; Diamond et al., 2017; Ryu et al., 2019; Balatsoukas-Stimming and Studer, 2019). Our method is designed to integrate with these methods, optimizing and calibrating generalization guarantees for *any* learned optimizer.

Generalization bounds in learned optimizers. Despite strong empirical outcomes in certain settings, learned optimizers lack generalization guarantees (Chen et al., 2022; Amos, 2023; Yang et al., 2022). To address this, Sucker and Ochs (2023) and Sucker et al. (2024) optimize PAC-Bayesian guarantees based on exponential families, but they assume exponential moment bounds, a condition difficult to verify in practice. In addition, they assume a specific update function: a multi-layer perceptron (Sucker et al., 2024) or a gradient step with learned hyperparameters (Sucker and Ochs, 2023). On the other hand, our method can be used *in conjunction* with any learned optimizer, including ones with learned initializations. Other works provide guarantees through Rademacher complexity (Chen et al., 2020; Sambharya et al., 2023), the PAC-Bayes framework (Bartlett et al., 2022; Gupta and Roughgarden, 2017; Sambharya et al., 2024), and pseudo-dimension bounds (Balcan et al., 2021). Yet, these bounds tend to be loose or difficult to compute. We construct numerical bounds by optimizing the PAC-Bayes bounds themselves, a strategy previously used for classification (Dziugaite and Roy, 2017) and control (Majumdar et al., 2021).

Meta-learning. Meta-learning (Hospedales et al., 2021; Vilalta and Drissi, 2001) overlaps with learning to optimize when the parametric problem is a learning task (Chen et al., 2022). Both learned initializations (Finn et al., 2017) and algorithm updates (Li and Malik, 2016; Andrychowicz et al., 2016; Metz et al., 2022) have been effectively used in meta-learning. Methods have been developed to improve generalization in practice (Almeida et al., 2021; Yang et al., 2023) and to provide theoretical generalization bounds (Amit and Meir, 2018; Balcan et al., 2019). Yet existing bounds tend to be challenging to evaluate or loose. Addressing this issue, Farid and Majumdar (2021) derive a novel PAC-Bayes bounds, focusing on practically useful guarantees. Our method is more general in that it can be applied to not only learning tasks, but also optimization and inverse problems.

3. Classical and learned optimizers

In this section, we delve into the mechanics of classical and learned optimizers, laying the groundwork for the bounds we provide later. In Section 3.1 we explain how to run and evaluate classical optimizers, focusing on fixed-point optimization algorithms. For learned optimizers, we first explain how to run and evaluate them given fixed weights in Section 3.2, and then how to train them to learn the weights in Section 3.3.

3.1 Running and evaluating classical optimizers

As it turns out, problem (1) can often be written as an equivalent fixed-point problem

$$\text{find } z \text{ subject to } z = T(z, x), \quad (3)$$

where $T : \mathbf{R}^n \times \mathbf{R}^d \rightarrow \mathbf{R}^n$ is the fixed-point operator. Indeed, nearly all convex optimization problems can be reformulated as a finding the fixed-point of an operator (Ryu and Yin, 2022) which often represents the optimality conditions (Garstka et al., 2019). We denote the set of fixed-points for the fixed-point problem parametrized by x to be $\mathbf{fix} T_x$. Note that the ground truth solution $z^*(x)$ satisfies the fixed-point condition $z^*(x) \in \mathbf{fix} T_x$. For classical optimizers, we focus on the parametric fixed-point problem (3) as it is a convenient way of analyzing worst-case performance (Banjac et al., 2019; Giselsson and Boyd, 2014). This in turn allows for a direct comparison of our guarantees with those previously established.

Initializations. In classical optimizers, the initialization is not learned and is typically set to the zero vector, *i.e.*, $z^0(x) = 0$, which is often referred to as a *cold start*. In contexts where we have an estimate of the solution, it is common to *warm-start* the problem from this point. For example, in model predictive control (Borrelli et al., 2017), where similar instances of the same problem are solved sequentially, the problems are often warm-started from the previous solution shifted by one time index (Diehl et al., 2009).

Algorithm steps. The iterates $z^{k+1}(x) = T(z^k(x), x)$ in (2) are a popular way to solve problem (3). Many classical optimizers consist of fixed-point iterations, *e.g.*, gradient descent, proximal gradient descent (Parikh and Boyd, 2014), and ADMM (Boyd et al., 2011).

Evaluation metrics. A variety of metrics can be used to evaluate the performance of algorithms. A standard metric is the fixed-point residual (Ryu and Yin, 2022, Section 2.4)

$$\phi^{\text{fp}}(z, x) = \|T(z, x) - z\|_2,$$

which quantifies the gap between successive iterations. Such metrics, assess the quality of candidate solutions for problems parametrized by x . To determine if an optimization algorithm meets specific performance benchmarks, we introduce the 0–1 error function

$$e(x) = \mathbf{1} \left(\phi \left(z^k(x), x \right) \geq \epsilon \right), \quad (4)$$

assigning a value of 1, if the performance metric $\phi(z, x)$ exceeds a specified threshold ϵ after k steps, indicating a failure to meet the desired criteria, and 0 otherwise. We later provide guarantees for this error function $e(x)$, for *any* underlying metric $\phi(z, x)$ in Section 4.

Convergence. Under certain conditions on the operator T , the fixed-point iterates in (2) are known to converge to a fixed-point, *i.e.*, $\lim_{k \rightarrow \infty} \|z^k(x) - z^*(x)\|_2 = 0$ for some $z^*(x)$ in the set of fixed-points $\text{fix } T_x$. For instance, if the operator T is contractive, linearly convergent, or averaged (see Appendix D for their definitions), and the set of fixed-points is non-empty, then the iterates are guaranteed to converge (Ryu and Yin, 2022, Section 2.4). We refer the reader to Appendix D for the rates of convergence for these cases.

3.2 Running and evaluating learned optimizers

The goal of learning to optimize methods is to accelerate an algorithm to quickly find a high-quality candidate solution $\hat{z}_\theta(x)$ for problem (1). Learned optimizers typically learn either the initial point or the steps for a given algorithm, by adjusting some weights $\theta \in \mathbf{R}^p$.

Learned initializations. Some learned optimizers focus on learning the initializations for algorithms (Sambharya et al., 2024; Finn et al., 2017). Typically, this involves predicting an initial point $z^0 \in \mathbf{R}^n$ from the parameter x with a function $h_\theta : \mathbf{R}^d \rightarrow \mathbf{R}^n$:

$$z_\theta^0(x) = h_\theta(x).$$

Learned algorithm steps. Another common strategy in learned optimizers is to learn the steps of the algorithm, which can be represented as

$$z_\theta^{k+1}(x) = T_\theta(z_\theta^k(x), x).$$

Here, the function $T_\theta : \mathbf{R}^n \times \mathbf{R}^d \rightarrow \mathbf{R}^n$ is the learned update rule. Note that the iterates $z_\theta^k(x)$ depend on the parameter x and the weights θ .

Evaluation metrics. The evaluation metric ϕ depends on the task at hand. For inverse problems, a common metric of interest is the squared distance to the ground truth solution $\phi^{\text{mse}}(z, x) = \|z - z^*(x)\|_2^2$. In meta-learning, a common measure is the performance on a learning task over an unseen dataset $\mathcal{D}^{\text{test}}$ and learning objective \mathcal{L} (Finn et al., 2017; Li and Malik, 2016), which fits into our framework with the performance metric $\phi^{\text{meta}}(z, x) = \mathcal{L}(z, \mathcal{D}^{\text{test}})$. As in the classical optimizers case, we consider the 0–1 error function associated with an underlying metric ϕ . In this case, the error function depends on the weights θ :

$$e_\theta(x) = \mathbf{1} \left(\phi \left(z_\theta^k(x), x \right) \geq \epsilon \right). \quad (5)$$

It is important to remark that the metric ϕ can also be different from the objective f from problem (1). Our generalization guarantees are designed to provide bounds for the error function $e_\theta(x)$ with *any* underlying metric ϕ .

Convergence for learned optimizers. When the algorithm steps are replaced with learned variants, convergence may not be guaranteed (Chen et al., 2022; Amos, 2023).

3.3 Training learned optimizers

In this subsection, we formulate the learning to optimize training problem, beginning with the loss functions. Depending on the task at hand, the loss can take varying forms, generally falling into two categories: regression-based and objective-based (Amos, 2023).

Regression-based loss. The *regression-based loss* measures the distance to a ground truth solution $z^*(x)$, *i.e.*,

$$\ell_\theta^{\text{reg}}(x) = \|\hat{z}_\theta(x) - z^*(x)\|_2^2. \quad (6)$$

Objective-based loss. The *objective-based loss* directly penalizes the objective f :

$$\ell_\theta^{\text{obj}}(x) = f(\hat{z}_\theta(x), x). \quad (7)$$

Unlike the regression-based loss, the objective-based loss does not require access to ground truth solutions.

The learning to optimize training problem. Given the loss function, algorithm steps, and initialization we formulate the training problem as

$$\begin{aligned} & \text{minimize} && \mathbf{E}_{x \sim \mathcal{X}} \ell_\theta(x) \\ & \text{subject to} && z_\theta^{k+1}(x) = T_\theta(z_\theta^k(x), x), \quad k = 0, 1, \dots, K-1 \\ & && z_\theta^0(x) = h_\theta(x). \end{aligned} \quad (8)$$

Here, K is the number of algorithm steps used during training, and the loss function $\ell_\theta(x)$ is either chosen to be the regression-based loss $\ell_\theta^{\text{reg}}(x)$ or the objective-based loss $\ell_\theta^{\text{obj}}(x)$. The loss function is applied to the K -th iterate $\hat{z}_\theta(x) = z_\theta^K(x)$, but, in principle, it could be a (weighted) sum of a loss function applied all the iterates $z_\theta^0(x), \dots, z_\theta^K(x)$. Since in general we do not know the distribution \mathcal{X} to solve problem (8), we approximate the expectation over N independent and identically distributed (i.i.d.) training samples $S = \{x_i\}_{i=1}^N$. In Section 5, we modify the empirical training problem to provide guarantees to unseen data.

4. Probabilistic guarantees for classical parametric optimization

In this section, we use statistical learning theory to provide probabilistic guarantees for classical parametric optimization. In particular, we focus on the fixed-point problem setting (3), and obtain performance guarantees on the quality of the iterates from (2), $z^{k+1}(x) = T(z^k(x), x)$. Recall that the parameter x is drawn in an i.i.d. fashion from distribution \mathcal{X} . We first provide bounds for algorithms initialized to the zero vector (*i.e.*, cold-started) and then consider how to adapt the bounds to include warm starts.

Obtaining guarantees via statistical learning theory. Given an underlying metric ϕ , a number of algorithm steps k , and a tolerance ϵ , we consider the 0–1 error $e(x)$ given by Equation (4) which takes a value of 1 if ϕ is above ϵ after k steps and 0 otherwise. There are three steps to obtain bounds on the risk $r_{\mathcal{X}}$ given N sample parameters S as depicted in Figure 1. First, for each sample x we run k fixed-point steps starting from the zero vector to obtain $z^k(x)$. Second, we compute the empirical risk \hat{r}_S , the fraction of problems that fail to reach the desired tolerance in k steps. Last, we apply the sample convergence bound from Theorem 3 in Appendix Section A to bound the risk $r_{\mathcal{X}}$ with probability at least $1 - \delta$:

$$r_{\mathcal{X}} \leq \text{kl}^{-1} \left(\hat{r}_S \left| \frac{\log(2/\delta)}{N} \right. \right). \quad (9)$$

We remark that other concentration bounds could be used in (9), and that we could instead bound $\mathbf{E}\phi(z^k(x), x)$ directly instead of the risk $\mathbf{E}e(x)$. Typically, using a concentration bound requires an upper bound on the metric of interest, a condition trivially satisfied by the error function. The choice to bound the error function is driven by this convenience, which also proves particularly beneficial in the analysis of learned optimizers, as discussed in Section 5.

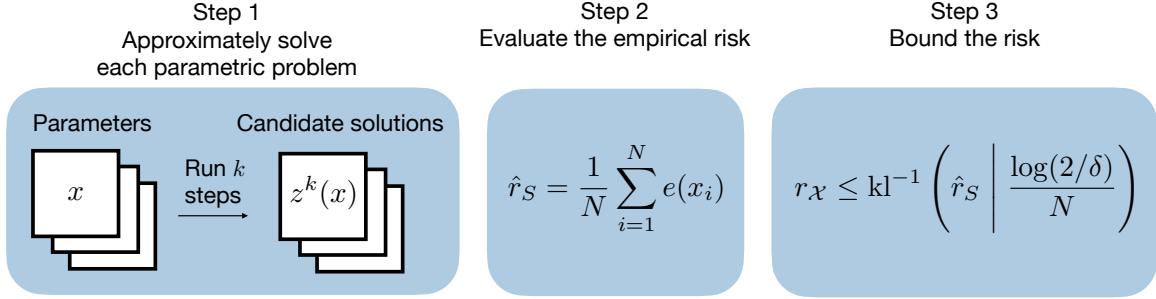


Figure 1: The procedure to generate probabilistic guarantees for classical optimizers. Given N parameter samples, we first approximately solve each parametric problem by running k fixed-point steps in step 1. Then given an error function $e(x)$ with an underlying metric ϕ , number of algorithm steps k , and tolerance ϵ , we evaluate the empirical risk \hat{r}_S in step 2. Lastly in step 3, we apply the sample convergence bound to bound the risk $r_{\mathcal{X}}$ with high probability.

Incorporating warm starts. It is natural to wonder if the bound (9) can be adapted to algorithms initialized from warm starts rather than from the zero vector. Indeed this adaptation is feasible, as long as the errors $e(x)$ are i.i.d. random variables. One setting where this condition is met, what we call the nearest-neighbor warm start (Sambharya et al., 2024) setting, assumes access to a base set of N^{base} problem parameters and a corresponding optimal solution for each one. The nearest-neighbor warm start initializes the sample problem with the given optimal solution of the nearest of the base problems measured by distance in terms of its parameter $x \in \mathbf{R}^d$. Since the metric $e(x)$ is still i.i.d., the sample convergence bound from inequality (9) holds.

Strengthening the bound with worst-case guarantees. One downside of the guarantee given by the sample convergence bound is that a non-zero term $(1/N) \log(2/\delta)$ in inequality (9) prevents the risk $r_{\mathcal{X}}$ from ever reaching zero even if the empirical risk \hat{r}_S is zero and k is very large. If the underlying metric is the fixed-point residual, the worst-case guarantees from (40) and (41) can provide a stronger result: a risk of exactly zero that holds with probability one, for a large enough number of iterations (provided an upper bound on the distance from the initialization to the set of optimal solutions). In this case, we simply adapt our bounds to take the better of the worst-case guarantee and the probabilistic bound in Equation (9).

5. Generalization bounds for learned optimizers

In this section, we derive generalization bounds for learned optimizers using tools from PAC-Bayes theory. In particular, we adapt Maurer’s bound from Theorem 4 presented

in Appendix Section A to allow for a data-dependent prior and then apply it for learned optimizers.

Maurer’s bound from (34) is used to provide bounds where the weights are drawn from a distribution, while learned optimizers (as outlined in Section 3) are deterministic. To reconcile this, we adapt the learning to optimize framework so that the weights of the learned optimizers θ are drawn from a posterior distribution P . Then, considering the 0–1 error metric $e_\theta(x)$ from Equation (5) as a function of the loss, *i.e.*, $e_\theta(x) = \mathbf{1}(\ell_\theta(x) \geq \epsilon)$, we aim to bound the expected risk of the posterior $R_{\mathcal{X}}(P)$ defined in Equation (29). As in the case of classical optimizers, we choose to bound the error function rather than the loss. This approach is particularly useful for learned optimizers, as obtaining an upper bound on the loss, an important assumption in the PAC-Bayes framework, can be difficult due to the lack of convergence guarantees (Amos, 2023). Bounding the error allows us to bypass this complication, as an upper bound of one is trivially given.

The posterior. To obtain the KL divergence in closed form from Equation (34), we consider posterior and prior distributions on the algorithm weights that are multivariate normal distributions. We further enforce a diagonal covariance structure for both. Our posterior takes the form $\mathcal{N}(w, \mathbf{diag}(s))$ where the mean is $w \in \mathbf{R}^p$ and the covariance is $\mathbf{diag}(s) \in \mathbf{S}_+^p$. We use the notation $\mathcal{N}_{w,s} = \mathcal{N}(w, \mathbf{diag}(s))$ for convenience.

The prior. In the next section, we would like to optimize over the bounds themselves; however, recall that the vanilla Maurer bound from Theorem 4 requires that the prior is fixed and independent of the training samples. Our strategy is to consider a data-dependent prior where the mean is fixed and the variance is optimized over. We then round the variance to a pre-defined grid where we use a union-bound argument to satisfy the assumptions of Theorem 4. This strategy has been taken in the literature, for example in Dziugaite and Roy (2017) and Langford and Caruana (2001) where the covariance matrix takes the form $\Lambda = \lambda I$ for a scalar λ . We generalize this approach, by instead partitioning the weights into J groups and optimizing over a *vector* $\lambda \in \mathbf{R}_+^J$ rather than a scalar. For the j -th group (where $j \in \{1, \dots, J\}$), we let \mathcal{I}_j be the corresponding index set of weights. We construct the diagonal prior variance $\Lambda \in \mathbf{S}_+^p$ by assigning the value λ_j to the indices in group \mathcal{I}_j , *i.e.*,

$$\mathbf{diag}(\Lambda)_{\mathcal{I}_j} = \lambda_j \mathbf{1}_{|\mathcal{I}_j|}, \quad \text{for } j = 1, \dots, J.$$

Here, $|\mathcal{I}_j|$ is the cardinality of the set \mathcal{I}_j . This partitioning approach allows for a more nuanced adaptation to weights associated with different groups. Consider, for instance, LISTA-type algorithms, where distinct weights are used for shrinkage thresholds and step sizes (Gregor and LeCun, 2010; Liu et al., 2019) (see Section 7.2.1 for more details). Intuitively, accommodating different priors for each group can be advantageous because different weight groups can have different orders of magnitudes. Hence, it is natural that allowing for different variances across partitions is beneficial.

Main generalization bound for learned optimizers. We now give our main generalization bound theorem which uses the union-bound argument to allow for a data-dependent prior. Specifically, we enforce that the prior variance term $\lambda \in \mathbf{R}_+^J$ takes the form $\lambda = \lambda^{\max} \exp(-a/b)$ for some $a \in \mathbf{N}^J$. We design Maurer’s bound to hold for a given

a with probability

$$\delta_a = \left(\frac{6}{\pi^2} \right)^J \frac{\delta}{\prod_{j=1}^J a_j^2}, \quad (10)$$

for some pre-determined $\delta \in (0, 1)$. Then with probability at least $1 - \delta$, Maurer’s bound holds uniformly for all $a \in \mathbf{N}^J$. This strategy generalizes the union-bound arguments made in the literature (Dziugaite and Roy, 2017; Langford and Caruana, 2001) to allow for J to be larger than one. We formalize this result with the following theorem.

Theorem 1 *Consider a set of N i.i.d. samples S . Let the prior mean $w_0 \in \mathbf{R}^p$, and the prior variance hyperparameters $\lambda^{\max} \in \mathbf{R}_+$ and $b \in \mathbf{R}_+$, be independent of the samples. Then for any $\delta \in (0, 1)$, posterior distribution $\mathcal{N}_{w,s}$, and vector $a \in \mathbf{N}_+^J$, with probability at least $1 - \delta$ the following bound holds:*

$$R_{\mathcal{X}}(\mathcal{N}_{w,s}) \leq \text{kl}^{-1} \left(\hat{R}_S(\mathcal{N}_{w,s}) \mid B(w, s, \lambda) \right). \quad (11)$$

Here, $\lambda = \lambda^{\max} \exp(-a/b)$ and the regularization term is

$$B(w, s, \lambda) = \frac{1}{N} \left(\text{KL}(\mathcal{N}_{w,s} \parallel \mathcal{N}(w_0, \Lambda)) + \sum_{j=1}^J 2 \log \left(b \log \frac{\lambda^{\max}}{\lambda_j} \right) + J \log \frac{\pi^2}{6} + \log \frac{2\sqrt{N}}{\delta} \right). \quad (12)$$

Using Equation (30), the KL term $\text{KL}(\mathcal{N}_{w,s} \parallel \mathcal{N}(w_0, \Lambda))$ simplifies to

$$-\frac{1}{2} (p + \mathbf{1}_p^T \log s) + \frac{1}{2} \sum_{j=1}^J \left(\frac{1}{\lambda_j} \|s_{\mathcal{I}_j}\|_1 + \frac{1}{\lambda_j} \|w_{\mathcal{I}_j} - (w_0)_{\mathcal{I}_j}\|^2 + |\mathcal{I}_j| \log \lambda_j \right),$$

where $\log s$ is applied element-wise. See Appendix C.1 for the proof.

6. Optimizing the generalization bounds

In this section, we show how to optimize the PAC-Bayes bounds obtained in Section 5. In Section 6.1 we present the penalized training problem whose objective aligns with the PAC-Bayes generalization bound. The objective of this problem includes a KL inverse term which involves solving a one-dimensional convex optimization problem. In Section 6.2 we show how to use implicit differentiation to differentiate through the KL inverse. This technique allows us to implement a gradient-based learning algorithm, which we present in Section 6.3. In Section 6.4, we show how to calibrate the PAC-Bayes bounds after training, thereby providing generalization guarantees on the expected risk.

6.1 The penalized training problem

Our overarching strategy to obtain strong generalization guarantees is to use gradient-based methods to optimize the PAC-Bayes bounds from Theorem 1. Gradient-based methods emerge as a natural choice to optimize our PAC-Bayes bounds as they are often used to train learned optimizers (Chen et al., 2022; Monga et al., 2021). Indeed, the learned

optimizers that we provide generalization guarantees for in our numerical experiments in Section 7 are all trained with gradient-based methods in their original works. Nevertheless, this approach gives rise to several obstacles that we will address in this section, culminating in the formulation of a *penalized training problem*.

The first obstacle is that the decision variable λ , which corresponds to the prior variance, must belong to a discrete set as described in Section 5. To simplify the training, we treat λ as a continuous variable, and then after training, round its value to the discrete set (Dziugaite and Roy, 2017). The second obstacle is that the 0–1 loss function $e_\theta(x)$ is non-differentiable. To address this, we replace the loss in $e_\theta(x)$ with its logistic transformation

$$\ell_\theta^{\text{logistic}}(x) = \frac{1}{1 + \exp(-\ell_\theta(x))}. \quad (13)$$

The transformed loss $\ell_\theta^{\text{logistic}}(x)$ achieves two desired properties; it is differentiable and lies in the range $(0, 1)$ (hence, a good proxy for the error function). We denote the expected empirical risk of the logistic loss over a distribution P as

$$\hat{R}_S^{\text{logistic}}(P) = \mathbf{E}_{\theta \sim P} \frac{1}{N} \sum_{i=1}^N \ell_\theta^{\text{logistic}}(x_i).$$

At this point, the optimization problem can be formulated as

$$\begin{aligned} & \text{minimize} && \text{kl}^{-1} \left(\hat{R}_S^{\text{logistic}}(\mathcal{N}_{w,s}) \mid B(w, s, \lambda) \right) \\ & \text{subject to} && 0 \leq \lambda \leq \lambda^{\max} \\ & && s \geq 0, \end{aligned} \quad (14)$$

where the decision variables are $w \in \mathbf{R}^p$, $s \in \mathbf{R}_+^p$, and $\lambda \in \mathbf{R}_+^J$. In practice, a third and particularly practical obstacle arises with the initial formulation of our optimization problem. There is an imbalance between the expected empirical risk of the logistic loss $\hat{R}_S^{\text{logistic}}(\mathcal{N}_{w,s})$ and the regularizer $B(w, s, \lambda)$. The regularizer $B(w, s, \lambda)$ can be disproportionately large while the quantity $\hat{R}_S^{\text{logistic}}(\mathcal{N}_{w,s})$ is always in the range $(0, 1)$. We observe that in many cases, applying gradient-based methods to solve problem (14) tends to reduce the regularizer $B(w, s, \lambda)$ to zero, typically by making w close to zero, resulting in suboptimal solutions.

The penalized training problem. To remedy this problem, we add a penalty term to the objective to penalize the distance between $B(w, s, \lambda)$ and a hyperparameter $B^{\text{target}} \in \mathbf{R}_{++}$. Now, we are ready to define our *penalized training problem*:

$$\begin{aligned} & \text{minimize} && \text{kl}^{-1} \left(\hat{R}_S^{\text{logistic}}(\mathcal{N}_{w,s}) \mid B(w, s, \lambda) \right) + \mu (B(w, s, \lambda) - B^{\text{target}})^2 \\ & \text{subject to} && 0 \leq \lambda \leq \lambda^{\max} \\ & && s \geq 0. \end{aligned} \quad (15)$$

Here, $\mu \in \mathbf{R}_{++}$ is a large constant term that weights the penalty. The value of B^{target} will control the gap between the expected empirical risk and the expected risk. Specifically, if $B(w, s, \lambda) = B^{\text{target}}$, then the expected risk can be upper bounded with $R_{\mathcal{X}}(\mathcal{N}_{w,s}) \leq \hat{R}_S(\mathcal{N}_{w,s}) + \sqrt{B^{\text{target}}/2}$. In practice, we cross-validate over values for B^{target} , as detailed in Appendix B.1.

6.2 Differentiating through the KL inverse

In order to use gradient-based methods to solve the penalized training problem (15), we need to compute gradients through the KL inverse $p = \text{kl}^{-1}(q \mid c)$. However, the output p is not an explicit function of the inputs q and c . Rather, p is *implicitly* defined by q and c , and is obtained by solving the geometric program (31). Previous approaches that use gradient-based methods to minimize a PAC-Bayes bound, such as those employed by Dziugaite and Roy (2017) and Majumdar et al. (2021), sidestep this challenge by applying Pinsker’s inequality from Equation (32) to transform p into an explicit function of q and c . However, using Pinsker’s inequality can lead to a less precise bound compared to directly solving the KL inverse problem. In contrast to these methods, our approach leverages the technique of implicit differentiation, supported by the implicit function theorem (Dontchev and Rockafellar, 2009, Theorem 1B.1). Given the KL inverse $\text{kl}^{-1}(q \mid c)$, the implicit derivatives can be written as (Reeb et al., 2018)

$$\begin{aligned} \frac{\partial \text{kl}^{-1}(q \mid c)}{\partial q} &= \frac{\text{kl}^{-1}(q \mid c)(1 - \text{kl}^{-1}(q \mid c))}{\text{kl}^{-1}(q \mid c) - q} \left(\log \frac{q}{\text{kl}^{-1}(q \mid c)} + \log \frac{1 - \text{kl}^{-1}(q \mid c)}{1 - q} \right) \\ \frac{\partial \text{kl}^{-1}(q \mid c)}{\partial c} &= \frac{\text{kl}^{-1}(q \mid c)(1 - \text{kl}^{-1}(q \mid c))}{\text{kl}^{-1}(q \mid c) - q}. \end{aligned}$$

Smoothness of the KL inverse. We note that for $q \in (0, 1)$ and $c \in \mathbf{R}_{++}$, the KL inverse $\text{kl}^{-1}(q \mid c)$ lies in the range $(q, 1)$ (Reeb et al., 2018, Appendix A). Under these conditions, these derivatives exist, *i.e.*, $\text{kl}^{-1}(q \mid c)$ is differentiable with respect to both q and c . Since we use the logistic loss from Equation (13), the value for q is always in the range $(0, 1)$ (as opposed to taking a value strictly in $\{0, 1\}$). Additionally, the regularizer $c = B(w, s, \lambda)$ is always strictly positive. Hence our implicit layer is always differentiable and thus amenable to gradient-based optimization methods.

6.3 PAC-Bayes learning algorithm

In this subsection we present a learning algorithm based on gradient descent to solve the penalized training problem (15). We cannot apply vanilla gradient descent to solve problem (15) yet because we cannot compute the expected empirical logistic risk $\hat{R}_S^{\text{logistic}}(\mathcal{N}_{w,s})$ nor its gradients efficiently. We can, however, compute the gradient of its unbiased estimate $(1/N) \sum_{i=1}^N \ell_{w'}^{\text{logistic}}(x_i)$, where $w' = w + \xi \odot \sqrt{s}$ for $\xi \sim \mathcal{N}(0, I_p)$. In each iteration we take an i.i.d. copy of ξ and a step in the direction of the negative gradient of the function

$$C_S(w, s, \lambda, w') = \text{kl}^{-1} \left(\hat{R}_S^{\text{logistic}}(w') \mid B(w, s, \lambda) \right) + \mu(B(w, s, \lambda) - B^{\text{target}})^2.$$

To ensure the non-negativity of the variable s and λ , we optimize over variables $\zeta \in \mathbf{R}^d$ and $\eta \in \mathbf{R}^J$, and set $s = \exp(\zeta)$ and $\lambda = \exp(\eta)$. As mentioned in Section 5, after the gradient descent algorithm terminates, we must round the prior λ to fit into the pre-determined grid. To do so, we compute $a^* = \mathbf{round}(b \log(\lambda^{\max}/\lambda))$ and then $\lambda^* = \lambda^{\max} \exp(-a^*/b)$. We summarize this discretization via the function

$$\mathbf{roundPrior}(\lambda, \lambda^{\max}, b) = \lambda^{\max} \exp \left(\frac{-\mathbf{round}(b \log(\lambda^{\max}/\lambda))}{b} \right), \quad (16)$$

and set the rounded prior with $\lambda^* = \mathbf{roundPrior}(\lambda, \lambda^{\max}, b)$.

Algorithm 1 PAC-Bayes Learning to solve problem (15)

-
- 1: **Inputs:**
 - 2: Target penalty: $B^{\text{target}} \in \mathbf{R}_{++}$
 - 3: Prior hyperparameters: $\lambda^{\max} \in \mathbf{R}_{++}$, $b \in \mathbf{R}_{++}$
 - 4: Initial weights: $w_0 \in \mathbf{R}^p$, $s_0 \in \mathbf{R}_+^p$, $\lambda_0 \in (0, \lambda^{\max})^J$ ▷ Random initialization
 - 5: Desired probability: $\delta \in (0, 1)$
 - 6: Learning rate: $\gamma \in \mathbf{R}_{++}$
 - 7: Number of epochs: $M \in \mathbf{N}$
 - 8: **Procedure:**
 - 9: $(w, \zeta, \nu) = (w_0, \log(s_0), \log(\lambda_0))$
 - 10: **for** $i = 1$ **to** M **do** ▷ Loop over epochs
 - 11: sample $\xi \sim \mathcal{N}(0, I_p)$
 - 12: $w' = w + \xi \odot \sqrt{\exp(\zeta)}$ ▷ Sample from $\mathcal{N}_{w,s}$
 - 13:
$$\begin{bmatrix} w \\ \zeta \\ \nu \end{bmatrix} = \begin{bmatrix} w \\ \zeta \\ \nu \end{bmatrix} - \gamma \begin{bmatrix} \nabla_w C_S(w, \exp(\zeta), \exp(\nu), w') \\ \nabla_\zeta C_S(w, \exp(\zeta), \exp(\nu), w') \\ \nabla_\nu C_S(w, \exp(\zeta), \exp(\nu), w') \end{bmatrix}$$
 ▷ Gradient step
 - 14: $(w^*, s^*, \lambda^*) = (\exp(\zeta), \exp(\nu), \text{roundPrior}(\exp(\nu), \lambda^{\max}, b))$ ▷ Round prior: Eq. (16)
 - 15: **Outputs:**
 - 16: Learned weights (w^*, s^*, λ^*)
-

6.4 Calibrating the PAC-Bayes bounds

The training procedure returns the learned weights: the posterior mean w^* , the posterior variance s^* , and the prior variance λ^* . Together, these determine the posterior distribution $P = \mathcal{N}_{w^*, s^*}$ and the regularizer $B(w^*, s^*, \lambda^*)$. To obtain the final generalization bounds after the training procedure terminates, we need to *calibrate* the PAC-Bayes bounds for a given metric ϕ , number of algorithm steps k , and tolerance ϵ .

Conceptually, we would like to apply the McAllester bound to bound the expected risk $R_{\mathcal{X}}(P)$ in terms of the expected empirical risk $\hat{R}_S(P)$. However, this is not immediately possible since evaluating the expected empirical risk $\hat{R}_S(P)$ is intractable. To circumvent this issue, we generate \hat{P} a Monte Carlo approximation of P , compute the Monte Carlo estimate of the expected empirical risk $\hat{R}_S(\hat{P})$, and bound the expected empirical risk $\hat{R}_S(P)$ using inequality (33). We fully detail and enumerate the steps needed to calibrate the bounds below.

First, we draw H i.i.d. samples, denoted by $\{\theta_i\}_{i=1}^H$, from the posterior distribution P . We then construct the Monte Carlo approximation $\hat{P} = (1/H) \sum_{j=1}^H \delta_{\theta_j}$, where δ_{θ_j} represents the Dirac delta function centered at θ_j . We then run k steps of the learned optimizer for each of the H samples for each of the N training problems. Second, we compute the Monte Carlo approximation of the expected empirical risk $\hat{R}_S(P)$ as follows:

$$\hat{R}_S(\hat{P}) = \frac{1}{NH} \sum_{i=1}^N \sum_{j=1}^H e_{\theta_j}(x_i), \quad (17)$$

Algorithm 2 Calibrating the PAC-Bayes bounds

- 1: **Inputs:**
 - 2: Learned weights: w^*, s^*, λ^* ▷ Output of Algorithm 1
 - 3: Desired probabilities: $\delta, \omega \in (0, 1)$
 - 4: Metric: ϕ
 - 5: Number of algorithm steps: k
 - 6: Desired tolerance: ϵ
 - 7: Number of samples: H ▷ For Monte Carlo approx.
 - 8: **Procedure:**
 - 9: Generate H samples $\{\theta_j\}_{j=1}^H$ from \mathcal{N}_{w^*, s^*} ▷ Monte Carlo samples
 - 10: $\hat{R} = (1/(NH)) \sum_{j=1}^H \sum_{i=1}^N e_{\theta_j}(x_i)$ ▷ Empirical est. Eq. (17) (metric ϕ , k steps, tol. ϵ)
 - 11: $\bar{R} = \text{kl}^{-1}(\hat{R} \mid (1/H) \log(2/\omega))$ ▷ Sample convergence bound: Eq. (18)
 - 12: $R^* = \text{kl}^{-1}(\bar{R} \mid B(w^*, s^*, \lambda^*))$ ▷ main Thm. 1 bound: Eq. (19)
 - 13: **Outputs:**
 - 14: R^* ▷ The final bound on the expected risk
-

where the error function e is based on the underlying metric ϕ , number of steps k , and tolerance ϵ . Last, we apply two PAC-Bayes bounds to obtain the final bounds on the expected risk. Using the sample convergence bound from Theorem 3, the following inequality holds with probability at least $1 - \omega$ for $\omega \in (0, 1)$:

$$\hat{R}_S(P) \leq \bar{R}_S(P) = \text{kl}^{-1} \left(\hat{R}_S(\hat{P}) \mid \frac{1}{H} \log \frac{2}{\omega} \right). \quad (18)$$

We then apply a union bound and our Theorem 1 to obtain the final bound on the expected risk

$$R_{\mathcal{X}}(P) \leq R_S^*(P) = \text{kl}^{-1}(\bar{R}_S(P) \mid B(w^*, s^*, \lambda^*)), \quad (19)$$

which holds with probability $1 - \delta - \omega$.

We outline this calibration procedure in Algorithm 2, and we also depict the entire process to obtain generalization bounds for learned optimizers, including the training and calibration phases, in Figure 2. We note that the number of steps k that the bounds are computed for need not be the same as the number of steps K that are used to train the weights. Moreover, the metric ϕ does not need to be the same as the metric used in the loss function. We remark that some of choices made to facilitate training (*e.g.*, using the logistic regression loss instead of the 0–1 loss) do not affect the validity of our bounds. This is because our main generalization bound is applied *after* training is complete. Furthermore, by applying union bounds appropriately (*i.e.*, for the prior variance, the hyperparameter B^{target} , and the Monte Carlo approximation of the expected empirical risk), we ensure the validity of the final bound. Indeed, in the numerical experiments in Section 7, we will calibrate the bounds for many different tolerances and algorithm steps, and sometimes, multiple metrics.

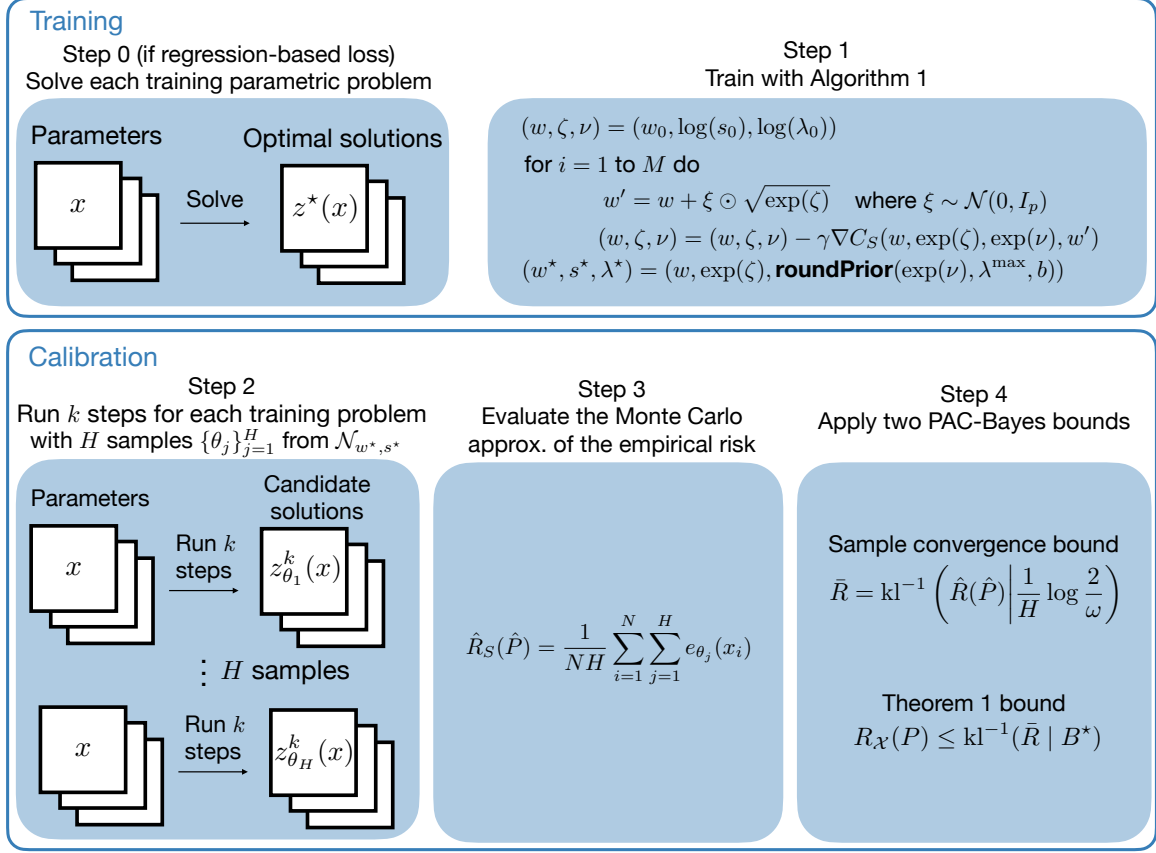


Figure 2: The two-phase procedure to generate generalization guarantees for learned optimizers for a metric ϕ , number of algorithm steps k , and tolerance ϵ . The first phase is the training phase. If the loss function is the regression-based loss, then we solve each parametric problem in step 0 as these are needed in order to train. In step 1, we train the architecture to optimize the PAC-Bayes guarantee over M epochs using Algorithm 1. We also round the prior according to Equation (16). Then we enter the second, calibration phase. In step 2 we sample weights $\{\theta_j\}_{j=1}^H$ from the distribution \mathcal{N}_{w^*, s^*} and run k algorithm steps for each training problem and each weight sample θ_j . In step 3 we compute the Monte Carlo approximation of the empirical expected risk $\hat{R}_S(\hat{P})$. In step 4, we bound the expected risk $R_{\mathcal{X}}(P)$ by applying a sample convergence bound from Equation (18) and then Theorem 1 where the regularization term is $B^* = B(w^*, s^*, \lambda^*)$.

7. Experiments

In this section, we illustrate the effectiveness of our guarantees for both classical and learned optimizers with numerical experiments. The code to reproduce our results is available at

https://github.com/stellatogrp/data_driven_optimizer_guarantees.

In Section 7.1 we apply our framework from Section 4 to provide guarantees for classical fixed-point optimization algorithms in the context of parametric optimization. In Section 7.2 we apply our training algorithm and generalization guarantees from Section 6 to obtain strong bounds for a variety of learned optimizers.

7.1 Guarantees for classical parametric optimization

In this subsection, we apply our method to obtain generalization guarantees to image deblurring in Section 7.1.1 and robust Kalman filtering in Section 7.1.2. We focus on solving convex QPs and convex conic programs, for which we use the Operator Splitting Quadratic Program (OSQP) solver (Stellato et al., 2020) and the Splitting Conic Solver (SCS) (O’Donoghue, 2021) respectively as the fixed-point algorithm. The fixed-point vector z consists of both primal and dual variables; see (Sambharya et al., 2024, Table 1) for more details on how this vector is constructed for OSQP and SCS. In each of the examples, we vary the number of algorithm steps and tolerances and report a lower bound on the success rate $1 - r_{\mathcal{X}}$. Then, by combining these bounds for the risk across many tolerances, we construct upper *quantile bounds* on the fixed-point residual at each algorithm step and compare them against the empirical quantile performance. See Section B.2 for more details on how we construct the quantiles. We show that our probabilistic guarantees are much tighter than bounds given through worst-case theoretical analysis. Finally, where relevant, we repeat our analysis to include task-specific metrics instead of the fixed-point residual, again providing risk and quantile bounds. The probabilistic results hold with probability at least 0.9999 for the risk and with probability at least 0.9919 for the quantiles. See Appendix Section B.3 for more details. We provide guarantees for 10, 100, and 1000 samples in each example.

Worst-case guarantees. The worst-case guarantee is determined by our best estimate; although, we remark that a better numerical result may be possible to obtain. Indeed, it can be difficult to check when certain conditions are met to guarantee a given convergence rate (*e.g.*, linear convergence for ADMM) (Yuan et al., 2020). To generate our best estimate of the worst-case guarantee, we proceed as follows. We first estimate a value for $\text{dist}_{\text{fix } T_x}(z^0)$ (where $z^0 = 0$ is the initial point) by taking the largest optimal solution in terms of its 2-norm across problem instances and multiplying it by 1.1. Then we generate the worst-case guarantees based on the property of the fixed-point operator given in the rates from (40) and (41). Both OSQP and SCS are algorithms based on Douglas-Rachford splitting (Banjac et al., 2019; O’Donoghue, 2021). For both algorithms, we pick hyperparameters so that the fixed-point algorithm is $(1/2)$ -averaged (Banjac et al., 2019; O’Donoghue, 2021). For OSQP, we set the penalty and relaxation parameters to be one; see Banjac et al. (2019) for more details. For SCS, we enforce identity scaling and set the relaxation parameter to be one; see O’Donoghue (2021) for more details. Hence, the sublinear rate with $\alpha = 1/2$ from (41) holds as a worst-case guarantee in all three instances. This rate can be improved upon if additional conditions (*e.g.*, strong convexity) are satisfied. We verified our theoretical bounds with PEP (Drori and Teboulle, 2014) using the PEPit toolbox (Goujaud et al., 2022). However, that approach does not scale well to more than 100 number of iterations because, in contrast to ours, it requires solving a semidefinite program (Drori and Teboulle, 2014). It takes nearly 59 minutes to provide guarantees for 70 iterations for algorithms where the iterations are averaged, for which the guarantees are within 6% of the theoretical bound. Because of the lack of scalability and how close the PEP bounds are to the theoretical guarantees, we omit them in our tables and plots. We emphasize that our probabilistic analysis is not meant to replace worst-case analyses; rather, it is meant to offer complementary insights. The comparisons illustrate the significant gap between worst-case bounds and the behavior observed on average over a parametric problem family.

7.1.1 IMAGE DEBLURRING

The first task we consider is image deblurring. Given a blurry image $x \in \mathbf{R}^n$, the goal is to recover the original image $y \in \mathbf{R}^n$. The vectors b and y are created by stacking the columns of the matrix representations of their images. We formulate the image deblurring problem as the QP

$$\begin{aligned} & \text{minimize} && \|Ay - x\|_2^2 + \rho \|y\|_1 \\ & \text{subject to} && 0 \leq y \leq 1, \end{aligned}$$

where $y \in \mathbf{R}^n$ is the decision variable. In this problem, the matrix $A \in \mathbf{R}^{n \times n}$ functions as a Gaussian blur operator embodying a two-dimensional convolutional operator. The regularizer coefficient $\rho \in \mathbf{R}_{++}$, balances the importance of the fidelity term $\|Ay - x\|_2^2$, relative to the ℓ_1 penalty.

Task-specific metric. In signal recovery tasks, it is common to report the normalized mean squared error (NMSE) in decibel (dB) units (Chen et al., 2022) between the z and the original signal \tilde{z} given by

$$\text{NMSE}(z, \tilde{z}) = 10 \log_{10} \frac{\|z - \tilde{z}\|_2^2}{\|\tilde{z}\|_2^2}. \quad (20)$$

Numerical example. We consider handwritten letters from the EMNIST dataset (Cohen et al., 2017). We apply a Gaussian blur of size 8 to each letter and then add i.i.d. Gaussian noise with standard deviation 0.001. The hyperparameter weighting term is $\rho = 10^{-4}$.

Results. Figure 3 shows our results. In this case, the objective is strongly convex, which can be used to guarantee linear convergence (Giselsson and Boyd, 2014). Therefore, we calculate the most optimistic linear convergence factor possible based on the performance in the samples and combine it with (41) to estimate the worst-case guarantee.

Table 1: The quantile results for image deblurring. Left: fixed-point residual. Right: NMSE. We report the number of iterations to reach given tolerances. For different quantiles (Qtl.) and tolerances (Tol.), we compare the empirical (Emp.) and estimated worst-case quantities against our probabilistic bounds with a varying number of samples N . The worst-case bound holds independently of the quantile.

Fixed-point residual							NMSE					
Qtl.	Tol.	Worst-Case	Emp.	Bound			Qtl.	Tol.	Emp.	Bound		
				$N = 10$	$N = 100$	$N = 1000$				$N = 10$	$N = 100$	$N = 1000$
30	0.01	80932	217	383	289	270	30	-20	152	985	290	210
	0.001	95975	1342	2903	1938	1674		-30	701	2707	1285	916
	0.0001	108779	8343	15765	11597	10512		-40	2688	5919	4016	3236
90	0.01	80932	285	-	383	362	90	-20	698	-	1497	985
	0.001	95975	2166	-	3274	2816		-30	2392	-	3885	2950
	0.0001	108779	12415	-	17018	15161		-40	5922	-	8050	6964
99	0.01	80932	320	-	-	-	99	-20	1338	-	-	2731
	0.001	95975	2615	-	-	4046		-30	3765	-	-	6707
	0.0001	108779	14523	-	-	17283		-40	8241	-	-	12077

7.1.2 ROBUST KALMAN FILTERING

Kalman filtering (Kalman, 1960) is a popular method to predict system states in the presence of noise in dynamic systems. In this example, we consider robust Kalman filtering (Xie

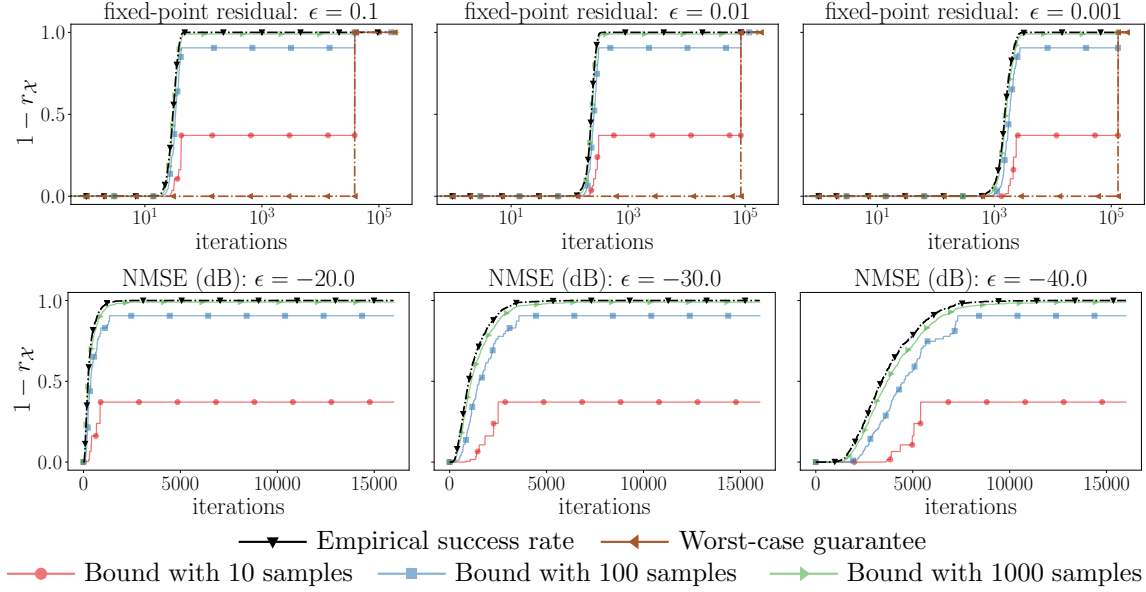


Figure 3: Probabilistic lower bounds of the success rate for image deblurring. The top row shows results for the fixed-point residual (fp. res.) and the bottom row shows bounds for the quantile. For both metrics, the lower bounds on the success rate are tight for $N = 1000$ samples.

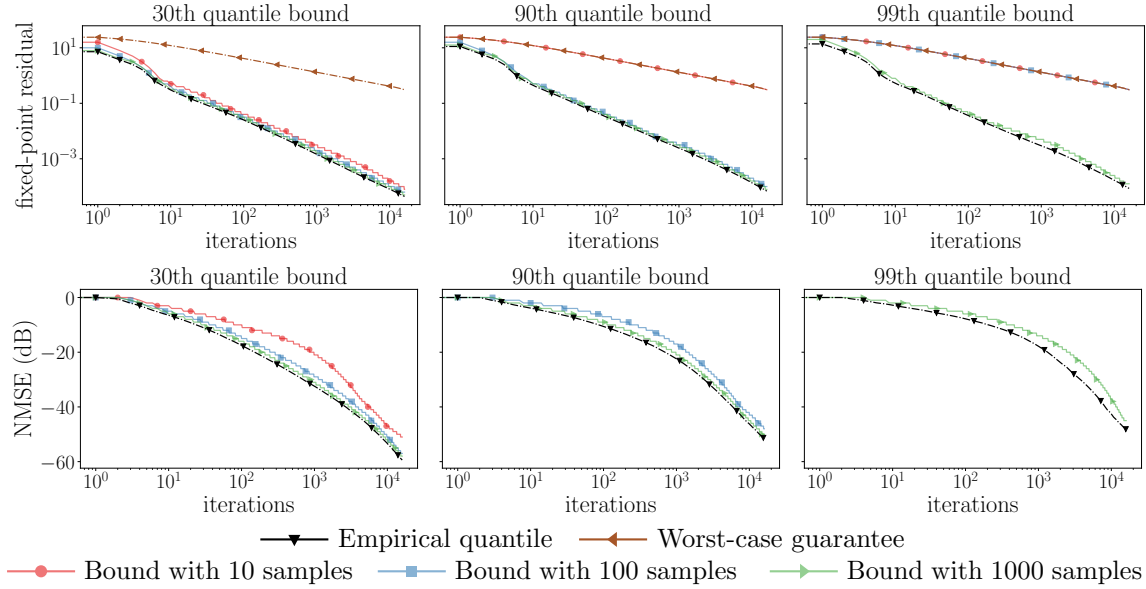


Figure 4: Probabilistic guarantees for OSQP to solve the image deblurring problem. The top row shows results for the fixed-point residual (fp. res.) and the bottom row shows results for the NMSE. The quantile bounds for both quantities improve as the number of samples increases.

and Soh, 1994) which mitigates the impact of outliers and model misspecifications to track a moving vehicle from noisy data location as in Venkataraman and Amos (2021). The linear

Table 2: The quantile results for robust Kalman filtering. Left: fixed-point residual. Right: max Euclidean distance. We report the number of iterations to reach given tolerances. For different quantiles (Qtl.) and tolerances (Tol.), we compare the empirical (Emp.) and estimated worst-case quantities against our probabilistic bounds with a varying number of samples N . The worst-case bound holds independently of the quantile.

Fixed-point residual							Max Euclidean distance					
Qtl.	Tol.	Worst-Case	Emp.	$N = 10$	$N = 100$	$N = 1000$	Qtl.	Tol.	Emp.	$N = 10$	$N = 100$	$N = 1000$
30	0.01	6.4e7	135	156	146	144	30	0.01	81	117	92	80
	0.001	6.4e9	220	239	230	229		0.001	147	193	164	156
	0.0001	6.4e11	308	327	321	319		0.0001	228	270	236	235
90	0.01	6.4e7	144	-	158	154	90	0.01	103	-	144	121
	0.001	6.4e9	228	-	243	238		0.001	175	-	222	194
	0.0001	6.4e11	317	-	335	328		0.0001	250	-	307	267
99	0.01	6.4e7	150	-	-	162	99	0.01	116	-	-	150
	0.001	6.4e9	233	-	-	247		0.001	176	-	-	222
	0.0001	6.4e11	324	-	-	336		0.0001	251	-	-	311

dynamical system with matrices $A \in \mathbf{R}^{n_s \times n_s}$, $B \in \mathbf{R}^{n_s \times n_u}$, and $C \in \mathbf{R}^{n_o \times n_s}$ is given by

$$s_{t+1} = As_t + Bw_t, \quad y_t = Cs_t + v_t, \quad \text{for } t = 0, 1, \dots, \quad (21)$$

where $s_t \in \mathbf{R}^{n_s}$ is the state, $y_t \in \mathbf{R}^{n_o}$ is the observation, $w_t \in \mathbf{R}^{n_u}$ is the input, and $v_t \in \mathbf{R}^{n_o}$ is a perturbation to the observation. We aim to recover the state x_t from the noisy measurements y_t by solving the following problem:

$$\begin{aligned} & \text{minimize} && \sum_{t=1}^{T-1} \|w_t\|_2^2 + \mu \psi_\rho(v_t) \\ & \text{subject to} && s_{t+1} = As_t + Bw_t \quad t = 0, \dots, T-1 \\ & && y_t = Cs_t + v_t \quad t = 0, \dots, T-1. \end{aligned} \quad (22)$$

Here, the Huber penalty function (Huber, 1964) with parameter $\rho \in \mathbf{R}_{++}$ that robustifies against outliers is given by

$$\psi_\rho(a) = \begin{cases} \|a\|_2 & \|a\|_2 \leq \rho \\ 2\rho\|a\|_2 - \rho^2 & \|a\|_2 \geq \rho. \end{cases}$$

The given quantity $\mu \in \mathbf{R}_{++}$ weights this penalty term. The decision variables are the s_t 's, w_t 's, and v_t 's, while the parameters are the observed y_t 's: $x = (y_0, \dots, y_{T-1})$. We formulate problem (22) as a second-order cone program, and use SCS (O'Donoghue, 2021) to solve it.

Task-specific metric. In Kalman filtering, traditional metrics such as the fixed-point residual, which encompasses both primal and dual variables, may not fully capture specific aspects of state recovery accuracy. In light of this context, we propose a task-specific metric aimed at quantifying the fidelity of state estimation. This metric measures the deviation of the algorithmically recovered states, s_1, \dots, s_T , (extracted from the fixed-point vector z) after k iterations, from their corresponding optimal states, $s_1^*(x), \dots, s_T^*(x)$:

$$\phi(z, x) = \max_{t=1, \dots, T} \|s_t - s_t^*(x)\|_2. \quad (23)$$

The associated error metric indicates success when each recovered state lies within an ϵ -radius ball centered at its optimal counterpart.

Numerical example. We follow the setup from Venkataraman and Amos (2021) where $n_s = 4$, $n_o = 2$, $n_u = 2$, $\mu = 2$, $\rho = 2$, and $T = 50$. The dynamics matrices are

$$A = \begin{bmatrix} 1 & 0 & (1 - (\gamma/2)\Delta t)\Delta t & 0 \\ 0 & 1 & 0 & (1 - (\gamma/2)\Delta t)\Delta t \\ 0 & 0 & 1 - \gamma\Delta t & 0 \\ 0 & 0 & 0 & 1 - \gamma\Delta t \end{bmatrix}, B = \begin{bmatrix} 1/2\Delta t^2 & 0 \\ 0 & 1/2\Delta t^2 \\ \Delta t & 0 \\ 0 & \Delta t \end{bmatrix}, C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

where $\Delta t = 0.5$ and $\gamma = 0.05$ are fixed to be respectively the sampling time and the velocity dampening parameter. We generate true trajectories $\{x_0^*, \dots, x_{T-1}^*\}$ of the vehicle by first letting $x_0^* = 0$. Then we sample the inputs as $w_t \sim \mathcal{N}(0, 0.01)$ and $v_t \sim \mathcal{N}(0, 0.01)$. The trajectories are then fully defined via the dynamics equations in Equation (21) with the sampled w_t 's and v_t 's.

Results. To the best of our knowledge, the tightest guarantees that can be obtained for this problem on the fixed-point residual are given the bound from the averaged iterations (41). We visualize our bounds on the maximum Euclidean distance metric from Equation (23) in Figure 7 with a ball of radius 0.1 around the optimal state. We obtain probabilistic guarantees on the error metric that says that all of the recovered states are within their respective balls.

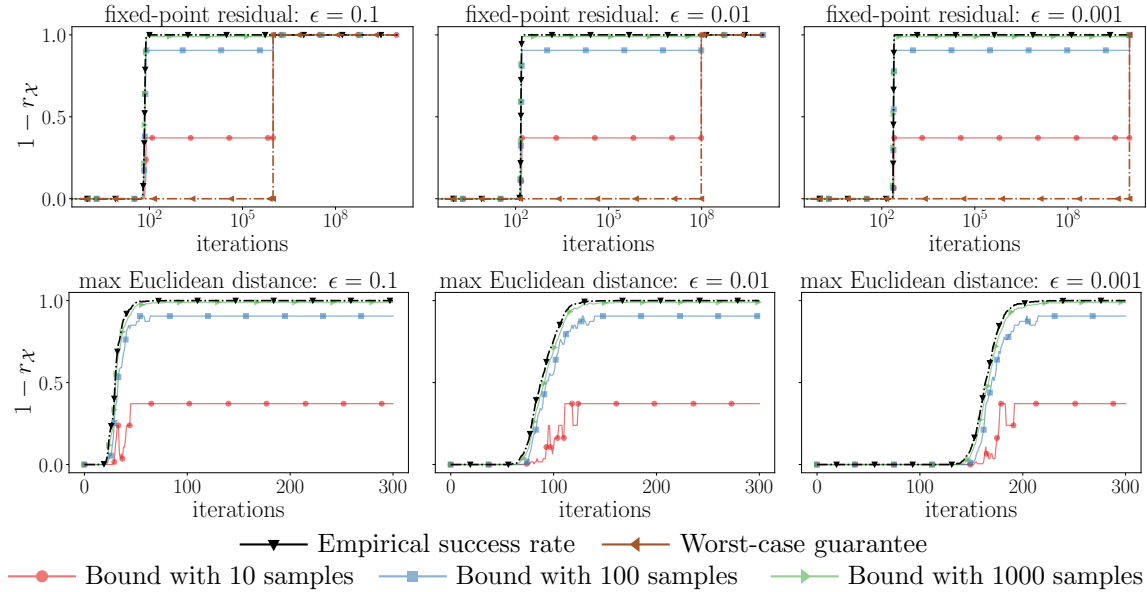


Figure 5: Probabilistic lower bounds of the success rate for robust Kalman filtering. Top: fixed-point residual. Bottom: maximum Euclidean distance from Equation (23). Note that the x-axes are different for the top and bottom rows. The bounds get tighter as the number of samples increases.

7.2 Learned optimizers

In this subsection we apply our method to obtain generalization guarantees for a variety of learned optimizers: LISTA (Gregor and LeCun, 2010) and several of its variants (Liu

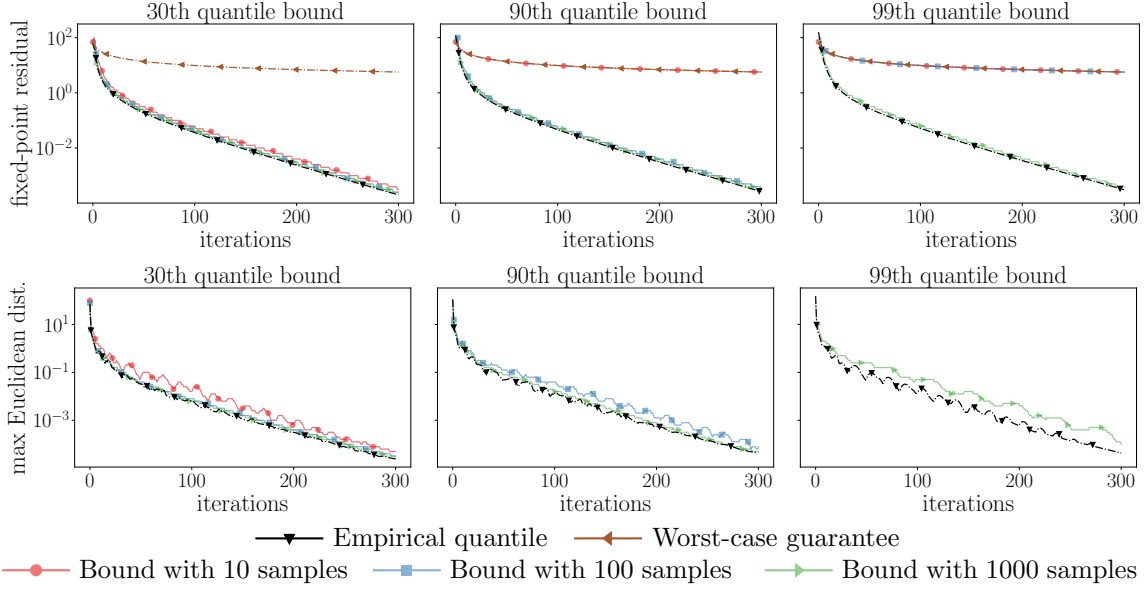


Figure 6: Probabilistic guarantees for SCS to solve the robust Kalman filtering problem. Top: fixed-point residual. Bottom: maximum Euclidean distance from Equation (23). Our bounds resemble linear convergence, while the worst-case guarantee gives sublinear convergence.

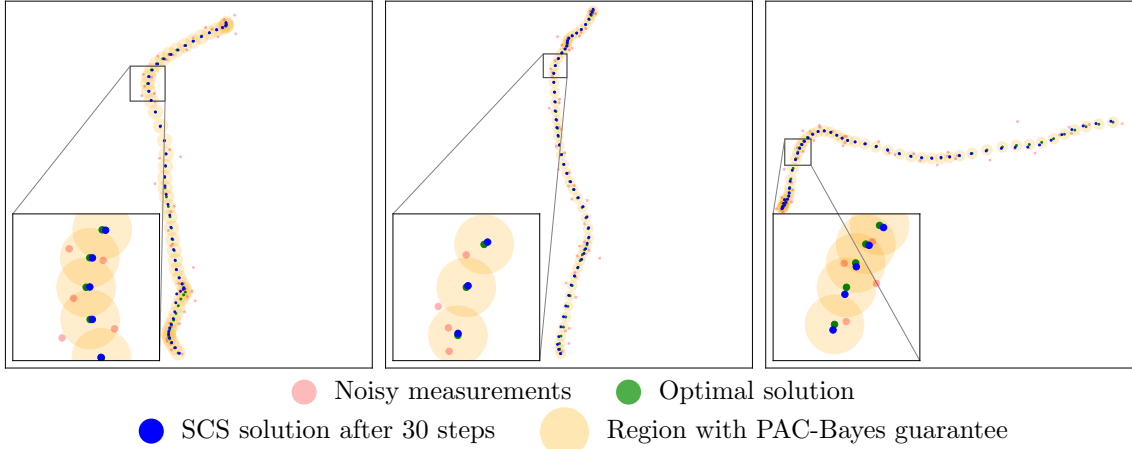


Figure 7: Visualizing the guarantees for robust Kalman filtering with the maximum Euclidean distance metric (23). Each plot is a separate parametric problem. The noisy, observed trajectory is made up of the pink points which parametrize the problem. The robust Kalman filtering recovery, the optimal solution of problem (22), is shown as green points. The shaded beige regions are centered at the optimal points with radius 0.1. With high probability, all of the extracted states after 30 steps will fall within their respective beige regions across 86% of problem instances.

et al., 2019; Wu et al., 2020) in Section 7.2.1, learning warm starts (L2WS) for fixed-point optimization problems (Sambharya et al., 2024) in Section 7.2.2, and model-agnostic meta-learning (MAML) (Finn et al., 2017) in Section 7.2.3.

We implement our learning algorithm for the different learned optimizers in JAX (Bradbury et al., 2018), using the JAXOPT library (Blondel et al., 2021) with the ADAM optimizer (Kingma and Ba, 2015). To do the implicit differentiation during training, we use a bisection method to solve the KL inverse problems. For each of the learned optimizers, we describe how we partition the weights into groups as mentioned in Section 5, and report the number of partitions J . We also describe how the prior mean is set for each learned optimizer. We use 50000 training samples and evaluate on 1000 test problems in each example.

After training, we calibrate our bounds, and report a lower bound on the success rate $1 - R_{\mathcal{X}}(P)$ for the learned optimizer with posterior distribution $P = \mathcal{N}_{w^*, s^*}$. As in the results for the classical optimizers, we report bounds across many algorithm steps and tolerances, construct quantile bounds at each step, and report the results for task-specific metrics where applicable. The probabilistic results hold with probability at least 0.99988 for the risk and with probability at least 0.99028 for the quantiles; see Appendix Section B.3 for details.

7.2.1 LISTA VARIANTS FOR SPARSE CODING PROBLEMS

In the sparse coding problem, the goal is to recover a sparse vector $z \in \mathbf{R}^n$ given a dictionary $D \in \mathbf{R}^{m \times n}$ from noisy linear measurements

$$b = Dz + \epsilon,$$

where $b \in \mathbf{R}^m$ is the noisy measurement and $\epsilon \in \mathbf{R}^m$ is additive Gaussian white noise. A popular approach to solve this problem is to formulate it as the lasso problem

$$\text{minimize } (1/2)\|Dz - b\|_2^2 + \rho\|z\|_1, \quad (24)$$

where $\rho \in \mathbf{R}_{++}$ is a hyperparameter, and then run the iterative shrinkage thresholding algorithm (ISTA) with algorithm steps

$$z^{k+1} = \eta_{\rho/L} \left(z^k - \frac{1}{L} D^T (Dz^k - b) \right).$$

Here η_{ψ} is the soft-thresholding function $\eta_{\psi}(z) = \text{sign}(z) \max(0, |z| - \psi)$ and $L \in \mathbf{R}_{++}$ is less than or equal to the largest eigenvalue of $D^T D$. Seeking faster convergence, learned ISTA (LISTA) (Gregor and LeCun, 2010) and its variants learn some of the components of the update function. All of these learned optimizers seek a good set of weights θ to solve problem (8) where the initial iterate is set to zero, *i.e.*, $h_{\theta}(x) = 0$. In this subsection, we apply our method to LISTA and several of its variants enumerated below, and compare the performance against classical algorithms: ISTA and its accelerated version, Fast ISTA (FISTA) (Beck and Teboulle, 2009).

LISTA. The LISTA updates from the seminal work of Gregor and LeCun (2010) are

$$z^{k+1} = \eta_{\psi^k} \left(W_1^k z^k + W_2^k b \right), \quad (25)$$

where the learned parameters are $\theta = (\{\psi^k, W_1^k, W_2^k\}_{k=0}^{K-1}) \in \mathbf{R}^{K(1+mn+n^2)}$. We partition the weights into $J = 3$ groups: the shrinkage thresholds $\{\psi^k\}_{k=0}^{K-1}$, the first set of weight matrices $\{W_1^k\}_{k=0}^{K-1}$, and the second set of weight matrices $\{W_2^k\}_{k=0}^{K-1}$. We set the prior means for the weights to the values of ISTA with $\rho = 0.1$.

TiLISTA. TiLISTA (Liu et al., 2019), a variant of LISTA, couples the two weight matrices and ties the matrix updates over the iterates so that they only differ by a learned scalar factor. The TiLISTA updates are given by

$$z^{k+1} = \eta_{\psi^k} \left(z^k - \gamma^k \tilde{W}^T (Dz^k - b) \right),$$

where the weights are $\theta = (\tilde{W}, \{\psi^k, \gamma^k\}_{k=0}^{K-1}) \in \mathbf{R}^{2K+mn}$. We partition the weights into $J = 3$ groups: the shrinkage thresholds $\{\psi^k\}_{k=0}^{K-1}$, the step sizes $\{\gamma^k\}_{k=0}^{K-1}$, and the matrix \tilde{W} . We set the prior mean for \tilde{W} to be the pre-computed value given by solving problem (27) and zero for the other groups.

ALISTA. Liu et al. (2019) also propose ALISTA, which significantly reduces the number of learned algorithm parameters by determining the matrix \tilde{W} from TiLISTA in a data-free manner. The ALISTA updates are given by

$$z^{k+1} = \eta_{\psi^k} (z^k - \gamma^k \tilde{W}^T (Dz^k - b)), \quad (26)$$

where $\tilde{W} \in \mathbf{R}^{m \times n}$ is pre-computed in a data-free manner by solving the convex QP

$$\begin{aligned} & \text{minimize} \quad \|W^T D\|_F^2 \\ & \text{subject to} \quad W_{:,i}^T D_{:,i} = 1 \quad i = 1, \dots, m. \end{aligned} \quad (27)$$

Then the parameters $\theta = (\{\psi^k, \gamma^k\}_{k=0}^{K-1}) \in \mathbf{R}^{2K}$ are learned in an end-to-end fashion. We partition the weights into $J = 2$ groups: the shrinkage thresholds $\{\psi^k\}_{k=0}^{K-1}$ and the step sizes $\{\gamma^k\}_{k=0}^{K-1}$. Since the matrix \tilde{W} is pre-computed for ALISTA in a data-free manner, we do not train over this variable. We set all of the prior means for the weights to zero.

GLISTA. In GLISTA (Wu et al., 2020), we incorporate gain gates and overshoot gates to the ALISTA model. The GLISTA updates are given by

$$\begin{aligned} \tilde{z}^{k+1} &= \eta_{\psi^k} \left(1 + \mu^k \psi^{k-1} \exp(\nu_k |z^k|) - \gamma^k \tilde{W}^T \left(D(1 + \mu^k \psi^{k-1} \exp(\nu_j |z^k|)) - b \right) \right) \\ z^{k+1} &= \left(1 + \frac{a^k}{|\tilde{z}^{k+1} - z^k| + \epsilon} \right) \odot \tilde{z}^{k+1} - \left(\frac{a^k}{|\tilde{z}^{k+1} - z^k| + \epsilon} \right) \odot z^k, \end{aligned}$$

where $\epsilon \in \mathbf{R}_{++}$ is a small positive value. The weight matrix \tilde{W} is pre-computed in a data-free manner as in ALISTA. We partition the weights into $J = 5$ groups: the shrinkage thresholds $\{\psi^k\}_{k=0}^{K-1}$, the step sizes $\{\gamma^k\}_{k=0}^{K-1}$, the two sets of gated parameters $\{\mu^k\}_{k=0}^{K-1}$ and $\{\nu^k\}_{k=0}^{K-1}$, and the overshoot parameters $\{a^k\}_{k=0}^{K-1}$. Thus the learned parameters are $\theta = (\{\psi^k, \gamma^k, \mu^k, \nu^k, a^k\}_{k=0}^{K-1}) \in \mathbf{R}^{5K}$. We set all of the prior means for the weights to zero.

Task-specific metric. In this example, we only report normalized mean squared error in decibel (dB) units from Equation (20) as this is common in the literature for sparse coding (Chen et al., 2022).

Numerical example. We follow the setup from Chen et al. (2022) for this example. We sample a dictionary $D \in \mathbf{R}^{m \times n}$ with i.i.d. entries from the distribution $\mathcal{N}(0, 1/m)$. Then we normalize D so that each column has Euclidean norm of one. To generate each sample, we generate the ground truth from the distribution $\mathcal{N}(0, 1)$ and zero out each entry with a probability of 0.9. The noise ϵ is set to a signal to noise ratio of 40dB. Then the measurement is $b = Dz + \epsilon$. We take a matrix with dimensions $m = 256$, $n = 512$, and pick the number of algorithm steps to be $K = 10$. We compare against ISTA and FISTA, setting $\rho = 0.1$, a value picked in Chen et al. (2018b). We calibrate with 20000 Monte Carlo samples of the weights.

Results. Figure 8 along with Table 3 show the behavior of our method. The classical optimizers ISTA and FISTA hardly make progress within 10 iterations as is commonly observed in the literature (Chen et al., 2018b). Our method with all of the learned optimizers except for LISTA provides generalization guarantees that are much stronger than the baseline performance. Moreover, the guarantees are close to the empirical results showing that our bounds are tight. Our method with LISTA performs poorly because there are a very large number of weights, which in turn makes the regularizer $B(w, s, \lambda)$ significantly larger and more difficult to optimize.

Table 3: The quantile results for sparse coding on the NMSE after 10 iterations. We report the empirical average of test problems (Emp.) and bounds (Bnd.) for each of the learned optimizers.

Quantile	ISTA Emp.	FISTA Emp.	LISTA		TiLISTA		ALISTA		GLISTA	
			Emp.	Bnd.	Emp.	Bnd.	Emp.	Bnd.	Emp.	Bnd.
10	-0.93	-1.99	-2.2	-1.0	-36.79	-34.0	-38.09	-37.0	-43.96	-43.0
30	-0.53	-1.86	-1.82	-1.0	-34.48	-32.0	-36.38	-35.0	-42.74	-42.0
50	-0.39	-1.78	-1.57	-1.0	-33.05	-31.0	-35.06	-33.0	-41.78	-41.0
60	-0.31	-1.73	-1.46	0.0	-32.31	-30.0	-34.52	-32.0	-41.33	-40.0
80	-0.16	-1.66	-1.22	0.0	-30.08	-27.0	-31.99	-30.0	-40.15	-39.0
90	-0.09	-1.61	-1.04	0.0	-28.58	-22.0	-29.82	-27.0	-39.19	-38.0
95	-0.0	-1.54	-0.92	0.0	-27.38	-18.0	-29.06	-23.0	-38.37	-36.0

7.2.2 LEARNING TO WARM STARTS FOR FIXED-POINT PROBLEMS

In our second example of learned optimizers, we consider the L2WS framework (Sambharya et al., 2024) which seeks to learn a high-quality initialization to solve the fixed-point problem (3). The training problem is problem (8) where the initialization $h_\theta(x)$ is learned rather than the algorithm steps (*i.e.*, $T_\theta(z, x) = T(z, x)$). The objective is the fixed-point residual $f(z, x) = \|z - T(z, x)\|_2$. Here, $h_\theta : \mathbf{R}^d \rightarrow \mathbf{R}^n$ is a neural network with ReLU activation functions, and the warm start is computed as

$$h_\theta(x) = W_{L-1}\psi(W_{L-2}\psi(\dots\psi(W_0x + b_0)) + b_{L-2}) + b_{L-1},$$

where $\psi(z) = \max(0, z)$ element-wise, the matrix in the i -th layer is $W_i \in \mathbf{R}^{m_i \times n_i}$, and the bias term in the i -th layer is $b_i \in \mathbf{R}^{n_i}$. The learnable weights consist of all of the weight and bias terms in the neural network, *i.e.*, $\theta = (W_0, b_0, \dots, W_{L-1}, b_{L-1})$. For this approach, $\ell_\theta : \mathbf{R}^n \rightarrow \mathbf{R}_+$ can take either the form of the regression loss from Equation (6) or objective loss from Equation (7). We partition the weights into $J = 2L$ groups corresponding to each bias and weight term in each layer. We set the prior means to be zero for all of the weights.

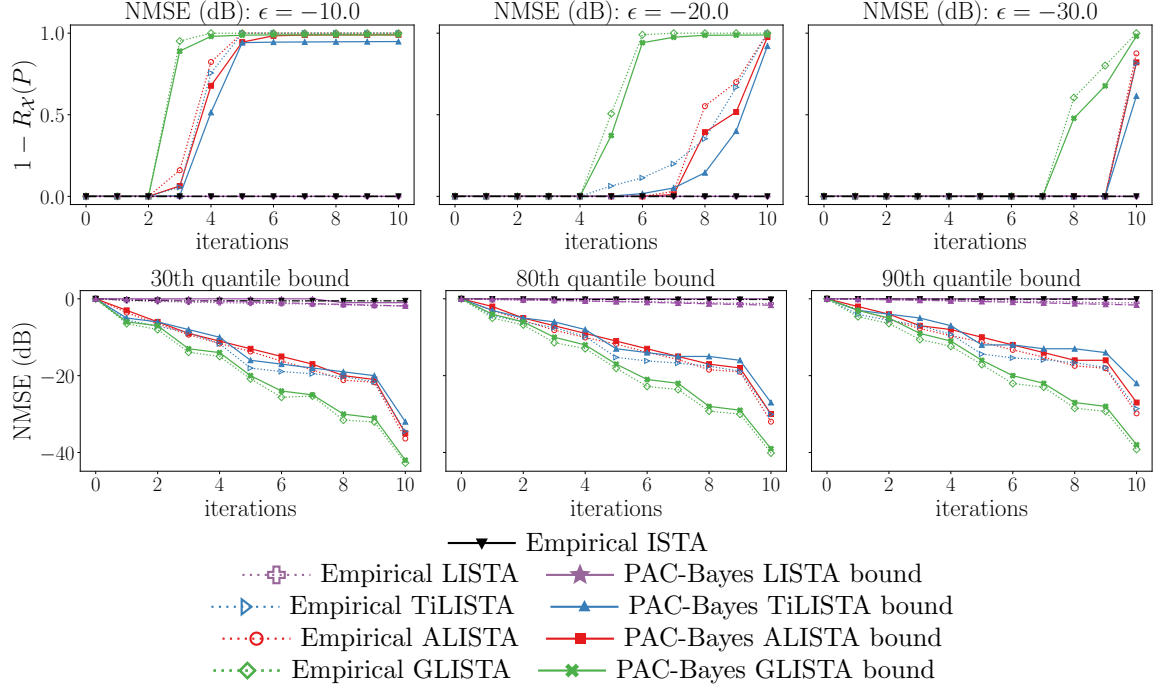


Figure 8: Sparse coding results. Top: lower bound on the success rate. Bottom: upper bound on the quantile. The PAC-Bayes guarantees for ALISTA, TiLISTA, and GLISTA significantly outperform the empirical results given by ISTA. The guarantees are close to the corresponding empirical values for the learned optimizers.

Strengthening the bounds. As in the case of classical optimizers, one downside of our approach is that the bound on the expected risk cannot reach exactly zero (even for a very large number of iterations) due to a non-zero regularization term. For the L2WS framework specifically, we bypass this problem and show how the expected risk can be bounded to zero with high probability. We first bound the distance from the warm start to optimality $\text{dist}_{\text{fix } T_x}(h_\theta(x))$ with the following theorem.

Theorem 2 Let $w = (W_0, b_0, \dots, W_{L-1}, b_{L-1})$ and $s = (\Sigma_0, \sigma_0, \dots, \Sigma_{L-1}, \sigma_{L-1})$ be the mean and variance terms of the weights of an L -layer stochastic neural network. Let \bar{x} and \bar{z} be upper bounds on $\|x\|_2$ and $\|z^*(x)\|_2$ for any x drawn from the distribution \mathcal{X} . Let $a_0^* = \bar{x}$,

$$a_{i+1}^* = \left(\|W_i\|_2 + \|b_i\|_2 + v_i \sqrt{2(m_i + n_i + 1) \log((L - \delta)/(2Lh))} \right) (a_i^* + 1), \quad \tilde{\Sigma}_i = \begin{bmatrix} \Sigma_i \\ \sigma_i^T \end{bmatrix},$$

for $i = 0, \dots, L - 1$, where $v_i^2 = \max\{\max_j \|(\tilde{\Sigma}_i)_{j:}^{1/2}\|_2^2, \max_k \|(\tilde{\Sigma}_i)_{:,k}^{1/2}\|_2^2\}$. Then with probability at least $1 - \delta$ the following bound holds for any x drawn from the distribution \mathcal{X} :

$$\text{dist}_{\text{fix } T_x}(h_\theta(x)) \leq \bar{z} + a_L^*.$$

See Appendix C.2 for the proof. This upper bound on the distance from the warm start to an optimal solution given by $\bar{z} + a_L^*$ can be easily input into inequalities (40) and (41)

to bound the fixed-point residual for a given number of iterations. To see this, recall that inequalities (40) and (41) include a term $\|z^0(x) - z^*(x)\|_2$ and that the bounds hold for *any* $z^*(x) \in \mathbf{fix} T_x$. Therefore, replacing these terms with $\mathbf{dist}_{\mathbf{fix} T_x}(h_\theta(x))$ yields valid inequalities on the fixed-point residual.

Unconstrained quadratic optimization. The learned warm starts example we consider is an unconstrained quadratic optimization problem

$$\text{minimize} \quad (1/2)z^T P z + c^T z,$$

where $P \in \mathbf{S}_{++}^n$, and $c \in \mathbf{R}^n$ are the problem data and $z \in \mathbf{R}^n$ is the decision variable. The parameter is $x = c$ and the fixed-point algorithm is gradient descent.

Numerical example. We take the first example, from Sambharya et al. (2024) where $n = 20$, and the neural network has a single hidden layer with 10 neurons. Let $P \in \mathbf{S}_{++}^n$ be a diagonal matrix where the first 10 diagonals take the value 100 and the last ten take the value of 1. Let $x = c \in \mathbf{R}^n$. Here, the i -th index of x is sampled according to the uniform distribution $\mu_i \mathcal{U}[-10, 10]$, where $\mu_i = 10000$ if $i \leq 10$ else 1. We pick $K = 15$ steps for training and use the fixed-point residual loss. We calibrate with 1000 Monte Carlo samples of the weights.

Results. Figures 9 along with Table 4 show the behavior of our method. The PAC-Bayes guarantees outperform both the cold start and the nearest neighbor.

Table 4: The quantile results for L2WS on unconstrained QP results on the number of iterations required to reach a given tolerance. For different quantiles and tolerances (Tol.), we compare the cold start and nearest neighbor empirical performances against our learned warm starts for which we report the empirical (Emp.) quantile and the bound (Bnd.).

Quantile	Tol.	Cold Start	Nearest Neighbor	L2WS Emp.	L2WS Bnd.
30	0.01	280	234	1	1
	0.001	509	463	169	206
	0.0001	738	692	398	435
80	0.01	300	258	1	31
	0.001	529	487	207	260
	0.0001	758	716	436	490
90	0.01	304	264	1	158
	0.001	533	493	221	388
	0.0001	763	723	450	617

7.2.3 MODEL-AGNOSTIC META-LEARNING

In this subsection, we apply our method to obtain generalization guarantees for the MAML framework (Finn et al., 2017), which aims to learn a model that quickly generalize to new tasks from minimal training examples. Each task \mathcal{T} is associated with a dataset \mathcal{D} , split into two disjoint sets: the training set $\mathcal{D}^{\text{train}}$ and the test set $\mathcal{D}^{\text{test}}$. The dataset $\mathcal{D}^{\text{train}}$ consists of K^{train} input-output pairs $\{a_i, y_i\}_{i=1}^{K^{\text{train}}}$. Similarly, $\mathcal{D}^{\text{test}}$ consists of K^{test} input-output pairs. At its heart, MAML seeks to optimize a model’s initial parameters θ , so it can quickly adapt to unseen tasks. MAML fits into the learning to optimize framework

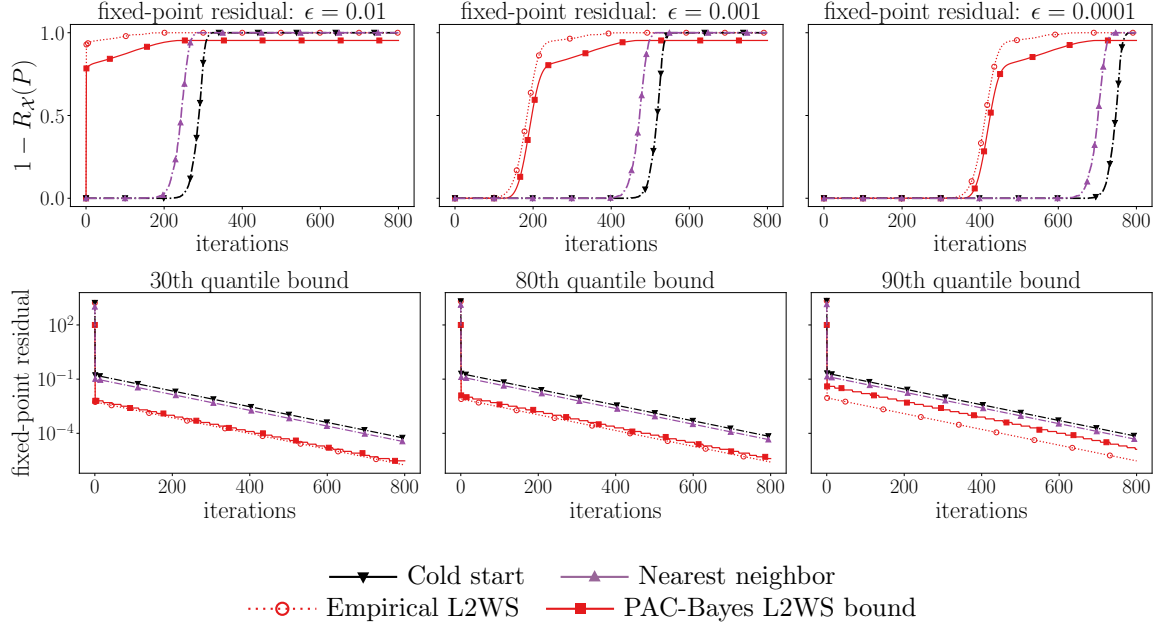


Figure 9: L2WS unconstrained QP fixed-point residual results. Top: lower bounds on the success rate. Bottom: upper bounds on the quantiles. The PAC-Bayes bound is very close to the empirical curve and outperforms both the cold start and the nearest neighbor curves.

from Section 3.2 where the parameter is the training set, *i.e.*, $x = \mathcal{D}^{\text{train}}$. The pre-defined (*i.e.*, not learned) update function is a step in the direction of the negative of the gradient of the loss over the training set $\mathcal{D}^{\text{train}}$, *i.e.*,

$$z^{k+1}(x) = z^k(x) - \gamma \nabla_z \mathcal{L}(z^k(x), x).$$

Here, γ is a pre-determined positive number indicating the step size. MAML learns the initial parameters $h_\theta(x) = \theta$, which is shared across tasks. The loss for the learned optimizer is computed on the test set and is calculated as

$$\ell_\theta(x) = \mathcal{L}(\hat{z}_\theta(x), \mathcal{D}^{\text{test}}).$$

We consider regression tasks where \mathcal{L} gives the mean squared error (MSE) in a dataset \mathcal{D} :

$$\mathcal{L}(z, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} (g_z(a_i) - y_i)_2^2.$$

Here, g_z is the neural network predictor with weights z . We partition the weights into $2L$ groups as in Section 7.2.2. We set the prior means to be zero for all of the weights.

Sinusoid curves. We consider the meta-learning task of regressing inputs to outputs of sine waves using a few datapoints as in Finn et al. (2017). We generate each task by first sampling an amplitude A and a phase b . We then generate the datasets $\mathcal{D}^{\text{train}}$ and $\mathcal{D}^{\text{test}}$ in the following manner. The inputs a are uniformly sampled from an interval, and the corresponding outputs are given by $y = A \sin(a - b)$. The neural network consists of two hidden layers of size 40 each with ReLU activations.

Task-specific metric. To help visualize our results, we consider the task-specific metric that is the ℓ_∞ norm of the errors over the dataset $\mathcal{D}^{\text{test}} = \{(a_i, y_i)\}_{i=1}^{K^{\text{test}}}$:

$$\max_{i=1, \dots, K^{\text{test}}} |g_z(a_i) - y_i|. \quad (28)$$

Numerical example. We follow the exact setup from Finn et al. (2017). For each task \mathcal{T} , we sample an amplitude A from the uniform distribution $\mathcal{U}[0.1, 5.0]$ and a phase from the uniform distribution $\mathcal{U}[0, \pi]$. All of the a datapoints are sampled i.i.d. from the uniformly from $[-5.0, 5.0]$. We pick the number of datapoints in the training and test sets to be $K^{\text{train}} = 5$ and $K^{\text{test}} = 100$ respectively. The step size γ is 0.01, and we unroll 2 steps during training. We calibrate with 20000 Monte Carlo samples of the weights.

Results. Figures 10 and 11 along with Table 5 show the behavior of our method with two unrolled steps. In this example, the baseline that we compare against is the pretrained model from Finn et al. (2017) which trains the network on the sinusoid curves without unrolling any algorithm steps. For both metrics, our bounds are much stronger than the pretrained model. We visualize our results in Figure 12. We obtain probabilistic guarantees that the solution returned after 10 steps initialized with MAML will fall within a band of width two centered around the true sine curve. The deterministic pretrained model fails to completely fall within the band of error for many of the problems.

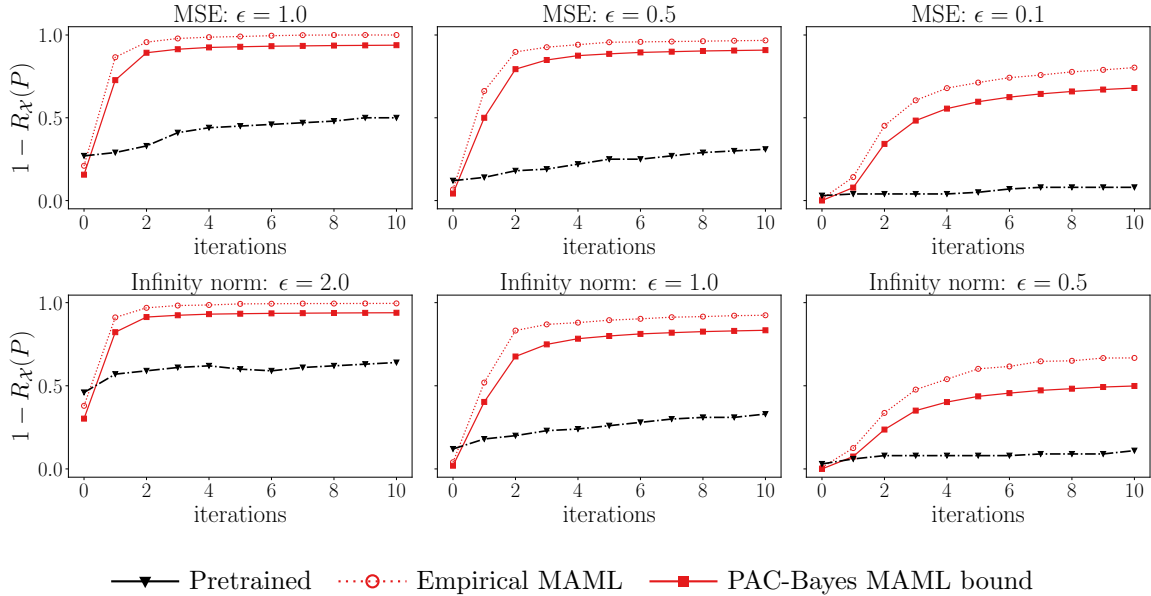


Figure 10: MAML success rate results for sinusoid curves. Top: MSE. Bottom: infinity norm from Equation (28). Our lower bounds on the success rate $1 - R_{\mathcal{X}}(P)$ for both metrics are much higher than the empirical success rate of the pretrained model across many tolerances.

8. Conclusion

We present a data-driven framework to provide guarantees for the performance of classical and learned optimizers in the setting of parametric optimization. For classical optimizers,

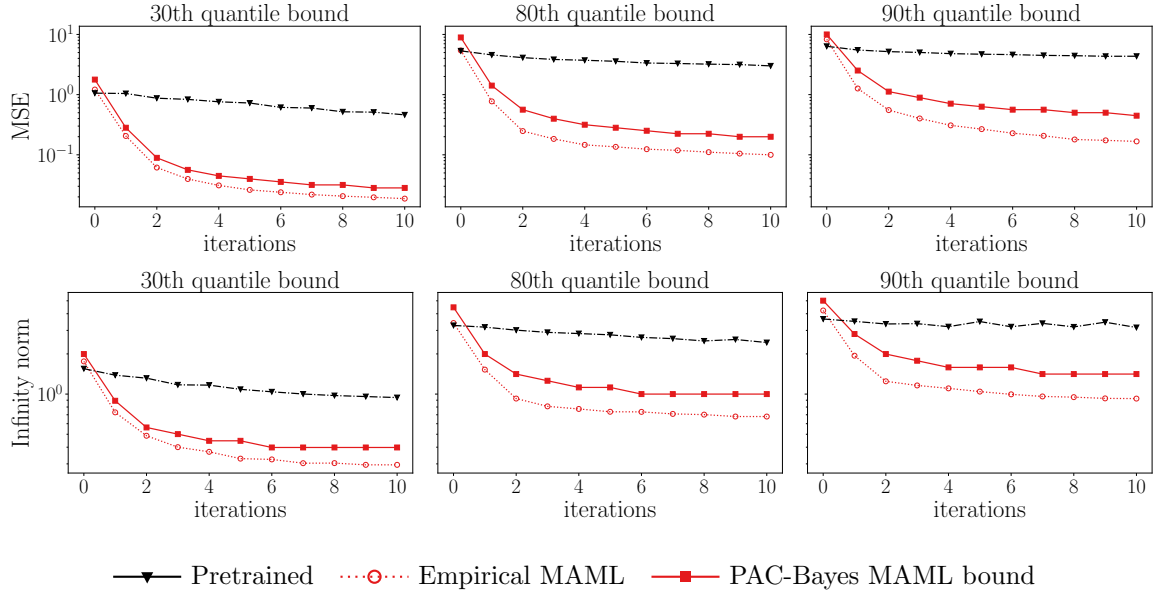


Figure 11: MAML quantile results for sinusoid curves. Top: MSE. Bottom: infinity norm from Equation (28). For the quantiles 30, 80, and 90, our MAML upper bounds are significantly lower than the pretrained empirical curve after a few iterations.

Table 5: The quantile results for MAML on sinusoidal regression tasks after 10 iterations for both the mean square error (MSE) and the infinity norm. Since the expected risk is never bounded to a value below 0.05, we cannot provide guarantees for the 95th quantile.

MSE				Infinity norm			
Quantile	Pretrained	MAML		Quantile	Pretrained	MAML	
		Emp.	Bnd.			Emp.	Bnd.
10	0.496	0.207	0.251	10	0.153	0.009	0.014
30	0.942	0.295	0.398	30	0.461	0.019	0.028
50	1.429	0.391	0.562	50	0.972	0.035	0.056
60	1.829	0.452	0.631	60	1.429	0.051	0.079
80	2.434	0.679	1.0	80	3.004	0.1	0.2
90	3.155	0.925	1.413	90	4.342	0.167	0.447
95	3.849	1.191	-	95	6.6	0.351	-

we provide strong guarantees using a sample convergence bound. For learned optimizers, we provide generalization guarantees using the PAC-Bayes framework and a learning algorithm designed to optimize these guarantees. We showcase the effectiveness of our approach for both classical and learned optimizers on many examples including ones from control, signal processing, and meta-learning.

We see a few directions for interesting future research. The bounds in this paper all required the problem parameters to be drawn of a distribution in an i.i.d. fashion. While in many applications this assumption is common, *e.g.*, in sparse coding, or machine learning problems, there are other applications where this assumption is not met. For example, in control problems, where the problems need to be solved sequentially. Extending our work

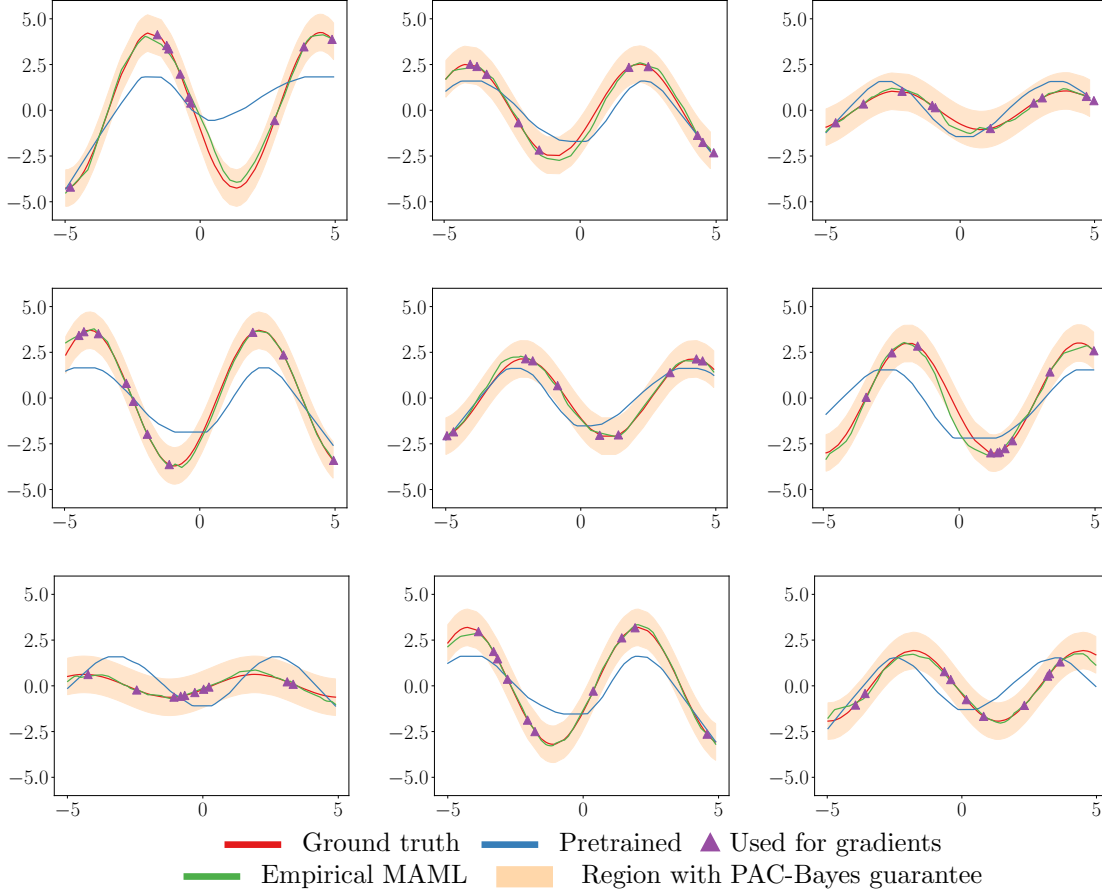


Figure 12: MAML visualizations for regressing on sine curves. The purple triangles are the K^{train} datapoints used for computing the gradients. With high probability, after 10 steps, MAML is guaranteed to produce a curve that remains entirely in the banded region 81% of the time, while the pretrained model produces a curve that only entirely lies in the banded region around 33% of the time.

beyond the i.i.d. assumption is an avenue for future work. Another avenue is to scale our approach to tackle larger-scale problems for learned optimizers.

Acknowledgments

We thank Ernest Ryu, Anirudha Majumdar, Jingyi Huang, and two anonymous reviewers for helpful and detailed comments that improved the quality of this work. Bartolomeo Stellato and Rajiv Sambharya are supported by the NSF CAREER Award ECCS-223977. Bartolomeo Stellato is also supported by the ONR YIP Award N000142512147. We are pleased to acknowledge that the work reported on in this paper was substantially performed using the Princeton Research Computing resources at Princeton University which is consor-

tium of groups led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology's Research Computing.

References

- Diogo Almeida, Clemens Winter, Jie Tang, and Wojciech Zaremba. A generalizable approach to learning optimizers. *arXiv preprint arXiv:2106.00958*, 2021.
- Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv e-prints*, 2023.
- Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *International Conference on Machine Learning*, pages 205–214, 2018.
- Brandon Amos. Tutorial on amortized optimization. *Foundations and Trends in Machine Learning*, 16(5):592–732, 2023.
- Marcin Andrychowicz, Misha Denil, Sergio Gómez Colmenarejo, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Neural Information Processing Systems*, 2016.
- Shaojie Bai, Vladlen Koltun, and J Zico Kolter. Neural deep equilibrium solvers. In *International Conference on Learning Representations*, 2022.
- Kyri Baker. Learning warm-start points for ac optimal power flow. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.
- Alexios Balatsoukas-Stimming and Christoph Studer. Deep unfolding for communications systems: A survey and some new directions. In *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, pages 266–271, 2019.
- Maria-Florina Balcan. Data-driven algorithm design. *arXiv preprint arXiv:2011.07177*, 2020.
- Maria-Florina Balcan, Vaishnavh Nagarajan, Ellen Vitercik, and Colin White. Learning-theoretic foundations of algorithm configuration for combinatorial partitioning problems. In *Conference on Learning Theory*, pages 213–274. PMLR, 2017.
- Maria-Florina Balcan, Travis Dick, and Colin White. Data-driven clustering via parameterized lloyd’s families. *Advances in Neural Information Processing Systems*, 31, 2018.
- Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, pages 424–433, 2019.
- Maria-Florina Balcan, Dan DeBlasio, Travis Dick, Carl Kingsford, Tuomas Sandholm, and Ellen Vitercik. How much data is sufficient to learn high-performing algorithms? generalization guarantees for data-driven algorithm design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 919–932, 2021.
- Sebastian Banert, Jevgenija Rudzusika, Ozan Öktem, and Jonas Adler. Accelerated forward-backward optimization using deep learning. *arXiv preprint arXiv:2105.05210*, 2021.

- Goran Banjac, Paul Goulart, Bartolomeo Stellato, and Stephen Boyd. Infeasibility detection in the alternating direction method of multipliers for convex optimization. *Journal of Optimization Theory and Applications*, 183, 2019.
- Peter L. Bartlett, Piotr Indyk, and Tal Wagner. Generalization bounds for data-driven numerical linear algebra. In *Conference on Learning Theory*, volume 178, pages 2013–2040, 2022.
- Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.
- Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 693–696, 2009.
- Dimitris Bertsimas and Bartolomeo Stellato. Online Mixed-Integer Optimization in Milliseconds. *INFORMS Journal on Computing*, 34(4):2229–2248, 2022.
- Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *arXiv preprint arXiv:2105.15183*, 2021.
- Daniel Boley. Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. *SIAM Journal on Optimization*, 23(4):2183–2207, 2013.
- Francesco Borrelli, Alberto Bemporad, and Manfred Morari. *Predictive Control for Linear and Hybrid Systems*. Cambridge University Press, 2017.
- Stephen Boyd, Seung-Jean Kim, Lieven Vandenbergh, and Arash Hassibi. A tutorial on geometric programming. *Optimization and engineering*, 8:67–127, 2007.
- Stephen P. Boyd, Neal Parikh, Eric King wah Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.
- James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.
- Julia Briden, Changrak Choi, Kyongsik Yun, Richard Linares, and Abhishek Cauligi. Constraint-informed learning for warm starting trajectory optimization. *arXiv preprint arXiv:2312.14336*, 2023.
- Steven Chen, Kelsey Saulnier, Nikolay Atanasov, Daniel D. Lee, Vijay Kumar, George J. Pappas, and Manfred Morari. Approximating explicit model predictive control using constrained neural networks. In *American Control Conference*, pages 1520–1527, 2018a.
- Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022.

- Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Theoretical linear convergence of unfolded ista and its practical weights and thresholds. *Advances in Neural Information Processing Systems*, 31, 2018b.
- Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Hyperparameter tuning is all you need for lista. In *Advances in Neural Information Processing Systems*, 2021.
- Xinshi Chen, Yufei Zhang, Christoph Reisinger, and Le Song. Understanding deep architectures with reasoning layer. In *Neural Information Processing Systems*, 2020.
- Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: an extension of mnist to handwritten letters, 2017.
- Steven Diamond, Vincent Sitzmann, Felix Heide, and Gordon Wetzstein. Unrolled optimization with deep priors. *arXiv preprint arXiv:1705.08041*, 2017.
- Moritz Diehl, Hans Joachim Ferreau, and Niels Haverbeke. *Efficient Numerical Methods for Nonlinear MPC and Moving Horizon Estimation*. 2009.
- Asen L Dontchev and Ralph Tyrrell Rockafellar. *Implicit functions and solution mappings: A view from variational analysis*, volume 616. Springer, 2009.
- Priya Donti, David Rolnick, and Zico Kolter. Dc3: A learning method for optimization with hard constraints. In *International Conference on Learning Representations*, 2021.
- Yoel Drori and Marc Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145(1):451–482, 2014.
- John C. Duchi. Derivations for linear algebra and optimization. 2016.
- Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.
- Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.
- Alec Farid and Anirudha Majumdar. Generalization bounds for meta-learning via pac-bayes and uniform stability. In *Neural Information Processing Systems*, 2021.
- Mahyar Fazlyab, Alejandro Ribeiro, Manfred Morari, and Victor M. Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal of Optimization*, 28:2654–2689, 2017.
- Mahyar Fazlyab, Manfred Morari, and George Pappas. Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. *IEEE Transactions on Automatic Control*, 67(1):1–15, 2022.

- Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017.
- Michael Garstka, Mark Cannon, and Paul Goulart. COSMO: A conic operator splitting method for large convex problems. In *European Control Conference*, 2019.
- Pontus Giselsson and Stephen P. Boyd. Linear convergence and metric selection for douglas-rachford splitting and admm. *IEEE Transactions on Automatic Control*, 62:532–544, 2014.
- Baptiste Goujaud, Céline Moucer, François Glineur, Julien Hendrickx, Adrien Taylor, and Aymeric Dieuleveut. Pepit: computer-assisted worst-case analyses of first-order optimization methods in python. *arXiv preprint arXiv:2201.04040*, 2022.
- Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *International Conference on Machine Learning*, Madison, WI, USA, 2010. Omnipress.
- Rishi Gupta and Tim Roughgarden. A PAC approach to application-specific algorithm selection. *SIAM Journal on Computing*, 46(3):992–1017, 2017.
- Howard Heaton, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Safeguarded learned convex optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- Mingyi Hong and Zhi-Quan Tom Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162:165 – 199, 2012.
- Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.
- Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.
- Jeffrey Ichnowski, Paras Jain, Bartolomeo Stellato, Goran Banjac, Michael Luo, Francesco Borrelli, Joseph. E Gonzales, Ian Stoica, and Ken Goldberg. Accelerating quadratic optimization with reinforcement learning. In *Advances in Neural Information Processing Systems 35*, 2021.
- Haewon Jung, Junyoung Park, and Jinkyoo Park. Learning context-aware adaptive solvers to accelerate quadratic programming. *arXiv preprint arXiv:2211.12443*, 2022.
- Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME—Journal of Basic Engineering*, 82(Series D):35–45, 1960.
- Benjamin Karg and Sergio Lucia. Efficient representation and approximation of model predictive control laws via deep learning. *IEEE Transactions on Cybernetics*, PP, 2020.

- Ethan King, James Kotary, Ferdinando Fioretto, and Jan Drgona. Metric learning to accelerate convergence of operator splitting methods for differentiable parametric programming. *CoRR*, abs/2404.00882, 2024.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, and Bryan Wilder. End-to-end constrained optimization learning: A survey. In *International Joint Conference on Artificial Intelligence, IJCAI-21*, 2021.
- S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.
- John Langford and Rich Caruana. (not) bounding the true error. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.
- John Langford and Matthias Seeger. *Bounds for averaging classifiers*. School of Computer Science, Carnegie Mellon University, 2001.
- Laurent Lessard, Ben Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1): 57–95, jan 2016.
- Ke Li and Jitendra Malik. Learning to optimize. *arXiv preprint arXiv:1606.01885*, 2016.
- Felix Lieder. *Projection Based Methods for Conic Linear Programming — Optimal First Order Complexities and Norm Constrained Quasi Newton Methods*. Phd thesis, HHU Düsseldorf, 2018.
- Jialin Liu, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. ALISTA: Analytic weights are as good as learned weights in LISTA. In *International Conference on Learning Representations*, 2019.
- Anirudha Majumdar, Alec Farid, and Anoopkumar Sonar. PAC-Bayes control: learning policies that provably generalize to novel environments. *The International Journal of Robotics Research*, 40(2-3):574–593, 2021.
- Terrence WK Mak, Minas Chatzos, Mathieu Tanneau, and Pascal Van Hentenryck. Learning regionally decentralized ac optimal power flows with ADMM. *IEEE Transactions on Smart Grid*, 2023.
- Andreas Maurer. A note on the PAC Bayesian theorem. *arXiv preprint cs/0411099*, 2004.
- David A. McAllester. Some PAC-Bayesian theorems. In *Conference on Computational Learning Theory*. Association for Computing Machinery, 1998.
- Luke Metz, James Harrison, C Daniel Freeman, Amil Merchant, Lucas Beyer, James Bradbury, Naman Agrawal, Ben Poole, Igor Mordatch, Adam Roberts, et al. Velo: Training versatile learned optimizers by scaling up. *arXiv preprint arXiv:2211.09760*, 2022.

- Sidhant Misra, Line Roald, and Yeesian Ng. Learning for constrained optimization: Identifying optimal active constraint sets. *INFORMS Journal on Computing*, 34(1):463–480, 2022.
- Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.
- Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. 1983.
- Brendan O’Donoghue. Operator splitting for a homogeneous embedding of the linear complementarity problem. *SIAM Journal on Optimization*, 31(3):1999–2023, 2021.
- Courtney Paquette, Bart van Merriënboer, Elliot Paquette, and Fabian Pedregosa. Halting time is predictable for large models: A universality property and average-case analysis. *Foundations of Computational Mathematics*, 23(2):597–673, feb 2022.
- Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.
- Fabian Pedregosa and Damien Scieur. Acceleration through spectral density estimation. In *International Conference on Machine Learning*, 2020.
- Isabeau Prémont-Schwarz, Jaroslav Vitku, and Jan Feyereisl. A simple guard for learned optimizers. In *International Conference on Machine Learning*, 2022.
- Vinit Ranjan and Bartolomeo Stellato. Verification of first-order methods for parametric quadratic optimization. *arXiv preprint arXiv:2403.03331*, 2024.
- David Reeb, Andreas Doerr, Sebastian Gerwinn, and Barbara Rakitsch. Learning gaussian processes by minimizing PAC-Bayesian generalization bounds. *Advances in Neural Information Processing Systems*, 31, 2018.
- Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, 2019.
- Ernest Ryu, Adrien B. Taylor, Carolina Bergeling, and Pontus Giselsson. Operator splitting performance estimation: Tight contraction factors and optimal parameter selection. *SIAM Journal on Optimization*, 30(3):2251–2271, 2020.
- Ernest K. Ryu and Wotao Yin. *Large-Scale Convex Optimization: Algorithms and Analyses via Monotone Operators*. Cambridge University Press, 2022.
- R. Sambharya, G. Hall, B. Amos, and B. Stellato. Learning to warm-start fixed-point optimization algorithms. *Journal of Machine Learning Research*, 25(166):1–46, 2024.
- Rajiv Sambharya, Georgina Hall, Brandon Amos, and Bartolomeo Stellato. End-to-End Learning to Warm-Start for Real-Time Quadratic Optimization. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference*, 2023.

- Jens Sjölund and Maria Bänkestad. Graph-based neural acceleration for nonnegative matrix factorization. *arXiv preprint arXiv:2202.00264*, 2022.
- Bartolomeo Stellato, Goran Banjac, Paul Goulart, Alberto Bemporad, and Boyd Stephen. OSQP: An Operator Splitting Solver for Quadratic Programs. *Mathematical Programming Computation*, 12(4):637–672, 2020.
- Michael Sucker and Peter Ochs. PAC-Bayesian learning of optimization algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2023.
- Michael Sucker, Jalal Fadili, and Peter Ochs. Learning-to-optimize with PAC-Bayesian guarantees: Theoretical considerations and practical implementation. *arXiv preprint arXiv:2404.03290*, 2024.
- Hong Ye Tan, Subhadip Mukherjee, Junqi Tang, and Carola-Bibiane Schönlieb. Data-driven mirror descent with input-convex neural networks. *SIAM Journal on Mathematics of Data Science*, 5(2):558–587, 2023.
- Adrien Taylor, Julien Hendrickx, and François Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161, 02 2015.
- Adrien B. Taylor, Bryan Van Scoy, and Laurent Lessard. Lyapunov functions for first-order methods: Tight automated convergence guarantees. In *International Conference on Machine Learning*, 2018.
- Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2011.
- Shobha Venkataraman and Brandon Amos. Neural fixed-point acceleration for convex optimization. *arXiv preprint arXiv:2107.10254*, 2021.
- Ricardo Vilalta and Youssef Drissi. A perspective view and survey of meta-learning. *Artificial Intelligence Review*, 18, 2001.
- Stephen J. Wright. *Primal-dual interior-point methods*. SIAM, 1997.
- Kailun Wu, Yiwen Guo, Ziang Li, and Changshui Zhang. Sparse coding with gated learned ista. In *International Conference on Learning Representations*, 2020.
- Lihua Xie and Yeng Chai Soh. Robust kalman filtering for uncertain systems. *Systems & Control Letters*, 22(2):123–129, 1994.
- Junjie Yang, Xuxi Chen, Tianlong Chen, Zhangyang Wang, and Yingbin Liang. M-l2o: Towards generalizable learning-to-optimize by test-time fast self-adaptation. In *International Conference on Learning Representations*, 2022.
- Junjie Yang, Tianlong Chen, Mingkan Zhu, Fengxiang He, Dacheng Tao, Yingbin Liang, and Zhangyang Wang. Learning to generalize provably in learning to optimize. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. Discerning the linear convergence of admm for structured convex optimization through the lens of variational analysis. *Journal of Machine Learning Research*, 21(83):1–75, 2020.

Junzi Zhang, Brendan O’Donoghue, and Stephen Boyd. Globally convergent type-I anderson acceleration for nonsmooth fixed-point iterations. *SIAM Journal on Optimization*, 30(4): 3170–3197, 2020.

Appendix A. PAC-Bayes background

In this section, we introduce the PAC-Bayes background needed to construct generalization guarantees given a set of N i.i.d. samples S . We first introduce the Kullback-Leibler (KL) divergence, an important component in our bounds, and show how to compute its inverse in Section A.1. In Section A.2, we present two PAC-Bayes bounds: a sample convergence bound and Maurer’s bound. Specifically, for classical optimizers, we will use the sample convergence bound to bound the *risk*

$$r_{\mathcal{X}} = \mathbf{E}_{x \sim \mathcal{X}} e(x),$$

in terms of the *empirical risk*

$$\hat{r}_S = \frac{1}{N} \sum_{i=1}^N e(x_i).$$

Recall from Equation (4) that the error term $e(x)$ for a given parameter x is always equal to 0 or 1. For learned optimizers, we consider weights θ drawn from a distribution P and use Maurer’s bound to bound the *expected risk*

$$R_{\mathcal{X}}(P) = \mathbf{E}_{\theta \sim P} \mathbf{E}_{x \sim \mathcal{X}} e_{\theta}(x), \quad (29)$$

in terms of its *expected empirical risk*

$$\hat{R}_S(P) = \mathbf{E}_{\theta \sim P} \frac{1}{N} \sum_{i=1}^N e_{\theta}(x_i).$$

We use randomized weights to represent a distribution of learned optimizers, which is a key component of the PAC-Bayes methods. Using randomized weights does not limit which types of learned optimizers we can apply our method to.

A.1 KL divergence

The KL divergence, a measure of distance between two probability distributions, features prominently in the PAC-Bayes guarantees that we use. To derive our generalization bounds, it is sufficient to examine the KL divergence in two scenarios: between Normal distributions and between Bernoulli distributions.

Normal distributions. The KL-divergence between continuous distributions with density functions q and p over the Euclidean space \mathbf{R}^m is defined as

$$\text{KL}(q \parallel p) = \int_{-\infty}^{\infty} q(y) \log \left(\frac{q(y)}{p(y)} \right) dy.$$

We are particularly interested in the case where both p and q are densities of multivariate normal distributions: $\mathcal{N}_p = \mathcal{N}(\mu_p, \Sigma_p)$ and $\mathcal{N}_q = \mathcal{N}(\mu_q, \Sigma_q)$ over \mathbf{R}^m . In this case, the KL divergence can be obtained in closed-form (Duchi, 2016):

$$\text{KL}(\mathcal{N}_q \parallel \mathcal{N}_p) = \frac{1}{2} \left(\text{tr}(\Sigma_p^{-1} \Sigma_q) + (\mu_q - \mu_p)^T \Sigma_p^{-1} (\mu_q - \mu_p) + \log \frac{\det \Sigma_p}{\det \Sigma_q} - m \right). \quad (30)$$

Bernoulli distributions. Our goal is to bound the risk $r_{\mathcal{X}}$ for classical optimizers and the expected risk $R_{\mathcal{X}}(P)$ for learned optimizers with posterior distribution P in terms of their empirical counterparts. Importantly, we remark that these quantities are the expected values of 0–1 error functions in equations (4) and (5), which correspond to the key parameters of Bernoulli distributions. We denote the KL divergence between two Bernoulli distributions, $\mathcal{B}(q)$ with mean q and $\mathcal{B}(p)$ with mean p , as (Kullback and Leibler, 1951)

$$\text{kl}(q \parallel p) = \text{KL}(\mathcal{B}(q) \parallel \mathcal{B}(p)) = q \log \frac{q}{p} + (1 - q) \log \frac{1 - q}{1 - p}.$$

In the next subsection, we will bound the gap between the key parameter of a Bernoulli distribution denoted as p and its estimated value $q \in [0, 1]$ as

$$\text{kl}(q \parallel p) \leq c,$$

where $c > 0$. This implies the inequality

$$p \leq \text{kl}^{-1}(q \mid c) = \sup\{p \in [0, 1] \mid \text{kl}(q \parallel p) \leq c\},$$

where $\text{kl}^{-1}(q \mid c)$ can be computed by solving the following one-dimensional convex geometric program (Boyd et al., 2007),

$$\begin{aligned} & \text{maximize} && p \\ & \text{subject to} && q \log \left(\frac{q}{p} \right) + (1 - q) \log \left(\frac{1 - q}{1 - p} \right) \leq c \\ & && 0 \leq p \leq 1. \end{aligned} \tag{31}$$

Precise solutions to this problem can be obtained through convex optimization algorithms (*e.g.*, through interior point methods (Wright, 1997)). Problem (31) is used to compute our performance guarantees for both classical and learned optimizers. We note that an upper bound to the KL inverse can be explicitly computed using Pinsker’s inequality

$$\text{kl}^{-1}(q \mid c) \leq q + \sqrt{c/2}. \tag{32}$$

However, the gap between Pinsker’s bound and the KL inverse can be large; the KL inverse is always upper bounded by one, but Pinsker’s bound can be infinitely large.

A.2 Probabilistic Bounds

In this subsection, we present the probabilistic bounds that we use to obtain our generalization guarantees: a sample convergence bound and Maurer’s bound.

Sample convergence bound. The sample convergence bound below will be used to bound the risk $r_{\mathcal{X}}$ in terms of the empirical risk \hat{r}_S for classical optimizers.

Theorem 3 (*Langford and Caruana, 2001*). *Given $\delta \in (0, 1)$ and N samples S , with probability at least $1 - \delta$ the following bound holds:*

$$\text{kl}(\hat{r}_S \parallel r_{\mathcal{X}}) \leq \frac{\log(2/\delta)}{N}. \tag{33}$$

Proof We seek to prove the following inequality for $\epsilon > 0$

$$\mathbf{P}(\text{kl}(\hat{r}_S \parallel r_{\mathcal{X}}) \geq \epsilon) \leq 2e^{-N\epsilon}.$$

To get the final bound, we set $\delta = 2\exp(-N\epsilon)$. The proof proceeds by breaking the case that the KL divergence exceeds ϵ into the case where $\hat{r}_S > r_{\mathcal{X}}$ and the other case where $\hat{r}_S < r_{\mathcal{X}}$. For a given value of the risk $r_{\mathcal{X}}$, we define the quantity $r_{\epsilon}^1 > r_{\mathcal{X}}$ implicitly so that $\text{kl}(r_{\epsilon}^1 \parallel r_{\mathcal{X}}) = \epsilon$. Similarly, we define $r_{\epsilon}^2 < r_{\mathcal{X}}$ implicitly so that $\text{kl}(r_{\epsilon}^2 \parallel r_{\mathcal{X}}) = \epsilon$. We continue as follows for the case where $\hat{r}_S > r_{\mathcal{X}}$:

$$\mathbf{P}(\text{kl}(\hat{r}_S \parallel r_{\mathcal{X}}) \geq \epsilon, \hat{r}_S > r_{\mathcal{X}}) = \mathbf{P}(\hat{r}_S \geq r_{\epsilon}^1) \leq e^{-N\epsilon}.$$

The equality comes from the definition of r_{ϵ}^1 . The inequality follows from the upper tail Chernoff bound for $\Delta > 0$:

$$\mathbf{P}(\hat{r}_S \geq r_{\mathcal{X}} + \Delta) \leq \exp(-N \text{kl}(r_{\mathcal{X}} + \Delta \parallel r_{\mathcal{X}})).$$

For the case where $\hat{r}_S < r_{\mathcal{X}}$ we have

$$\mathbf{P}(\text{kl}(\hat{r}_S \parallel r_{\mathcal{X}}) \geq \epsilon, \hat{r}_S < r_{\mathcal{X}}) = \mathbf{P}(\hat{r}_S \leq r_{\epsilon}^2) \leq e^{-N\epsilon}.$$

The equality comes from the definition of r_{ϵ}^2 . The inequality follows from the lower tail Chernoff bound for $\Delta > 0$:

$$\mathbf{P}(\hat{r}_S \leq r_{\mathcal{X}} - \Delta) \leq \exp(-N \text{kl}(r_{\mathcal{X}} - \Delta \parallel r_{\mathcal{X}})).$$

The proof concludes by summing the probabilities over both cases

$$\mathbf{P}(\text{kl}(\hat{r}_S \parallel r_{\mathcal{X}}) \geq \epsilon) = \mathbf{P}(\text{kl}(\hat{r}_S \parallel r_{\mathcal{X}}) \geq \epsilon, \hat{r}_S > r_{\mathcal{X}}) + \mathbf{P}(\text{kl}(\hat{r}_S \parallel r_{\mathcal{X}}) \geq \epsilon, \hat{r}_S < r_{\mathcal{X}}) \leq 2e^{-N\epsilon}.$$

■

Maurer's bound. The derivation of our generalization bounds for learned optimizers is based on Maurer's bound (itself an adaptation of Seeger's bound (Langford and Seeger, 2001)), which allows us to provide bounds when the weights θ are drawn from a distribution $P \in \mathcal{P}$. Here, \mathcal{P} is the space of all probability distributions in \mathbf{R}^p . Specifically, Maurer's bound provides a bound on the expected risk $R_{\mathcal{X}}(P)$ in terms of its expected empirical risk $\hat{R}_S(P)$.

Theorem 4 (Maurer, 2004). *Given a set of N samples S where $N \geq 8$, a prior distribution independent of the training data $P_0 \in \mathcal{P}$, and $\delta \in (0, 1)$, with probability at least $1 - \delta$ the following bound holds for all distributions $P \in \mathcal{P}$:*

$$R_{\mathcal{X}}(P) \leq \text{kl}^{-1} \left(\hat{R}_S(P) \mid \frac{1}{N} \left(\text{KL}(P \parallel P_0) + \log \frac{2\sqrt{N}}{\delta} \right) \right). \quad (34)$$

The PAC-Bayes framework typically adopts the following steps. First, we select the prior $P_0 \in \mathcal{P}$ before observing any training data. Then, we observe the training data S and we choose the posterior distribution P (e.g., through a learning algorithm (Dziugaite and Roy, 2017)). Lastly, we use the inequality (34) to bound the expected risk of the posterior distribution $R_{\mathcal{X}}(P)$. This posterior is allowed to depend on the prior and the samples.

Proof The relative entropy $\text{KL}(P \parallel P_0)$ of two probability measures P and P_0 on a set \mathcal{H} is defined to be infinite if P is not absolutely continuous with respect to P_0 . Otherwise, $\text{KL}(P \parallel P_0) = \mathbf{E}_P[\log \frac{dP}{dP_0}]$ where dP/dP_0 is the density of P with respect to P_0 . In the proof, we let S be a set of N training samples of the parameters drawn i.i.d. from the distribution \mathcal{X} . The proof continues as

$$\begin{aligned}
& \mathbf{E}_S[\exp(N \text{kl}(R_{\mathcal{X}}(P) \parallel \hat{R}_S(P)) - \text{KL}(P \parallel P_0))] \\
& \leq \mathbf{E}_S \left[\exp \left(\mathbf{E}_{\theta \sim P} \left[N \text{kl} \left(\frac{1}{N} \sum_{i=1}^N e_{\theta}(x_i) \parallel \mathbf{E}_{x \sim \mathcal{X}} e_{\theta}(x) \right) - \log \frac{dP}{dP_0}(\theta) \right] \right) \right] \\
& \leq \mathbf{E}_S \left[\mathbf{E}_{\theta \sim P} \left[\exp \left(N \text{kl} \left(\frac{1}{N} \sum_{i=1}^N e_{\theta}(x_i) \parallel \mathbf{E}_{x \sim \mathcal{X}} e_{\theta}(x) \right) - \log \frac{dP}{dP_0}(\theta) \right) \right] \right] \\
& = \mathbf{E}_S \mathbf{E}_{\theta \sim P_0} \exp \left(N \text{kl} \left(\frac{1}{N} \sum_{i=1}^N e_{\theta}(x_i) \parallel \mathbf{E}_{x \sim \mathcal{X}} e_{\theta}(x) \right) \right) \left(\frac{dP}{dP_0} \right)^{-1} \left(\frac{dP}{dP_0} \right) \\
& = \mathbf{E}_{\theta \sim P_0} \mathbf{E}_S \exp \left(N \text{kl} \left(\frac{1}{N} \sum_{i=1}^N e_{\theta}(x_i) \parallel \mathbf{E}_{x \sim \mathcal{X}} e_{\theta}(x) \right) \right) \\
& \leq 2\sqrt{N}.
\end{aligned}$$

The first inequality follow's from Jensen's inequality and the convexity of the KL divergence. The second inequality follow's from Jensen's inequality and the convexity of the exponential function. The third line applies the Radon-Nikodyn derivative to change the expectation of θ over the posterior P to be the expectation of θ over the prior P_0 . The second to last line uses Tonelli's theorem to switch the order of the expectations of a non-negative random variable. The last inequality uses inequality 1 in Maurer (2004). Then, by Markov's inequality the proof finishes with

$$\begin{aligned}
\delta & \geq \mathbf{P}_S \left(\exp(N \text{kl}(R_{\mathcal{X}}(P) \parallel \hat{R}_S(P)) - \text{KL}(P \parallel P_0)) > \frac{2\sqrt{N}}{\delta} \right) \\
& = \mathbf{P}_S \left(\text{kl}(R_{\mathcal{X}}(P) \parallel \hat{R}_S(P)) > \frac{\text{KL}(P \parallel P_0) + \log \left(\frac{2\sqrt{N}}{\delta} \right)}{N} \right).
\end{aligned}$$

■

Appendix B. Experimental details

B.1 Cross-validating B^{target}

In our experiments, we cross-validate over six B^{target} hyperparameter values. If a particular bound on the expected risk holds with probability $1 - \delta$ for a given B^{target} , then all six bounds hold with probability $1 - 6\delta$ by a union bound. Table 6 enumerates the B^{target} values chosen for cross-validation, alongside the corresponding upper bound on the generalization gap as determined by Pinsker’s inequality from Equation (32). After training, if $B(w^*, s^*, \lambda^*) = B^{\text{target}}$ (which we observe, approximately holds true due to the penalty form from problem (15)), then we can bound the generalization gap: $R_{\mathcal{X}}(P) - \hat{R}_S(P) \leq \sqrt{B^{\text{target}}/2}$ where the posterior is $P = \mathcal{N}_{w^*, s^*}$.

Table 6: The different B^{target} values used during cross-validation and their associated upper bounds on the generalization gap.

B^{target}	$\sqrt{B^{\text{target}}/2}$
0.01	0.071
0.03	0.122
0.05	0.158
0.1	0.223
0.2	0.316
0.3	0.387

B.2 Quantile bounds

The results from Sections 4 and 5 provide probabilistic bounds on the risk and the expected risk respectively for a number of algorithm steps k and tolerance ϵ . Recall that the (expected) risk is equivalent to the probability of failing to reach a given tolerance (due to the use of the error function). Therefore an upper bound on the (expected) risk corresponds to an upper bound on the quantile. For instance, if after k steps, the risk is bounded with probability (w.p.) $1 - \delta$ by 0.1 with some underlying metric ϕ and tolerance ϵ , then, w.p. $1 - \delta$, the tolerance ϵ upper bounds the metric ϕ after k steps at least 90% of the time. Using our notation, this is equivalent to the following statement; if $r_{\mathcal{X}} = \mathbf{E}_{x \sim \mathcal{X}}[\mathbf{1}(\phi(z^k(x), x) \geq \epsilon)] \leq 0.1$ w.p. $1 - \delta$, then $\mathbf{P}_{x \sim \mathcal{X}}(\phi(z^k(x), x) \geq \epsilon) \leq 0.1$ w.p. $1 - \delta$. Therefore the tolerance ϵ is a valid, probabilistic 90-th quantile bound.

To obtain the *tightest* quantile bounds in Section 7 for a given number of steps k , we proceed as follows. We first obtain bounds on the risk for N^{tol} pre-determined tolerances. If each bound on the risk with a specific tolerance holds with probability $1 - \delta$, then all of the bounds across all of the tolerances hold simultaneously with probability $1 - \delta N^{\text{tol}}$ by virtue of a union bound. Then for a given k and quantile Q , we find the lowest tolerance such that the bound on the (expected) risk is at most $1 - Q$. For example, say we want to bound the 90th quantile bound of the fixed-point residual at k steps. We first take all of the bounds on the risk for $\epsilon_1, \dots, \epsilon_{N^{\text{tol}}}$. Then we find the lowest value ϵ_i such that the (expected) risk with tolerance ϵ_i is at most 0.1; this value of ϵ_i bounds the 90th quantile with probability at least $1 - \delta N^{\text{tol}}$. Note that the bounds do not hold simultaneously across different values of k . However, if desired, they can be obtained by applying another union bound over the algorithm steps.

B.3 Other numerical details

To obtain the quantile bounds, we discretize the metric into 81 pre-determined tolerances. For the metrics in the MAML problem, the discretization is 81 points evenly spaced out on a log scale between 10^{-3} and 10^1 . For the NMSE metric, we discretize between -80 and 0 evenly on a linear scale. For all other metrics, the discretization is 81 points evenly spaced out on a log scale between 10^{-6} and 10^2 . For classical optimizers, we set the desired probability value to be $\delta = 10^{-4}$. The bounds on the risk for the classical optimizers holds with probability $1 - \delta = 0.9999$ and each of the quantile bounds holds with probability 0.9919 due to the union bound over the 81 tolerances. For learned optimizers, the desired probability values are $\delta = 10^{-5}$ and $\omega = 10^{-5}$. For the learned optimizers, there are two additional considerations: the additional sample convergence bounds which holds with probability $1 - \omega$ and the cross-validation over the set of B^{target} values which requires a union bound. After taking a union bound over the cross-validated B^{target} values, the bound on the expected risk holds with probability at least $1 - 6(\delta + \omega) = 0.99988$. The bounds on the quantiles each hold with probability at least 0.99028 . For all learned optimizers, we set the prior hyperparameters to be $\lambda^{\max} = 100$ and $b = 100$.

Appendix C. Proofs

C.1 Proof of Theorem 1

The vector $a \in \mathbf{N}_+^J$ and constant $\delta \in (0, 1)$ defines the quantity δ_a from Equation (10), $\delta_a = \delta (6/(\pi^2))^J (\prod_{j=1}^J a_j^2)^{-1}$. Note that δ_a is in the range $(0, 1)$ since all of a_j terms and J are at least one and $\delta \in (0, 1)$. Next, we apply Maurer's bound from Theorem 4 which states that with probability at least $1 - \delta_a$, the following inequalities hold:

$$\begin{aligned} \text{kl}(\hat{R}_S(\mathcal{N}_{w,s}) \parallel R_{\mathcal{X}}(\mathcal{N}_{w,s})) &\leq \frac{1}{N} \left(\text{KL}(\mathcal{N}_{w,s} \parallel \mathcal{N}(w_0, \Lambda)) + \log \frac{2\sqrt{N}}{\delta_a} \right) \\ &= \frac{1}{N} \left(\text{KL}(\mathcal{N}_{w,s} \parallel \mathcal{N}(w_0, \Lambda)) + \log \left(\prod_{j=1}^J a_j^2 \right) + J \log \frac{\pi^2}{6} + \log \frac{2\sqrt{N}}{\delta} \right) \\ &= \frac{1}{N} \left(\text{KL}(\mathcal{N}_{w,s} \parallel \mathcal{N}(w_0, \Lambda)) + 2 \sum_{j=1}^J \log \left(b \log \frac{\lambda^{\max}}{\lambda_j} \right) + J \log \frac{\pi^2}{6} + \log \frac{2\sqrt{N}}{\delta} \right). \end{aligned} \tag{35}$$

In the final line, we use the equality from the theorem $\lambda_j = \lambda^{\max} \exp(-a_j/b)$ where b and λ^{\max} are pre-defined. Now, to get the main result, we take a union bound over all possible vectors $a \in \mathbf{N}_+^J$. By taking a union bound with probability at least

$$1 - \sum_{a_1=1}^{\infty} \cdots \sum_{a_J=1}^{\infty} \left(\frac{6}{\pi^2} \right)^J \frac{\delta}{a_1^2 a_2^2 \cdots a_J^2}, \tag{36}$$

inequality (35) holds uniformly for all $a \in \mathbf{N}_+^J$. Since

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6},$$

this probability given by line (36) simplifies to $1 - \delta$.

C.2 Proof of Theorem 2

Lemma 5 *If $0 \leq A \leq C$ element-wise for $A \in \mathbf{R}^{m \times n}$ and $C \in \mathbf{R}^{m \times n}$, then $\|A\|_2 \leq \|C\|_2$.*

Proof The proof of the lemma proceeds as follows:

$$\begin{aligned} \|A\|_2 &= \max_{\|v\|_2=1, v \geq 0} \|Av\|_2 \\ &= \|Av^*\|_2 \\ &\leq \|Cv^*\|_2 \\ &\leq \|C\|_2. \end{aligned}$$

The first line follows from the definition of the spectral norm, and noting that since $A \geq 0$, a maximizer occurs where $v \geq 0$. To see this, observe that if a vector \bar{v} is a maximizer, then so is $|\bar{v}|$. In the second line, we let v^* be the maximizer. The third line comes from $A \leq C$. The last line follows from the definition of the spectral norm. \blacksquare

Bounding the spectral norm of the weight matrix. We first let $U_i \sim \mathcal{N}(0, \Sigma_i)$ and $u_i \sim \mathcal{N}(0, \sigma_i)$ and define the following matrices:

$$\tilde{U}_i = \begin{bmatrix} U_i \\ u_i^T \end{bmatrix}, \quad \tilde{\Sigma}_i = \begin{bmatrix} \Sigma_i \\ \sigma_i^T \end{bmatrix}.$$

We now state a result from Tropp (2011, Section 4.3) that will allow us to bound $\|\tilde{U}_i\|_2$ with high probability.

Theorem 6 *Consider a fixed matrix $B \in \mathbf{R}^{d_1 \times d_2}$ and a random matrix $\Gamma \in \mathbf{R}^{d_1 \times d_2}$ whose entries are independent standard normal variables. Define the variance parameter*

$$v^2 = \max\{\max_j \|B_{j:}\|_2^2, \max_k \|B_{:k}\|_2^2\}, \quad (37)$$

where $B_{j:}$ and $B_{:k}$ are the j -th row and k -th column of the matrix B . Then for all $t \geq 0$,

$$\mathbf{P}(\|\Gamma \odot B\|_2 \geq t) \leq (d_1 + d_2)e^{-t^2/2v^2}.$$

We let v_i^2 be the variance parameter of the i -th layer from Equation (37):

$$v_i^2 = \max\{\max_j \|(\tilde{\Sigma}_i)_{j:}^{1/2}\|_2^2, \max_k \|(\tilde{\Sigma}_i)_{:k}^{1/2}\|_2^2\}.$$

Using Theorem 6, we bound the spectral norm of \tilde{U}_i as

$$\mathbf{P}(\|\tilde{U}_i\|_2 \geq \tau_i) \leq (m_i + n_i + 1)e^{-\tau_i^2/2v_i^2}. \quad (38)$$

We set the right hand side to be δ/L to get

$$\tau_i = v_i \sqrt{2 \log(L(m_i + n_i + 1)/\delta)},$$

thereby bounding the spectral norm of \tilde{U}_i by τ_i with probability at least $1 - \delta/L$. We take a union bound across all layers so that with probability $1 - \delta$, the inequalities $\|\tilde{U}_i\|_2 \leq \tau_i$ for $i = 0, \dots, L - 1$ hold simultaneously.

Bounding the output of each layer. We turn our attention to bounding the output of the i -th layer, which we denote as $y_i(x)$ (where $y_0 = x$). Due to the bias terms, it is helpful to include the notation $\bar{y}_i(x) = (y_i(x), 1)$. We then have the following bound for $i = 0, \dots, L - 2$:

$$\begin{aligned} \|y_{i+1}(x)\|_2 &= \|\psi((\tilde{W}_i + \tilde{U}_i)\bar{y}_i(x))\|_2 \\ &\leq \|\tilde{W}_i + \tilde{U}_i\|_2 \|\bar{y}_i(x)\|_2 \\ &\leq (\|W_i\|_2 + \|b_i\|_2 + \|\tilde{U}_i\|_2)(\|y_i(x)\|_2 + 1). \end{aligned} \quad (39)$$

The first inequality follows from the ReLU activation function and Cauchy-Schwarz inequality. The second inequality follows from the triangle inequality. Note that the final layer does not have a ReLU activation, but the inequality holds nonetheless to bound $\|y_L(x)\|_2$. Now, we let $a_0^* = \bar{x}$ and

$$a_{i+1}^* = (\|W_i\|_2 + \|b_i\|_2 + \tau_i)(a_i^* + 1).$$

It follows from inequalities (38) and (39) that with probability at least $1 - \delta$, the quantity a_i^* upper bounds the output of the i -th layer. Since $h_\theta(x)$ is the output of the L -th layer, the bound $\|h_\theta(x)\|_2 \leq a_L^*$ holds with probability $1 - \delta$.

Appendix D. Operator theory definitions

First, recall that the set of fixed-points of operator T is denoted as $\mathbf{fix} T$.

Definition 7 (β -contractive operator) *An operator T is β -contractive for $\beta \in (0, 1)$ if*

$$\|Tx - Ty\|_2 \leq \beta \|x - y\|_2 \quad \forall x, y \in \mathbf{dom} T.$$

Definition 8 (β -linearly convergent operator) *An operator T is β -linearly convergent for $\beta \in [0, 1)$ if*

$$\mathbf{dist}_{\mathbf{fix} T}(Tx) \leq \beta \mathbf{dist}_{\mathbf{fix} T}(x) \quad \forall x \in \mathbf{dom} T.$$

Definition 9 (Non-expansive operator) *An operator T is non-expansive if*

$$\|Tx - Ty\|_2 \leq \|x - y\|_2, \quad \forall x, y \in \mathbf{dom} T.$$

Definition 10 (α -averaged operator) *An operator T is α -averaged for $\alpha \in (0, 1)$ if there exists a non-expansive operator R such that $T = (1 - \alpha)I + \alpha R$.*

Rates of convergence. The convergence rate of the fixed-point iterations can be summarized as follows for any $z^*(x) \in \mathbf{fix} T_x$. If operator T , with parameter x , is β -linearly convergent, with $\beta \in (0, 1)$, then (Sambharya et al., 2024)

$$\|z^{k+1}(x) - z^k(x)\|_2 \leq 2\beta^k \|z^*(x) - z^0(x)\|_2. \quad (40)$$

This rate also applies to β -contractive operators as they are a subset of β -linearly-convergent operators. If operator T with parameter x is α -averaged then the averaged iteration, also called the Krasnosel'skiĭ-Mann iteration, satisfies the following bound (Lieder, 2018):

$$\frac{\|z^{k+1}(x) - z^k(x)\|_2}{\|z^\star(x) - z^0(x)\|_2} \leq \begin{cases} \sqrt{\frac{1}{k+1} \left(\frac{k}{k+1}\right)^k \frac{1}{\alpha(1-\alpha)}} & \text{if } \frac{1}{2} \leq \alpha \leq \frac{1}{2} \left(1 + \sqrt{\frac{k}{k+1}}\right) \\ \frac{1}{2}(2\alpha - 1)^k & \text{if } \frac{1}{2} \left(1 + \sqrt{\frac{k}{k+1}}\right) \leq \alpha \leq 1. \end{cases} \quad (41)$$