Beyond Discretization: Learning the Optimal Solution Path

Qiran Dong University of California, Berkeley
Paul Grigas University of California, Berkeley
Vishal Gupta University of Southern California

Abstract

Many applications require minimizing a family of optimization problems indexed by some hyperparameter $\lambda \in \Lambda$ to obtain an entire solution path. Traditional approaches proceed by discretizing Λ and solving a series of optimization problems. We propose an alternative approach that parameterizes the solution path with a set of basis functions and solves a *single* stochastic optimization problem to learn the entire solution path. Our method offers substantial complexity improvements over discretization. When using constant-step size SGD, the uniform error of our learned solution path relative to the true path exhibits linear convergence to a constant related to the expressiveness of the basis. When the true solution path lies in the span of the basis, this constant is zero. We also prove stronger results for special cases common in machine learning: When $\lambda \in [-1, 1]$ and the solution path is ν-times differentiable, constant step-size SGD learns a path with ϵ uniform error after at most $O(\epsilon^{\frac{1}{1-\nu}} \log(1/\epsilon))$ iterations, and when the solution path is analytic, it only requires $O\left(\log^2(1/\epsilon) \log\log(1/\epsilon)\right)$. By comparison, the best-known discretization schemes in these settings require at least $O(\epsilon^{-1/2})$ discretization points (and even more gradient calls). Finally, we propose an adaptive variant of our method that sequentially adds basis functions and demonstrate strong numerical performance through experiments.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

Many decision-making applications entail solving a family of *parametrized* optimization problems:

$$\boldsymbol{\theta}^*(\lambda) \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} h(\boldsymbol{\theta}, \lambda), \quad \lambda \in \Lambda, \tag{1}$$

where Λ is an arbitrary set of parameters indexing the problems. (We assume Problem (1) admits an optimal solution for each $\lambda \in \Lambda$.) In such applications, we often seek to compute the entire solution path $\{\boldsymbol{\theta}^*(\lambda) : \lambda \in \Lambda\}$ in order to present experts with a portfolio of possible solutions to compare and assess tradeoffs.

As an example, consider the p-norm fair facility location problem (Gupta et al., 2023), which minimizes facility opening costs while incorporating l_p-regularization to promote fairness across socioeconomic groups. Since there is no obvious choice for p, we might prefer computing solutions for all p and allowing experts to (qualitatively) assess the resulting solutions. Many other applications entail navigating similar tradeoffs, including upweighting the minority class in binary classification to tradeoff Type I and II errors or aggregating features to increase interpretability at the expense of accuracy. In each case, we seek the entire solution path because selecting the "best" solution requires weighing a variety of criteria, some of which may be qualitative. These settings differ from classical hyperparameter tuning where there is a clear auxiliary criterion (like out-of-sample performance), and it would be enough to identify the single $\lambda^* \in \Lambda$ such that $\boldsymbol{\theta}^*(\lambda^*)$ optimizes this criteria.

The most common approach to learning the entire solution path is discretization: discretize Λ, solve Problem (1) at each grid point, and interpolate the resulting solutions. With enough grid points, interpolated solutions are approximately optimal along the entire path. Several authors have sought to determine the minimal the number of discretization points needed to achieve a target level of accuracy, usually under minimal assumptions on $h(\boldsymbol{\theta}, \lambda)$. Giesen et al. (2012a) considers convex optimization problems over the unit simplex when $\Lambda \subseteq \mathbb{R}$ and show that learning the solution path to accuracy ϵ requires at least $O(1/\epsilon)$ grid points.

Giesen et al. (2012b) consider the case where $h(\boldsymbol{\theta}, \lambda)$ is concave in λ and $\Lambda \subseteq \mathbb{R}$ and show only $O(1/\sqrt{\epsilon})$ points are needed. More recently, Ndiaye et al. (2019) relate the required number of points to the regularity of $h(\boldsymbol{\theta}, \lambda)$, arguing that if h is uniformly convex of order d, one requires $O(1/\sqrt[d]{\epsilon})$ points.

A potential criticism of this line of research is that the computational complexity of these methods depends not only on the number of grid points but also depends on the amount of work per grid point, e.g., as measured by the number of gradient calls used by a first-order method. Generally speaking, these methods do not share much information across grid points, at most warm-starting subsequent optimization problems. However, gradient evaluations at nearby grid points contain useful information for optimizing the current grid point, and leveraging this information presents an opportunity to reduce the total work.

1.1 Our Contributions

We propose a novel, simple algorithmic procedure to learn the solution path that applies to an arbitrary set Λ. The key idea is to replace the family of problems in (1) with a *single* stochastic optimization problem. This stochastic optimization problem depends on two user-specified components: a distribution $\mathbb{P}_\lambda$ over values of $\lambda \in \Lambda$ and a collection of basis functions $\boldsymbol{\Phi}_j(\cdot) : \Lambda \to \mathbb{R}^d$, $j = 1, \ldots, p$. We then seek to approximate $\boldsymbol{\theta}^*(\lambda)$ as a linear combination of basis functions, $\boldsymbol{\Phi}_{1:p}(\lambda)\hat{\boldsymbol{\beta}}$, where $\boldsymbol{\Phi}_{1:p} := [\boldsymbol{\Phi}_1 \ \boldsymbol{\Phi}_2 \ \cdots \ \boldsymbol{\Phi}_p]$ and $\hat{\boldsymbol{\beta}}$ is an approximate solution to

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \ \mathbb{E}_{\tilde{\lambda} \sim \mathbb{P}_\lambda} \left[h(\boldsymbol{\Phi}_{1:p}(\tilde{\lambda})\boldsymbol{\beta}, \tilde{\lambda}) \right]. \tag{2}$$

In contrast to discretization schemes which only leverage local information, (stochastic) gradient evaluations of Problem (2) inform *global* structure. Moreover, through a suitable choice of basis functions, we can naturally accomodate complex sets Λ in contrast to earlier work that only treats $\Lambda \subseteq \mathbb{R}$. Finally, any stochastic optimization routine can be used on Problem (2) beyond just SGD (see, e.g., Lan (2020); Bottou et al. (2018), among others).

Despite its simplicity, our approach can approximate the solution path to higher accuracy than discretization with fewer gradient evaluations. See Figure 1 for a sample of our numerical results on weighted binary classification using SGD to solve Problem (2). We prove this behavior is typical. Loosely,

i) When using constant step-size SGD to solve Problem (2), we prove that the uniform error of our learned path $\boldsymbol{\Phi}_{1:p}(\lambda)\hat{\boldsymbol{\beta}}$ to the true path $\boldsymbol{\theta}^*(\lambda)$ exhibits linear convergence to an irreducible constant that is proportional to the expressiveness of the basis (Theorem 3.4). This behavior is already visible in Figure 1.

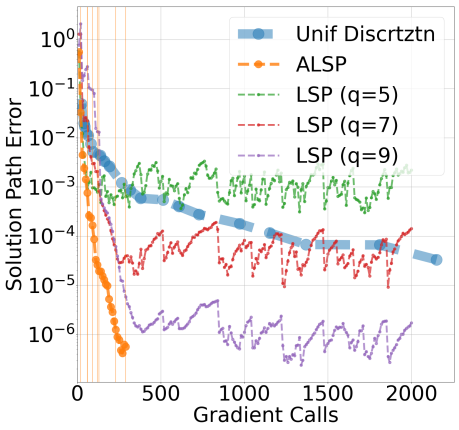


Figure 1: **Learning Solution Path of Weighted Binary Classification.** See Section 6 for setup. We compare our method with $q = 5, 7, 9$ *Legendre* polynomials as our basis with uniform discretization. Orange line is our adaptive method (c.f. Algorithm 1). Vertical lines indicate when new basis functions are added.

The proof of this result utilizes ideas from the convergence of SGD under various "growth conditions" of the gradient (Bottou et al., 2018; Nguyen et al., 2018; Vaswani et al., 2019; Liu et al., 2024; Bertsekas, 1996). See Khaled and Richtárik (2020) for a summary and comparison of these various conditions. Our contribution to this literature is to relate the expressiveness of the basis $\boldsymbol{\Phi}_{1:p}(\lambda)$ to a relaxed, weak-growth condition of Gower et al. (2019, Lemma 2.4). Indeed, we show that under some assumptions, Problem (2) always satisfies this relaxed weak-growth condition, and if the solution path lies in the span of the basis, it satisfies the weak-growth condition of Vaswani et al. (2019). This allows us to leverage results from those works to establish Theorem 3.4.

In special cases, we can leverage a priori knowledge of the structure of $\boldsymbol{\theta}^*(\lambda)$ to prove stronger results. For example, suppose $\Lambda = [-1, 1]$, and we use Legendre polynomials as our basis. We prove that

ii) If the solution path $\boldsymbol{\theta}^*(\lambda)$ is ν-times differentiable, then using $p = O\left(\epsilon^{\frac{1}{2(1-\nu)}}\right)$ polynomials ensures that after $T = O\left(\epsilon^{\frac{1}{1-\nu}} \log(1/\epsilon)\right)$ gradient calls, we obtain an ϵ-approximation to the solution path (Theorem 4.3).

For comparison, Ndiaye et al. (2019) implies that when $h(\boldsymbol{\theta}, \lambda)$ is strongly-convex, discretization requires at least $O(\epsilon^{-1/2})$ points. Hence, even if the optimiza-

Algorithm 1 Adaptively Learn the Solution Path (ALSP)

1: **Initialize:** $\boldsymbol{\Phi}_{1:0}(\cdot) \leftarrow [\,]$ and $\hat{\boldsymbol{\beta}}_0 \leftarrow [\,]$
2: **Return:** A sequence of coefficients: $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots$
3: **At iteration** $p = 1, 2, \ldots$ **do**
4: Append a new basis function:

$$\boldsymbol{\Phi}_{1:p}(\cdot) \leftarrow [\boldsymbol{\Phi}_{1:p-1}(\cdot) \quad \boldsymbol{\Phi}_p(\cdot)]$$

5: Run a first-order method, starting from
6: $[\hat{\boldsymbol{\beta}}_{p-1}, 0] \in \mathbb{R}^p$, to obtain

$$\hat{\boldsymbol{\beta}}_p \;\approx\; \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \; \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[h(\boldsymbol{\Phi}_{1:p}(\tilde{\lambda})\boldsymbol{\beta}, \tilde{\lambda})\right].$$

tion problem at each grid point can be solved with 1 gradient evaluation, our approach requires asymptotically fewer evaluations whenever $\nu > 3$. If ν is large, the savings is substantive.

iii) If the solution path $\boldsymbol{\theta}^*(\lambda)$ is analytic ($\nu = \infty$), then using $p = O(\log(1/\epsilon))$ basis polynomials ensures that after $T = O(\log^2(1/\epsilon)\log\log(1/\epsilon))$ iterations, we obtain an ϵ-approximation to the solution path (Theorem 4.5).

This is almost *exponentially* less work than discretization for unidimensional hyperparameter space.

These specialized results strongly suggest that in the general setting our approach should be competitive as long as we use enough basis functions. This observation motivates a natural heuristic in which we adaptively add basis functions whenever the stochastic optimization routine (e.g., SGD) stalls (see Algorithm 1 and Section 5 for details). We illustrate this idea in Figure 1 (orange line) where we can see that by progressively adding functions we drive the uniform error to zero rapidly.

1.2 Other Related Work

There is a rich literature on specialized methods for computing the solution path under *regularization*, i.e., where $\lambda \in \mathbb{R}_+$ is the weight on a convex regularizer. These methods generally employ a path-following algorithm. See, for example, Rosset (2004); Friedman et al. (2010); Park and Hastie (2007) for cases where $\boldsymbol{\theta}^*(\lambda)$ is smooth, and Rosset and Zhu (2007) for when it is piecewise linear. The LASSO, or ℓ_1 regularized regression, has received special emphasis (Osborne et al., 2000; Efron et al., 2004; Tibshirani and Taylor, 2011), as has support vector machines (Hastie et al., 2004) and certain structured regularizers (Bao et al., 2019). Liu and Grigas (2023) adopts a slightly different perspective than this existing literature, using ordinary differential equations to motivate a second-order method to compute the path. Overall, like our work, these works generally consider the total computational time to compute the path, not just the number of discretization points.

However, our work differs from this literature in two important ways. First, our proposed method applies to general $\lambda \in \Lambda$ which may be multidimensional, whereas the previous literature largely focuses on the case of $\lambda \in \mathbb{R}_+$. This limitation of the previous literature is perhaps fundamental as *path*-following algorithms do not generalize easily to multidimensional settings. Second, we treat a general function $h(\cdot,\cdot)$, not just the case of regularization. Despite our generality, our paper performs a relatively precise computational analysis by bounding total computation time in terms of the number of gradient calls, i.e., we treat the gradient as a block-box oracle instead of a more complicated black-box consisting of minimizing the objective $h(\cdot,\cdot)$ at fixed parameter values. Furthermore, our numerical experiments go beyond simple regularization examples that are common in the literature by considering the upweighing of a minority class in classification and weighting different components of a multi-objective optimization problem. Thus, our method applies more broadly than the aforementioned works.

2 MODEL SETUP AND PRELIMINARIES

We denote the ℓ_2-norm by $\|\cdot\|$ throughout. We focus on the case of smooth, strongly convex functions:

Assumption 2.1 (**Uniform Smoothness and Strong Convexity**). There are constants $0 < \mu \le L$ such that, for all $\lambda \in \Lambda$, $h(\cdot,\lambda)$ is μ-strongly convex and L-smooth, i.e., for all $\boldsymbol{\theta}, \bar{\boldsymbol{\theta}} \in \mathbb{R}^d$,

$$\frac{\mu}{2}\|\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\|^2 \;\le\; h(\boldsymbol{\theta})-h(\bar{\boldsymbol{\theta}})-\nabla h(\bar{\boldsymbol{\theta}})^\top(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}) \;\le\; \frac{L}{2}\|\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\|^2.$$

For any candidate solution path $\boldsymbol{\theta}(\cdot)$, we define the *solution path error* of $\boldsymbol{\theta}(\cdot)$ by

$$\epsilon_{\text{sp}}(\boldsymbol{\theta}(\cdot)) := \sup_{\lambda\in\Lambda}\left\{h(\boldsymbol{\theta}(\lambda),\lambda) - h(\boldsymbol{\theta}^*(\lambda),\lambda)\right\}.$$

An *ϵ-solution path* is a solution path $\boldsymbol{\theta}(\cdot)$ such that $\epsilon_{\text{sp}}(\boldsymbol{\theta}(\cdot)) < \epsilon$. Finally, given any vector of coefficients $\boldsymbol{\beta}$ and basis functions $\boldsymbol{\Phi}_{1:p}(\cdot)$, we define $\epsilon_{\text{sp}}(\boldsymbol{\beta}) := \epsilon_{\text{sp}}(\boldsymbol{\Phi}_{1:p}(\cdot)\boldsymbol{\beta})$ to be the solution path error of $\boldsymbol{\Phi}_{1:p}(\cdot)\boldsymbol{\beta}$.

As mentioned, our method depends on two user-chosen parameters: a distribution $\mathbb{P}_\lambda$ over $\lambda \in \Lambda$ and a series of basis functions $\boldsymbol{\Phi}_j : \Lambda \mapsto \mathbb{R}^d$, for $j = 1, \ldots, p$. We require some minor assumptions on these choices:

Assumption 2.2. (**Distribution, Basis Functions, and Linear Independence**).

i) It is easy to generate i.i.d. samples from $\mathbb{P}_\lambda$.

ii) It is easy to compute $\boldsymbol{\Phi}_j(\lambda) \in \mathbb{R}^d$ for any $\lambda \in \Lambda$, and $1 \leq j \leq p$.

iii) There does not exist $\boldsymbol{\beta} \in \mathbb{R}^p$ with $\boldsymbol{\beta} \neq 0$ such that $\boldsymbol{\Phi}_{1:p}(\tilde{\lambda})\boldsymbol{\beta} = \mathbf{0}$ holds $\mathbb{P}_\lambda$-almost everywhere.

This last condition is a linear independence assumption. If violated, one can remove a function without affecting the span of the basis on the support of $\mathbb{P}_\lambda$.

Given $\boldsymbol{\Phi}_{1:p}$, we define the minimal solution path error for this basis by

$$\epsilon^*_{\text{sp}} := \inf_{\boldsymbol{\beta} \in \mathbb{R}^p} \epsilon_{\text{sp}}(\boldsymbol{\beta}).$$

We also define the following auxiliary constants:

$$\begin{aligned} C &:= \sup_{\lambda \in \Lambda} \sigma_{\max}\left(\boldsymbol{\Phi}(\lambda)^\top \boldsymbol{\Phi}(\lambda)\right), \\ c &:= \sigma_{\min}\left\{\mathbb{E}_{\tilde{\lambda} \sim \mathbb{P}_\lambda}\left[\boldsymbol{\Phi}(\tilde{\lambda})^\top \boldsymbol{\Phi}(\tilde{\lambda})\right]\right\}. \end{aligned} \tag{3}$$

In words, C is a uniform bound on the largest eigenvalue of the positive semidefinite matrix $\boldsymbol{\Phi}(\lambda)^\top \boldsymbol{\Phi}(\lambda)$, and c is the smallest eigenvalue of the corresponding expected matrix. By construction, $0 \leq c \leq C$. Under Assumption 2.2, both constants are strictly positive.

Lemma 2.3 (**Positive Spectral Values**). *If Assumption 2.2 holds, then* $0 < c \leq C$.

We can now state our first key result which relates the suboptimality of a feasible solution $\boldsymbol{\beta}$ in Problem (2) to its solution path error. Define

$$\boldsymbol{\beta}^*_{\text{avg}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{E}_{\tilde{\lambda} \sim \mathbb{P}_\lambda}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}, \tilde{\lambda})\right].$$

Theorem 2.4 (**Relating Suboptimality to Solution Path Error**). *Under Assumptions 2.1 and 2.2, for any* $\boldsymbol{\beta} \in \mathbb{R}^p$, *we have*

$$\epsilon_{\text{sp}}(\boldsymbol{\beta}) \leq 2CL\|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\text{avg}}\|^2 + \left(\frac{8CL}{c\mu}\right)\epsilon^*_{\text{sp}}.$$

This bound decomposes into two terms, one proportional to $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\text{avg}}\|^2$, which represents the suboptimality of $\boldsymbol{\beta}$ and the other proportional to ϵ^*_{sp}, which measures the maximal expressiveness of the basis. By solving Problem (2) to greater accuracy, we can drive down the first term, but we will not affect the second term. To reduce the second term, we must add basis functions to obtain a better quality approximation.

Theorem 2.4 shows any algorithm for solving Problem (2) can be used to approximate the solution path. In the next section, we argue that when ϵ^*_{sp} is small, constant step-size SGD for Problem (2) exhibits linear convergence to a constant proportional to ϵ^*_{sp}, making it an ideal algorithm to study.

3 SGD TO LEARN THE SOLUTION PATH

In this section, we apply constant step-size SGD to solve Problem (2). The key idea is to show that Problem (2) satisfies a certain growth condition and apply Gower et al. (2019, Theorem 3.1). One minor detail is that Gower et al. (2019) proves their result under a stronger "expected smoothness" condition for the setting where the objective Problem (2) is a finite sum. However, the result holds more generally under a weaker condition (Eq. (9) of their work). Hence, for clarity, we first restate this condition and the more general result.

The following definition is motivated by Gower et al. (2019, Lemma 2.4):

Definition 3.1 (**Relaxed Weak Growth Condition (RWGC)**). Consider a family of functions $g(\cdot, z) : \mathbb{R}^d \to \mathbb{R}$, $z \sim \mathbb{P}_z$, and $G(\cdot) := \mathbb{E}_{z \sim \mathbb{P}_z}[g(\cdot, z)]$. Let $G^* := \min_{\boldsymbol{w} \in \mathbb{R}^d} G(\boldsymbol{w})$. Then, g and G are said to satisfy the *relaxed weak growth condition* with constants $\rho \geq 0$ and $\sigma \geq 0$, if for all $\boldsymbol{w} \in \mathbb{R}^d$,

$$\mathbb{E}_{z \sim \mathbb{P}_z}\left[\|\nabla g(\boldsymbol{w}, z)\|^2\right] \leq 2\rho(G(\boldsymbol{w}) - G^*) + \sigma^2.$$

When $\sigma = 0$, Definition 3.1 recovers the *weak growth condition* of Vaswani et al. (2019). In this case, the variance of the stochastic gradients goes to zero as we approach an optimal solution. Hence, at optimality, not only is the expectation of the gradient zero, but it is zero almost surely in z. For regression problems, this condition corresponds to interpolation.

Definition 3.1 is implied by a standard second moment condition on the gradients (take $\rho = 0$). However, the most interesting cases are when σ^2 is small relative to ρ and $\rho > 0$. (This will be the case for Problem (2). See Section 3.1.)

We then have the following theorem.

Theorem 3.2 (Gower et al. (2019)). *Suppose that the RWGC holds for the family of functions* $g(\cdot, z) : \mathbb{R}^d \to \mathbb{R}$, *and that* $G(\cdot) := \mathbb{E}_{z \sim \mathbb{P}_z}[g(\cdot, z)]$ *is* μ_g*-strongly convex for some* $\mu_g > 0$. *Consider the SGD algorithm, initialized at* $\boldsymbol{w}_0 \in \mathbb{R}^d$, *with iterations*

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \frac{\bar{\eta}}{\rho}\nabla g(\boldsymbol{w}_t, z_t),$$

where $z_t \sim \mathbb{P}_z$ *is a random sample and* $\bar{\eta} \in \left(0, \min\left\{1, \frac{\rho}{\mu_g}\right\}\right]$ *parametrizes the step-size. Let* $\boldsymbol{w}^* := \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} G(\boldsymbol{w})$. *Then for all* $t \geq 0$, *we have*

$$\mathbb{E}[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2] \leq \left(1 - \frac{\bar{\eta}\mu_g}{\rho}\right)^t \|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + \frac{\bar{\eta}\sigma^2}{\rho\mu_g}.$$

In other words, constant step-size SGD under the RWGC exhibits a "fast" linear convergence in expectation up to a constant that is directly proportional to σ^2, after which it stalls. When $\sigma^2 = 0$, we exactly recover the result of Vaswani et al. (2019). To keep the exposition self-contained, we provide a proof in the appendix.

3.1 Solving Problem (2) with SGD

We next prove that Problem (2) satisfies RWGC, and in particular, that the "σ^2" term depends on ϵ^*_{sp}, the minimal solution path error of the basis. To our knowledge, we are the first to make this observation.

Define

$$f(\boldsymbol{\beta}, \lambda) := h(\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}, \lambda), \tag{4a}$$
$$F(\boldsymbol{\beta}) := \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[f(\boldsymbol{\beta}, \tilde{\lambda})\right] = \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}, \tilde{\lambda})\right],$$

Lemma 3.3 (Problem (2) satisfies RWGC). *Suppose that Assumptions 2.1 and 2.2 hold and recall the constants defined in Equation (3). Then, the family of functions $f(\cdot, \lambda)$ and the function $F(\cdot)$, defined in Equation (4), satisfy the relaxed weak growth condition (RWGC) with constants $\rho = CL$ and $\sigma^2 = 2CL\epsilon^*_{\text{sp}}$. Namely, for all $\boldsymbol{\beta} \in \mathbb{R}^d$,*

$$\mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[\|\nabla f(\boldsymbol{\beta}, \tilde{\lambda})\|^2\right] \;\leq\; 2CL(F(\boldsymbol{\beta}) - F^*) + 2CL\epsilon^*_{\text{sp}}.$$

In addition, $F(\cdot)$ is $c\mu$-strongly convex.

We now combine Theorems 2.4 and 3.2 and Lemma 3.3 to yield our main result: a bound on the expected solution path error for constant step-size SGD. Recall constant step-size SGD in this setting yields the iteration

$$\boldsymbol{\beta}_{t+1} \leftarrow \boldsymbol{\beta}_t - \eta\nabla_{\boldsymbol{\beta}} h(\boldsymbol{\Phi}(\tilde{\lambda}_t)\boldsymbol{\beta}_t, \tilde{\lambda}_t),$$

where

$$\nabla_{\boldsymbol{\beta}} h(\boldsymbol{\Phi}(\tilde{\lambda}_t)\boldsymbol{\beta}_t, \tilde{\lambda}_t) = \boldsymbol{\Phi}(\tilde{\lambda}_t)^\top \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\Phi}(\tilde{\lambda}_t)\boldsymbol{\beta}_t, \tilde{\lambda}_t),$$

and $\tilde{\lambda}_t \sim \mathbb{P}_\lambda$. We assume that the gradient $\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}, \lambda)$ is easily computable for any $\boldsymbol{\theta}$ and λ.

We then have:

Theorem 3.4 (Expected Solution Path Error Convergence for SGD). *Under Assumptions 2.1 and 2.2, consider applying SGD to Problem (2) with a constant step-size $\eta = \frac{\bar{\eta}}{CL}$ parameterized by $\bar{\eta} \in (0, 1]$. Then, after T iterations, the expected solution path error is at most*

$$\mathbb{E}\left[\epsilon_{\text{sp}}(\boldsymbol{\beta}_T)\right] \;\leq\; 2CL\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2\left(1 - \frac{\bar{\eta}c\mu}{CL}\right)^T + \left(\frac{4CL(\bar{\eta}+2)}{c\mu}\right)\epsilon^*_{\text{sp}}.$$

In particular, when $\epsilon^*_{\text{sp}} > 0$, then for $T \geq \left\lceil \frac{CL}{\bar{\eta}c\mu}\log\left(\frac{c\mu\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2}{2\bar{\eta}\epsilon^*_{\text{sp}}}\right)\right\rceil$,

$$\mathbb{E}\left[\epsilon_{\text{sp}}(\boldsymbol{\beta}_T)\right] \;\leq\; \left(\frac{8CL(\bar{\eta}+1)}{c\mu}\right)\epsilon^*_{\text{sp}}.$$

When $\epsilon^*_{\text{sp}} = 0$, then for any $\epsilon > 0$ and $T \geq \left\lceil \frac{CL}{\bar{\eta}c\mu}\log\left(\frac{2CL\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2}{\epsilon}\right)\right\rceil$,

$$\mathbb{E}\left[\epsilon_{\text{sp}}(\boldsymbol{\beta}_T)\right] \;\leq\; \epsilon.$$

Theorem 3.4 highlights the role of the basis functions in our results. First we see that as $T \to \infty$, the expected solution path error plateaus at a constant proportional to $\frac{CL}{c\mu}\epsilon^*_{\text{sp}}$. As mentioned, ϵ^*_{sp} measures the expressiveness of the basis, and we expect ϵ^*_{sp} to be small as we add more basis functions. We interpret $\frac{CL}{c\mu}$ as a condition number for Problem (2), which also depends on the choice of basis through Equation (3). This constant increases as we grow the basis. Finally, the iteration complexity scales linearly with this condition number. Thus, an "ideal" basis must navigate this tradeoff between ϵ^*_{sp} and C/c.

Fortunately, there exists a rich theory on function approximation that studies the relationship between basis functions, uniform error, and eigenspectra. In the next section we leverage this theory to provide a comparison of our method with existing discretization techniques in a specialized setting.

4 SPECIALIZED RESULTS FOR $\Lambda = [-1, 1]$

We next leverage results from function approximation theory to bound the number of basis functions needed to achieve a target solution path error. We focus on the case $\Lambda = [-1, 1]$ as it facilitates simple comparisons to existing results and elucidates key intuition.

We use a simple basis: we approximate each component $i = 1, \ldots, d$ of $\theta^*_i(\cdot)$ by the first q Legendre polynomials. Hence the total number of basis functions is $p = qd$. Recall, the Legendre polynomials form an orthogonal basis on $[-1, 1]$ with respect to the uniform distribution, i.e., $\mathbb{E}_{\tilde{\lambda}\sim\text{Unif}[-1,1]}\left[P_n(\tilde{\lambda})P_m(\tilde{\lambda})\right] = \frac{2}{2n+1}\mathbb{I}\{n = m\}$, where P_n and P_m are the n^{th} and m^{th} Legendre polynomial, respectively. Since we approximate each component separately, the matrix $\boldsymbol{\Phi}(\lambda) \in \mathbb{R}^{d\times qd}$ is block-diagonal, with d blocks of size $1 \times q$.

For this basis, depending on the value of $p = qd$ implied by the choice of q, let C_p, c_p refer to the constants (3). We can bound the constant C_p/c_p:

Lemma 4.1 (C_p/c_p **for Legendre Polynomials**). *Take* $\Lambda = [-1, 1]$ *and* $\mathbb{P}_\lambda$ *to be the uniform distribution on* $[-1, 1]$. *Then, for the above basis,* $C_p/c_p \leq q^2$.

Notice, that this constant is independent of d and grows mildly with q. This is not true of all polynomial bases. One can check empirically that for the monomial basis, C_p/c_p grows exponentially fast in q.

Bounding ϵ^*_{sp} depends on the properties $\boldsymbol{\theta}^*(\lambda)$. As a first example,

Lemma 4.2 (ϵ^*_{sp} **for** ν**-Differentiable Solution Paths**). *Let* $\Lambda = [-1, 1]$. *Suppose Assumption 2.1 holds. Further assume that there exists an integer* $\nu \geq 0$ *and constant* $V > 0$ *such that for all* $i = 1, \ldots, d$, $\boldsymbol{\theta}^*_i(\cdot)$ *has* ν *derivatives, where* $\boldsymbol{\theta}^*_i(\lambda)$, ..., $\boldsymbol{\theta}^{(\nu-1)*}_i(\lambda)$ *are absolutely continuous and* $\boldsymbol{\theta}^{(\nu)*}_i(\lambda)$ *has total variation bounded by* V. *Then, for any* $q \geq \nu + 1$, *for the basis described above,* $\epsilon^*_{\text{sp}} \leq \frac{dL}{2}\left(\frac{2V}{\pi\nu(q-\nu)^\nu}\right)^2$.

The proof of Lemma 4.2 is constructive; we exhibit a polynomial with the given solution path error. The bound confirms the intuition that if $\boldsymbol{\theta}^*(\cdot)$ is smooth (has many derivatives), then adding basis function drives down the approximation error rapidly. Problems arising in many application domains, including machine learning problems like ridge regression and relatives, often exhibit highly or even infinitely differentiable solution paths.

Combining these lemmas with Theorem 3.4 allows us to calculate the requisite basis size and number of iterations needed to achieve a target solution-path error:

Theorem 4.3 (**SGD for** ν**-differentiable Solution Paths**). *Let* $\Lambda = [-1, 1]$, $\mathbb{P}_\lambda$ *be the uniform distribution on* $[-1, 1]$. *Then, under the conditions of Lemma 4.2 and assume* $\nu \geq 2$, *if we use* $q = O(\epsilon^{\frac{1}{2(1-\nu)}})$ *polynomials in the previous basis, and run constant-step size SGD for* $O(\epsilon^{\frac{1}{1-\nu}}\log(1/\epsilon))$ *iterations, the resulting iterate* $\boldsymbol{\beta}_T$ *satisfies* $\mathbb{E}\left[\epsilon_{\text{sp}}(\boldsymbol{\beta}_T)\right] \leq \epsilon$.

For clarity, while Lemma 4.2 holds for any $\nu \geq 0$, Theorem 4.3 requires $\nu \geq 2$. Moreover, both big "Oh" terms should be interpreted as $\epsilon \to 0$, and both suppress constants not depending on ϵ (but possibly depending on h). The theorem establishes that the "smoother" $\boldsymbol{\theta}^*(\cdot)$ is (i.e. larger ν), the fewer iterations required by our method to achieve a target tolerance.

Recall, Ndiaye et al. (2019) established that for strongly convex functions, discretization requires at least $O(\epsilon^{-1/2})$ points. One might expect the number of gradient evaluations per point to scale like $O(\log(1/\epsilon))$. Hence, if $\nu \geq 3$, our approach requires asymptotically less work. The larger ν, the larger the savings.

This gap becomes more striking as $\nu \to \infty$. Hence, we next consider the case where $\theta^*_i(\lambda)$ is analytic on $[-1, 1]$, i.e, its Taylor Series is absolutely convergent on this interval.

Lemma 4.4 (ϵ^*_{sp} **for Analytic Solution Paths**). *Suppose Assumption 2.1 holds and that for each* $i = 1, \ldots, d$, $\boldsymbol{\theta}^*_i(\cdot)$ *is analytic on the interval* $[-1, 1]$. *Then, there exist constants* $\omega > 1$ *and* $M > 0$ *such that for any* $q > 0$, *with the basis described above,* $\epsilon^*_{\text{sp}} \leq \frac{dL}{2}\left(\frac{2M\omega^{-q}}{\omega-1}\right)^2$.

The proof is again constructive. The constants ω and M pertain to the analytic continuation of $\theta^*_i(\cdot)$ to the complex plane. Importantly, the solution path error now dies geometrically fast (like ω^{-2q}). Using this faster decay rate yields,

Theorem 4.5 (**SGD for Analytic Solution Paths**). *Let* $\Lambda = [-1, 1]$, *and* $\mathbb{P}_\lambda$ *be the uniform distribution on* $[-1, 1]$. *Then, under the conditions of Lemma 4.4, if we use* $q = \log(1/\epsilon)/\log(\omega)$ *polynomials in the previous basis, and run constant-step size SGD for* $O\left(\log^2(1/\epsilon)\log\log(1/\epsilon)\right)$ *iterations, the resulting iterate* $\boldsymbol{\beta}_T$ *satisfies* $\mathbb{E}\left[\epsilon_{\text{sp}}(\boldsymbol{\beta}_T)\right] \leq \epsilon$.

Again, the big "Oh" hides constants that do not depend on ϵ. Compared to Ndiaye et al. (2019), the amount of work required is almost exponentially smaller for $\dim(\Lambda) = 1$.

In higher dimensions, the curse of dimensionality causes the required discretization points to grow exponentially with $\dim(\Lambda)$. In contrast, our method's performance depends not on $\dim(\Lambda)$ but on the minimal solution path error ϵ^*_{sp}. The key advantage arises when we either select a basis ensuring a small ϵ^*_{sp} using prior knowledge or adopt a highly flexible, adaptive approach, as we next discuss in Section 5.

5 IMPLEMENTATION GUIDELINES

Although Section 4 provides insight on how to select a basis (in certain cases), choosing the "right" basis apriori from approximation theory remains a difficult challenge. To circumvent this issue, we re-introduce Algorithm 1. Algorithm 1 is more practical, as we only need to specify: i) a sequence of basis functions, ii) a distribution $\mathbb{P}_\lambda$ and, iii) a criterion for deciding when to add a new function to the basis. We next provide intuition and practical guidance on these choices.

Intuition for an "Ideal" Basis. Consider

$$h(\boldsymbol{\theta}, \boldsymbol{\lambda}) \;=\; \boldsymbol{\theta}^\top \boldsymbol{Q}(\boldsymbol{\lambda})\boldsymbol{\theta} + \boldsymbol{b}(\boldsymbol{\lambda})^\top \boldsymbol{\theta}.$$

In this special case, Problem (2) becomes

$$\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top \mathbb{E}\left[\boldsymbol{\Phi}(\tilde{\boldsymbol{\lambda}})^\top \boldsymbol{Q}(\tilde{\boldsymbol{\lambda}})\boldsymbol{\Phi}(\tilde{\boldsymbol{\lambda}})\right]\boldsymbol{\beta} + \mathbb{E}\left[\boldsymbol{b}(\tilde{\boldsymbol{\lambda}})^\top \boldsymbol{\Phi}(\tilde{\boldsymbol{\lambda}})\right]\boldsymbol{\beta}.$$

The complexity of this problem depends strongly on its condition number (in our theory, this is bounded by $\frac{CL}{c\mu}$), which equals the condition number of the matrix $\mathbb{E}\left[\boldsymbol{\Phi}(\tilde{\boldsymbol{\lambda}})^\top \boldsymbol{Q}(\tilde{\boldsymbol{\lambda}})\boldsymbol{\Phi}(\tilde{\boldsymbol{\lambda}})\right]$ and is directly computable in this case. Hence, an ideal basis would ensure this condition number is as close to 1 (its lower bound) as possible. In particular, if the basis functions $\boldsymbol{\Phi}_j(\cdot)$ are orthonormal with respect to the inner product

$$\langle \boldsymbol{\Phi}_j(\cdot), \boldsymbol{\Phi}_k(\cdot)\rangle \;\equiv\; \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[\boldsymbol{\Phi}_j(\tilde{\boldsymbol{\lambda}})^\top \boldsymbol{Q}(\tilde{\boldsymbol{\lambda}})\boldsymbol{\Phi}_k(\tilde{\boldsymbol{\lambda}})\right],$$

then, the resulting condition number is 1. For general h, insofar as $F(\boldsymbol{\beta}) = \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}, \tilde{\lambda})\right]$ might be approximated by a second-order Taylor series expansion around the optimum $\boldsymbol{\beta}^*_{\text{avg}}$, a good basis might be orthonormal w.r.t. to the Hessian of $F(\cdot)$ at optimality. We verify this in Section 6.3. Absent knowledge about the Hessian at optimality, the best approach is to use a family of flexible and expressive basis functions. In Sections 6.1 and 6.2, we use orthogonal polynomials for the following reasons: i) Adding new basis functions does not require altering the existing basis, and ii) Intuitively, orthogonality suggests that if $\hat{\boldsymbol{\beta}}_{p-1}$ is near optimal in the $(p-1)^{\text{th}}$ iteration, then $[\hat{\boldsymbol{\beta}}_{p-1}, 0]$ is likely to be a near optimal (and good warm-start) solution in the p^{th} iteration. The first benefit is not shared, for example, by other approaches like cubic splines or neural networks.

Interplay Between Basis and Distribution. Although *any* sequence of basis functions and *any* distribution can be used in Algorithm 1, we suggest making these choices in concert. In our experiments, we focus on sequences of polynomials that are orthogonal with respect to $\mathbb{P}_\lambda$. The Legendre polynomials and the uniform distribution on $[-1, 1]$ is one such pair, but there are many canonical examples including Hermite polynomials with the normal distribution and Laguerre polynomials with the exponential distribution. There are performance-optimized implementations of these families in standard software (see, e.g., `scipy.special`). Although polynomials map to $\mathbb{R}$, by approximating each dimension separately as in Section 4, we can extend polynomials to a basis for $d > 1$. Finally, polynomials are highly expressive as they are uniform approximators for Lipschitz functions.

Deciding the Number of Basis Functions "On the Fly". As reflected by our theoretical convergence guarantees, choosing a large, flexible basis induces a tradeoff. A larger basis increases the condition number of Problem (2) (as measured by the ratio $\frac{CL}{c\mu}$), and hence convergence will be slower. This observation motivates our adaptive approach that only adds basis functions as necessary until we reach a desired accuracy. One approach is to empirically approximate the objective of Problem (2) using a hold-out validation set, and add a basis function when performance stalls. In our experiment with low dimensional hyperparameters (Section 6.1 and 6.2), we use Gauss-Legendre quadrature to evaluate the stochastic objective exactly over the entire hyperparameter space. In the high dimensional hyperparameter experiment (Section 6.3), we instead compute the performance of our learned $\boldsymbol{\beta}$ through an ERM version of the stochastic objective on a validation set of size 1000.

6 NUMERICAL EXPERIMENTS

We next empirically compare i) our approach using a fixed, large basis (denoted *LSP* for "Learning the Solution Path") ii) our adaptive Algorithm 1 (denoted *ALSP* for "Adaptive LSP" and iii) uniform discretization, a natural benchmark. We aim to show that both our approaches not only outperform the benchmark but that the qualitative insights from our theoretical results hold for more general optimization procedures than constant step-size SGD. We also provide preliminary results on how the choice of basis affects performance. Our repository can be found at `https://github.com/Cumberkid/Learning-the-Optimal-Solution-Path`.

We choose uniform discretization to be our benchmark in lieu of other schemes (like geometric spacing) because i) it matches the theoretical lower bound from (Ndiaye et al., 2019) and ii) we see it as most intuitive for learning the solution path with small *uniform* error. Specifically, for various ϵ, we consider a uniform spacing of size $\sqrt{\epsilon}$ in each dimension of λ, and run (warm-started) gradient descent for $O(\log(1/\epsilon))$ iterations. The constant hidden by big "Oh" here is calibrated in an oracle fashion to achieve a solution-path error of $O(\epsilon)$ (see Appendix B.1 for details) giving the benchmark a small advantage.

6.1 Weighted Binary Classification

We consider a binary classification problem using a randomly selected subset of 1000 cases from the highly imbalanced Law School Admission Bar Passage dataset (Wightman, 1998). Of the 1000 cases, there are 956 positive instances and 44 negatives. Standard logistic regression predicts 992 positives with a false positive rate of 0.86. When identifying students likely to fail is key, the default classifier may not be useful. Reweighting cases is a standard approach to improve false positive rate at the cost of overall accuracy.

We take

$$h(\boldsymbol{\theta}, \lambda) = (1-\lambda) l_{\text{pos}}(\boldsymbol{\theta}) + \lambda l_{\text{neg}}(\boldsymbol{\theta}) + 0.125\|\boldsymbol{\theta}\|^2,$$

where $l_{\text{pos}}(\boldsymbol{\theta})$ and $l_{\text{neg}}(\boldsymbol{\theta})$ denote the negative log-likelihood on the positive and negative classes respectively. Specifically, letting $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{45} \times \{0,1\}$ for $i = 1, \ldots, n$ denote the data,

$$l_{\text{pos}}(\boldsymbol{\theta}) = \frac{1}{|\{i : \ y_i = 1\}|} \sum_{i:y_i=1} \log(1 + e^{(-2y_i+1)\boldsymbol{x}_i^\top \boldsymbol{\theta}}),$$

and similarly for $l_{\text{neg}}(\boldsymbol{\theta})$.

We consider two different choices of distribution for $\lambda \in [0,1]$ together with two different orthogonal polynomial bases:

i) A Unif$[0,1]$ distribution. Here we use (scaled and shifted) Legendre polynomials.
ii) A Beta$(b+1, a+1)$ distribution. Here we use (shifted) Jacobi polynomials with parameters $(a,b) = (-.3, -.7)$.

The ground truth $\boldsymbol{\theta}^*(\lambda)$ is computed via 5000 iterations of (warm-started) gradient descent over a uniform grid of 2^{10} points. Solution path error is approximated by the uniform error over this grid.

For LSP, we run SGD using `torch.optim.SGD`. In lieu of a constant learning rate, we dynamically reduce the learning rate according to the Distance Diagnostic of Pesme et al. (2020, Section 4). This dynamic updating is more reflective of practice. We use the suggested parameters from Pesme et al. (2020) with the exception of `q`[1], which we tune by examining the performance after 200 iterations. (We take `q` $= 1.3$ for both the Legendre and Jacobi bases.) For ALSP, we initialize the algorithm with 5 polynomials and stop it after reaching 12 polynomials ($q = 5$ to 12).

Figure 1 in the introduction and Figure 2 panel (a) show the performance for the Legendre and Jacobi bases respectively. As we can see, LSP methods converge rapidly to an irreducible error, and this error decreases as we add basis functions. Our adaptive method converges rapidly. This performance is resonated by Figure 2 panels (c) and (d), where we plot the coefficient profile for $\theta_3^*(\lambda)$ when using 7 Legendre and Jacobi polynomials. The approximation improves as we increase the number of iterations, and eventually become closely aligned with the true solution path.

Figure 2 panel (b) shows that the largest errors occur almost periodically at a few places across $\Lambda = [0,1]$, confirming the equioscillation theorem.

Practically, this experiment suggests that when the region of Λ is compact, taking a polynomial basis paired with a distribution supported everywhere over Λ is favorable. The polynomial basis can then be taken to be orthogonal to $\mathbb{P}_\lambda$ as discussed in Section 5.

[1]Note that `q`, defined as a diagnostic frequency control parameter in Pesme et al. (2020, Section 4), is different from q, the number of polynomials in our paper.

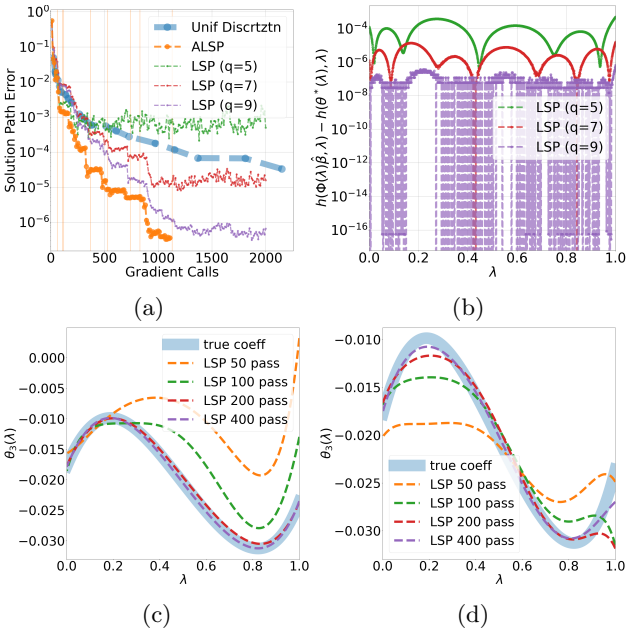


Figure 2: **LSP and ALSP for Weighted Binary Classification.** (a) Compares methods using Jacobi polynomials. (b) Error in solution path as a function of λ using Jacobi polynomials. (c) and (d) compare $\boldsymbol{\theta}_3(\lambda)$ across λ in the Legendre and Jacobi bases respectively with 7 polynomials.

6.2 Portfolio Allocation

We next consider a portfolio allocation problem calibrated to real data where $\boldsymbol{\theta} \in \mathbb{R}^d$ represents the weights on $d = 10$ different asset classes. Namely, let $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ be the mean and covariance matrix of the returns of the different asset classes. We fit these parameters to the monthly return data from Aug. 2014 to July 2024 using the Fama-French 10 Industry index dataset.[2]

We then solve

$$\min_{\boldsymbol{\theta}} \quad -\lambda_1 \mu^\top \boldsymbol{\theta} + \lambda_2 \boldsymbol{\theta}^\top \Sigma \boldsymbol{\theta} + \bar{\ell}_1(\boldsymbol{\theta}).$$

Here the first term represents the (negative) expected return, the second represents the risk (variance) and the third is a smoothed version ℓ_1 regularization to induce sparsity. Namely, $\bar{\ell}_1(\boldsymbol{\theta}) := \sum_{i=1}^d \sqrt{\theta_i^2 + .01^2} - .01$. The parameters $\lambda_1 \in [0,1]$ and $\lambda_2 \in [0.2, 1]$ control the tradeoffs in these multiple objectives.

Ground truth $\boldsymbol{\theta}^*(\lambda)$ is computed over a 100×100 grid.

We focus on bivariate-Legendre polynomials for our basis, scaled and shifted to be orthogonal to uniform distribution on $[0,1] \times [.2, 1]$. We initialize with ALSP with 2 polynomials in each dimension, and iterate by

[2]`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

adding one polynomial in each dimension, so that $q = 4, 9, 16, 25$, after which we stop.

Unlike our previous experiment, to showcase that the qualitative insights of our theory hold for other algorithms other than SGD, we use `torch.optim.LBFGS` for LSP, ALSP and uniform discretization (for a fair comparison). Unlike SGD, L-BFGS uses both function and gradient evaluations. We restrict it to use only 10 function calls per gradient step so that the total work is still proportional to the number of gradient calls.

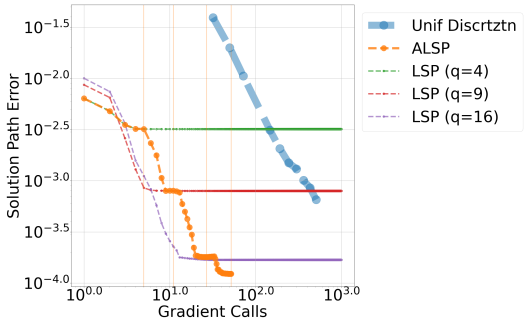


Figure 3: **LSP and ALSP for Portfolio Allocation.** Compares methods using the first $q = 4, 9, 16$ bivariate-Legendre polynomials.

As in our previous experiment, we see that LSP rapidly converges to an irreducible error and then plateaus. By contrast, ALSP seems to make continued progress as we add polynomials. Both methods substantively outperform uniform discretization.

The improved performance over uniform discretization is partially attributable to the increased dimension of Λ (because discretization suffers from the curse of dimensionality), but is also because (traditionally) discretization interpolates solutions in a piecewise constant fashion. In this example, $\boldsymbol{\theta}^*(\lambda)$ is very smooth. See Figure 6 in Appendix. Hence, polynomial can learn to interpolate values very fast, while uniform discretization needs a great deal more resolution.

6.3 Portfolio Allocation with Moderate Dimensional Λ

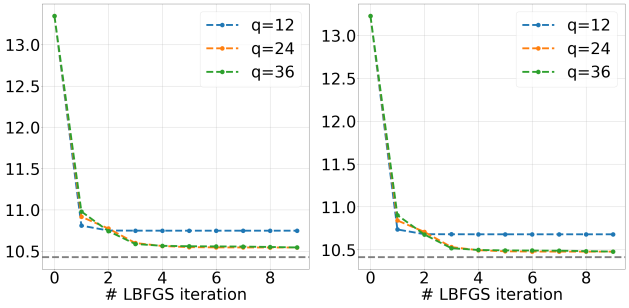
We next modify the portfolio allocation problem in Section 6.2 to include transaction costs relative to the current portfolio as follows:

$$h(\theta, \lambda) \;=\; -\lambda_1 \cdot \mu^\top \theta + \lambda_2 \cdot \theta^\top \Sigma \theta + \|\theta - \lambda_{3:12}\|_2^2.$$

Here $\lambda_{3:12} \in \mathbb{R}^{10}$ represents the current portfolio holding, and $\lambda \in \mathbb{R}^{12}$. We use the following basis

$$\lambda_{(j \bmod 12)} \cdot \lambda_2^{\lfloor j/12 \rfloor}, \quad j = 1, \ldots, q,$$

for each component of $\hat{\boldsymbol{\theta}}_i(\cdot)$, for $i = 1, \ldots, 10$, and let $q = 12, 24, 36$ in our experiments. As $q \to \infty$, this basis contains the optimal solution path in its span for any choice of μ, Σ (c.f. Appendix B.2).



(a) (b)

Figure 4: **Portfolio Allocation,** $\dim(\Lambda) = 10$**.** Comparing the ERM objective for varying q. Panel a) training, Panel b) validation.

For simplicity, we ran LSP using `torch.optim.LBFGS` to solve an ERM version of the problem in Step 6 of Algorithm 1 on a training set of size 1000, and present results for an independent validation set of size 1000.

Discretization on a 12-dimensional grid is computationally challenging. Hence, to assess our method we instead compute an ERM approximation to $\mathbb{E}\left[h(\boldsymbol{\theta}^*(\tilde{\lambda}), \tilde{\lambda})\right]$ over the training/validation sets, i.e., compute $\frac{1}{1000}\sum_{i=1}^{1000} h(\boldsymbol{\theta}^*(\lambda_i), \lambda_i)$ by solving 1000 separate optimization problems. This value is the dotted grey line in Figure 4. The smaller the gap, the closer $\boldsymbol{\theta}^*(\lambda_i)$ is to the span of the basis.

Similar to previous experiments, as the basis size increases, performance improves, and even with a fairly small basis, the error is quite small.

7 CONCLUSION

We propose a new method for learning the optimal solution path of a family of problems by reframing it as a single stochastic optimization problem over a linear combination of pre-specified basis functions. Compared to discretization schemes, our method offers flexibility and scalability by taking a global perspective on the solution path. We prove that our problem satisfies a certain relaxed weak-growth condition that allows us to solve the single optimization problem very efficiently when using sufficiently rich bases. Theoretical results in special cases and numerical experiments in more general settings support these findings. Future research might more carefully examine the interplay between the parameterization of the family of problems and the choice of basis. One might also consider other universal function approximators (e.g. deep neural networks, forests of trees) within this context.

Acknowledgements

PG acknowledges the support of the NSF AI Institute for Advances in Optimization, Award 2112533. The authors thank the anonymous reviewers for their thoughtful and constructive comments.

References

R. Bao, B. Gu, and H. Huang. Efficient approximate solution path algorithm for order weight l_1-norm with accuracy guarantee. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 958–963. IEEE, 2019.

D. Bertsekas. Neuro-dynamic programming. *Athena Scientific*, 1996.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, pages 407–451, 2004.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33 (1):1, 2010.

J. Giesen, M. Jaggi, and S. Laue. Approximating parameterized convex optimization problems. *ACM Transactions on Algorithms (TALG)*, 9(1):1–17, 2012a.

J. Giesen, S. Laue, J. Mueller, and S. Swiercy. Approximating concavely parameterized optimization problems. *Advances in Neural Information Processing Systems*, 3:2105–2113, 01 2012b.

R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. Sgd: General analysis and improved rates. In *International conference on machine learning*, pages 5200–5209. PMLR, 2019.

S. Gupta, J. Moondra, and M. Singh. Which lp norm is the fairest? approximations for fair facility location across all" p". In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 817–817, 2023.

T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5 (Oct):1391–1415, 2004.

A. Khaled and P. Richtárik. Better theory for sgd in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*, volume 1. Springer, 2020.

C. Liu, D. Drusvyatskiy, M. Belkin, D. Davis, and Y. Ma. Aiming towards the minimizers: fast convergence of sgd for overparametrized problems. *Advances in neural information processing systems*, 36, 2024.

H. Liu and P. Grigas. New methods for parametric optimization via differential equations. *arXiv preprint arXiv:2306.08812*, 2023.

E. Ndiaye, T. Le, O. Fercoq, J. Salmon, and I. Takeuchi. Safe grid search with optimal complexity. In *International Conference on Machine Learning*, pages 4771–4780. PMLR, 2019.

L. Nguyen, P. H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, and M. Takác. Sgd and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758. PMLR, 2018.

M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20 (3):389–403, 2000.

M. Y. Park and T. Hastie. L1-Regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):659–677, 08 2007.

S. Pesme, A. Dieuleveut, and N. Flammarion. On convergence-diagnostic based step sizes for stochastic gradient descent. In *International Conference on Machine Learning*, pages 7641–7651. PMLR, 2020.

S. Rosset. Following curved regularized optimization solution paths. *Advances in Neural Information Processing Systems*, 17, 2004.

S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, pages 1012–1030, 2007.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39 (3):1335 – 1371, 2011.

L. N. Trefethen. *Approximation Theory and Approximation Practice*. SIAM, 2013.

S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.

L. F. Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Omitted Proofs

A.1 Proofs from Section 2

Proof for Lemma 2.3. Suppose by contradiction that $c = 0$. Then there exists a corresponding eigenvector $\boldsymbol{v} \in \mathbb{R}^p$ with $\boldsymbol{v} \neq \mathbf{0}$ such that

$$\mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[\boldsymbol{\Phi}(\tilde{\lambda})^\top \boldsymbol{\Phi}(\tilde{\lambda})\right]\boldsymbol{v} = \mathbf{0} \implies \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[\|\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{v}\|_2^2\right] = 0.$$

This further implies that $\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{v} = 0$ almost everywhere. By the linear independence assumption in Assumption 2.2, we then must have $\boldsymbol{v} = \mathbf{0}$, but this is a contradiction since $\boldsymbol{v}$ is an eigenvector.

For the second statement, notice

$$c \le \sigma_{\max}\left\{\mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[\boldsymbol{\Phi}(\tilde{\lambda})^\top \boldsymbol{\Phi}(\tilde{\lambda})\right]\right\} \le \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[\sigma_{\max}\left\{\boldsymbol{\Phi}(\tilde{\lambda})^\top \boldsymbol{\Phi}(\tilde{\lambda})\right\}\right] \le C,$$

where the penultimate inequality follows from Jensen's inequality and the convexity of the maximal eigenvalue function. □

Next, recall the family of functions $f(\cdot, \lambda)$ and the function $F(\cdot)$ defined in (4):

$$f(\boldsymbol{\beta}, \lambda) \;:=\; h(\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}, \lambda), \qquad F(\boldsymbol{\beta}) \;:=\; \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[f(\boldsymbol{\beta}, \tilde{\lambda})\right] = \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}, \tilde{\lambda})\right].$$

Before proceeding, we establish the smoothness and strong convexity of these functions.

Proposition A.1 (Uniform Smoothness and Strong Convexity). *The family of functions $f(\cdot, \lambda)$, over all $\lambda \in \Lambda$, is CL-smooth. Moreover, $F(\cdot)$ is CL-smooth and $c\mu$-strongly convex.*

Proof for Proposition A.1. Fix an arbitrary $\lambda \in \Lambda$. Then, for any $\boldsymbol{\beta}$, $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$, smoothness of $f(\cdot, \lambda)$ is certified by

$$\begin{aligned}
\|\nabla f(\boldsymbol{\beta}, \lambda) - \nabla f(\bar{\boldsymbol{\beta}}, \lambda)\| &= \|\nabla_{\boldsymbol{\beta}} h(\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}, \lambda) - \nabla_{\boldsymbol{\beta}} h(\boldsymbol{\Phi}(\lambda)\bar{\boldsymbol{\beta}}, \lambda)\| \\
&\le \|\boldsymbol{\Phi}(\lambda)^\top (\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}, \lambda) - \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\Phi}(\lambda)\bar{\boldsymbol{\beta}}, \lambda))\| && \text{(chain rule)} \\
&\le \|\boldsymbol{\Phi}(\lambda)^\top\| \cdot L\|\boldsymbol{\Phi}(\lambda)(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\| && (L\text{-smoothness}) \\
&\le CL\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\|
\end{aligned}$$

for any $\boldsymbol{\beta}$, $\bar{\boldsymbol{\beta}} \in \mathbb{R}^d$. To obtain the same conclusions for $F(\cdot)$, we may take expectation over $\lambda \sim \mathbb{P}_\lambda$ on the above inequalities, and invoke the linearity of expectations as well as the property that

$$\|\mathbb{E}\left[g(\cdot)\right]\| \;\le\; \mathbb{E}\left[\|g(\cdot)\|\right].$$

Next, we verify the strong convexity of $F(\cdot)$. Fixing $\boldsymbol{\beta}$, $\bar{\boldsymbol{\beta}} \in \mathbb{R}^d$, for any $\lambda \in \Lambda$, define $\boldsymbol{\theta}(\lambda) := \boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}$, $\bar{\boldsymbol{\theta}}(\lambda) := \boldsymbol{\Phi}(\lambda)\bar{\boldsymbol{\beta}}$. Then, by strong convexity of h, we have

$$h(\bar{\boldsymbol{\theta}}(\lambda), \lambda) \ge h(\boldsymbol{\theta}(\lambda), \lambda) + \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}(\lambda), \lambda)^\top \left(\bar{\boldsymbol{\theta}}(\lambda) - \boldsymbol{\theta}(\lambda)\right) + \frac{\mu}{2}\|\bar{\boldsymbol{\theta}}(\lambda) - \boldsymbol{\theta}(\lambda)\|^2.$$

Take expectation w.r.t. $\tilde{\lambda} \sim \mathbb{P}_\lambda$ on both sides, we obtain

$$\begin{aligned}
F(\bar{\boldsymbol{\beta}}) &= \mathbb{E}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\bar{\boldsymbol{\beta}}, \tilde{\lambda})\right] \\
&\ge \mathbb{E}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}, \tilde{\lambda})\right] + \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}(\tilde{\lambda}), \tilde{\lambda})^\top \left(\boldsymbol{\Phi}(\tilde{\lambda})(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\right] + \frac{\mu}{2}\mathbb{E}\left[\|\boldsymbol{\Phi}(\tilde{\lambda})(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2\right] \\
&= F(\boldsymbol{\beta}) + \mathbb{E}\left[\boldsymbol{\Phi}(\tilde{\lambda})^\top \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}(\tilde{\lambda}), \tilde{\lambda})\right]^\top (\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{\mu}{2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbb{E}\left[\boldsymbol{\Phi}(\tilde{\lambda})^\top \boldsymbol{\Phi}(\tilde{\lambda})\right](\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&\ge F(\boldsymbol{\beta}) + \nabla F(\boldsymbol{\beta})^\top (\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{c\mu}{2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})
\end{aligned}$$

The second term in the last inequality comes from the Leibniz integral rule and the third term invokes a well-known property of the smallest eigenvalue of a positive definite matrix. □

Proof for Theorem 2.4. First, we prove that $\epsilon^*_{\rm sp}$ is attained at some $\boldsymbol{\beta}^*_{\rm sp} \in \mathbb{R}^p$. Observe that $\epsilon_{\rm sp}(\cdot)$ is continuous. Thus, its level set $M(1) := \{\boldsymbol{\beta} \in \mathbb{R}^p : \epsilon_{\rm sp}(\boldsymbol{\beta}) \leq 1\}$ is closed. Furthermore, $M(1)$ is bounded. To see this, for any $\boldsymbol{\beta} \in M(1)$, by strong convexity of F,

$$\begin{aligned}
\frac{c\mu}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\rm avg}\|^2 \;&\leq\; F(\boldsymbol{\beta}) - F(\boldsymbol{\beta}^*_{\rm avg}) \\
&=\; \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}, \tilde{\lambda}) - h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}^*_{\rm avg}, \tilde{\lambda})\right] \\
&\leq\; \mathbb{E}_{\tilde{\lambda}\sim\mathbb{P}_\lambda}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}, \tilde{\lambda}) - h(\boldsymbol{\theta}^*(\tilde{\lambda}), \tilde{\lambda})\right] \\
&\leq\; \epsilon_{\rm sp}(\boldsymbol{\beta}) \\
&\leq\; 1.
\end{aligned} \tag{4}$$

Since $\boldsymbol{\beta}$ is arbitrary, for any $\boldsymbol{\beta}$, $\bar{\boldsymbol{\beta}} \in M(1)$,

$$\|\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}\| \;\leq\; \|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\rm avg}\| + \|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_{\rm avg}\| \;\leq\; \frac{2\sqrt{2}}{\sqrt{c\mu}}.$$

Applying the Weierstrass Theorem, there exists $\boldsymbol{\beta}^*_{\rm sp} \in M(1) \subset \mathbb{R}^p$ s.t.

$$\epsilon_{\rm sp}(\boldsymbol{\beta}^*_{\rm sp}) \;=\; \epsilon^*_{\rm sp}.$$

Next, we prove the main result of the theorem. Again, fix an arbitrary λ. By L-smoothness of $h(\cdot, \lambda)$,

$$h(\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}, \lambda) - h(\boldsymbol{\theta}^*(\lambda), \lambda) \;\leq\; \frac{L}{2}\|\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta} - \boldsymbol{\theta}^*(\lambda)\|^2 \tag{5}$$

Using the identity $(a+b+c)^2 \leq 4a^2 + 4b^2 + 4c^2$ and triangle inequality on the right-hand side, this inequality becomes

$$h(\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}, \lambda) - h(\boldsymbol{\theta}^*(\lambda), \lambda) \;\leq\; 2L(\|\boldsymbol{\Phi}(\lambda)(\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\rm avg})\|^2 + \|\boldsymbol{\Phi}(\lambda)(\boldsymbol{\beta}^*_{\rm avg} - \boldsymbol{\beta}^*_{\rm sp})\|^2 + \|\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}^*_{\rm sp} - \boldsymbol{\theta}^*(\lambda))\|^2).$$

By strong convexity of $h(\cdot, \lambda)$ and definition of $\boldsymbol{\beta}^*_{\rm sp}$,

$$\|\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}^*_{\rm sp} - \boldsymbol{\theta}^*(\lambda)\|^2 \;\leq\; \frac{2}{\mu}\sup_{\lambda\in\Lambda}\left\{h(\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}^*_{\rm sp}, \lambda) - h(\boldsymbol{\theta}^*(\lambda), \lambda)\right\} \;=\; \frac{2}{\mu}\epsilon_{\rm sp}(\boldsymbol{\beta}^*_{\rm sp}) \;=\; \frac{2}{\mu}\epsilon^*_{\rm sp},$$

So

$$\epsilon_{\rm sp}(\boldsymbol{\beta}) \;\leq\; 2CL(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\rm avg}\|^2 + \|\boldsymbol{\beta}^*_{\rm avg} - \boldsymbol{\beta}^*_{\rm sp}\|^2) + \frac{4L}{\mu}\epsilon^*_{\rm sp}. \tag{6}$$

Next, we bound $\|\boldsymbol{\beta}^*_{\rm avg} - \boldsymbol{\beta}^*_{\rm sp}\|^2$. Observe that due to the optimality of $\boldsymbol{\theta}^*(\lambda)$ for each λ,

$$\mathbb{E}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}^*_{\rm sp}, \tilde{\lambda})\right] - \mathbb{E}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}^*_{\rm avg}, \tilde{\lambda})\right] \;\leq\; \mathbb{E}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}^*_{\rm sp}, \tilde{\lambda}) - h(\boldsymbol{\theta}^*(\tilde{\lambda}), \tilde{\lambda})\right] \;\leq\; \epsilon^*_{\rm sp}.$$

On the other hand, from the optimality of $\boldsymbol{\beta}^*_{\rm avg}$ and strong-convexity of $F(\cdot)$,

$$F(\boldsymbol{\beta}^*_{\rm sp}) - F(\boldsymbol{\beta}^*_{\rm avg}) \;=\; \mathbb{E}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}^*_{\rm sp}, \tilde{\lambda})\right] - \mathbb{E}\left[h(\boldsymbol{\Phi}(\tilde{\lambda})\boldsymbol{\beta}^*_{\rm avg}, \tilde{\lambda})\right] \;\geq\; \frac{c\mu}{2}\|\boldsymbol{\beta}^*_{\rm avg} - \boldsymbol{\beta}^*_{\rm sp}\|^2.$$

Combining this inequality with the previous one shows $\frac{c\mu}{2}\|\boldsymbol{\beta}^*_{\rm avg} - \boldsymbol{\beta}^*_{\rm sp}\|^2 \;\leq\; \epsilon^*_{\rm sp}$, which implies that

$$\|\boldsymbol{\beta}^*_{\rm avg} - \boldsymbol{\beta}^*_{\rm sp}\|^2 \;\leq\; \frac{2\epsilon^*_{\rm sp}}{c\mu}.$$

Substitute the above into Equation (6). Take expectations over sample paths on both sides, we conclude that

$$\epsilon_{\rm sp}(\boldsymbol{\beta}) \;\leq\; 2CL\left(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*_{\rm avg}\|^2 + \frac{2\epsilon^*_{\rm sp}}{c\mu}\right) + \frac{4L}{\mu}\epsilon^*_{\rm sp}.$$

Rearranging and using the fact that $C/c \geq 1$ verifies the statement. □

A.2 Proofs from Section 3

Proof for Theorem 3.2. The proof follows the proof of Theorem 3.1 of Gower et al. (2019); we include it for completeness under the assumption of the RWGC Definition 3.1. Let $\eta := \frac{\bar{\eta}}{\rho}$ denote the step-size, $\boldsymbol{w}^* := \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} G(\boldsymbol{w})$ denote the optimality input of G. For $t \geq 0$, we have

$$\begin{aligned}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 &= \|\boldsymbol{w}_t - \boldsymbol{w}^* - \eta\nabla g(\boldsymbol{w}_t, z_t)\|^2 \\ &= \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta\nabla g(\boldsymbol{w}_t, z_t)^\top(\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2\|\nabla g(\boldsymbol{w}_t, z_t)\|^2.\end{aligned}$$

Taking conditional expectations yields

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{w}_{t+1} - \boldsymbol{w}^*\|^2 \,\middle|\, \boldsymbol{w}_t\right] &= \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta\nabla G(\boldsymbol{w}_t)^\top(\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2\mathbb{E}\left[\|\nabla g(\boldsymbol{w}_t, z_t)\|^2 \,\middle|\, \boldsymbol{w}_t\right] \\ &\leq \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 - 2\eta\nabla G(\boldsymbol{w}_t)^\top(\boldsymbol{w}_t - \boldsymbol{w}^*) + \eta^2[2\rho(G(\boldsymbol{w}_t) - G^*) + \sigma^2] \\ &\leq \|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 + 2\eta\left(G^* - G(\boldsymbol{w}_t) - \tfrac{\mu_g}{2}\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2\right) + \eta^2[2\rho(G(\boldsymbol{w}_t) - G^*) + \sigma^2] \\ &= (1 - \eta\mu_g)\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 + 2\left(\eta^2\rho - \eta\right)(G(\boldsymbol{w}_t) - G^*) + \eta^2\sigma^2 \\ &\leq (1 - \eta\mu_g)\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2 + \eta^2\sigma^2 \qquad (7)\end{aligned}$$

The first inequality above uses the RWGC and the second uses strong convexity. The final inequality holds since $\eta^2\rho - \eta = \left(\frac{\bar{\eta}}{\rho}\right)^2\rho - \frac{\bar{\eta}}{\rho} = \left(\frac{\bar{\eta}}{\rho}\right)(\bar{\eta} - 1)$ and $\bar{\eta} \leq 1$. Furthermore, since $\bar{\eta} \leq \frac{\rho}{\mu_g}$, we have $\eta\mu_g = \left(\frac{\bar{\eta}}{\rho}\right)\mu_g \leq 1$; therefore, by iterating Equation (7) and taking overall expectations, we get

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{w}_t - \boldsymbol{w}^*\|^2\right] &\leq (1 - \eta\mu_g)^t\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + \eta^2\sigma^2\sum_{k=0}^{t-1}(1 - \eta\mu_g)^k \\ &\leq (1 - \eta\mu_g)^t\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + \frac{\eta^2\sigma^2}{1 - (1 - \eta\mu_g)} \\ &= (1 - \eta\mu_g)^t\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|^2 + \frac{\eta\sigma^2}{\mu_g},\end{aligned}$$

where the second inequality uses the geometric series bound. Recalling that $\eta = \frac{\bar{\eta}}{\rho}$ completes the proof.

□

Proof for Lemma 3.3. Recall that Proposition A.1 already establishes the smoothness of f and smoothness and strong convexity of F.

Define

$$\bar{\boldsymbol{\beta}}(\lambda) \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} f(\boldsymbol{\beta}, \lambda) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} h(\boldsymbol{\Phi}(\lambda)\boldsymbol{\beta}, \lambda).$$

By CL-smoothness, we have

$$f(\boldsymbol{\beta}, \lambda) \geq f(\bar{\boldsymbol{\beta}}(\lambda), \lambda) + \nabla f(\bar{\boldsymbol{\beta}}(\lambda), \lambda)^\top(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}(\lambda)) + \frac{1}{2CL}\|\nabla f(\boldsymbol{\beta}, \lambda) - \nabla f(\bar{\boldsymbol{\beta}}(\lambda), \lambda)\|^2.$$

Using $\nabla f(\bar{\boldsymbol{\beta}}(\lambda), \lambda) = 0$ for all λ, rearranging and recalling the definitions of $\boldsymbol{\beta}^*_{\text{avg}}$ (Section 2) and $\boldsymbol{\beta}^*_{\text{sp}}$ (proof of Theorem 2.4) yields

$$\begin{aligned}\frac{1}{2CL}\|\nabla f(\boldsymbol{\beta}, \lambda)\|^2 &\leq f(\boldsymbol{\beta}, \lambda) - f(\bar{\boldsymbol{\beta}}(\lambda), \lambda) \\ &\leq f(\boldsymbol{\beta}, \lambda) - h(\boldsymbol{\theta}^*(\lambda), \lambda) \\ &= (f(\boldsymbol{\beta}, \lambda) - f(\boldsymbol{\beta}^*_{\text{avg}}, \lambda)) + (f(\boldsymbol{\beta}^*_{\text{avg}}, \lambda) - f(\boldsymbol{\beta}^*_{\text{sp}}, \lambda)) + (f(\boldsymbol{\beta}^*_{\text{sp}}, \lambda) - h(\boldsymbol{\theta}^*(\lambda), \lambda)) \\ &\leq (f(\boldsymbol{\beta}, \lambda) - f(\boldsymbol{\beta}^*_{\text{avg}}, \lambda)) + (f(\boldsymbol{\beta}^*_{\text{avg}}, \lambda) - f(\boldsymbol{\beta}^*_{\text{sp}}, \lambda)) + \epsilon^*_{\text{sp}}\end{aligned}$$

where the second inequality uses optimality of $\boldsymbol{\theta}^*(\lambda)$ and the third uses that ϵ^*_{sp} is attained at $\boldsymbol{\beta}^*_{\text{sp}}$. Taking expectations on both sides, we get

$$\frac{1}{2CL}\mathbb{E}\left[\|\nabla f(\boldsymbol{\beta}, \tilde{\lambda})\|^2\right] \leq F(\boldsymbol{\beta}) - F(\boldsymbol{\beta}^*_{\text{avg}}) + F(\boldsymbol{\beta}^*_{\text{avg}}) - F(\boldsymbol{\beta}^*_{\text{sp}}) + \epsilon^*_{\text{sp}} \leq F(\boldsymbol{\beta}) - F(\boldsymbol{\beta}^*_{\text{avg}}) + \epsilon^*_{\text{sp}},$$

where we used that the definition of $\boldsymbol{\beta}^*_{\text{avg}}$ implies that $F(\boldsymbol{\beta}^*_{\text{avg}}) \leq F(\boldsymbol{\beta}^*_{\text{sp}})$. Multiplying through by $2CL$ completes the proof. □

Proof for Theorem 3.4. Our strategy will be to invoke Lemma 3.3 and apply Theorem 3.2, in conjunction with Theorem 2.4. To that end, notice that the upper bound on the step-size parameter $\bar{\eta}$ given in Theorem 3.2 becomes $\frac{CL}{c\mu} \geq 1$. Hence, we need only constrain $0 < \bar{\eta} \leq 1$.

Applying Theorem 3.2 yields

$$\mathbb{E}\left[\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*_{\text{avg}}\|^2\right] \;\leq\; \left(1 - \bar{\eta}\frac{c\mu}{CL}\right)^T \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2 + \frac{2\bar{\eta}\epsilon^*_{\text{sp}}}{c\mu}.$$

Taking the expectation of Theorem 2.4 on both sides and substituting the above, we have

$$\begin{aligned}\mathbb{E}\left[\epsilon_{\text{sp}}(\boldsymbol{\beta}_T)\right] \;&\leq\; 2CL\mathbb{E}\left[\|\boldsymbol{\beta}_T - \boldsymbol{\beta}^*_{\text{avg}}\|^2\right] + \frac{8CL}{c\mu}\epsilon^*_{\text{sp}} \\ &\leq\; 2CL\left(1 - \bar{\eta}\frac{c\mu}{CL}\right)^T \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2 + 4\frac{\bar{\eta}CL}{c\mu}\epsilon^*_{\text{sp}} + \frac{8CL}{c\mu}\epsilon^*_{\text{sp}}. \qquad (8)\end{aligned}$$

Rearranging yields the first result. Next, we prove the two bounds of the iteration complexity.

First, consider the case $\epsilon^*_{\text{sp}} > 0$. Our goal will be to choose T large enough to drive the first term on the right side of (8) to below $4\frac{CL}{c\mu}\bar{\eta}\epsilon^*_{\text{sp}}$. To simplify exposition, let $\kappa = \frac{CL}{c\mu}$. Then, we need to ensure that

$$T\log(1 - \bar{\eta}/\kappa) \leq \log\left(\frac{2\kappa\bar{\eta}\epsilon^*_{\text{sp}}}{CL\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2}\right).$$

We can upper bound the right side using the identity $\log(1 + x) \leq x$. Hence, it would be sufficient if

$$T \geq \frac{\kappa}{\bar{\eta}}\log\left(\frac{CL\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2}{2\kappa\bar{\eta}\epsilon^*_{\text{sp}}}\right).$$

Replacing κ with its definition yields the first case.

For the second case where $\epsilon^*_{\text{sp}} = 0$, (8) reduces to

$$\mathbb{E}\left[\epsilon_{\text{sp}}(\boldsymbol{\beta}_T)\right] \;\leq\; 2CL\left(1 - \bar{\eta}\frac{c\mu}{CL}\right)^T \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2.$$

Given $\epsilon > 0$, we need to ensure that

$$T\log(1 - \bar{\eta}/\kappa) \leq \log\left(\frac{\epsilon}{2CL\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2}\right).$$

Again using $\log(1 + x) \leq x$, it would be sufficient if

$$T \geq \frac{\kappa}{\bar{\eta}}\log\left(\frac{2CL\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*_{\text{avg}}\|^2}{\epsilon}\right).$$

Replacing κ with its definition yields the second case.

□

A.3 Proofs from Section 4

Proof of Lemma 4.1. As described in the main text, $\boldsymbol{\Phi}(\lambda) \in \mathbb{R}^{d \times qd}$ is block-diagonal with d blocks of size $1 \times q$. Denote this block by $\psi(\lambda) \in \mathbb{R}^{1 \times q}$ and note the elements of $\psi(\lambda)$ are precisely the first q Legendre polynomials evaluated at λ.

Now the matrix $\Phi(\lambda)^\top \Phi(\lambda) \in \mathbb{R}^{qd \times qd}$ is also block diagonal with d copies of the matrix $\psi(\lambda)^\top \psi(\lambda)$. As a consequence, the eigenvectors of $\Phi(\lambda)^\top \Phi(\lambda)$ are the stacked copies of the eigenvectors of $\psi(\lambda)^\top \psi(\lambda)$ with the same eigenvalues.

Since $\psi(\lambda)^\top \psi(\lambda)$ is a rank-one matrix, it has at most one non-zero eigenvalue, and by inspection, this eigenvalue is $\|\psi(\lambda)\|^2$ with eigenvector $\psi(\lambda)^\top$. Then,

$$C_p = \sup_{\lambda \in [-1,1]} \|\psi(\lambda)\|^2 = q,$$

because the Legendre polynomials achieve their maxima at 1 with a value of 1.

By a nearly identical argument, we can see that

$$\begin{aligned} c_p &= \sigma_{\min} \left\{ \mathbb{E}\left[\psi(\tilde{\lambda})^\top \psi(\tilde{\lambda}) \right] \right\} \\ &= \min_{n=0,\ldots,q-1} \frac{2}{2n+1} \\ &= \frac{2}{2q-1}, \end{aligned}$$

because by the orthogonality of the Legendre polynomials, the above matrix is diagonal. Hence

$$C_p / c_p \;=\; \frac{2q^2 - q}{2} \leq q^2.$$

□

Proof for Lemma 4.2. Our proof is constructive and we will show a slightly stronger result. We will approximate $\boldsymbol{\theta}_i^*(\lambda)$ by its Chebyshev truncation up to degree q for each i. Letting $\bar{\boldsymbol{\beta}}$ be the coefficients corresponding to the resulting polynomials, we will show that $\epsilon_{\text{sp}}(\bar{\boldsymbol{\beta}})$ satisfies the bound in Lemma 4.2, which implies that ϵ_{sp}^* also satisfies the bound.

Let $T_n(\lambda)$ denote the n^{th} Chebyshev polynomial of the first kind. For the i^{th} dimension of the solution path, let $a_{i,n}$ denote the coefficient of T_n in the Chebyshev truncation of $\boldsymbol{\theta}_i^*$ up to degree q.

We assumed that $\boldsymbol{\theta}_i^*(\lambda)$, ..., $\boldsymbol{\theta}_i^{(\nu-1)*}(\lambda)$ are absolutely continuous and $\boldsymbol{\theta}_i^{(\nu)*}(\lambda)$ has bounded variation V. When $\nu = 0$, we are assuming simply that $\boldsymbol{\theta}_i^*(\lambda)$ is of bounded variation V. Thus, for any $q \geq \nu + 1$ and $i \in [d]$, Trefethen (2013, Theorem 7.2) guarantees that

$$\sup_{\lambda \in [-1,1]} \left| \sum_{n=0}^{q} a_{i,n} T_n(\lambda) - \boldsymbol{\theta}_i^*(\lambda) \right| \;\leq\; \frac{2V}{\pi \nu (q-\nu)^\nu}.$$

Then,

$$\epsilon_{\text{sp}}(\bar{\boldsymbol{\beta}}) \;=\; \sup_{\lambda \in \Lambda} \left\{ h\left(\left(\sum_{n=0}^{q} a_{i,n} T_n(\lambda) \right)_{i=[d]}, \lambda \right) - h(\boldsymbol{\theta}^*(\lambda), \lambda) \right\}.$$

Moreover, smoothness of $h(\cdot, \lambda)$ ensures that for any $\lambda \in \Lambda$,

$$h\left(\left(\sum_{n=0}^{q} a_{i,n} T_n(\lambda) \right)_{i=[d]}, \lambda \right) - h(\boldsymbol{\theta}^*(\lambda), \lambda) \;\leq\; \frac{L}{2} \left\| \left(\sum_{n=0}^{q} a_{i,n} T_n(\lambda) \right)_{i=[d]} - \boldsymbol{\theta}^*(\lambda) \right\|^2.$$

Combine the above and using the definition of l_2-norm, we deduce that

$$\begin{aligned}\epsilon_{\text{sp}}(\bar{\boldsymbol{\beta}}) &\leq \sup_{\lambda\in\Lambda}\left\{\frac{L}{2}\left\|\left(\sum_{n=0}^{q}a_{i,n}T_n(\lambda)\right)_{i=[d]}-\boldsymbol{\theta}^*(\lambda)\right\|^2\right\}\\ &\leq \sup_{\lambda\in[-1,1]}\left\{\frac{dL}{2}\cdot\sup_{i\in[d]}\left|\sum_{n=0}^{q}a_{i,n}T_n(\lambda)-\boldsymbol{\theta}_i^*(\lambda)\right|^2\right\}\\ &\leq \frac{dL}{2}\left(\frac{2V}{\pi\nu(q-\nu)^\nu}\right)^2.\end{aligned}$$

We note that this result is presented and holds for any $\nu \geq 0$. (When $\nu = 0$, the bound is infinity and hence trivially valid.) As will be seen in the proof of Theorem 4.3 below, however, we will only apply the result when $\nu \geq 2$.

□

Proof of Theorem 4.3. As we are only interested in asymptotic behavior as $\epsilon \to 0$, we will often suppress any constants that do not depend on ϵ below. In particular, we will write $a \lesssim b$ whenever there exists a constant C (not depending on ϵ but perhaps depending on h and Λ) such that $a \leq Cb$.

Note that in order to attain a small solution path error, we will need to use a large number of polynomials q. Based on Theorem 3.4, our first goal is to choose q large enough that

$$\epsilon \;\geq\; \frac{8L(\bar{\eta}+1)}{\mu}\frac{C_q}{c_q}\cdot\epsilon_{\text{sp}}^*(q). \tag{9}$$

First observe that Lemma 4.2 establishes that the solution path error is at most

$$\epsilon_{\text{sp}}^*(q) \;\leq\; \frac{4dLV^2}{2\pi^2\nu^2(q-\nu)^{2\nu}} \;\lesssim q^{-2\nu}. \tag{10}$$

Using Lemma 4.1, we can upper bound the right side of Equation (9) by

$$\frac{8L(\bar{\eta}+1)}{\mu}\frac{C_q}{c_q}\cdot\epsilon_{\text{sp}}^*(q) \lesssim q^{2(1-\nu)}.$$

Hence it suffices to take

$$\epsilon^{-\frac{1}{2(\nu-1)}} \lesssim q$$

polynomials to achieve our target error.

We now seek to bound the iteration count. By Theorem 3.4, the iteration count should exceed

$$\frac{L}{\bar{\eta}\mu}\frac{C_q}{c_q}\log\left(\frac{c_q\mu\|\boldsymbol{\beta}_{\text{avg}}^*\|^2}{2\bar{\eta}\epsilon_{\text{sp}}^*}\right) \;\lesssim\; q^2\left(\log\|\boldsymbol{\beta}_{\text{avg}}^*\|^2+\log(q)\right)$$

where we have assumed SGD was initialized at $\boldsymbol{\beta} = 0$ and used Equation (10) and $c_q = \frac{2}{2q-1}$ (cf. the proof of Lemma 4.1) to simplify.

To complete the proof we need to bound $\|\boldsymbol{\beta}_{\text{avg}}^*\|^2$. To this end, we first bound $\|\bar{\boldsymbol{\beta}}\|$, where $\bar{\boldsymbol{\beta}}$ is constructed by approximating each component $\theta_i^*(\lambda)$ by its Chebyshev truncation to degree q for each i.

Then,

$$\begin{aligned}\|\bar{\boldsymbol{\beta}}\|^2 &= \sum_{i=1}^{d}\sum_{k=1}^{q}|a_{ik}|^2 \\ &= \sum_{i=1}^{d}\sum_{k=1}^{\nu}|a_{ik}|^2 + \sum_{i=1}^{d}\sum_{k=\nu+1}^{q}|a_{ik}|^2. \\ &\lesssim \sum_{k=\nu+1}^{q}\frac{1}{k^{2(\nu+1)}}, \\ &\lesssim \sum_{k=\nu+1}^{\infty}\frac{1}{k^{2(\nu+1)}},\end{aligned}$$

where the penultimate inequality collects constants and invokes Trefethen (2013, Theorem 7.1) to bound the second summation. Note, for $\nu \geq 0$, this last sum is summable, so that $\|\bar{\boldsymbol{\beta}}\| \lesssim 1$.

Furthermore, from the proof of Lemma 4.2, $\epsilon_{\text{sp}}(\bar{\boldsymbol{\beta}}) \lesssim q^{-2\nu}$. Hence, since the solution-path error upper bounds the stochastic error, we also have $F(\bar{\boldsymbol{\beta}}) - F(\boldsymbol{\beta}^*_{\text{avg}}) \lesssim q^{-2\nu}$. From strong convexity, this implies that $\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_{\text{avg}}\| \lesssim q^{-\nu}$.

Putting it together, we have that

$$\|\boldsymbol{\beta}^*_{\text{avg}}\| \;\leq\; \|\boldsymbol{\beta}^*_{\text{avg}} - \bar{\boldsymbol{\beta}}\| + \|\bar{\boldsymbol{\beta}}\| \;\lesssim\; q^{-\nu} + 1 \;\lesssim\; 1.$$

Substituting above shows that it suffices to take $O(q^2 \log q)$ iterations as $\epsilon \to 0$. Given our previous calculation of q, this amount to $O(\bar{\epsilon}^{\frac{1}{1-\nu}} \log(1/\bar{\epsilon}))$ iterations.

□

Proof for Lemma 4.4. The proof is very similar to the proof of Lemma 4.2, with the only difference being that we use Theorem 8.2 from (Trefethen, 2013) instead of Theorem 7.2.

Recall that a Bernstein ellipse $\mathcal{E}_\omega$ with radius ω in the complex plain is the ellipse

$$\mathcal{E}_\omega = \left\{ z = \frac{1}{2}\left(\omega e^{\theta\sqrt{-1}} + \omega^{-1} e^{-\theta\sqrt{-1}}\right) : \theta \in [0, 2\pi) \right\}.$$

Since $\theta^*_i(\lambda)$ is analytic, there exists an analytic continuation of $\theta^*_i(\lambda)$ to a Bernstein ellipse of radius $\omega_i > 1$. The value of ω_i generally depends on if $\theta^*_i(\lambda)$ has any singularities in the complex plane, but is guaranteed to be larger than 1. Let M_i be $\max_{z \in \mathcal{E}_{\omega_i}} |\theta^*_i(z)|$. Finally, let $\omega \equiv \min_{i=1,\ldots d} \omega_i > 1$ and $M = \max_{i=1,\ldots,d} M_i < \infty$.

Now, define $T_n(\lambda)$ and $a_{i,n}$ as in the proof for Lemma 4.2.

For any $q > 0$ and $i \in [d]$, Theorem 8.2 in (Trefethen, 2013) guarantees that

$$\sup_{\lambda \in [-1,1]} \left| \sum_{n=0}^{q} a_{i,n} T_n(\lambda) - \boldsymbol{\theta}^*_i(\lambda) \right| \;\leq\; \frac{2M\omega^{-q}}{\omega - 1}.$$

Plug this result into the last inequality of the proof for Lemma 4.2, we obtain

$$\epsilon_{\text{sp}}(\bar{\boldsymbol{\beta}}) \;\leq\; \sup_{\lambda \in [-1,1]} \left\{ \frac{dL}{2} \cdot \sup_{i \in [d]} \left| \sum_{n=0}^{q} a_{i,n} T_n(\lambda) - \boldsymbol{\theta}^*_i(\lambda) \right|^2 \right\} \;\leq\; \frac{dL}{2}\left(\frac{2M\omega^{-q}}{\omega - 1}\right)^2.$$

□

Proof of Theorem 4.5. The proof is quite similar to the proof of Theorem 4.3. Again, we only consider asymptotic behavior as $\epsilon \to 0$ and suppress all other constants. Hence, $a \lesssim b$ means there exists a constant C (not depending on ϵ) such that $a \leq Cb$.

Again, our first goal is to identify a q sufficiently large to achieve our target error. By Theorem 3.4 we seek q large enough that

$$\epsilon \geq \frac{8L(\bar{\eta}+1)}{\mu}\frac{C_q}{c_q}\epsilon^*_{\text{sp}}(q).$$

Lemma 4.4 shows that there exists an $\omega > 1$ such that

$$\epsilon^*_{\text{sp}}(q) \lesssim \omega^{-2q},$$

hence it suffice to take q large enough that

$$q^2\omega^{-2q} \lesssim \bar{\epsilon}.$$

Solving this equation exactly requires the Lambert-W function. Instead, we take $q = \frac{\log(1/\bar{\epsilon})}{\log \omega}$. Then,

$$q^2\omega^{-2q} = \frac{\epsilon^2 \log^2\left(\frac{1}{\epsilon}\right)}{\log^2(\omega)} \lesssim \bar{\epsilon}$$

for $\bar{\epsilon}$ sufficiently small.

Again, to bound the number of iterations, we will need to bound $\|\boldsymbol{\beta}^*_{\text{avg}} - \boldsymbol{\beta}_0\|$. We again will assume that $\boldsymbol{\beta}_0 = \mathbf{0}$, and first consider bounding $\bar{\boldsymbol{\beta}}$. Recall, $\bar{\boldsymbol{\beta}}$ is obtained by approximating each component i by the Chebyshev truncation of $\boldsymbol{\theta}^*(\lambda)$ to degree q. Then,

$$\begin{aligned}
\|\bar{\boldsymbol{\beta}}\|^2 &= \sum_{i=1}^{d}\sum_{k=1}^{q}|a_{ik}|^2 \\
&\lesssim \sum_{k=1}^{q}\omega^{-k} \\
&\leq \sum_{k=1}^{\infty}\omega^{-k}.
\end{aligned}$$

Here, the second to last inequality uses (Trefethen, 2013, Theorem 8.1). Since the last summation is summable, we again conclude that $\|\bar{\boldsymbol{\beta}}\|_2 \lesssim 1$.

The proof of Lemma 4.4 establishes that $\epsilon_{\text{sp}}(\bar{\boldsymbol{\beta}}) \lesssim \omega^{-2q}$. Since solution path error upper-bounds the stochastic error, we conclude that $F(\bar{\boldsymbol{\beta}}) - F(\boldsymbol{\beta}^*_{\text{avg}}) \lesssim \omega^{-2q}$, and by strong convexity, $\|\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_{\text{avg}}\| \lesssim \omega^{-q}$.

Putting it together, we have that

$$\|\boldsymbol{\beta}^*_{\text{avg}}\| \leq \|\boldsymbol{\beta}^*_{\text{avg}} - \bar{\boldsymbol{\beta}}\| + \|\bar{\boldsymbol{\beta}}\| \lesssim \omega^{-q} + 1 \lesssim 1.$$

Substituting into the iteration complexity shows that it suffices to take

$$q^2\left(\log\|\boldsymbol{\beta}^*_{\text{avg}}\| + \log q\right) \lesssim \log^2(1/\bar{\epsilon})\log\log(1/\bar{\epsilon})$$

iterations, by using our condition on q. □

B Implementation Details

B.1 Calibrating Uniform Discretization

To apply uniform discretization in practice one must decide i) the number of grid points to use, and ii) the number of gradient calls to make at each grid point. Loosely speaking, our approach to these two decisions is to first fix a set of desired "target" solution path errors denoted Δ. Then, for each $\delta \in \Delta$, we determine the number of grid points and gradient calls to approximately achieve a solution path error of δ.

More specifically, Ndiaye et al. (2019) suggests that to achieve a solution path error of δ, we require $O(1/\sqrt{\delta})$ grid points. Thus, for each $\delta \in \Delta$, we construct a uniform discretization with $1/\sqrt{\delta}$ grid points across every

Weighted Binary Classification	$c_1 = 1$	$c_2 = 0.5$	$\Delta = \{2^{-6}, 2^{-6.5}, 2^{-7}, \ldots, 2^{-18}\}$
Portfolio Allocation	$c_1 = 0.65$	$c_2 = 1$	$\Delta = \{4^{-2}, 5^{-2}, \ldots, 19^{-2}\}$

Table 1: **Parameters Used in Experiments.**

dimension of the hyperparameter space. We denote this grid by $G(\delta)$. Recall that deterministic gradient descent requires $O(\log(1/\delta))$ steps to achieve an error of $O(\delta)$. Motivated by this result, we use $c_1 \log(c_2/\delta)$ gradient calls per grid point. The total number of gradient calls is thus $c_1 \log(c_2/\delta)/\sqrt{\delta}$.

The specific values of c_1, c_2, and Δ used in our experiments are recorded in Table 1.

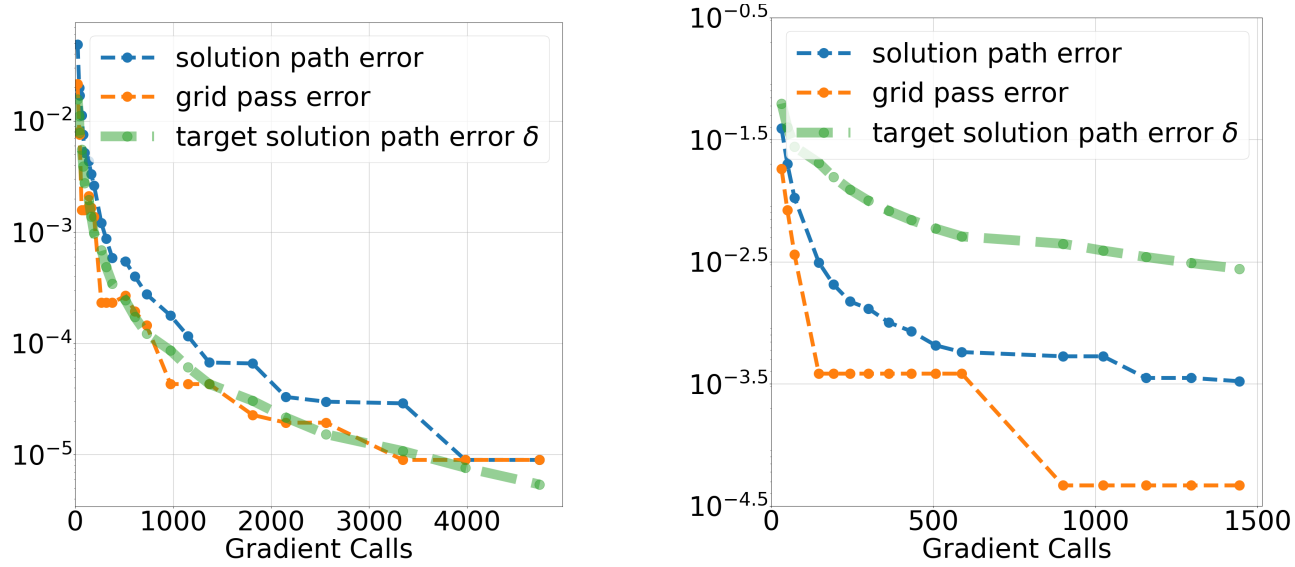
We next argue that this procedure with these constants is roughly efficient. Define the *grid pass error* as

$$\sup_{\lambda \in G(\delta)} h(\hat{\boldsymbol{\theta}}(\lambda), \lambda) - h(\boldsymbol{\theta}^*(\lambda), \lambda).$$

For comparison, the solution-path error is

$$\epsilon_{\text{sp}}(\hat{\boldsymbol{\theta}}(\cdot)) = \sup_{\lambda \in \Lambda} h(\hat{\boldsymbol{\theta}}(\lambda), \lambda) - h(\boldsymbol{\theta}^*(\lambda), \lambda).$$

Since $G(\delta) \subseteq \Lambda$, the grid pass error is always less than the solution path error.



(a) (b)

Figure 5: **Compare: Ideal Stopping Criterion, Grid Pass Error, and Solution Path Error.** (a) weighted binary classification; (b) portfolio allocation. Solution path error and grid pass error are plotted against the corresponding total number of gradient calls incurred in practice. Target solution path error δ is plotted against $c_1 \log(c_2/\delta)/\sqrt{\delta}$, which also equals the total number of gradient calls incurred in practice. Each discrete point in the line plots above corresponds to a discretization grid point $\delta \in G(\delta)$; thus smaller δ is associated with more gradient calls.

By comparing the grid pass error, the solution path error, and the target solution path error, we can calibrate the amount of work done at each grid point and the total number of grid points. Specifically, if the grid pass error is much smaller than the solution pass error, it suggests we have allocated too much work to gradient calls at grid points and have insufficient grid points. On the other hand, if the grid pass error is very far from the target solution path error, it suggests we have not allocated enough work to gradient calls at the grid points and have too many grid points. Based on this trade-off, we tune the constants c_1 and c_2.

Figure 5 compares the solution path error and grid pass error to the target solution path error for our experiments. Figure 5a illustrates that in the weighted binary classification experiment, the discretization scheme performs

well, satisfying both objectives as the solution path error and grid pass error closely align with the target solution path error. Figure 5b demonstrates that for the portfolio allocation problem, both solution path error and grid pass error are still close to each other, but fall below the target solution path error. This is primarily due to the high precision of L-BFGS, though only very few gradient calls are performed at each grid.

B.2 Basis for Moderate Dimensional Portfolio Allocation

Recall that this experiment uses the following objective function:

$$h(\theta, \lambda) \;=\; -\lambda_1 \cdot \mu^\top \theta + \lambda_2 \cdot \theta^\top \Sigma \theta + \|\theta - \lambda_{3:12}\|_2^2,$$

where $\lambda_{3:12}$ represents the current holdings.

Since this is a quadratic objective, let us first directly compute the optimal solution path

$$\boldsymbol{\theta}^*(\lambda) \;=\; \frac{1}{2}\left(\boldsymbol{I} + \lambda_2 \Sigma\right)^{-1}\left(\lambda_1 \mu + 2\lambda_{3:12}\right).$$

Consider the eigendecomposition of Σ, which takes the form $\Sigma = P^\top D P$ for some orthonormal matrix P and diagonal matrix $D = \mathrm{diag}(d_1, \ldots, d_{10})$. Plugging into the above closed form optimal solution path yields

$$\boldsymbol{\theta}^*(\lambda) \;=\; \frac{1}{2} P^\top \mathrm{diag}\left(\frac{1}{1+\lambda_2 d_1}, \ldots, \frac{1}{1+\lambda_2 d_{10}}\right) P\left(\lambda_1 \mu + 2\lambda_{3:12}\right).$$

Using the Taylor series expansion for $\frac{1}{1+a\lambda_2}$ around 0

$$\frac{1}{1+a\lambda_2} = \sum_{i=0}^{\infty} (-1)^i (a\lambda_2)^i, \qquad a \in \mathbb{R},$$

we conclude that, as $q \to \infty$, each component of the optimal solution path $\boldsymbol{\theta}_i^*(\lambda)$ is in the span of our chosen basis:

$$\lambda_{(j \bmod 12)} \cdot \lambda_2^{\lfloor j/12 \rfloor}, \qquad j = 1, ..., q.$$

C Additional Figures

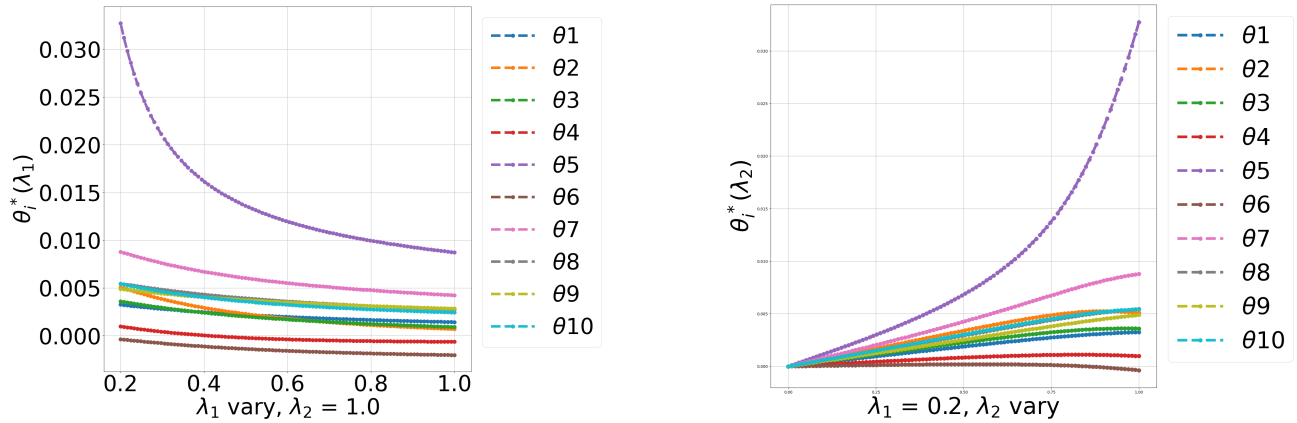


Figure 6: $\boldsymbol{\theta}^*(\lambda)$ **for Portfolio Allocation from Section 6.2.** The true solution paths are very smooth in each dimension, suggesting that this problem is highly interpolable by a polynomial basis.

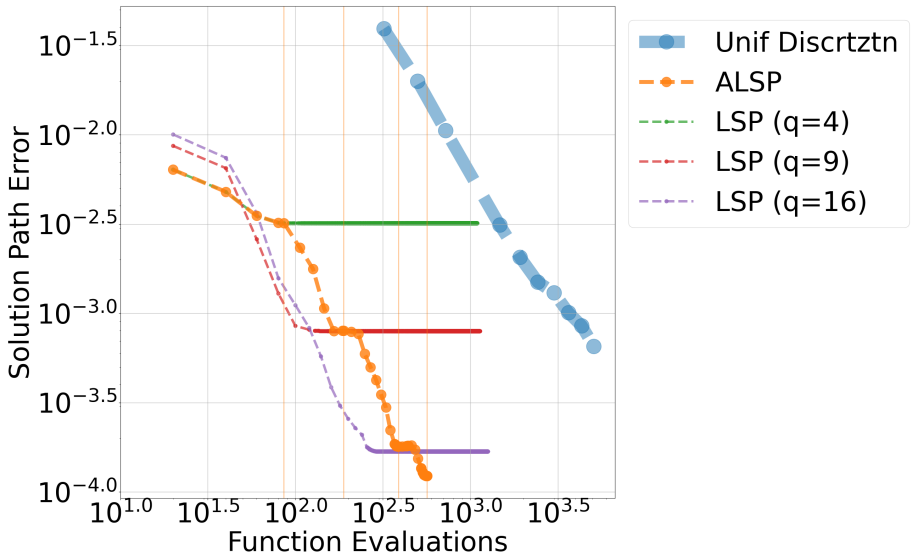


Figure 7: **LSP and ALSP for Portfolio Allocation.** Compares methods using the first $q = 4, 9, 16$ bivariate-Legendre polynomials. Differs from Figure 3 as the y-axis records the number of function evaluations during line-search of L-BFGS instead of gradient calls.

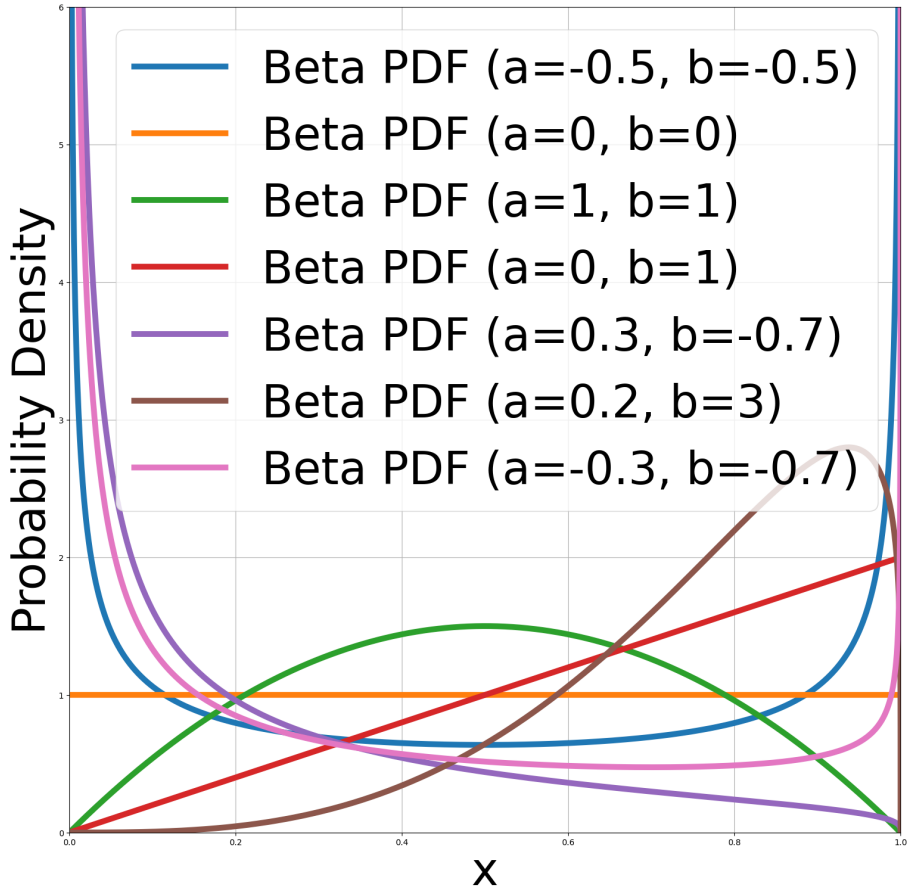


Figure 8: **Probability Density Function(PDF) of Beta Distribution over** $[0, 1]$ **with Various** a**,** b**.**