Data-Pooling in Stochastic Optimization

Vishal Gupta
Data Science and Operations, USC Marshall School of Business, Los Angles, CA 90089,
guptavis@usc.edu

Nathan Kallus
School of Operations Research and Information Engineering and Cornell Tech, Cornell University, New York, NY 10044,
kallus@cornell.edu

Managing large-scale systems often involves simultaneously solving thousands of unrelated stochastic optimization problems, each with limited data. Intuition suggests one can decouple these unrelated problems and solve them separately without loss of generality. We propose a novel data-pooling algorithm called Shrunken-SAA that disproves this intuition. In particular, we prove that combining data across problems can outperform decoupling, even when there is no a priori structure linking the problems and data are drawn independently. Our approach does not require strong distributional assumptions and applies to constrained, possibly non-convex, non-smooth optimization problems such as vehicle-routing, economic lot-sizing or facility location. We compare and contrast our results to a similar phenomenon in statistics (Stein's Phenomenon), highlighting unique features that arise in the optimization setting that are not present in estimation. We further prove that as the number of problems grows large, Shrunken-SAA learns *if* pooling can improve upon decoupling *and* the optimal amount to pool, even if the average amount of data per problem is fixed and bounded. Importantly, we highlight a simple intuition based on stability that highlights *when* and *why* data-pooling offers a benefit, elucidating this perhaps surprising phenomenon. This intuition further suggests that data-pooling offers the most benefits when there are many problems, each of which has a small amount of relevant data. Finally, we demonstrate the practical benefits of data-pooling using real data from a chain of retail drug stores in the context of inventory management.

Key words: Data-driven optimization. Small-data, large-scale regime. Shrinkage. James-Stein Estimation.
History: This paper was first submitted in May 2019. A revision was submitted in August 2020.

1. Introduction

The stochastic optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \quad \mathbb{E}^{\mathbb{P}}[c(\boldsymbol{x},\boldsymbol{\xi})] \tag{1.1}$$

is a fundamental model with applications ranging from inventory management to personalized medicine. In typical data-driven settings, the measure $\mathbb{P}$ governing the random variable $\boldsymbol{\xi}$ is unknown. Instead, we have access to a dataset $\mathcal{S}=\{\hat{\boldsymbol{\xi}}_1,\ldots,\hat{\boldsymbol{\xi}}_N\}$ drawn i.i.d. from $\mathbb{P}$ and seek a decision $\boldsymbol{x}\in\mathcal{X}$ depending on these data. This model and its data-driven variant have been extensively studied in the literature (see Shapiro et al. 2009 for an overview).

Managing real-world, large-scale systems, however, frequently involves solving thousands of potentially unrelated stochastic optimization problems like Problem (1.1) simultaneously. For example, inventory management often requires optimizing stocking levels for many distinct products across categories, not just a single product. Firms typically determine staffing and capacity for many warehouses and fulfillment centers across the supply-chain, not just at a single location. Logistics companies often divide large territories into many small regions and solve separate vehicle routing problems, one for each region, rather than solving a single monolithic problem. In such applications, a more natural model than Problem (1.1) might be

$$\frac{1}{K}\sum_{k=1}^{K}\frac{\lambda_k}{\lambda_{\text{avg}}}\min_{\boldsymbol{x}_k\in\mathcal{X}_k}\ \mathbb{E}^{\mathbb{P}_k}[c_k(\boldsymbol{x}_k,\boldsymbol{\xi}_k)], \tag{1.2}$$

where we solve a separate subproblem of the form (1.1) for each k, e.g., setting a stocking level for each product. Here, $\lambda_k > 0$ represents the frequency with which the decision-maker incurs costs from problems of type k, and $\lambda_{\text{avg}} = \frac{1}{K}\sum_{k=1}^{K}\lambda_k$. Thus, this formulation captures the fact that our total costs in such systems are driven by the frequency-weighted average of the costs of many distinct optimization problems.

Of course, intuition strongly suggests that since there are no coupling constraints across the feasible regions $\mathcal{X}_k$ in Problem (1.2), one can and should decouple the problem into K unrelated subproblems and solve them separately. Indeed, when the measures $\mathbb{P}_k$ are known, this procedure is optimal. When the $\mathbb{P}_k$ are unknown and unrelated, but one has access to a dataset $\mathcal{S}_k = \{\hat{\boldsymbol{\xi}}_{k,1},\ldots,\hat{\boldsymbol{\xi}}_{k,\hat{N}_k}\}$ drawn i.i.d. from $\mathbb{P}_k$ independently across k, intuition *still* suggests decoupling is without loss of generality and that data-driven procedures can be applied separately by subproblem.

A key message of this paper is that this intuition is <u>false</u>.

In the data-driven setting, when solving many stochastic optimization problems, we show there exist algorithms which pool data across sub-problems that outperform decoupling *even* when the underlying problems are unrelated, and data are *independent.* This phenomenon holds despite the fact that the k^{th} dataset $\mathcal{S}_k$ tells us nothing about $\mathbb{P}_l$ for $l \neq k$, and there is no a priori relationship between the $\mathbb{P}_k$. We term this phenomenon the *data-pooling phenomenon in stochastic optimization.*

Figure 1 illustrates the data-pooling phenomenon with a simulated example for emphasis. Here $K = 10,000$, and the k^{th} subproblem is a newsvendor problem with critical quantile 90%, i.e., $c_k(x;\xi) = \max\{9(\xi - x), (x - \xi)\}$. The measures $\mathbb{P}_k$ are fixed and in each run we simulate $\hat{N}_k = 20$ data points per subproblem. For the decoupled benchmark, we use a standard method, Sample Average Approximation (SAA; Definition 2.1) which is particularly well-suited to the data-driven newsvendor problem (Levi et al. 2015). For comparison, we use our novel Shrunken-SAA algorithm which exploits the data-pooling phenomenon. We motivate and formally define Shrunken-SAA in

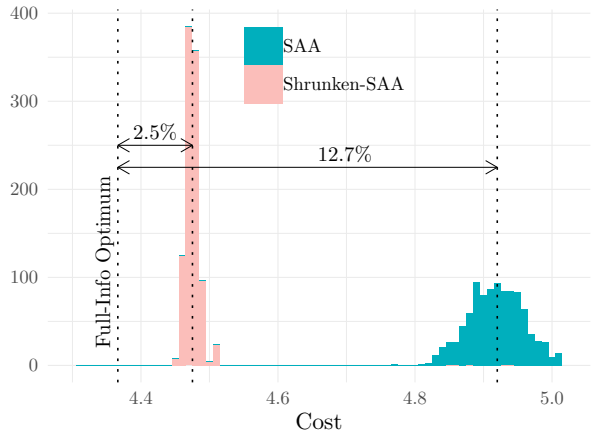


Figure 1 **The Data-Pooling Phenomenon** Consider $K = 10,000$ data-driven newsvendor problems each with critical fractile 90% and 20 data points drawn independently across problems. SAA decouples the problems and orders the 90^{th}-sample quantile in each. Shrunken-SAA (cf. Algorithm 1 in Section 3), leverages data-pooling. Indicated percentages are losses to the full-information optimum. Additional details in Appendix E.1.

Section 3, but, loosely speaking Shrunken-SAA proceeds by replacing the k^{th} dataset $\mathcal{S}_k$ with a "pooled" dataset which is a weighted average of the original k^{th} dataset and all of the remaining $l \neq k$ datasets. It then applies SAA to these each of these new pooled datasets. Perhaps surprisingly, by pooling data across the unrelated subproblems, Shrunken-SAA reduces the loss to full-information optimum by over 80% compared to SAA in this example.

Our Contributions: We describe and study the data-pooling phenomenon in stochastic optimization in context of Problem (1.2). Our analysis applies to constrained, potentially non-convex, non-smooth optimization problems under fairly mild assumptions on the data-generating process. In particular, we assume only that each $\mathbb{P}_k$ has finite support (potentially differing across k); in some cases, we can even relax this assumption. We contrast the data-pooling phenomenon to a similar phenomenon in statistics (Stein's phenomenon), highlighting unique features that arise in the optimization setting (cf. Theorem 2.2 and Example 2.3). In particular, and in contrast to traditional statistical settings, we show that the potential benefits of data-pooling depend strongly on the structure of the underlying optimization problems, and, in some cases, data-pooling may offer no benefit over decoupling.

This observation raises important questions: Given a particular data-driven instance of Problem (1.2), should we data-pool, and, if so, how? More generally, does data-pooling *typically* offer a significant benefit over decoupling, or are instances like Fig. 1 somehow the exception to the rule?

To help resolve these questions, we propose a simple, novel algorithm we call Shrunken Sample Average Approximation (Shrunken-SAA). Shrunken-SAA generalizes the classical SAA algorithm and, consequently, inherits many of its excellent large-sample asymptotic properties (cf. Remark 4.1). Moreover, Shrunken-SAA is incredibly versatile and can be tractably applied to a wide variety of optimization problems with computational requirements similar to traditional SAA (cf. Remark 3.1). Unlike traditional SAA, however, Shrunken-SAA exploits the data-pooling phenomenon to improve performance over SAA, as seen in Fig. 1. Moreover, Shrunken-SAA exploits

the structure of the optimization problems and strictly improves upon an estimate-then-optimize approach using traditional statistical shrinkage estimators (cf. Example 2.3 and Section 6).

Shrunken-SAA data-pools by combining data across subproblems in a particular fashion motivated by an empirical Bayesian argument. We prove that (under frequentist assumptions) for many classes of optimization problems, as the number of subproblems K grows large, Shrunken-SAA determines *if* pooling in this way can improve upon decoupling and, if so, also determines the optimal amount to pool (cf. Theorems 4.2, 4.3, 4.5 and 4.6). These theoretical results study Problem Eq. (1.2) when the random variables $\boldsymbol{\xi}_k$ have finite, discrete support and the amount of data available for the k^{th} subproblem is, itself, random (see Assumption 3.1). Some of our results do extend to the case of continuous $\boldsymbol{\xi}_k$ (cf. Section 4.6 and Theorems F.1 to F.3 in Appendix F), and numerical experiments suggest our results are generally robust to the assumption of a random amount of data.

More interestingly, our theoretical performance guarantees for Shrunken-SAA hold even when the expected amount of data per subproblem is small and fixed, and the number of problems K is large, as in Fig. 1, i.e., they hold in the so-called small-data, large-scale regime (Gupta and Rusmevichientong 2017). Indeed, since many traditional data-driven methods (including SAA) converge to the full-information optimum in the large-sample regime, the small-data, large-scale regime is arguably the more interesting regime in which to study the benefits of data-pooling.

In light of the above results, Shrunken-SAA provides an algorithmic approach to deciding if, and, by how much to pool. To develop an intuitive understanding of *when* and *why* data-pooling might improve upon decoupling, we also introduce the *Sub-Optimality-Instability Tradeoff*, a decomposition of the benefits of data-pooling. We show that the performance of a data-driven solution to Problem (1.2) (usually called its out-of-sample performance in machine learning settings) can be decomposed into a sum of two terms: a term that roughly depends on its in-sample sub-optimality, and a term that depends on its instability, i.e., how much does in-sample performance change when training with one fewer data points? As we increase the amount of data-pooling, we increase the in-sample sub-optimality because we "pollute" the k^{th} subproblem with data from other, unrelated subproblems. At the same time, however, we decrease the instability of the k^{th} subproblem, because the solution no longer relies on its own data so strongly. Shrunken-SAA works by navigating this tradeoff, seeking a "sweet spot" to improve performance. (See Section 5 for discussion.)

In many ways, the Sub-Optimality-Instability Tradeoff resembles the classical bias-variance tradeoff from statistics. However, they differ in that the Sub-Optimality-Instability tradeoff applies to general optimization problems, while the bias-variance tradeoff applies specifically to the case of mean-squared error. Moreover, even in the special case when Problem (1.2) models mean-squared

error, we prove that these two tradeoffs are distinct (cf. Appendix D). In this sense, the Sub-Optimality-Instability Tradeoff may be of independent interest outside data-pooling.

Stepping back, this simple intuition suggests that Shrunken-SAA, and data-pooling more generally, offer significant benefits whenever the decoupled solutions to the subproblems are sufficiently unstable, which typically happens when there is only a small amount of relevant data per subproblem. It is in this sense that the behavior in Fig. 1 is typical and not pathological. Moreover, this intuition also naturally extends beyond Shrunken-SAA, paving the way to developing and analyzing new algorithms which also exploit the, hitherto underutilized, data-pooling phenomenon.

Finally, we present numerical evidence in an inventory management context using real-data from a chain of European Drug Stores showing that Shrunken-SAA can offer significant benefits over decoupling when the amount of data per subproblem is small to moderate. These experiments also suggest that Shrunken-SAA's ability to identify an optimal amount of pooling and improve upon decoupling are relatively robust to violations of our assumptions on the data-generating process.

Connections to Prior Work: As shown in Section 3, our proposed algorithm Shrunken-SAA generalizes SAA. In many ways, SAA is *the* most fundamental approach to solving Problem (1.1) in a data-driven setting. SAA proxies $\mathbb{P}$ in (1.1) by the empirical distribution $\hat{\mathbb{P}}$ on the data and optimizes against $\hat{\mathbb{P}}$. It enjoys strong theoretical and practical performance in the large-sample limit, i.e., when N is large (Kleywegt et al. 2002, Shapiro et al. 2009). For data-driven newsvendor problems, specifically – an example we use throughout our work – SAA is the maximum likelihood estimate of the optimal solution and also is the distributionally robust optimal solution when using a Wasserstein ambiguity set (Esfahani and Kuhn 2018, pg. 151). SAA is incredibly versatile and applicable to a wide-variety of classes of optimization problems. This combination of strong performance and versatility has fueled SAA's use in practice.

When applied to Problem (1.2), SAA by construction decouples the problem into its K subproblems. Because of this strong theoretical and practical performance, we use SAA throughout as the natural, "apples-to-apples" decoupled benchmark to which we compare our data-pooling procedure Shrunken-SAA.

More generally, the data-pooling phenomenon for stochastic optimization is closely related to Stein's phenomenon in statistics (Stein 1956; see also Efron and Hastie 2016 for a modern overview). Stein (1956) considered estimating the mean of K normal distributions, each with known variance σ^2, from K datasets. The k^{th} dataset is drawn i.i.d. from the k^{th} normal distribution and draws are independent across k. The natural decoupled solution to the problem (and the maximum likelihood estimate) is to use the k^{th} sample mean as an estimate for the k^{th} distribution. Surprisingly,

while this estimate is optimal for each problem separately in a very strong sense (uniformly minimum variance unbiased and admissible), Stein (1956) describes a pooled procedure that *always* outperforms this decoupled procedure with respect to total mean-squared error whenever $K \geq 3$.

The proof of Stein's landmark result is remarkably short, but arguably opaque. Indeed, many textbooks refer to it as "Stein's Paradox," perhaps because it is not immediately clear what drives the result. Why does it always improve upon decoupling, and what is special about $K = 3$? Is this a feature of normal distributions? The known variance assumption? The structure of mean-squared error loss? All of the above?

Many authors have tried to develop simple intuition for Stein's result (e.g., Efron and Morris 1977, Stigler 1990, Brown et al. 2012, Brown 1971, Beran 1996) with mixed success. As a consequence, although Stein's phenomenon has had tremendous impact in statistics, it has, in our humble opinion, had fairly limited impact on data-driven optimization. It is simply not clear how to generalize Stein's original algorithm to optimization problems different from minimizing mean-squared error. Indeed, the few data-driven optimization methods that attempt to leverage shrinkage apply either to quadratic optimization (e.g., Davarnia and Cornuéjols 2017, Jorion 1986, DeMiguel et al. 2013) or else under Gaussian or near-Gaussian assumptions (Gupta and Rusmevichientong 2017, Mukherjee et al. 2015), both of which are very close to Stein's original setting.

By contrast, our analysis of the data-pooling phenomenon requires very mild distributional assumptions and applies to constrained, potentially non-convex, non-smooth optimization problems. Numerical experiments in Section 6 further suggest that even our few assumptions are not crucial to the data-pooling phenomenon. Moreover, our proposed algorithm, Shrunken-SAA, is extremely versatile, and can be applied in any setting in which SAA can be applied.

Finally, we note that (in)stability has been well-studied in the machine-learning community (see, e.g., Bousquet and Elisseeff 2002, Shalev-Shwartz et al. 2010, Yu 2013 and references therein). Shalev-Shwartz et al. (2010), in particular, argues that stability is the fundamental feature of data-driven algorithms that enables learning. Our Sub-Optimality-Instability Tradeoff connects the data-pooling phenomenon in stochastic optimization to this larger statistical concept. To the best of our knowledge, however, existing theoretical analyses of stability focus on the large-sample regime. Ours is the first work to leverage stability concepts in the small-data, large-scale regime. From a technical perspective, this analysis requires somewhat different tools.

Notation: Throughout the document, we use boldfaced letters ($\boldsymbol{p}, \boldsymbol{m}, \ldots$) to denote vectors and matrices, and ordinary type to denote scalars. We use "hat" notation ($\hat{\boldsymbol{p}}, \hat{\boldsymbol{m}}, \ldots$) to denote observed data, i.e., an observed realization of a random variable. We reserve the index k to denote parameters for the k^{th} subproblem. For any random variable X and $p \geq 1$, let $\|X\|_p \equiv \sqrt[p]{\mathbb{E}[|X|^p]}$ denote the p^{th} norm of X. Finally, $\boldsymbol{e}_i$ refers to the i^{th} unit vector and $\to_p$ denotes convergence in probability.

2. Model Setup and the Data-Pooling Phenomenon

As discussed in the introduction, we assume throughout that $\mathbb{P}_k$ has finite, discrete support, i.e., $\boldsymbol{\xi}_k \in \{\boldsymbol{a}_{k1}, \ldots, \boldsymbol{a}_{kd}\}$ with $d \geq 2$. Notice that while the support may in general be distinct across subproblems, without loss of generality d is common.[1] To streamline the notation, we write

$$p_{ki} \equiv \mathbb{P}_k(\boldsymbol{\xi}_k = \boldsymbol{a}_{ki}) \quad \text{and} \quad c_{ki}(\boldsymbol{x}) \equiv c_k(\boldsymbol{x}, \boldsymbol{a}_{ki}), \qquad i = 1 \ldots, d.$$

For each k, we let $\mathcal{S}_k = \{\hat{\boldsymbol{\xi}}_{kj} : j = 1, \ldots, \hat{N}_k\}$ be the k^{th} dataset with $\hat{\boldsymbol{\xi}}_{kj} \sim \mathbb{P}_k$ drawn i.i.d. Since $\mathbb{P}_k$ is discrete, we can equivalently represent the k^{th} dataset $\mathcal{S}_k$ via counts, $\hat{\boldsymbol{m}}_k = (\hat{m}_{k1}, \ldots, \hat{m}_{kd})$, where $\hat{m}_{ki}$ denotes the number of times that $\boldsymbol{a}_{ki}$ occurs in $\mathcal{S}_k$, and $\boldsymbol{e}^\top \hat{\boldsymbol{m}}_k = \hat{N}_k$. In what follows, we will use $\hat{\boldsymbol{m}}_k$ and $\mathcal{S}_k$ interchangeably to refer to the k^{th} dataset.

Note that because $\hat{\boldsymbol{\xi}}_{kj}$ are i.i.d.,

$$\hat{\boldsymbol{m}}_k \mid \hat{N}_k \sim \text{Multinomial}(\hat{N}_k, \boldsymbol{p}_k), \quad k = 1, \ldots K. \tag{2.1}$$

Let $\mathcal{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_K)$, or equivalently, $\hat{\boldsymbol{m}} = (\hat{\boldsymbol{m}}_1, \ldots, \hat{\boldsymbol{m}}_K)$, denote all the data across all K subproblems, and let $\hat{\boldsymbol{N}} = (\hat{N}_1, \ldots, \hat{N}_K)$ denote the total observation counts. For convenience, we let $\hat{N}_{\max} = \max_k \hat{N}_k$ and $\hat{N}_{\text{avg}} \equiv \frac{1}{K}\sum_{k=1}^K \hat{N}_k$. Finally, let $\hat{\boldsymbol{p}}_k \equiv \hat{\boldsymbol{m}}_k / \hat{N}_k$ denote the empirical distribution for the k^{th} subproblem.

Notice we have used $\hat{\cdot}$ notation when denoting $\hat{N}_k$ and conditioned on its value in specifying the distribution of $\hat{\boldsymbol{m}}_k$. This is because in our subsequent analysis, we will sometimes view the amount of data available for each problem as random (see Sec. 3.2 below). When the amount of data is fixed and *non-random*, we condition on $\hat{N}_k$ explicitly to emphasize this fact.

With this notation, we can rewrite our target optimization problem:

$$Z^* \equiv \min_{\boldsymbol{x}_1 \in \mathcal{X}_1, \ldots, \boldsymbol{x}_K \in \mathcal{X}_K} \quad \frac{1}{K}\sum_{k=1}^K \frac{\lambda_k}{\lambda_{\text{avg}}} \boldsymbol{p}_k{}^\top \boldsymbol{c}_k(\boldsymbol{x}_k) \tag{2.2}$$

Our goal is to identify a data-driven policy, i.e., a function $\boldsymbol{x}(\hat{\boldsymbol{m}}) = (\boldsymbol{x}_1(\hat{\boldsymbol{m}}), \ldots, \boldsymbol{x}_K(\hat{\boldsymbol{m}}))$ mapping $\hat{\boldsymbol{m}}$ to $\mathcal{X}_1 \times \cdots \times \mathcal{X}_K$ for which $\frac{1}{K}\sum_{k=1}^K \frac{\lambda_k}{\lambda_{\text{avg}}} \boldsymbol{p}_k{}^\top \boldsymbol{c}_k(\boldsymbol{x}_k(\hat{\boldsymbol{m}}))$ is small. We stress that the performance of a data-driven policy is random because it depends on the data.

As mentioned with full information of $\boldsymbol{p}_k$, Problem (2.2) decouples across k, and, after decoupling, no longer depends on the frequency weights $\frac{\lambda_k}{K\lambda_{\text{avg}}}$. Our proposed algorithms will also *not* require knowledge of the weights λ_k. For convenience we let $\lambda_{\min} = \min_k \lambda_k$, and $\lambda_{\max} = \max_k \lambda_k$.

A canonical policy to which we will compare is the *Sample Average Approximation* (SAA) policy which proxies the solution of these de-coupled problems by replacing $\boldsymbol{p}_k$ with $\hat{\boldsymbol{p}}_k$:

[1] Section 4.6 below discusses relaxing this discrete support assumption.

DEFINITION 2.1 (**Sample Average Approximation**). Let $\boldsymbol{x}_k^{\mathrm{SAA}}(\hat{\boldsymbol{m}}_k) \in \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \hat{\boldsymbol{p}}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k)$ denote the SAA policy for the k^{th} problem and let $\boldsymbol{x}^{\mathrm{SAA}}(\hat{\boldsymbol{m}}) = (\boldsymbol{x}_1^{\mathrm{SAA}}(\hat{\boldsymbol{m}}_1), \ldots, \boldsymbol{x}_K^{\mathrm{SAA}}(\hat{\boldsymbol{m}}_K))$.

As we will see, SAA is closely related to our proposed algorithm Shrunken-SAA, and hence provides a natural (decoupled) benchmark when assessing the value of data-pooling.

Finally, we use the newsvendor problem as a running example in what follows. We say the k^{th} subproblem is a *newsvendor problem* with critical fractile $0 < s < 1$ if $c_k(x;\xi) = \max\left\{\frac{s}{1-s}(\xi - x), (x-\xi)\right\}$. Its full-information solution is the s^{th} quantile of the k^{th} distribution.

2.1. A Bayesian Perspective of Data-Pooling

To motivate data-pooling, we first consider a Bayesian approximation to our problem. Specifically, suppose that each $\boldsymbol{p}_k$ were independently drawn from a common Dirichlet prior, i.e.,

$$\boldsymbol{p}_k \sim \mathrm{Dir}(\boldsymbol{p}_0, \alpha_0), \quad k = 1, \ldots, K,$$

with $\alpha_0 > 0$ and $\boldsymbol{p}_0 \in \Delta_d$, the d-dimensional simplex. The Bayes-optimal decision minimizes the posterior risk, which is $\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \frac{\lambda_k}{\lambda_{\mathrm{avg}}} \boldsymbol{p}_k{}^\top \boldsymbol{c}_k(\boldsymbol{x}_k) \mid \hat{\boldsymbol{m}}\right] = \frac{1}{K}\sum_{k=1}^K \frac{\lambda_k}{\lambda_{\mathrm{avg}}} \mathbb{E}\left[\boldsymbol{p}_k \mid \hat{\boldsymbol{m}}\right]^\top \boldsymbol{c}_k(\boldsymbol{x}_k)$, by linearity. Furthermore, by independence and conjugacy, respectively,

$$\mathbb{E}\left[\boldsymbol{p}_k \mid \hat{\boldsymbol{m}}\right] \;=\; \mathbb{E}\left[\boldsymbol{p}_k \mid \hat{\boldsymbol{m}}_k\right] \;=\; \frac{\alpha_0}{\hat{N}_k + \alpha_0}\boldsymbol{p}_0 + \frac{\hat{N}_k}{\hat{N}_k + \alpha_0}\hat{\boldsymbol{p}}_k.$$

Hence, a Bayes-optimal solution is $\boldsymbol{x}(\alpha_0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) = (\boldsymbol{x}_1(\alpha_0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_1), \ldots, \boldsymbol{x}_K(\alpha_0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_K))$, where

$$\hat{\boldsymbol{p}}_k(\alpha) = \left(\frac{\alpha}{\hat{N}_k + \alpha}\boldsymbol{p}_0 + \frac{\hat{N}_k}{\hat{N}_k + \alpha}\hat{\boldsymbol{p}}_k\right), \quad k = 1, \ldots, K \tag{2.3}$$

$$\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) \in \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \quad \hat{\boldsymbol{p}}_k(\alpha)^\top \boldsymbol{c}_k(\boldsymbol{x}_k), \quad k = 1, \ldots, K. \tag{2.4}$$

For any fixed (non-data-driven) α and $\boldsymbol{p}_0$, $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ only depends on the data through $\hat{\boldsymbol{m}}_k$, but not on $\hat{\boldsymbol{m}}_l$ for $l \neq k$.

This policy has an appealing, intuitive structure. Notice $\hat{\boldsymbol{p}}_k(\alpha)$ overloads notation slightly and is a convex combination between $\hat{\boldsymbol{p}}_k = \hat{\boldsymbol{p}}_k(0)$, a data-based estimated of $\boldsymbol{p}_k$, and $\boldsymbol{p}_0$, an a priori estimate of $\boldsymbol{p}_k$. In traditional statistical parlance, we say $\hat{\boldsymbol{p}}_k(\alpha)$ *shrinks* the empirical distribution $\hat{\boldsymbol{p}}_k$ toward the anchor $\boldsymbol{p}_0$. The Bayes-optimal solution is the plug-in solution when using this shrunken empirical measure, i.e., it optimizes $\boldsymbol{x}_k$ as though that were the known true measure. Note in particular, this differs from the SAA solution, which is the plug-in solution when using the "unshrunken" $\hat{\boldsymbol{p}}_k$.

The parameter α controls the degree of shrinkage. As $\alpha \to 0$, $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ converges to an SAA solution, and as $\alpha \to \infty$, $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ converges to the (non-random) solution to the fully-shrunken

k^{th} subproblem. In this sense the Bayes-optimal solution "interpolates" between the SAA solution and the fully-shrunken solution. The amount of data $\hat{N}_k$ attenuates the amount of shrinkage, i.e., subproblems with more data are shrunk less aggressively for the same α.

Alternatively, we can give a data-pooling interpretation of $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ via the Bayesian notion of pseudocounts. Observe $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) \in \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \left(\frac{\alpha \boldsymbol{p}_0 + \hat{\boldsymbol{m}}_k}{\hat{N}_k + \alpha}\right)^\top \boldsymbol{c}_k(\boldsymbol{x}_k)$ and that $\frac{\alpha \boldsymbol{p}_0 + \hat{\boldsymbol{m}}_k}{\hat{N}_k + \alpha}$ is a distribution on $\{\boldsymbol{a}_{k1}, \ldots, \boldsymbol{a}_{kd}\}$. In other words, we can interpret $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ as the solution obtained when we augment each of our original K datasets with α additional "synthetic" data points with counts $\alpha \boldsymbol{p}_0$. As we increase α, we add more synthetic data.

For $\alpha > 0$, $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \boldsymbol{0})$ is the solution to the fully shrunken k^{th} subproblem. For emphasis, let

$$\boldsymbol{x}_k(\infty, \boldsymbol{p}_0) \in \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \sum_{i=1}^d p_{0i} c_{ki}(\boldsymbol{x}_k),$$

so that $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \boldsymbol{0}) = \boldsymbol{x}_k(\infty, \boldsymbol{p}_0)$ for all $\alpha > 0$. For completeness, we also define $\boldsymbol{x}_k(0, \boldsymbol{p}_0, \boldsymbol{0}) = \boldsymbol{x}_k(\infty, \boldsymbol{p}_0)$, so that $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \cdot)$ is continuous in α.

In summary, $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ has an intuitive structure that is well-defined *regardless of the precise structure of the cost functions* $\boldsymbol{c}_k(\cdot)$ *or feasible region* $\mathcal{X}$. Importantly, this analysis shows that when the $\boldsymbol{p}_k$ follow a Dirichlet prior, data-pooling by α is never worse than decoupling, and will be strictly better whenever $\boldsymbol{x}_k^{\text{SAA}}(\hat{\boldsymbol{m}}_k)$ is not an optimal solution to the problem defining $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$.

2.2. Data-Pooling in a Frequentist Setting

It is perhaps not surprising that data-pooling (or shrinkage) improves upon the decoupled SAA solution in the Bayesian setting because problems $l \neq k$ contain information about α and $\boldsymbol{p}_0$ which in turn contain information about $\boldsymbol{p}_k$. What may be surprising is that even in frequentist settings, i.e., when the $\boldsymbol{p}_k$ are fixed constants that may have no relationship to one another and there is no "ground-truth" values for α or $\boldsymbol{p}_0$, policies like $\boldsymbol{x}(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ can still improve upon the decoupled SAA solution through a careful choice of α and $\boldsymbol{p}_0$ that depend on *all* the data. Indeed, this is the heart of Stein's result for Gaussian random variables and mean-squared error.

To build intuition, we first study the specific case of minimizing mean-squared error and show that data-pooling can improve upon the decoupled SAA solution in the frequentist framework of Eq. (2.1). This result is thus reminiscent of Stein's classical result, but does not require the Gaussian assumptions. Consider the following example:

Example 2.1 (A Priori-Pooling for Mean-Squared Error). Consider a special case of Problem (2.2) such that for all k that $\lambda_k = \lambda_{\text{avg}}$, $\hat{N}_k = \hat{N} \geq 2$, $\boldsymbol{p}_k$ is supported on $\{a_{k1}, \ldots, a_{kd}\} \subseteq \mathbb{R}$, $\mathcal{X}_k = \mathbb{R}$ and $c_{ki}(x) = (x - a_{ki})^2$. In words, the k^{th} subproblem estimates the unknown mean $\mu_k = \boldsymbol{p}_k^\top \boldsymbol{a}_k$ by minimizing the mean-squared error. Let $\sigma_k^2 = \boldsymbol{p}_k^\top (\boldsymbol{a}_k - \mu_k \boldsymbol{e})^2$.

Fix any $\boldsymbol{p}_0 \in \Delta_d$ and $\alpha \geq 0$ (not depending on the data). A direct computation shows that

$$x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) \;\equiv\; \hat{\mu}_k(\alpha) \;\equiv\; \tfrac{\hat{N}}{\hat{N}+\alpha}\hat{\mu}_k + \tfrac{\alpha}{\hat{N}+\alpha}\mu_{k0},$$

where $\hat{\mu}_k = \frac{1}{\hat{N}}\sum_{i=1}^{\hat{N}} \hat{\xi}_{ki}$ is the usual sample mean, and $\mu_{k0} = \boldsymbol{p}_0^\top \boldsymbol{a}_k$. Notice in particular that the decoupled SAA solution is $\boldsymbol{x}^{\mathsf{SAA}} = (\hat{\mu}_1, \ldots, \hat{\mu}_K)$, corresponding to $\alpha = 0$.

For any $\boldsymbol{p}_0$ and α, the objective value of $\boldsymbol{x}(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ is

$$\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{p}_k^\top \boldsymbol{c}_k(x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)) \;=\; \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[(\hat{\mu}_k(\alpha) - \xi_k)^2 \mid \hat{\boldsymbol{m}}\right] \;=\; \frac{1}{K}\sum_{k=1}^{K} \left(\sigma_k^2 + (\mu_k - \hat{\mu}_k(\alpha))^2\right),$$

by the usual bias-variance decomposition of mean-squared error (MSE). This objective is the average of K independent random variables. Hence, we might intuit that under appropriate regularity conditions (see Theorem 2.1 below) that, conditional on $\hat{N}$, as $K \to \infty$,

$$\tfrac{1}{K}\textstyle\sum_{k=1}^{K} (\sigma_k^2 + (\mu_k - \hat{\mu}_k(\alpha))^2) - \tfrac{1}{K}\left(\sum_{k=1}^{K} \sigma_k^2 + \mathbb{E}\left[(\mu_k - \hat{\mu}_k(\alpha))^2 \mid \hat{N}\right]\right) \;\to_p\; 0. \tag{2.5}$$

Moreover, $\frac{1}{K}\left(\sum_{k=1}^{K} \sigma_k^2 + \mathbb{E}\left[(\mu_k - \hat{\mu}_k(\alpha))^2 \mid \hat{N}\right]\right) \;=\; \frac{1}{K}\sum_{k=1}^{K}\left(\sigma_k^2 + \left(\frac{\alpha}{\hat{N}+\alpha}\right)^2 (\mu_k - \mu_{k0})^2 + \left(\frac{\hat{N}}{\hat{N}+\alpha}\right)^2 \frac{\sigma_k^2}{\hat{N}}\right)$, again using the bias-variance decomposition of MSE. We can minimize the righthand side over α explicitly, yielding the value

$$\alpha_{\boldsymbol{p}_0}^{\mathsf{AP}} \;=\; \frac{\sum_{k=1}^{K} \sigma_k^2}{\sum_{k=1}^{K} (\mu_k - \mu_{k0})^2} \;> 0,$$

where AP stands for *a priori*, meaning $\alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}$ is the on-average-best a priori choice of shrinkage before observing any data. In particular, substituting $\alpha = 0$ and $\alpha = \alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}$ into the second term of Eq. (2.5) shows that, up to a term that is vanishing as $K \to \infty$, shrinking by $\alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}$ decreases the MSE by

$$\left(\frac{1}{K}\sum_{k=1}^{K} \frac{\sigma_k^2}{\hat{N}}\right) \frac{\alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}}{\hat{N} + \alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}} \;=\; \frac{\left(\frac{1}{K\hat{N}}\sum_{k=1}^{K} \sigma_k^2\right)^2}{\frac{1}{K\hat{N}}\sum_{k=1}^{K} \sigma_k^2 + \frac{1}{K}\sum_{k=1}^{K} (\mu_k - \mu_{k0})^2} \;>\; 0. \tag{2.6}$$

This benefit is strictly positive for any values of $\boldsymbol{p}_k$ and $\boldsymbol{p}_0$, and increasing in $\alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}$.

Unfortunately, we cannot implement $x(\alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ in practice because $\alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}$ is not computable from the data; it depends on the unknown μ_k and σ_k^2. The next theorem shows that we can, however, estimate $\alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}$ from the data in a way that achieves the same benefit as $K \to \infty$, even if $\hat{N}$ is fixed and small. See Appendix A for proof.

THEOREM 2.1 **(Data-Pooling for MSE)**. *Consider a sequence of subproblems, indexed by $k = 1, 2, \ldots$. Suppose for each k, the k^{th} subproblem minimizes mean-squared error, i.e., $\boldsymbol{p}_k$ is supported on $\{a_{k1}, \ldots, a_{kd}\} \subseteq \mathbb{R}$, $\mathcal{X}_k = \mathbb{R}$ and $c_{ki}(x) = (x - a_{ki})^2$. Suppose further that there exists λ_{avg}, $\hat{N} \geq 2$ and $a_{\max} < \infty$ such that $\lambda_k = \lambda_{\text{avg}}$, $\hat{N}_k = \hat{N}$, and $\|\boldsymbol{a}_k\|_\infty \leq a_{\max}$ for all k. Fix any $\boldsymbol{p}_0 \in \Delta_d$, and let*

$$\alpha_{\boldsymbol{p}_0}^{\mathsf{JS}} = \frac{\frac{1}{K}\sum_{k=1}^{K} \frac{1}{\hat{N}-1}\sum_{i=1}^{\hat{N}} (\hat{\xi}_{ki} - \hat{\mu}_k)^2}{\frac{1}{K}\sum_{k=1}^{K} (\mu_{k0} - \hat{\mu}_k)^2 - \frac{1}{K\hat{N}}\sum_{k=1}^{K} \frac{1}{\hat{N}-1}\sum_{i=1}^{\hat{N}} (\hat{\xi}_{ki} - \hat{\mu}_k)^2}.$$

Then, conditional on $\hat{N}$, as $K \to \infty$,

$$\underbrace{\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{p}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k^{\mathsf{SAA}}) - \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{p}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha_{\boldsymbol{p}_0}^{\mathsf{JS}}, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))}_{\text{Benefit over decoupling of } \alpha = \alpha_{\boldsymbol{p}_0}^{\mathsf{JS}}} - \underbrace{\frac{\left(\frac{1}{K}\sum_{k=1}^{K} \sigma_k^2/\hat{N}\right)^2}{\frac{1}{K}\sum_{k=1}^{K} \sigma_k^2/\hat{N} + \frac{1}{K}\sum_{k=1}^{K}(\mu_k - \mu_{k0})^2}}_{\text{Expected benefit over decoupling of } \alpha = \alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}} \to_p 0.$$

Note that $x_k(\alpha_{\boldsymbol{p}_0}^{\mathsf{JS}}, \boldsymbol{p}_0, \hat{\boldsymbol{m}}) = (1-\theta)\hat{\mu}_k + \theta\hat{\mu}_{k0}$ where $\theta = \frac{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\hat{N}-1}\sum_{i=1}^{\hat{N}}(\hat{\xi}_{ki}-\hat{\mu}_k)^2}{\frac{1}{K}\sum_{k=1}^{K}(\mu_{k0}-\hat{\mu}_k)^2}$. In this form, we can see that the resulting estimator with pooling $\alpha_{\boldsymbol{p}_0}^{\mathsf{JS}}$ strongly resembles the classical James-Stein mean estimator (cf. Efron and Hastie 2016, Eq. (7.51)), with the exception that we have replaced the variance σ_k^2, which is assumed to be 1 in Stein's setting, with the usual, unbiased estimator of that variance. This resemblance motivates our "JS" notation. Theorem 2.1 is neither stronger nor weaker than the James-Stein theorem. Our result applies to non-Gaussian random variables and holds in probability, but is asymptotic; the James-Stein theorem requires Gaussian distributions and holds in expectation, but applies to any fixed $K \geq 3$.

Theorem 2.1 shows that data-pooling for mean-squared error always offers a benefit over decoupling for sufficiently large K, no matter what the $\boldsymbol{p}_k$ may be. Data-pooling for general optimization problems, however, exhibits more subtle behavior. In particular, as shown in the following example and theorem, there exist instances where data-pooling offers no benefit over decoupling, and instances where data-pooling may be worse than decoupling.

Example 2.2 (Data-Pooling for Simple Newsvendor). Consider a special case of Problem (2.2) such that for all k, $\lambda_k = \lambda_{\text{avg}}$, ξ_k is supported on $\{1, 0\}$, $\mathcal{X}_k = [0, 1]$ and $c_k(x, \xi_k) = |x - \xi_k|$ so that $\boldsymbol{p}_k^\top \boldsymbol{c}_k(x) = p_{k1} + x(1 - 2p_{k1})$. In words, the k^{th} subproblem estimates the median of a Bernoulli random variable by minimizing mean absolute deviation, or, equivalently, is a newsvendor problem with critical fractile 0.5 for Bernoulli demand. We order the support so that $p_{k1} = \mathbb{P}(\xi_k = 1)$, as is typical for a Bernoulli random variable. Suppose further for each k, $p_{k1} > \frac{1}{2}$, and fix any $p_{01} < \frac{1}{2}$.

Note $x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) = \mathbb{I}\left[\hat{p}_{k1} \geq \frac{1}{2} + \frac{\alpha}{\hat{N}_k}(\frac{1}{2} - p_{01})\right]$.[2] Further, for any α (possibly depending on $\hat{\boldsymbol{m}}$),

$$\begin{aligned}\boldsymbol{p}_k^\top \left(\boldsymbol{c}_k(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)) - \boldsymbol{c}_k(\boldsymbol{x}_k(0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))\right) &= (2p_{k1} - 1)\left(\mathbb{I}\left[\hat{p}_{k1} \geq 1/2\right] - \mathbb{I}\left[\hat{p}_{k1} \geq \frac{1}{2} + \frac{\alpha}{\hat{N}_k}\left(\frac{1}{2} - p_{01}\right)\right]\right) \\ &= (2p_{k1} - 1)\mathbb{I}\left[1/2 \;\leq\; \hat{p}_{k1} \;<\; \frac{1}{2} + \frac{\alpha}{\hat{N}_k}\left(\frac{1}{2} - p_{01}\right)\right],\end{aligned}$$

where the last equality follows since $\hat{p}_{k1} < 1/2 \implies \hat{p}_{k1} < \frac{1}{2} + \frac{\alpha}{2}(\frac{1}{2} - p_{01})$. Notice $p_{k1} > \frac{1}{2} \implies (2p_{k1} - 1) > 0$, so this last expression is nonnegative. It follows that path by path, shrinkage by any $\alpha > 0$ cannot improve upon the decoupled solution ($\alpha = 0$). Moreover, if $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) \neq \boldsymbol{x}_k(0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$, the performance is strictly worse.

One can check directly that if we had instead chosen $p_{01} \geq \frac{1}{2}$ and $p_{k1} < \frac{1}{2}$, a similar result holds.

[2] This solution is non-unique, and the solution $\mathbb{I}\left[\hat{p}_{k1} > \frac{1}{2} + \frac{\alpha}{\hat{N}_k}(\frac{1}{2} - p_{01})\right]$ is also valid. We adopt the former solution in what follows, but our comments apply to either solution.

We summarize this example in the following theorem:

THEOREM 2.2 **(Data-Pooling Does Not Always Offer Benefit)**. *Given any $\boldsymbol{p}_0$, there exist instances of Problem* (2.2) *such that shrinkage does not outperform the (decoupled) SAA solution. Moreover, if $\boldsymbol{x}(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ performs the same as SAA, then $\boldsymbol{x}(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ is, itself, an SAA solution.*

On the other hand, there exist examples where the James-Stein estimator and traditional statistical reasoning might suggest the benefits of pooling are marginal, but, by data-pooling in way that exploits the optimization structure, we can achieve significant benefits. Specifically, our Bayesian motivation in Section 2.1 suggests pooling offers little benefit when the $\boldsymbol{p}_k$ are very dispersed, i.e., the Dirichlet prior has high variance and α_0 is small. Similarly, Theorem 2.1 and Efron and Morris (1977) both suggest that the benefits of pooling over decoupling for MSE are marginal if the subproblem means are quite dispersed (cf. Eq. (2.6)). Nonetheless, for general optimization problems, we observe pooling might still offer substantive benefits in these situations:

EXAMPLE 2.3 (POOLING CAN OFFER BENEFIT EVEN WHEN $\boldsymbol{p}_k$ ARE DISPERSED). Let $d > 3$ and fix some $0 < s < 1$. Suppose the k^{th} subproblem is a newsvendor problem with critical fractile $f_k > s$ and demand distribution supported on the integers $1, \ldots, d$. For each k, let $p_{k1} = 0$, $p_{kd} = 1 - s$, and $p_{kj_k} = s$ for some $1 < j_k < d$. Consider the fixed anchor $p_{01} = s$, $p_{0d} = 1 - s$, and $p_{0j} = 0$ for $1 < j < d$. Notice typical $\boldsymbol{p}_k$'s are very far from $\boldsymbol{p}_0$ since $\|\boldsymbol{p}_k - \boldsymbol{p}_0\|_2 = \sqrt{2}s$. For s sufficiently close to 1, this value is close to $\sqrt{2}$, which is the maximal distance between two points on the simplex. In other words, the $\boldsymbol{p}_k$ are not very similar. Moreover, the means are also dispersed for s close to 1 since $\frac{1}{K}\sum_{k=1}^{K}(\mu_k - \mu_0)^2 = s^2\frac{1}{K}\sum_{k=1}^{K}(j_k - 1)^2 \approx s^2 d/2$ if the j_k are chosen uniformly.

Consequently, the James-Stein estimator does not shrink very much in this example. A straightforward computation shows that for K sufficiently large, $\alpha^{\mathsf{JS}}_{\boldsymbol{p}_0} \leq \frac{(1-s)d^2}{s}$ with high probability, which is close to 0 for s close to 1. However, the full-information solution for the k^{th} problem is $\boldsymbol{x}^*_k = d$, which *also* equals the fully-pooled ($\alpha = \infty$) solution, $\boldsymbol{x}_k(\infty, \boldsymbol{p}_0)$. Hence, pooling in an optimization-aware way can achieve full-information performance, while both decoupling and an "estimate-then-optimize" approach using James-Stein shrinkage *necessarily* perform worse. In other words, pooling offers significant benefits despite the $\boldsymbol{p}_k$ being as dispersed as possible, because of the optimization structure, and leveraging this structure is necessary to obtain the best shrinkage. □

Theorems 2.1 and 2.2 and Examples 2.2 and 2.3 highlight the fact that data-pooling for general optimization is more complex than Stein's phenomenon. In particular, in Stein's classical result for mean-squared error and Gaussian data, data-pooling *always* offers a benefit for $K \geq 3$. For other optimization problems and data distributions, data-pooling may *not* offer a benefit, or may offer a benefit but requires a new way of choosing the pooling amount. An interplay between $\boldsymbol{p}_0$, $\boldsymbol{p}_k$ and $\boldsymbol{c}_k$ determines if data-pooling can improve upon decoupling and how much pooling is best.

Algorithm 1 The Shrunken-SAA Algorithm.

Input: Data $\mathcal{S}_k = \{\hat{\boldsymbol{\xi}}_{k1}, \ldots, \hat{\boldsymbol{\xi}}_{k\hat{N}_k}\}$, $k = 1, \ldots, K$, and an anchor distribution $h(\mathcal{S})$

Fix a finite grid $\mathcal{A} \subseteq [0, \infty)$

for $\alpha \in \mathcal{A}, \quad k = 1, \ldots, K, \quad j = 1, \ldots, \hat{N}_k$ **define:**

$\boldsymbol{x}_{k,-j}(\alpha, h(\mathcal{S})) \leftarrow \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \sum_{\ell \neq j} c_k(\boldsymbol{x}_k, \hat{\boldsymbol{\xi}}_{k\ell}) + \alpha \mathbb{E}_{\boldsymbol{\xi}_k \sim h(\mathcal{S})}[c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k)]$ // Compute LOO solutions

end for

$\alpha_h^{\mathsf{S\text{-}SAA}} \leftarrow \arg\min_{\alpha \in \mathcal{A}} \sum_{k=1}^{K} \sum_{j=1}^{\hat{N}_k} c_k(\boldsymbol{x}_{k,-j}(\alpha, h(\mathcal{S})), \hat{\boldsymbol{\xi}}_{kj})$ // Modified LOO-Cross-Validation

for all $k = 1, \ldots, K$ **do**

$\boldsymbol{x}_k^{\mathsf{S\text{-}SAA}} \leftarrow \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \sum_{j=1}^{\hat{N}_k} c_k(\boldsymbol{x}_k, \hat{\boldsymbol{\xi}}_{kj}) + \alpha_h^{\mathsf{S\text{-}SAA}} \mathbb{E}_{\boldsymbol{\xi}_k \sim h(\mathcal{S})}[c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k)]$ // Compute Pooled solution

end for

return $\left(\boldsymbol{x}_1^{\mathsf{S\text{-}SAA}}, \ldots, \boldsymbol{x}_K^{\mathsf{S\text{-}SAA}}\right)$

This raises two important questions: First, how do we identify if an instance of Problem (2.2) would benefit from data-pooling? Second, if it does, how do we compute the "optimal" amount of pooling? In the next sections, we show how our Shrunken-SAA algorithm can be used to address both questions in the relevant regime, where K is large but the average amount of data per subproblem remains small. Indeed, we show that Shrunken-SAA achieves the best-possible shrinkage in an optimization-aware fashion for many types of problems and choices of anchor.

3. The Shrunken SAA Algorithm

Algorithm 1 formally defines Shrunken-SAA. The crucial step is the "Modified LOO-Cross-Validation," which we discuss in detail in Sections 3.2 and 3.3 below. To highlight similarities to SAA, we have stated the algorithm in terms of the datasets $\mathcal{S}_k$ and $\mathcal{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_K)$. Here $h(\mathcal{S})$ represents an arbitrary, possibly data-driven anchor distribution (see below for examples). Recall that we can equivalently express $\mathcal{S}_k$ in terms of the counts $\hat{\boldsymbol{m}}_k$. In that notation, we recognize that if the j^{th} data point of $\mathcal{S}_k$ is $\boldsymbol{a}_{ki}$, then $\boldsymbol{x}_{k,-j}(\alpha, h(\hat{\boldsymbol{m}})) = \boldsymbol{x}_k(\alpha, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)$ and $\boldsymbol{x}_k^{\mathsf{S\text{-}SAA}} = \boldsymbol{x}_k(\alpha^{\mathsf{S\text{-}SAA}}, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k)$. In other words, Shrunken-SAA retains the particular pooling structure of Eq. (2.4) suggested by our Bayesian argument, but allows for a data-dependent anchor $h(\mathcal{S})$ (equiv. $h(\hat{\boldsymbol{m}})$) and chooses the amount of pooling via a particular cross-validation scheme. We present Algorithm 1 using a finite grid of $\alpha \in \mathcal{A}$, but our theory below will study the algorithm with $\mathcal{A} = [0, \infty)$.

Remark 3.1 (Computational Complexity). Computationally, Algorithm 1 does not depend on d, the size of the support of $\boldsymbol{\xi}_k$. Its bottleneck is computing $\boldsymbol{x}_{k,-j}$ which is similar to solving the k^{th} subproblem by SAA with an augmented data set described by $h(\mathcal{S})$. More specifically, Algorithm 1 depends on the data only through $h(\mathcal{S})$ and averages of functions over subsets of $\mathcal{S}$, neither of which explicitly depend upon d. Consequently, although our setup and analysis

assumes $\boldsymbol{\xi}_k$ has finite discrete support, from an implementation perspective, we can apply Shrunken-SAA when $\boldsymbol{\xi}_k$ has continuous support *without* discretization so long as we can efficiently solve these augmented SAA problems (cf. our empirical study in Appendix E.7). From a theoretical perspective, some of our analysis extends to this continuous setting (see Section 4.6 below). In the remainder, we follow Section 2 and treat the data as discrete, referring to the data by $\hat{\boldsymbol{m}}_k$ and $\hat{\boldsymbol{m}}$.

We consider Shrunken-SAA to be *roughly* as tractable as SAA. We say "roughly" because, in the worst-case, one must solve at most $|\mathcal{A}|\sum_{k=1}^K \min(d, \hat{N}_k)$ problems in the LOO-cross-validation step, which, if we sample from $h(\hat{\boldsymbol{m}})$, have a similar structure to SAA. Fortunately, we can parallelize these problems in distributed computing environments and use previous iterations to "warm-start" solvers. Moreover, in Appendix E.9 we observe empirically that less computationally expensive κ-fold cross-validation procedures can be used in place of LOO with similar performance. □

For clarity, the $\alpha_h^{\text{S-SAA}}$ parameter (with $\mathcal{A} = [0, \infty)$) computed by Algorithm 1 is

$$\alpha_h^{\text{S-SAA}} \in \arg\min_{\alpha \geq 0} \ \sum_{k=1}^K \hat{\boldsymbol{m}}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)). \tag{3.1}$$

The Anchor Distribution $h(\hat{\boldsymbol{m}})$

As stated, the anchor in Algorithm 1, $h(\hat{\boldsymbol{m}})$, is an input. We think of $h(\hat{\boldsymbol{m}})$ as a function that selects an anchor distribution from a candidate set of distributions $\mathcal{P}$. In what follows, we will focus on two types of anchors and corresponding candidate sets $\mathcal{P}$:

- **Fixed Anchors**: In this case, $h(\hat{\boldsymbol{m}}) = \boldsymbol{p}_0$, $\mathcal{P} = \{\boldsymbol{p}_0\}$ for some fixed $\boldsymbol{p}_0$, e.g., the uniform distribution $\boldsymbol{p}_0 = \boldsymbol{e}/d$. In general, fixed-anchors might be used for computational/statistical simplicity or when there is strong a priori knowledge of a good anchor. In this special case, we abuse notation slightly, replacing the map $h : \hat{\boldsymbol{m}} \mapsto \boldsymbol{p}_0$ with the constant $\boldsymbol{p}_0$ when it is clear from context, e.g., we write $\alpha_{\boldsymbol{p}_0}^{\text{S-SAA}}$ for $\alpha_h^{\text{S-SAA}}$.
- **Data-Driven Anchors**: In this case $h(\hat{\boldsymbol{m}})$ is any procedure that uses the data $\hat{\boldsymbol{m}}$ to select a distribution, and $\mathcal{P}$ is the image of $h(\cdot)$. One example might be to use all the data to fit a parametric distribution, e.g., a lognormal distribution, via maximum likelihood and use this fitted distribution as the anchor. Then, $\mathcal{P}$ would be the set of lognormal distributions.

We also pay particular focus to two special cases of data-driven anchors in what follows:

- **LOO-Optimized Anchor**: For a given $\mathcal{P} \subseteq \Delta_d$, let

$$h_{\mathcal{P}}(\hat{\boldsymbol{m}}) \in \arg\min_{\boldsymbol{q} \in \mathcal{P}} \min_{\alpha \in \mathcal{A}} \ \sum_{k=1}^K \hat{\boldsymbol{m}}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i). \tag{3.2}$$

We will see below that $h_{\mathcal{P}}$ satisfies stronger optimality properties than general data-driven anchors and, hence, we treat it separately. From an implementation point of view, when applying Algorithm 1, we only ever require the value of $h_{\mathcal{P}}(\hat{\boldsymbol{m}})$, not the full-function $h_{\mathcal{P}}(\cdot)$. Thus,

Algorithm 1 with $h_{\mathcal{P}}(\cdot)$ amounts to replacing the "Modified LOO-Cross-Validation" step by a joint optimization over anchor and pooling amount:

$$(\alpha^{\mathsf{S\text{-}SAA}}_{h_{\mathcal{P}}}, h_{\mathcal{P}}(\hat{\boldsymbol{m}})) \leftarrow \arg\min_{\alpha\in\mathcal{A},\boldsymbol{q}\in\mathcal{P}} \sum_{k=1}^{K}\sum_{j=1}^{\hat{N}_k} c_k(\boldsymbol{x}_{k,-j}(\alpha,\boldsymbol{q}),\hat{\boldsymbol{\xi}}_{kj}). \tag{3.3}$$

We note that the multivariate optimization problem in Eq. (3.3) may be challenging depending on the structure of $\mathcal{P}$, motivating our second special case below.

- **GM-Anchor** We also consider a computationally simpler "grand-mean" anchor $h(\hat{\boldsymbol{m}}) = \hat{\boldsymbol{p}}^{\mathsf{GM}}$ where $\hat{\boldsymbol{p}}^{\mathsf{GM}} \equiv \sum_{k=1}^{K} \hat{\boldsymbol{p}}_k \mathbb{I}\left[\hat{N}_k > 0\right] / \sum_{k=1}^{K} \mathbb{I}\left[\hat{N}_k > 0\right]$ if $\hat{N}_{\max} > 0$ and $\boldsymbol{e}/d$ otherwise. (For this data-driven anchor, $\mathcal{P} = \Delta_d$.) This choice is motivated by our Bayesian perspective on data-pooling from Section 2.1. In the Bayesian setting $\hat{\boldsymbol{p}}^{\mathsf{GM}}$ is an unbiased estimator of the prior mean. We observe empirically in Section 6 that $\hat{\boldsymbol{p}}^{\mathsf{GM}}$ is a strong and computationally-efficient heuristic.

3.1. Oracle Benchmarks

From Theorem 2.2, data-pooling need not improve upon decoupling for a given $h(\cdot)$. To establish appropriate benchmarks, we first define the *oracle* pooling for given $h(\cdot)$, i.e.,

$$\alpha^{\mathsf{OR}}_h \in \arg\min_{\alpha\geq 0} \overline{Z}_K(\alpha, h(\hat{\boldsymbol{m}})), \quad \text{where} \quad \overline{Z}_K(\alpha,\boldsymbol{q}) = \tfrac{1}{K}\sum_{k=1}^{K} Z_k(\alpha,\boldsymbol{q}), \tag{3.4}$$
$$Z_k(\alpha,\boldsymbol{q}) = \tfrac{\lambda_k}{\lambda_{\text{avg}}} \boldsymbol{p}_k{}^{\top} \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k)).$$

Notice α^{OR}_h is random, depending on the entire data-sequence. By construction, $\overline{Z}_K(\alpha^{\mathsf{OR}}_h, h(\hat{\boldsymbol{m}}))$ lower bounds the performance of *any* other data-driven pooling policy with anchor $h(\hat{\boldsymbol{m}})$ path-by-path. Hence, it serves as a strong performance benchmark. However, α^{OR}_h also depends on the unknown $\boldsymbol{p}_k$ and λ_k, and hence, is not implementable in practice. In this sense, it is an oracle.

Given any α (possibly depending on the data), we measure the sub-optimality of pooling by α relative to the oracle pooling for $h(\cdot)$ on a particular data-realization by

$$\mathsf{SubOpt}_{h,K}(\alpha) = \overline{Z}_K(\alpha, h(\hat{\boldsymbol{m}})) - \overline{Z}_K(\alpha^{\mathsf{OR}}_h, h(\hat{\boldsymbol{m}})).$$

Good pooling procedures will have small sub-optimality with high-probability with respect to the data. Note we allow for the possibility that $\alpha^{\mathsf{OR}}_h = 0$, as is the case in Example 2.2. Thus, procedures that have small sub-optimality will still have good performance in instances where data-pooling is not beneficial. Moreover, studying when $\alpha^{\mathsf{OR}}_h > 0$ gives intuition into when and why data-pooling is helpful, a task we take up in Section 5.

The above oracle is defined with respect to a given anchor. One might also seek to benchmark performance relative to the best-possible anchor. Given any $\mathcal{P} \subseteq \Delta_d$, we define the oracle choice of anchor and pooling amount for anchors in $\mathcal{P}$ and for a particular data realization by

$$(\alpha^{\mathsf{OR}}_{\mathcal{P}},\, \boldsymbol{q}^{\mathsf{OR}}_{\mathcal{P}}) \in \arg\min_{\alpha\geq 0,\, \boldsymbol{q}\in\mathcal{P}} \overline{Z}_K(\alpha,\boldsymbol{q}). \tag{3.5}$$

Then, given any anchor $\boldsymbol{q} \in \mathcal{P}$ and pooling amount α (both possibly depending the data), we measure the sub-optimality of shrinking by α towards $\boldsymbol{q}$ by

$$\mathsf{SubOpt}_{\mathcal{P},K}(\alpha, \boldsymbol{q}) \;=\; \overline{Z}_K(\alpha, \boldsymbol{q}) - \overline{Z}_K(\alpha^{\mathsf{OR}}_{\mathcal{P}}, \boldsymbol{q}^{\mathsf{OR}}_{\mathcal{P}}).$$

For clarity, we observe that by construction $\alpha^{\mathsf{OR}}_{\mathcal{P}} = \alpha^{\mathsf{OR}}_{\boldsymbol{q}^{\mathsf{OR}}_{\mathcal{P}}}$.

3.2. Motivating $\alpha^{\mathsf{S\text{-}SAA}}$ through Unbiased Estimation

We first consider a fixed anchor $h(\hat{\boldsymbol{m}}) = \boldsymbol{p}_0$. Recall in this case, we abuse notation slightly, writing

$$\alpha^{\mathsf{OR}}_{\boldsymbol{p}_0} \in \arg\min_{\alpha \geq 0} \overline{Z}_K(\alpha, \boldsymbol{p}_0) \tag{3.6}$$

One approach to choosing $\alpha_{\boldsymbol{p}_0}$ might be to construct a suitable proxy for $\overline{Z}_K(\alpha, \boldsymbol{p}_0)$ in Eq. (3.6) based only on the data, and then choose the $\alpha_{\boldsymbol{p}_0}$ that optimizes this proxy.

If we knew the values of λ_k, a natural proxy might be to replace the unknown $\boldsymbol{p}_k$ with $\hat{\boldsymbol{p}}_k$, i.e., optimize $\frac{1}{K}\sum_{k=1}^K \frac{\lambda_k}{\lambda_{\text{avg}}} \hat{\boldsymbol{p}}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k))$. Unfortunately, even for a fixed, non-data-driven α, this proxy is *biased*, i.e. $\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \frac{\lambda_k}{\lambda_{\text{avg}}} \hat{\boldsymbol{p}}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))\right] \neq \mathbb{E}\left[\overline{Z}_K(\alpha, \boldsymbol{p}_0)\right]$, since both $\hat{\boldsymbol{p}}_k$ and $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ depend on the data $\hat{\boldsymbol{m}}_k$. Worse, this bias wrongly suggests $\alpha = 0$, i.e. decoupling, is always a good policy, because $\boldsymbol{x}_k(0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ always optimizes this proxy, by construction. By contrast, Theorem 2.1 shows data-pooling can offer significant benefits. This type of bias and its consequences are well-known in other contexts and are often termed the "optimizer's curse" – in-sample costs are optimistically biased and may not generalize well.

These features motivate us to seek an unbiased estimate of $\overline{Z}_K(\alpha, \boldsymbol{p}_0)$. At first glance, however, $Z_K(\alpha, \boldsymbol{p}_0)$, which depends on both the unknown $\boldsymbol{p}_k$ and unknown λ_k, seems particularly intractable unless $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ admits a closed-form solution as in Example 2.1. A key observation is that, in fact, $\overline{Z}_K(\alpha, \boldsymbol{p}_0)$ does more generally admit an unbiased estimator, *if* we also introduce an additional assumption on our data-generating mechanism, i.e., that the amount of data is random.

ASSUMPTION 3.1 **(Randomizing Amount of Data)**. *There exists an N such that $\hat{N}_k \sim \text{Poisson}(N\lambda_k)$ for each $k = 1, \ldots, K$.*

Under Assumption 3.1, (unconditional) expectations and probabilities should be interpreted as over both the random draw of $\hat{N}_k$ and the counts $\hat{\boldsymbol{m}}_k$.

Analytically, the benefit of Assumption 3.1 is that it breaks the dependence across i in $\hat{\boldsymbol{m}}_k$. Namely, by the Poisson-splitting property, under Assumption 3.1,

$$\hat{m}_{ki} \sim \text{Poisson}(m_{ki}) \;\; \text{where} \;\; m_{ki} \equiv N\lambda_k p_{ki}, \quad i = 1, \ldots, d, \quad k = 1, \ldots, K,$$

and, furthermore, the $\hat{m}_{ki}$ are independent across i and k. Notice if $\hat{N}_k$ were non-random, these $\hat{m}_{ki}$ would be dependent.

Beyond its analytical convenience, we consider Assumption 3.1 to be reasonable in many applications. Consider for instance a retailer optimizing the price of k distinct products, i.e., x_k represents the price of product k, ξ_k, represents the (random) valuation of a typical customer, and $c_k(x_k,\xi_k)$ is the (negative) profit earned. In such settings, one frequently ties data collection to time, i.e., one might collect $N=6$ months worth of data. To the extent that customers arrive seeking product k in a random fashion, the number of arrivals $\hat{N}_k$ that one might observe in N months is, itself, random, and reasonably modeled as Poisson with rate proportional to N. Similar statements apply whenever data for problem k is generated by an event which occurs randomly, e.g., when observing response time of emergency responders (disasters occur intermittently), effectiveness of a new medical treatment (patients with the relevant disease arrive sequentially), or any aspect of a customer service interaction (customers arrive randomly to service).

In some ways, this perspective tacitly underlies the formulation of Problem (2.2), itself. Indeed, one way to interpret the subproblem weights $\frac{\lambda_k}{K\lambda_{\text{avg}}}=\frac{\lambda_k}{\sum_{j=1}^K\lambda_j}$ is that the decision-maker incurs costs $c_k(x_k,\xi_k)$ at rate λ_k, so that problems of type k contribute a $\frac{\lambda_k}{\sum_{j=1}^K\lambda_j}$ fraction of the total long-run costs. However, if problems of type k occur at rate λ_k, it should be that observations of type k, i.e. realizations of $\boldsymbol{\xi}_k$, also occur at rate λ_k, supporting Assumption 3.1.

In settings where data-collection is not tied to randomly occurring events, modeling $\hat{N}_k$ as Poisson may still be a reasonable approximation if d is large relative to $\hat{N}_k$ and each of the individual p_{ki} are small. Indeed, under such assumptions, a Multinomial$(\hat{N}_k,\boldsymbol{p}_k)$ is well-approximated by independent Poisson random variables with rates $\hat{N}_k p_{ki}$, $i=1,\dots d$ (see McDonald 1980, Deheuvels and Pfeifer 1988 for a formal statement). In this sense, we can view the consequence of Assumption 3.1 as a useful approximation to the setting where $\hat{N}_k$ are fixed, even if it is not strictly true.

In any case, under Assumption 3.1, we develop an unbiased estimate for $\overline{Z}_K(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}})$. We use the following identity (Chen 1975). For any $f:\mathbb{Z}_+\to\mathbb{R}$, for which the expectations exist,

$$W\sim\text{Poisson}(\lambda)\implies\lambda\mathbb{E}[f(W+1)]=\mathbb{E}[Wf(W)]. \tag{3.7}$$

The proof of the identity is immediate from the Poisson probability mass function.[3]

Now, for any $\alpha\geq 0$ and $\boldsymbol{q}\in\Delta_d$, define

$$Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{q})\equiv\frac{1}{N\lambda_{\text{avg}}}\sum_{i=1}^d\hat{m}_{ki}c_{ki}(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i)),\quad\text{and}\quad\overline{Z}_K^{\mathsf{LOO}}(\alpha,\boldsymbol{q})\equiv\frac{1}{K}\sum_{k=1}^K Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0). \tag{3.8}$$

LEMMA 3.1 **(An Unbiased Estimator for $\overline{Z}_K(\alpha,\boldsymbol{p}_0)$).** *Under Assumption 3.1, we have for any $\alpha\geq 0$, and $\boldsymbol{q}\in\Delta_d$ that $\mathbb{E}\left[Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{q})\right]=\mathbb{E}\left[Z_k(\alpha,\boldsymbol{q})\right]$. In particular, $\mathbb{E}\left[\overline{Z}_K^{\mathsf{LOO}}(\alpha,\boldsymbol{q})\right]=\mathbb{E}\left[\overline{Z}_K(\alpha,\boldsymbol{q})\right]$.*

[3] In particular, $\mathbb{E}[Wf(W)]=\sum_{w=0}^\infty wf(w)e^{-\lambda}\frac{\lambda^w}{w!}=\lambda\sum_{w=0}^\infty f(w)e^{-\lambda}\frac{\lambda^{w-1}}{(w-1)!}=\lambda\mathbb{E}[f(W+1)]$.

Proof. Recall that $Z_k(\alpha, \boldsymbol{q}) = \frac{1}{N\lambda_{\text{avg}}} \sum_{i=1}^{d} m_{ki} c_{ki}(\boldsymbol{x}_k(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k))$ and that under Assumption 3.1 $\hat{m}_{ki} \sim \text{Poisson}(m_{ki})$ independently over $i = 1, \ldots, d$. Let $\hat{m}_{k,-i}$ denote $(\hat{m}_{k,j})_{j \neq i}$. Then, by Eq. (3.7),

$$\mathbb{E}\left[m_{ki} c_{ki}(\boldsymbol{x}_k(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k)) \mid \hat{m}_{k,-i}\right] \;=\; \mathbb{E}\left[\hat{m}_{ki} c_{ki}(\boldsymbol{x}_k(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)) \mid \hat{m}_{k,-i}\right].$$

Taking expectations of both sides, summing over $i = 1, \ldots, d$ and scaling by $N\lambda_{\text{avg}}$ proves $\mathbb{E}\left[Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{q})\right] = \mathbb{E}\left[Z_k(\alpha, \boldsymbol{q})\right]$. Finally, averaging this last equality over k completes the lemma. □

We therefore propose selecting α by minimizing the estimate $\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)$. As written, $\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)$ still depends on the unknown N and λ_{avg}, however, these values occur multiplicatively and are positive, and so do not affect the optimizer. Hence, the optimizer is exactly $\alpha_h^{\mathsf{S\text{-}SAA}}$ as in Eq. (3.1).

3.3. Motivating $\alpha^{\mathsf{S\text{-}SAA}}$ via Modified Leave-One-Out Cross-Validation

Although we motivated Eq. (3.1) via an unbiased estimator, we can alternatively motivate it through leave-one-out cross-validation. This latter perspective informs our "LOO" notation above. Indeed, consider again our decision-maker, and assume in line with Assumption 3.1 that subproblems of type k arrive randomly according to a Poisson process with rate λ_k, independently across k. When a problem of type k arrives, she incurs a cost $c_k(\boldsymbol{x}_k, \boldsymbol{\xi})$. Again, the objective of Problem (2.2) thus represents her expected, long-run costs.

We can alternatively represent her costs via the modified cost function $C(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K, \kappa, \boldsymbol{\xi}) = c_\kappa(\boldsymbol{x}_\kappa, \boldsymbol{\xi})$, where κ is a random variable indicating which of the k subproblems she is currently facing. In particular, letting $\mathbb{P}(\kappa = k) = \frac{\lambda_k}{K\lambda_{\text{avg}}}$ and $\mathbb{P}(\boldsymbol{\xi} = a_{ki} \mid \kappa = k) = p_{ki}$, the objective of Problem (2.2) can be more compactly written $\mathbb{E}\left[C(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K, \kappa, \boldsymbol{\xi})\right]$.

Now consider pooling all the data into a single "grand" data set of size $\hat{N}_1 + \cdots + \hat{N}_K$:

$$\left\{(k, \boldsymbol{\xi}_{kj}) : j = 1, \ldots, \hat{N}_k,\, k = 1, \ldots, K\right\}.$$

The grand dataset can be seen as i.i.d. draws of $(\kappa, \boldsymbol{\xi})$.

For a fixed α and $\boldsymbol{p}_0$, the leave-one-out estimate of $\mathbb{E}\left[C(\boldsymbol{x}_1(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}), \ldots, \boldsymbol{x}_K(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}), \kappa, \boldsymbol{\xi})\right]$ is given by removing one data point from the grand data set, training $\boldsymbol{x}_1(\alpha, \boldsymbol{p}_0, \cdot), \ldots, \boldsymbol{x}_K(\alpha, \boldsymbol{p}_0, \cdot)$ on the remaining data, and evaluating $C(\cdot)$ on the left-out point using these policies. We repeat this procedure for each point in the grand data set and average. After some bookkeeping, we can write this leave-one-out estimate as

$$\frac{1}{\sum_{k=1}^{K} \hat{N}_k} \sum_{k=1}^{K} \sum_{i=1}^{d} \hat{m}_{ki} c_{ki}(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)),$$

which agrees with the objective of Eq. (3.1) up to a positive multiplicative constant. Although this multiplicative constant does not affect the choice of $\alpha^{\mathsf{S\text{-}SAA}}$, it *does* cause the traditional leave-one-out estimator to be *biased*. This bias agrees with folklore results in machine learning that assert that leave-one-out does generally exhibit a small bias (Friedman et al. 2001).

For data-driven anchors, we stress that, unlike traditional leave-one-out validation, we do *not* use one fewer points when computing the anchor in Algorithm 1; we use $h(\hat{\boldsymbol{m}})$ for all iterations. Hence, Shrunken-SAA is *not* strictly a leave-one-out procedure, motivating our qualifier "Modified."

4. Performance Guarantees for Shrunken-SAA

In this section, we show that in the limit where the number of subproblems K grows, shrinking by $\alpha_h^{\mathsf{S\text{-}SAA}}$ is essentially best possible. More precisely, for any $K \geq 2$ and any $0 < \delta < 1/2$, with probability at least $1-\delta$, we prove that

$$\mathsf{SubOpt}_{h,K}(\alpha_h^{\mathsf{S\text{-}SAA}}) \leq \tilde{\mathcal{O}}\left(\frac{\log^{\beta}(1/\delta)}{\sqrt{K}}\right), \tag{4.1}$$

where the $\tilde{\mathcal{O}}(\cdot)$ notation suppresses logarithmic factors in K, and $1 < \beta < 2$ is a constant that depends on the particular class of optimization problems under consideration. Imporantly, by Borel-Cantelli lemma, Eq. (4.1) implies $\mathsf{SubOpt}_{h,K}(\alpha_h^{\mathsf{S\text{-}SAA}}) \to 0$, almost surely as $K \to \infty$, *even* if the expected amount of data per subproblem remains fixed.

Equation (4.1) asserts that for a *given* anchor $h(\cdot)$, Shrunken-SAA achieves the best possible shrinkage amount as $K \to \infty$. We will also prove similar bounds on $\mathsf{SubOpt}_{\mathcal{P},K}(\alpha_h^{\mathsf{S\text{-}SAA}}, h_{\mathcal{P}}(\hat{\boldsymbol{m}}))$. Such bounds assert that for a given class $\mathcal{P}$, Shrunken-SAA with $h_{\mathcal{P}}(\cdot)$ achieves the best possible anchor and shrinkage amount *simultaneously*.

4.1. Overview of Proof Technique

To prove performance guarantees like Eq. (4.1), we first bound the sub-optimality of Shrunken-SAA in terms of the maximal stochastic deviations of $\overline{Z}_K(\alpha, h)$ and $\overline{Z}_K^{\mathsf{LOO}}(\alpha, h)$ from their means.

LEMMA 4.1 **(Bounding Sub-Optimality)**. *Suppose Assumption 3.1 holds.*
For a non-data-driven anchor $h(\hat{\boldsymbol{m}}) = \boldsymbol{p}_0$,

$$\mathsf{SubOpt}_{\boldsymbol{p}_0,K}(\alpha_{\boldsymbol{p}_0}^{\mathsf{S\text{-}SAA}}) \leq 2 \underbrace{\sup_{\alpha \geq 0} \left|\overline{Z}_K(\alpha, \boldsymbol{p}_0) - \mathbb{E}\left[\overline{Z}_K(\alpha, \boldsymbol{p}_0)\right]\right|}_{\text{Maximal Stochastic Deviation in } \overline{Z}_K(\cdot, \boldsymbol{p}_0)} + 2 \underbrace{\sup_{\alpha \geq 0} \left|\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0,) - \mathbb{E}\left[\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)\right]\right|}_{\text{Maximal Stochastic Deviation in } \overline{Z}_K^{\mathsf{LOO}}(\cdot, \boldsymbol{p}_0)}.$$

Similarly, for a general data-driven anchor with $h(\hat{\boldsymbol{m}}) \in \mathcal{P}$,

$$\mathsf{SubOpt}_{h,K}(\alpha_h^{\mathsf{S\text{-}SAA}}) \leq 2 \underbrace{\sup_{\substack{\alpha \geq 0 \\ \boldsymbol{q} \in \mathcal{P}}} \left|\overline{Z}_K(\alpha, \boldsymbol{q}) - \mathbb{E}\left[\overline{Z}_K(\alpha, \boldsymbol{q})\right]\right|}_{\text{Maximal Stochastic Deviation in } \overline{Z}_K(\cdot,\cdot)} + 2 \underbrace{\sup_{\substack{\alpha \geq 0 \\ \boldsymbol{q} \in \mathcal{P}}} \left|\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{q}) - \mathbb{E}\left[\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{q})\right]\right|}_{\text{Maximal Stochastic Deviation in } \overline{Z}_K^{\mathsf{LOO}}(\cdot,\cdot)}. \tag{4.2}$$

Finally, for $h = h_{\mathcal{P}}$, $\mathsf{SubOpt}_{\mathcal{P},K}(\alpha_{h_{\mathcal{P}}}^{\mathsf{S\text{-}SAA}}, h_{\mathcal{P}}(\hat{\boldsymbol{m}}))$ is also bounded by the right-hand side of Eq. (4.2).

Proof. By definition of $\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0}$, $\overline{Z}^{\mathsf{LOO}}_K(\alpha^{\mathsf{OR}}_{\boldsymbol{p}_0},\boldsymbol{p}_0)-\overline{Z}^{\mathsf{LOO}}_K(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0},\boldsymbol{p}_0)\geq 0$. Therefore,

$$\begin{aligned}\mathsf{SubOpt}_{\boldsymbol{p}_0,K}(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0}) &\leq \overline{Z}_K(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0},\boldsymbol{p}_0)-\overline{Z}_K(\alpha^{\mathsf{OR}}_{\boldsymbol{p}_0},\boldsymbol{p}_0)+\overline{Z}^{\mathsf{LOO}}_K(\alpha^{\mathsf{OR}}_{\boldsymbol{p}_0},\boldsymbol{p}_0)-\overline{Z}^{\mathsf{LOO}}_K(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0},\boldsymbol{p}_0)\\ &\leq 2\sup_{\alpha\geq 0}\left|\overline{Z}_K(\alpha,\boldsymbol{p}_0)-\overline{Z}^{\mathsf{LOO}}_K(\alpha,\boldsymbol{p}_0)\right|\\ &\leq 2\sup_{\alpha\geq 0}\left|\overline{Z}_K(\alpha,\boldsymbol{p}_0)-\mathbb{E}\overline{Z}_K(\alpha,\boldsymbol{p}_0)\right|+2\sup_{\alpha\geq 0}\left|\overline{Z}^{\mathsf{LOO}}_K(\alpha,\boldsymbol{p}_0)-\mathbb{E}\overline{Z}^{\mathsf{LOO}}_K(\alpha,\boldsymbol{p}_0)\right|\\ &\quad+2\sup_{\alpha\geq 0}\left|\mathbb{E}\overline{Z}_K(\alpha,\boldsymbol{p}_0)-\mathbb{E}\overline{Z}^{\mathsf{LOO}}_K(\alpha,\boldsymbol{p}_0)\right|.\end{aligned}$$

By Lemma 3.1, the last term is zero, which establishes the first statement. The proof of the second statement is similar, but in the second inequality, we take an additional supremum over $\boldsymbol{q}\in\mathcal{P}$ to replace $h(\hat{\boldsymbol{m}})$. The proof of the third statement is similar, using $\overline{Z}^{\mathsf{LOO}}_K(\alpha^{\mathsf{S\text{-}SAA}}_{h_{\mathcal{P}}},h_{\mathcal{P}}(\hat{\boldsymbol{m}}))\leq \overline{Z}^{\mathsf{LOO}}_K(\alpha^{\mathsf{OR}}_{\mathcal{P}},\boldsymbol{q}^{\mathsf{OR}}_{\mathcal{P}})$, and taking a supremum over $\alpha\geq 0$, $\boldsymbol{q}\in\mathcal{P}$ in the second inequality. □

Proving a performance guarantee for $\alpha^{\mathsf{S\text{-}SAA}}_h$ thus reduces to bounding the maximal deviations in the lemma. Recall $\overline{Z}_K(\alpha,\boldsymbol{q})=\frac{1}{K}\sum_{k=1}^K Z_k(\alpha,\boldsymbol{q})$ and $\overline{Z}^{\mathsf{LOO}}_K(\alpha,\boldsymbol{q})=\frac{1}{K}\sum_{k=1}^K Z^{\mathsf{LOO}}_k(\alpha,\boldsymbol{q})$. Both processes have a special form: they are the empirical average of K independent stochastic processes (indexed by k). Fortunately, there exist standard tools to bound the maximal deviations of such empirical processes that rely on bounding their metric entropy.

To keep our paper self-contained, we summarize one such approach presented in Pollard (1990), specifically in Eq. (7.5) of that work. Recall, for any set $S\subseteq\mathbb{R}^d$, the ϵ-packing number of S, denoted by $D(\epsilon,S)$, is the largest number of elements of S that can be chosen so that the Euclidean distance between any two is at least ϵ. Intuitively, packing numbers describe the size of S at scale ϵ.

THEOREM 4.1 **(A Maximal Inequality; Pollard 1990)**. *Let $\mathbf{W}(t)=(W_1(t),\ldots,W_K(t))\in\mathbb{R}^K$ be a stochastic process indexed by $t\in\mathcal{T}$ and let $\overline{W}_K(t)=\frac{1}{K}\sum_{k=1}^K W_k(t)$. Let $\mathbf{F}\in\mathbb{R}^K_+$ be a random variable such that $|W_k(t)|\leq F_k$ for all $t\in\mathcal{T}$, $k=1,\ldots,K$. Finally, define the random variable*

$$J \;\equiv\; J\left(\{\mathbf{W}(t):t\in\mathcal{T}\},\boldsymbol{F}\right) \;\equiv\; 9\|\mathbf{F}\|_2\int_0^1\sqrt{\log D\left(\|\mathbf{F}\|_2 u,\;\left\{\mathbf{W}(t):t\in\mathcal{T}\right\}\right)}du. \tag{4.3}$$

Then, for any $p\geq 1$ and any $0<\delta<1$, with probability at least $1-\delta$,[4]

$$\sup_{t\in\mathcal{T}}\left|\overline{W}_K(t)-\mathbb{E}[\overline{W}_K(t)]\right| \;\leq\; 5^{1/p}\sqrt{p}\|J\|_p K^{-1}\delta^{-1/p}.$$

If $\mathcal{T}$ is finite, one can bound the maximal deviation with a union bound. Theorem 4.1 extends beyond this simple case to cases where $|\mathcal{T}|=\infty$. The random variable $\boldsymbol{F}$ in the theorem is called an *envelope* for the process $\mathbf{W}(t)$. The random variable J is often called the *Dudley integral*. While

[4] Strictly speaking, eq. (7.5) of Pollard (1990) shows that $\mathbb{E}\left[\left|\sup_{t\in\mathcal{T}}\left|\overline{W}_K(t)-\mathbb{E}[\overline{W}_K(t)]\right|\right|^p\right]\leq 2^p C_p^p\mathbb{E}\left[J^p\right]K^{-p}$, for some constant C_p that relates the ℓ_p norm of a random variable and a particular Orlicz norm. In Lemma B.4, we prove that it suffices to take $C_p=5^{1/p}\sqrt{\frac{p}{2e}}$. The result then follows from Markov's Inequality.

packing numbers describe the size of a set at scale ϵ, the Dudley integral roughly describes the size of the set at varying scales. We again refer the reader to Pollard (1990) for discussion.

Our overall proof strategy is to use Theorem 4.1 to bound the two suprema in Lemma 4.1, and thus obtain a bound on the sub-optimality. Specifically, define the following stochastic processes:

$$\mathbf{Z}(\alpha,\boldsymbol{q}) = (Z_1(\alpha,\boldsymbol{q}),\ldots,Z_K(\alpha,\boldsymbol{q})), \qquad \mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) = (Z_1^{\mathsf{LOO}}(\alpha,\boldsymbol{q}),\ldots,Z_K^{\mathsf{LOO}}(\alpha,\boldsymbol{q})).$$

Our proof strategy will be to 1) Compute envelopes for both processes 2) Compute the packing numbers and Dudley integrals for the relevant sets above 3) Apply Theorem 4.1 to bound the relevant maximal deviations and 4) Use these bounds in Lemma 4.1 to bound the sub-optimality. We execute this strategy for several special cases in the remainder of the section.

As a first step, we identify envelopes for each process. We restrict attention to the case where the optimal value of each subproblem is bounded for any choice of anchor and shrinkage.

Assumption 4.1 **(Bounded Optimal Values)**. *There exists C such that for all $i=1,\ldots,d$, and $k=1\ldots,K$, $\sup_{\boldsymbol{q}\in\Delta_d}|c_{ki}(\boldsymbol{x}_k(\infty,\boldsymbol{q}))|\le C$.*

Notice that $\sup_{\alpha\ge 0,\ \boldsymbol{q}\in\Delta_d}|c_{ki}(\boldsymbol{x}_k(\alpha,\boldsymbol{q}))| = \sup_{\boldsymbol{q}\in\Delta_d}|c_{ki}(\boldsymbol{x}_k(\infty,\boldsymbol{q})|$, so that the assumption bounds the optimal value associated to every policy. Assumption 4.1 is a mild assumption, and follows for example if $c_{ki}(\cdot)$ is continuous and $\mathcal{X}_k$ is compact. However, the assumption also holds, e.g, if $c_{ki}(\cdot)$ is unbounded but coercive. With it, we can easily compute envelopes. Recall, $\hat{N}_{\max}\equiv\max_k \hat{N}_k$.

Lemma 4.2 **(Envelopes for $\mathbf{Z},\mathbf{Z}^{\mathsf{LOO}}$)**. *Under Assumption 4.1,*

1. *The vector $\mathbf{F}^{\mathsf{Perf}}\equiv C\boldsymbol{\lambda}/\lambda_{\text{avg}}$ is an envelope for $\mathbf{Z}(\alpha,\boldsymbol{q})$ with $\|\mathbf{F}^{\mathsf{Perf}}\|_2 = \frac{C}{\lambda_{\text{avg}}}\|\boldsymbol{\lambda}\|_2$.*
2. *The random vector $\mathbf{F}^{\mathsf{LOO}} = C\frac{\hat{\boldsymbol{N}}}{N\lambda_{\text{avg}}}$ is an envelope for $\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q})$ with $\|\mathbf{F}^{\mathsf{LOO}}\|_2 = \frac{C}{N\lambda_{\text{avg}}}\|\hat{\boldsymbol{N}}\|_2$.*

The proof is immediate from the definitions and omitted.

Our next step is to bound the packing numbers (and Dudley integrals) for the sets $\{\mathbf{Z}(\alpha,\boldsymbol{p}_0):\alpha\ge 0\}\subseteq\mathbb{R}^K$, and $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0):\alpha\ge 0\right\}\subseteq\mathbb{R}^K$, for the case of fixed anchors and the sets $\{\mathbf{Z}(\alpha,\boldsymbol{q}):\alpha\ge 0,\ \boldsymbol{q}\in\mathcal{P}\}\subseteq\mathbb{R}^K$, and $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}):\alpha\ge 0,\ \boldsymbol{q}\in\mathcal{P}\right\}\subseteq\mathbb{R}^K$, for the case of data-driven anchors. Bounding these packing numbers is subtle and requires exploiting the specific structure of the optimization problem (2.2). We *separately* consider two general classes of optimization problems – strongly convex optimization problems and discrete optimization problems – in the remainder. Although we focus on these classes, we expect a similar proof strategy and technique might be employed to attack other classes of optimization problems.

Remark 4.1 (Performance of $\alpha^{\mathsf{S\text{-}SAA}}$ in the Large-Sample Regime). Although we focus on performance guarantees for $\alpha^{\mathsf{S\text{-}SAA}}$ in settings where K is large and the expected amount of data per problem is fixed, one could also ask how $\alpha^{\mathsf{S\text{-}SAA}}$ performs in the large-sample regime, i.e., where

K is fixed and $\hat{N}_k \to \infty$ for all k. Using similar techniques, i.e., reducing the problem to bounding a certain maximal stochastic deviation, one can show that $\boldsymbol{x}_k(\alpha^{\mathsf{S\text{-}SAA}}, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ performs comparably to the full-information solution in Problem (2.2) in this limit. The proof uses somewhat standard arguments for empirical processes. Moreover, the result is perhaps unsurprising; many data-driven methods converge to full-information performance in the large-sample regime (see, e.g., Kleywegt et al. (2002) for the case of SAA) since $\hat{\boldsymbol{p}}_k$ is consistent for $\boldsymbol{p}_k$ for all k in this regime. Consequently, we focus on the small-data, large-scale regime, where Shrunken SAA enjoys strong suboptimality guarantees not enjoyed by SAA. This small-data, large-scale focus, however, causes the N dependence in our bounds to be looser than that obtained from a direct large-sample analysis. Developing a unified analysis of data-pooling for *any* sequence of N, K remains an open question. □

4.2. Fixed Anchors and Strongly-Convex Optimization Problems

In this section, we treat the case where the K subproblems are smooth enough so that $\boldsymbol{x}_k(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k)$ is smooth in α and $\boldsymbol{q}$ for each k. Specifically, in this section we assume:

ASSUMPTION 4.2 **(Lipschitz, Strongly-Convex Optimization)**. *There exists L, γ such that $c_{ki}(\boldsymbol{x})$ are γ-strongly convex and L-Lipschitz over $\mathcal{X}_k$, and, moreover, $\mathcal{X}_k$ is non-empty and convex, for all $k = 1, \ldots, K$, and $i = 1, \ldots, d$.*

THEOREM 4.2 **(Shrunken-SAA with Fixed Anchors for Strongly Convex Problems)**. *Fix any $\boldsymbol{p}_0$. Suppose Assumptions 3.1, 4.1 and 4.2 hold, $K \geq 2$ and $N\lambda_{\min} \geq 1$. Then, there exists a universal constant* A *such that for any $0 < \delta < 1/2$, with probability at least $1 - \delta$, we have that*

$$\mathsf{SubOpt}_{\boldsymbol{p}_0,K}(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0}) \;\leq\; \mathrm{A} \cdot \max\left(C, L\sqrt{\frac{C}{\gamma}}\right) \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \cdot \frac{\log^2(1/\delta) \cdot \log^{3/2}(K)}{\sqrt{K}}.$$

The proof follows our strategy from Section 4.1. (See Appendix C.1.) We sketch the main ideas:

We first bound the packing numbers of $\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}$ and $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\right\}$. The key observation is that since the subproblems are strongly-convex, the optimal solutions $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ are continuous as functions of α. We utilize this continuity to construct a packing.

Specifically, consider $\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}$. Continuity in α implies that by evaluating $\boldsymbol{x}(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ on a sufficiently dense grid of α's, we can construct a covering of $\left\{(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))_{k=1}^K : \alpha \geq 0\right\}$, which in turn yields a covering of $\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}$. By carefully choosing the initial grid of α's, we can ensure that this last covering is a valid $(\epsilon/2)$-covering. By (Pollard 1990, pg. 10), the size of this covering bounds the ϵ-packing number as desired. Figure 2 illustrates this intuition and further argues the initial grid of α's should be of size $\mathcal{O}(1/\epsilon^2)$.

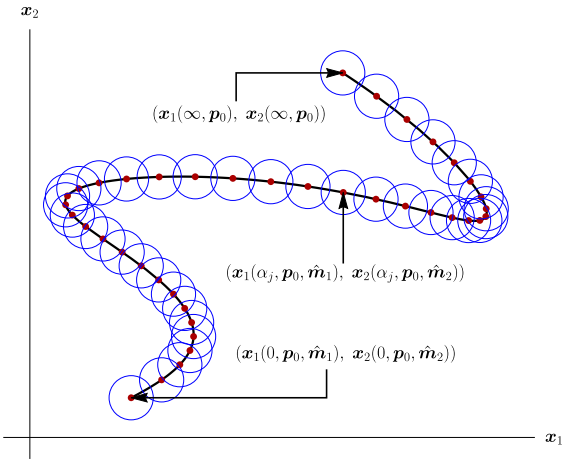


Figure 2 **Covering a continuous process.** The set $\{(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))_{k=1}^K : \alpha \geq 0\}$ can be thought of as a parametric curve indexed by α in the space $\prod_{k=1}^K \mathcal{X}_k$. Because of the continuity in α (cf. Lemma C.1, part iii)), to cover this curve for any compact set $\alpha \in [0, \alpha_{\max}]$ requires $\mathcal{O}(1/\epsilon)$ balls of size ϵ. Because of the continuity at $\alpha = \infty$ (cf. Lemma C.1, part iv)), it suffices to take $\alpha_{\max} = \mathcal{O}(1/\epsilon)$. This yields a packing number bound of $\mathcal{O}(1/\epsilon^2)$ (cf. Lemma C.2).

A similar argument holds for $D(\epsilon, \{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\})$, using a grid of α's to cover $\{(x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i) : i = 1, \ldots, d,\ k = 1, \ldots, K) : \alpha \geq 0\}$. The packing is also of size $\mathcal{O}(1/\epsilon^2)$.

To complete the proof, we use these packing numbers in Theorem 4.1 to bound the maximal deviations of $\overline{Z}_K(\cdot, \boldsymbol{p}_0), \overline{Z}_K^{\mathsf{LOO}}(\cdot, \boldsymbol{p}_0)$. Substituting into Lemma 4.1 proves Theorem 4.2 above. Again, please see Appendix C.1 for details.

4.3. Data-Driven Anchors and Strongly Convex Problems

We next consider the case of a data-driven anchor $h(\hat{\boldsymbol{m}}) \in \mathcal{P}$. Our performance guarantees will depend on the complexity of $\mathcal{P}$ as measured by the size of its ℓ_1-packing numbers. Namely, we let $D_1(\epsilon, \mathcal{P})$ be the largest number of elements of $\mathcal{P}$ that can be chosen so that the ℓ_1-distance between any two is at least ϵ.[5] Then,

THEOREM 4.3. **(Shrunken-SAA with Data-Driven Anchors for Strongly Convex Problems)** *Suppose Assumptions 3.1, 4.1 and 4.2 hold, $K \geq 2$. Let $d_0 \geq 1$ be such that for any $0 < \epsilon < 1/2$, $\log D_1(\epsilon, \mathcal{P}) \leq d_0 \log(1/\epsilon)$. Then, there exists a universal constant* A *such that for any $0 < \delta < 1/2$, with probability at least $1 - \delta$, we have that*

$$\mathsf{SubOpt}_{h,K}(\alpha_h^{\mathsf{S\text{-}SAA}}) \ \leq\ \mathrm{A} \cdot \max\left(C,\ \frac{L^2}{\gamma} \ +\ L\sqrt{\frac{C}{\gamma}}\right)\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \frac{d_0^2 \log^{7/2}(K) \log^2(1/\delta)}{\sqrt{K}}.$$

In the special case of $h_{\mathcal{P}}(\cdot)$, we can prove an even stronger result, i.e., that Shrunken-SAA with $h_{\mathcal{P}}$ performs comparably to pooling in an optimal way to the best anchor within the class $\mathcal{P}$.

THEOREM 4.4 **(Shrunken-SAA with $h_{\mathcal{P}}$ for Strongly Convex Problems)**. *Under the assumptions of Theorem 4.3, there exists a universal constant* A *such that for any $0 < \delta < 1/2$, with probability at least $1 - \delta$, we have that*

$$\mathsf{SubOpt}_{\mathcal{P},K}(\alpha_{h_{\mathcal{P}}}^{\mathsf{S\text{-}SAA}}, h_{\mathcal{P}}(\hat{\boldsymbol{m}})) \ \leq\ \mathrm{A} \cdot \max\left(C,\ \frac{L^2}{\gamma} \ +\ L\sqrt{\frac{C}{\gamma}}\right)\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \frac{d_0^2 \log^{7/2}(K) \log^2(1/\delta)}{\sqrt{K}}.$$

[5] Recall $D(\epsilon, S)$ is defined with respect to ℓ_2-distance.

In both theorems, the constant d_0 measures the complexity of $\mathcal{P}$. Without loss of generality, $d_0 \leq 3d$ since $\mathcal{P} \subseteq \Delta_d$ and $\log D_1(\epsilon, \Delta_d) \leq 3d \log(1/\epsilon)$ (Pollard 1990, Lemma 4.1). In practice, we might choose flexible, parametric families for $\mathcal{P}$ with small d_0 that do not scale with d. An example might be when $\mathcal{P}$ consists of all (truncated) Poisson distributions with mean at most Λ, in which case one can take $d_0 = 2\max(1, \log(\Lambda))$, *independently* of d (and the truncation). Another example is given in Section 6 using Beta-distributions. In general, we expect that our performance bounds must depend on the complexity of $\mathcal{P}$ in some way, because we impose no assumptions on the function $h(\hat{\boldsymbol{m}})$ that selects the anchor, and, hence, must control behavior across all of $\mathcal{P}$.

Both proofs follow the strategy of Section 4.1 (see Appendix C.2). The key idea to bounding the packing numbers is again to leverage continuity and cover the set $\{(\alpha, \boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \mathcal{P}\}$. Since both proofs leverage Lemma 4.1, the right hand sides of the bounds are the same.

By contrast, the left-hand sides of Theorems 4.3 and 4.4 are different: the first measures suboptimality relative to an oracle with a pre-specified anchor, while the second is relative to an oracle that can optimize the choice of anchor. This distinction mirrors the difference between "estimate-then-optimize" procedures and those which choose parameters in an optimization-aware fashion. Continuing our example where $\mathcal{P}$ is a set of Poisson distributions, Theorem 4.3 bounds the suboptimality of Shrunken-SAA when using (all) the data to fit a Poisson distribution without regard to the downstream optimization, e.g., by maximum likelihood, and then choosing α and $\boldsymbol{x}_k(\cdot)$ to optimize. By contrast, Theorem 4.4 bounds the performance of Shrunken-SAA when choosing the anchor, α and $\boldsymbol{x}_k(\cdot)$ simultaneously to optimize the downstream optimization.

4.4. Fixed Anchors and Discrete Optimization Problems

In this section we consider the case where the K subproblems are discrete optimization problems. Specifically, we require $|\mathcal{X}_k| < \infty$ for each $k = 1, \ldots, K$. This encompasses, e.g., binary linear or non-linear optimization and linear optimization over a polytope, since we may restrict to its vertices.

Unlike the case of strongly convex problems, the optimization defining $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ (cf. Eq. (2.4)) may admit multiple optima, and hence, $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ requires a tie-breaking rule. For our results below, we assume this tie-breaking rule is consistent in the sense that if the set of minimizers to Eq. (2.4) is the same for two distinct values of $(\alpha, \boldsymbol{p}_0)$, then the tie-breaking minimizer is also the same for both. We express this requirement by representing the tie-breaking rule as a function from a set of minimizers to a chosen minimizer:

ASSUMPTION 4.3 **(Consistent Tie-Breaking)**. *For each k, there exists $\sigma_k : 2^{\mathcal{X}_k} \to \mathcal{X}_k$ such that*

$$\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) = \sigma_k\left(\arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \ \hat{\boldsymbol{p}}_k(\alpha)^\top \boldsymbol{c}_k(\boldsymbol{x}_k)\right).$$

Then,

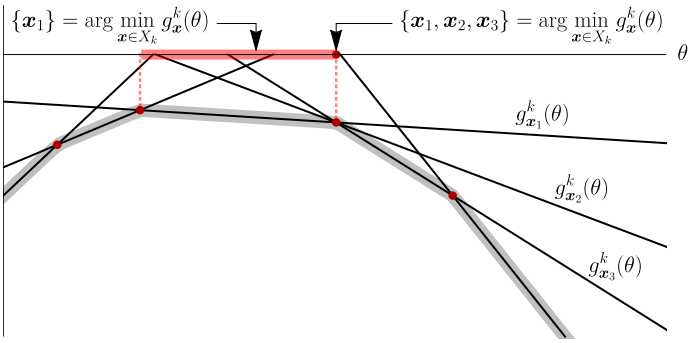


Figure 3 **Counting Discrete Solutions.** A concave piecewise-linear function consisting of $|\mathcal{X}_k|$ lines has at most $|\mathcal{X}_k|-1$ breakpoints, between which the set of active supporting lines is constant. Any function of this set of active supporting lines is piecewise constant with at most $|\mathcal{X}_k|-1$ discontinuities.

THEOREM 4.5 **(Shrunken-SAA with Fixed Anchors for Discrete Problems)**. *Suppose that $|\mathcal{X}_k|<\infty$ for each k, $K\geq 2$, and that Assumptions 3.1, 4.1 and 4.3 hold. Then, there exists a universal constant* A *such that for any $0<\delta<1/2$ we have that, with probability at least $1-\delta$,*

$$\mathsf{SubOpt}_{\boldsymbol{p}_0,K}(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0}) \;\leq\; \mathrm{A}\cdot C\frac{\lambda_{\max}}{\lambda_{\min}}\cdot\sqrt{\log\left(2N_{\max}\sum_{k=1}^{K}|\mathcal{X}_k|\right)}\cdot\frac{\log^{3/2}(K)\cdot\log^{3/2}(1/\delta)}{\sqrt{K}}.$$

We stress that $|\mathcal{X}_k|$ occurs logarithmically in the bound, so that the bound is reasonably tight even when the number of feasible solutions per subproblem may be large. For example, consider binary optimization. Then, $|\mathcal{X}_k|$ often scales exponentially in the number of binary variables, so that $\log(|\mathcal{X}_k|)$ scales like the number of binary variables. Thus, as long as the number of binary variables per subproblem is much smaller than K, the sub-optimality will be small with high probability.

We also note that, unlike Theorem 4.2, the above bound depends on $\log(N_{\max})$. This mild dependence stems from the fact that we have made *no assumptions of continuity* on the functions $\boldsymbol{c}_k(\boldsymbol{x},\boldsymbol{\xi})$ in $\boldsymbol{x}$ or $\boldsymbol{\xi}$. Since these functions could be arbitrarily non-smooth, we need to control their behavior separately across all of the LOO iterations, which introduces the $N_{\max}$ dependence. With stronger assumptions, it might be possible to remove this dependence. However, since we are mostly interested in the setting where N_k is moderate to small for all k, we do not pursue this idea.

To prove Theorem 4.5, we again follow the approach outlined in Section 4.1. Since the policy $\boldsymbol{x}(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}})$ need not be smooth in α, however, we adopt a different strategy than in Section 4.2. Specifically, we bound the cardinality of $\{\mathbf{Z}(\alpha,\boldsymbol{p}_0):\alpha\geq 0\}$, $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0):\alpha\geq 0\right\}$, directly. (Recall that the cardinality of a set bounds its ϵ-packing number for any ϵ.)

First note the cardinality of $\{\mathbf{Z}(\alpha,\boldsymbol{p}_0):\alpha\geq 0\}$ is at most that of $\left\{(\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k))_{k=1}^K:\alpha\geq 0\right\}$. A trivial bound on this latter set's cardinality is $\prod_{k=1}^K|\mathcal{X}_k|$. This bound is too crude for our purposes; it grows exponentially in K even if $|\mathcal{X}_k|$ is bounded for all k. Intuitively, this bound is crude because it supposes we can vary each solution $\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)$ independently of the others to achieve all $\prod_{k=1}^K|\mathcal{X}_k|$ possible combinations. In reality, we can only vary a single parameter, α, that

simultaneously controls all K solutions, rather than varying them separately. We use this intuition to show that a much smaller bound, i.e., $2\sum_{k=1}^K |\mathcal{X}_k|$, is valid.

To this end, we fix k and study the dependence of $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ on α. In the trivial case $\hat{N}_k = 0$, $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ takes only one value: $\boldsymbol{x}_k(\infty, \boldsymbol{p}_0)$. Hence we focus on the case $\hat{N}_k \geq 1$.

Consider reparameterizing the solution in terms of $\theta = \frac{\alpha}{\alpha+\hat{N}_k} \in [0,1)$ and let $\alpha(\theta) = \frac{\theta}{1-\theta}\hat{N}_k$. Then for any $\boldsymbol{x} \in \mathcal{X}_k$, define the linear function

$$g_{k\boldsymbol{x}}(\theta) = ((1-\theta)\hat{\boldsymbol{p}}_k + \theta \boldsymbol{p}^0)^\top \boldsymbol{c}_k(\boldsymbol{x}), \qquad \theta \in [0,1).$$

Since $g_{k\boldsymbol{x}}(\cdot)$ is linear, the function $\theta \mapsto \min_{\boldsymbol{x}\in\mathcal{X}_k} g_{k\boldsymbol{x}}(\theta)$ is concave, piecewise-linear with at most $|\mathcal{X}_k| - 1$ breakpoints. By construction, $\boldsymbol{x}_k(\alpha(\theta), \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) \in \arg\min_{\boldsymbol{x}_k\in\mathcal{X}_k} g_{k\boldsymbol{x}}(\theta)$. More precisely, for any θ, the set of active supporting hyperplanes of $\min_{\boldsymbol{x}\in\mathcal{X}_k} g_{k\boldsymbol{x}}(\cdot)$ at θ is $\left\{(\boldsymbol{p}^0 - \hat{\boldsymbol{p}}_k)^\top \boldsymbol{c}_k(\boldsymbol{x}) \;:\; \boldsymbol{x} \in \arg\min_{\boldsymbol{x}_k\in\mathcal{X}_k} g_{k\boldsymbol{x}}(\theta)\right\}$.

Since the set of active supporting hyperplanes is constant between breakpoints, the set of minimizers $\arg\min_{\boldsymbol{x}_k\in\mathcal{X}_k} g_{k\boldsymbol{x}}(\theta)$ is also constant between breakpoints. By Assumption 4.3, this implies $\theta \mapsto \boldsymbol{x}_k(\alpha(\theta), \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ is piecewise constant with at most $|\mathcal{X}_k| - 1$ points of discontinuity. (See also Fig. 3.) Viewed in the original parameterization in terms of α, it follows that $\alpha \mapsto \boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ is also piecewise constant with at most $|\mathcal{X}_k| - 1$ points of discontinuity. Thus,

LEMMA 4.3. *Suppose Assumption 4.3 holds. Fix any $\boldsymbol{p}_0$ and $\hat{\boldsymbol{m}}_k$. Then, the function $\alpha \mapsto \boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ is piecewise constant with at most $|\mathcal{X}_k| - 1$ points of discontinuity.*

Taking the union of all these points of discontinuity over k proves that $(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))_{k=1}^K$ is also piecewise constant with at most $\sum_{k=1}^K (|\mathcal{X}_k| - 1)$ points of discontinuity. Therefore, it takes at most $2\sum_{k=1}^K |\mathcal{X}_k| - 2K + 1$ different values – a distinct value for each of the $\sum_{k=1}^K (|\mathcal{X}_k| - 1)$ breakpoints plus a distinct value for the $\sum_{k=1}^K (|\mathcal{X}_k| - 1) + 1$ regions between breakpoints. This gives the desired cardinality bound on $|\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}|$. A similar argument considering the larger $(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i))_{i\in\mathcal{I}_k, k=1,\ldots,K}$, where $\mathcal{I}_k = \{i = 1, \ldots, d : \hat{m}_{ki} > 0\}$, gives a corresponding cardinality bound on $\left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\right\}\right|$. Noting $|\mathcal{I}_k| \leq \min(d, \hat{N}_k)$ gives the following (proof omitted):

COROLLARY 4.1 **(Size of Discrete Solutions Sets)**. *Suppose Assumption 4.3 holds. Then,*

$$|\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}| \;\leq\; 2\textstyle\sum_{k=1}^K |\mathcal{X}_k|, \quad \left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\right\}\right| \;\leq\; 1 + 2\textstyle\sum_{k=1}^K \min(d, \hat{N}_k)\, |\mathcal{X}_k|.$$

The additional "1" in the case of $\left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\right\}\right|$ covers the case where $\hat{N}_{\max} = 0$ and $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\right\} = \{\mathbf{0}\}$. Although these bounds may appear large, an important feature is that they are only linear in K as long as $|\mathcal{X}_k|$ are bounded over k.

We use these cardinality bounds to bound the packing numbers and then apply our usual strategy via Theorem 4.1 and Lemma 4.1 to prove Theorem 4.5. The details are in Appendix C.3.

4.5. Data-Driven Anchors and Discrete Optimization Problems

We next extend the results of Section 4.4 to the case of a data-driven anchor, $h(\hat{\boldsymbol{m}})$. As in Section 4.3, our bounds will depend on a measure of complexity of $\mathcal{P}$, namely, the dimension of $\text{span}(\mathcal{P}) \equiv \{\sum_{\ell=1}^{d} \theta_\ell \boldsymbol{q}_\ell \ : \theta_\ell \in \mathbb{R},\ \boldsymbol{q}_\ell \in \mathcal{P},\ \ell = 1,\ldots,d\}$ when viewed as a linear subspace. Denote this dimension by d_0 and note $1 \le d_0 \le d$. A canonical example might be when $\mathcal{P}$ consists of mixture distributions with d_0 (specified) components. We prove that:

THEOREM 4.6 **(Shrunken-SAA with Data-Driven Anchors for Discrete Problems)**. *Suppose that $|\mathcal{X}_k| < \infty$ for each k, that $\text{span}(\mathcal{P})$ has dimension d_0, and that Assumptions 4.1 and 4.3 hold. Then, there exists a universal constant A such that for all $0 < \delta < 1/2$, we have that, with probability at least $1-\delta$,*

$$\mathsf{SubOpt}_{h,K}(\alpha_h^{\mathsf{S\text{-}SAA}}) \ \le \ \mathrm{A} \cdot C \frac{\lambda_{\max}}{\lambda_{\min}} \sqrt{d_0 \log\left(N_{\max} \sum_{k=1}^{K} |\mathcal{X}_k|\right)} \cdot \frac{\log^{3/2}(K)\log^2(1/\delta)}{\sqrt{K}}.$$

THEOREM 4.7 **(Shrunken-SAA with $h_{\mathcal{P}}$ for Discrete Problems)**. *Under the assumptions of Theorem 4.6, there exists a universal constant A such that for any $0 < \delta < 1/2$, with probability at least $1-\delta$, we have that*

$$\mathsf{SubOpt}_{\mathcal{P},K}(\alpha_{h_{\mathcal{P}}}^{\mathsf{S\text{-}SAA}}, h_{\mathcal{P}}(\hat{\boldsymbol{m}})) \ \le \ \mathrm{A} \cdot C \frac{\lambda_{\max}}{\lambda_{\min}} \sqrt{d_0 \log\left(N_{\max} \sum_{k=1}^{K} |\mathcal{X}_k|\right)} \cdot \frac{\log^{3/2}(K)\log^2(1/\delta)}{\sqrt{K}}.$$

Both proofs follow the strategy from Section 4.1 (see Appendix C.4) and, hence, lead to the same right hand sides. However, the left hand sides are distinct. We sketch the main ideas of the proof:

We first bound the cardinality of $\{\mathbf{Z}(\alpha,\boldsymbol{q}) : \alpha \ge 0, \boldsymbol{q} \in \mathcal{P}\}$, $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) : \alpha \ge 0, \boldsymbol{q} \in \mathcal{P}\right\}$. The key is to generalize the argument of Section 4.4 from counting breakpoints in a univariate piecewise affine function to counting the pieces in a multivariate piecewise affine function. First, we reparameterize our policies. Let the columns of $\boldsymbol{V} \in \mathbb{R}^{d \times d_0}$ be a basis of $\text{span}(\mathcal{P})$. Then, intrepreting $\mathbf{0}/0$ as an arbitrary point in Δ_d (e.g., $\boldsymbol{e}/d$),

$$\begin{aligned}
|\{\mathbf{Z}(\alpha,\boldsymbol{q}) : \alpha \ge 0, \boldsymbol{q} \in \mathcal{P}\}| \ &\le \ \left|\left\{(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k))_{k=1}^K \ : \ \boldsymbol{q} \in \mathcal{P}, \alpha \ge 0\right\}\right| \\
&\le \ \left|\left\{(\boldsymbol{x}_k(\|\boldsymbol{w}\|_1, \boldsymbol{w}/\|\boldsymbol{w}\|_1, \hat{\boldsymbol{m}}_k))_{k=1}^K \ : \ \boldsymbol{w} \in \text{span}(\mathcal{P}) \cap \mathbb{R}^d_+\right\}\right| \\
&= \ \left|\left\{(\boldsymbol{x}_k(\|\boldsymbol{V}\boldsymbol{\theta}\|_1, \boldsymbol{V}\boldsymbol{\theta}/\|\boldsymbol{V}\boldsymbol{\theta}\|_1, \hat{\boldsymbol{m}}_k))_{k=1}^K \ : \ \boldsymbol{\theta} \in \mathbb{R}^{d_0}, \boldsymbol{V}\boldsymbol{\theta} \in \mathbb{R}^d_+\right\}\right|. \qquad (4.4)
\end{aligned}$$

Hence, it suffices to bound the right most side of Eq. (4.4). An advantage of this $\boldsymbol{\theta}$-parameterization over the original $(\alpha,\boldsymbol{q})$-parameterization is that, for $\hat{N}_k > 0$,

$$\boldsymbol{x}_k(\|\boldsymbol{V}\boldsymbol{\theta}\|_1, \boldsymbol{V}\boldsymbol{\theta}/\|\boldsymbol{V}\boldsymbol{\theta}\|_1, \hat{\boldsymbol{m}}_k) \in \arg\min_{\boldsymbol{x} \in \mathcal{X}_k} \ (\boldsymbol{V}\boldsymbol{\theta} + \hat{\boldsymbol{m}}_k)^\top \boldsymbol{c}_k(\boldsymbol{x}), \qquad (4.5)$$

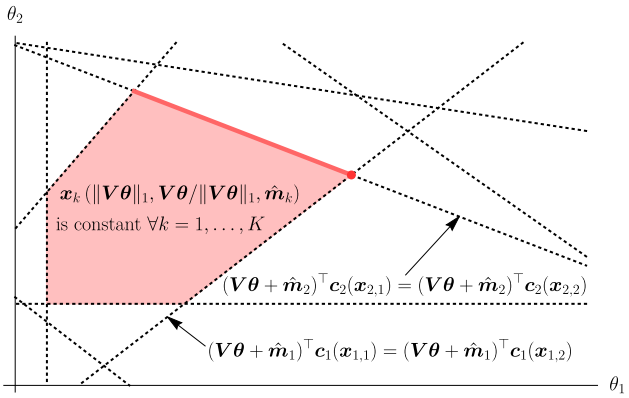


Figure 4 **Solution Induced Hyperplane Arrangement.** The hyperplanes H_{kij} (cf. Eq. (4.6)) in $\mathbb{R}^d$ are indifference curves between solutions $\boldsymbol{x}_{ki}$ and $\boldsymbol{x}_{kj}$ in Eq. (4.5). The total ordering on each set $\mathcal{X}_k$ induced by the objective of Eq. (4.5) is thus constant on the interior of the fully-specified polyhedra defined by the hyperplanes.

and $\boldsymbol{\theta}$ occurs linearly in this representation.

The set of $\boldsymbol{\theta}$ where we are indifferent between $\boldsymbol{x}_{ki}, \boldsymbol{x}_{kj} \in \mathcal{X}_k$ in Eq. (4.5) is the hyperplane

$$H_{kij} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{d_0} : \left(\boldsymbol{V\theta} + \hat{\boldsymbol{m}}_k\right)^\top \left(\boldsymbol{c}_k(\boldsymbol{x}_{ki}) - \boldsymbol{c}_k(\boldsymbol{x}_{kj})\right) = \boldsymbol{0} \right\}. \tag{4.6}$$

Consider drawing all $\sum_{k=1}^K \binom{|\mathcal{X}_k|}{2}$ such hyperplanes, as in Fig. 4. Then, for any $\boldsymbol{\theta} \in \mathbb{R}^{d_0}$, consider the polyhedron given by the equality constraints of those hyperplanes containing $\boldsymbol{\theta}$, and the inequality constraints defined by the side on which $\boldsymbol{\theta}$ lies for the remaining hyperplanes. The relative ordering of $\{(\boldsymbol{V\theta} + \hat{\boldsymbol{m}}_k)^\top \boldsymbol{c}_k(\boldsymbol{x}_k) : \boldsymbol{x}_k \in \mathcal{X}_k\}$ is constant for all $\boldsymbol{\theta}$ in this polyhedron's interior. Hence, $(\boldsymbol{x}_k(\|\boldsymbol{V\theta}\|_1, \boldsymbol{V\theta}/\|\boldsymbol{V\theta}\|_1, \hat{\boldsymbol{m}}_k))_{k=1}^K$ is also constant. Thus, to bound $\{\mathbf{Z}(\alpha, \boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \mathcal{P}\}$, it suffices to count the number of such polyhedra. We do this counting in Appendix C.4. A similar argument (with a different hyperplane arrangement) can be used to bound the cardinality of $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \mathcal{P}\right\}$. We summarize the results as:

Lemma 4.4 **(Size of Discrete Solutions Sets).** *Under the assumptions of Theorem 4.6,*

$$\left|\{\mathbf{Z}(\alpha, \boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \mathcal{P}\}\right| \leq \left(\sum_{k=1}^K |\mathcal{X}_k|^2\right)^{d_0}, \quad \left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \mathcal{P}\right\}\right| \leq 1 + \hat{N}_{\max}^{d_0} \left(\sum_{k=1}^K |\mathcal{X}_k|^2\right)^{d_0}.$$

Importantly, both bounds are polynomial in K if $|\mathcal{X}_k|$ are bounded over k. We then apply Theorem 4.1 to bound the maximal deviations in Lemma 4.1, proving the theorems. Again, see Appendix C.4 for details.

4.6. Performance Guarantees for Continuous Distributions

Notice that none of our previous theorems (cf. Theorems 4.2 to 4.7) depend explicitly on d, the size of the support of $\boldsymbol{p}_k$. Recall also that Algorithm 1 does not depend on d. These observations beg the question of whether similar performance guarantees hold for Shrunken-SAA when $\boldsymbol{\xi}_k$ are not discrete with finite support.

For the case of strongly-convex optimization problems, the short answer is "yes." One simply applies Algorithm 1 as written to the potentially continous $\boldsymbol{\xi}_k$, but *analyzes* a discretized system

where the discretization is chosen sufficiently fine that the two systems behave similarly. The details are somewhat tedious. See Appendix F in the appendix for a formal statement and proof.

Unfortunately, for the case of discrete optimization problems, the answer is more subtle, and it is not clear that similar performance guarantees hold without additional assumptions. Again, see Appendix F for a discussion of the key issues.

5. The Sub-Optimality-Stability Tradeoff: An Intuition for Data-Pooling

In the previous section, we established that for various classes of optimization problems, Shrunken SAA pools the data in the best possible way for a given anchor, or, when used with $h_{\mathcal{P}}$, pools the data in the best possible way to the best-in-class anchor, asymptotically as $K \to \infty$. In this section, we show how Shrunken SAA can also be used to build a strong intuition into *when* and *why* data-pooling improves upon decoupling.

We focus first on the case of a non-data-driven anchor $\boldsymbol{p}_0$ for simplicity. Lemma 3.1 shows that (under Assumption 3.1) $\mathbb{E}\left[\overline{Z}_K(\alpha, \boldsymbol{p}_0)\right] = \mathbb{E}\left[\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)\right]$. Theorems 4.2 and 4.5 establish that under mild conditions, we often have the stronger statement

$$\underbrace{\overline{Z}_K(\alpha, \boldsymbol{p}_0)}_{\text{True Performance of } \alpha} \quad = \quad \underbrace{\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)}_{\text{LOO Performance of } \alpha} \quad + \quad \underbrace{\tilde{\mathcal{O}}_p(1/\sqrt{K})}_{\text{Stochastic Error}},$$

where the error term is uniformly small in α. In these two senses, optimizing $\overline{Z}_K(\alpha, \boldsymbol{p}_0)$ over α is roughly equivalent to optimizing $\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)$ over α, especially for large K.

A simple algebraic manipulation then shows that

$$\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) \;=\; \tfrac{1}{N\lambda_{\text{avg}}}\big(\text{SAA-SubOpt}(\alpha) \;+\; \text{Instability}(\alpha) \;+\; \text{SAA}(0)\big),$$

$$\begin{aligned}
\text{where} \qquad \text{SAA-SubOpt}(\alpha) &\equiv \tfrac{1}{K}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{d} \hat{m}_{ki}\Big(c_{ki}\big(x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)\big) - c_{ki}\big(x_k(0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)\big)\Big) \\
\text{Instability}(\alpha) &\equiv \tfrac{1}{K}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{d} \hat{m}_{ki}\Big(c_{ki}\big(x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)\big) - c_{ki}\big(x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)\big)\Big), \\
\text{SAA}(0) &\equiv \tfrac{1}{K}\textstyle\sum_{k=1}^{K}\sum_{i=1}^{d} \hat{m}_{ki} c_{ki}\big(x_k(0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)\big).
\end{aligned}$$

Note SAA(0) does not depend on α. In other words, optimizing $\overline{Z}_K(\alpha, \boldsymbol{p}_0)$ over α is roughly equivalent to optimizing $\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)$, which in turn is equivalent to optimizing

$$\min_{\alpha \geq 0} \quad \text{SAA-SubOpt}(\alpha) + \text{Instability}(\alpha). \qquad \textbf{(Sub-Optimality-Instability Tradeoff)}$$

We term this last optimization the "Sub-Optimality-Instability Tradeoff."

To develop some intuition, notice SAA-SubOpt(α) is nonnegative, and measures the average degree to which each $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ is sub-optimal with respect to a (scaled) SAA objective. In particular, SAA-SubOpt(α) is minimized at $\alpha = 0$, and we generally expect it is increasing in α. By contrast, Instability(α) measures the average degree to which the (scaled) performance of

$\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ changes on the training sample if we were to use one fewer data points. It is minimized at $\alpha = \infty$, since the fully-shrunken solution $\boldsymbol{x}_k(\infty, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ does not depend on the data and is, hence, completely stable. Intuitively, we might expect Instability(α) to be decreasing since as α increases, the shrunken measure $\hat{\boldsymbol{p}}_k(\alpha)$ depends less and less on the data. In reality, Instability(α) is often decreasing for large enough α, but for smaller α can have subtle behavior depending on the optimization structure. (See below for examples.)

This tradeoff is intuitive in light of our data-pooling interpretation of $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ from Section 2.1. Recall, we interpret $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ as the solution when we augment our original dataset with a synthetic dataset of size α drawn from $\boldsymbol{p}_0$. As we increase α, we introduce more SAA-sub-optimality into $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ because we "pollute" the k^{th} dataset with draws from a distinct distribution. However, we also increase the stability of $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ because we reduce its dependence on $\hat{\boldsymbol{m}}_k$. Shrunken-SAA seeks an α in the "sweet spot" that balances these two effects.

Importantly, this tradeoff also illuminates *when* data-pooling offers an improvement, i.e., when $\alpha^{\mathsf{S\text{-}SAA}} > 0$. Intuitively, $\alpha^{\mathsf{S\text{-}SAA}} > 0$ only if Instability(0) is fairly large and decreasing. Indeed, in this setting, the SAA-sub-optimality incurred by choosing a small positive α is likely outweighed by the increased stability. However, if Instability(0) is already small, the marginal benefit of additional stability likely won't outweigh the cost of sub-optimality.

More precisely, we intuit that data-pooling offers a benefit whenever i) the SAA solution is unstable, ii) the fully-shrunken solution $\boldsymbol{x}_k(\infty, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ is not too sub-optimal, and iii) K is sufficiently large. In particular, when $\hat{N}_k$ is relatively small for most k, the SAA solution is likely to be *very* unstable. Hence, intuition suggests data-pooling likely provides a benefit whenever $\hat{N}_k$ is small but K is large, i.e., the small-data, large-scale regime.

The intuition for a data-driven anchor $h(\hat{\boldsymbol{m}})$ is essentially the same. The proofs of Theorems 4.3 and 4.6 show that the approximation $\overline{Z}_K(\alpha, \boldsymbol{p}_0) \approx \overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)$ holds uniformly in α *and* $\boldsymbol{p}_0$. Consequently, the Sub-Optimality-Instability Tradeoff also holds for all $\boldsymbol{p}_0$. Hence, it holds for the specific realization of $h(\hat{\boldsymbol{m}})$, and changing α balances these two sources of error for this anchor. We recall in contrast to traditional leave-one-out validation, however, Shrunken-SAA does not remove a data point and retrain the anchor. This detail is important because it ensures the fully-shrunken solution $\boldsymbol{x}_k(\infty, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}})$ is still completely stable per our definition, i.e., has instability equal to zero, despite depending on the data.

The Sub-Optimality-Instability Tradeoff resembles the classical bias-variance tradeoff for MSE. Both tradeoffs decompose performance into a systematic loss (bias or SAA-sub-optimality) and a measure of dispersion (variance or instability). An important distinction, however, is that the Sub-Optimality-Instability tradeoff applies to general optimization problems, not just mean-squared error. Even if we restrict to the case of MSE (cf. Example 2.1), however, the two tradeoffs still differ and are two different ways to split the "whole" into "pieces." See Appendix D.

5.1. Sub-Optimality-Instability Tradeoff as a Diagnostic Tool

Our comments above are qualitative, focusing on developing intuition. However, the Sub-Optimality-Instability Tradeoff also provides a quantitative diagnostic tool for studying data-pooling. Indeed, for simple optimization problems such as minimizing MSE, it may be possible to analytically study the effects of pooling (cf. Theorem 2.1), but for more complex optimization problems where $\boldsymbol{x}_k(\alpha, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k)$ is not known analytically, such a study is not generally possible. Fortunately, both SAA-SubOpt(α) and Instability(α) can be evaluated *directly from the data*. Studying their dependence on α for a particular instance provides insight into how data-pooling improves (or does not improve) solution quality. We illustrate with Example 2.2:

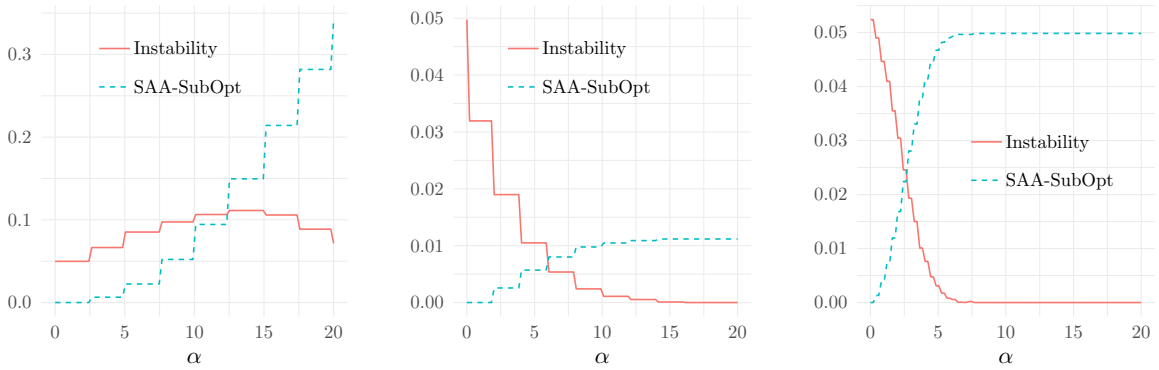
Example 5.1 (Simple Newsvendor Revisited). We revisit Example 2.2 and simulate an instance with $K = 1000$, p_{k1} distributed uniformly on $[.6, .9]$ and $p_{01} = .3$. One can confirm that as in Example 2.2, data-pooling offers no benefit over decoupling (regardless of the choice of $\hat{N}_k$) for these parameters. We take $\hat{N}_k \sim \text{Poisson}(10)$ for all k, and simulate a single data realization $\hat{\boldsymbol{m}}$.

Using the data, we can evaluate SAA-SubOpt(α) and Instability(α) explicitly. We plot them in the first panel of Fig. 5. Notice that as expected, SAA-SubOpt(α) increases steadily in α, however, perhaps surprisingly, Instability(α) *increases* at first, before ultimately decreasing. The reason is that as in Example 2.2, $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) = \mathbb{I}[\hat{p}_{k1}(\alpha) \geq 1/2]$. For small positive α, $\hat{p}_{k1}(\alpha)$ is generally closer to $\frac{1}{2}$ than $\hat{p}_{k1}$, and since $\frac{1}{2}$ is the critical threshold where $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ changes values, the solution is less stable. Hence, Instability(α) *increases* for small α. Because of this initial increasing behavior, the "gains" in stability never outweigh the costs of sub-optimality, and hence decoupling is best. Indeed, the first panel of Fig. EC.1 in the appendix shows $\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0} = \alpha^{\mathsf{OR}}_{\boldsymbol{p}_0} = 0.0$.

We earlier observed that the benefits of pooling depend on the anchor. We next consider the same parameters and data as above but let $p_{01} = .75$. The second panel of Fig. 5 shows the Sub-Optimality-Instability tradeoff. We see here that again Sub-Optimality(α) is increasing, and, perhaps more intuitively, Instability(α) is decreasing. Hence, there is a positive α that minimizes their sum, and the second panel Fig. EC.1 shows $\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0} \approx \alpha^{\mathsf{OR}}_{\boldsymbol{p}_0} \approx 16.16$.

Finally, as mentioned previously, the potential benefits of data-pooling also depends on the problem structure. The Sub-Optimality-Instability tradeoff allows us to study this dependence. Consider again letting $p_{01} = .3$, but now consider newsvendor problems with critical fractile $s = .2$. We again see a benefit to pooling. The Sub-Optimality-Instability tradeoff is in the last panel of Fig. 5. The last panel of Fig. EC.1 shows $\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0} \approx 2.42$ and $\alpha^{\mathsf{OR}}_{\boldsymbol{p}_0} \approx 2.22$.

In summary, while $\alpha^{\mathsf{S\text{-}SAA}}_h$ identifies a good choice of shrinkage in many settings, Sub-Optimality and Instability graphs as above often illuminate *why* this is a good choice of shrinkage, providing insight. This is particularly helpful for complex optimization problems for which it may be hard to reason about $\boldsymbol{x}_k(\alpha, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k)$.



(a) $p_{01} = .3,\ s = .5$ (b) $p_{01} = .75,\ s = .5$ (c) $p_{01} = .3,\ s = .2$

Figure 5 **Sub-Optimality-Instability Curves**. We consider $K = 10,000$ newsvendors where $p_{k1} \sim \text{Uniform}[.6, .9]$, $\hat{N}_k \sim \text{Poisson}(10)$, and a single data draw. The values of p_{01} and the critical fractile s is given in each panel. In the first panel, instability initially increases, and there is no benefit to pooling. In the second and third, instability is decreasing and there is a benefit to pooling.

6. Computational Experiments

In this section we study the empirical performance of Shrunken-SAA on synthetic and real data. All code for reproducing these experiments and plots is available at `https://github.com/vgupta1/JS_SAA` We focus on assessing the degree to which Shrunken-SAA is robust to violations of the assumptions underlying Theorems 4.2 to 4.7. Specifically, we ask how Shrunken-SAA performs when i) K is small to moderate, and not growing to infinity; ii) Assumption 3.1 is violated, i.e., each $\hat{N}_k$ is fixed and non-random; iii) the true $\mathbb{P}_k$ do not have finite, discrete support; or iv) N grows large.

For simplicity, we take each subproblem to be a newsvendor problem with critical fractile $s = 95\%$. Since the performance of Shrunken-SAA depends on the true distributions $\boldsymbol{p}_k$, we use real sales data from a chain of European pharmacies. (See Section 6.1 for more details.)

We compare several policies: The first two, **SAA** and **KS**, are decoupled-benchmarks. Recall that for the newsvendor problem, SAA, i.e., $\boldsymbol{x}(0, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$, is also the optimal solution to a distributionally robust formulation using a Wasserstein ambiguity set (Esfahani and Kuhn (2018)). We define KS to be an optimal solution to a distributionally robust formulation of the newsvendor problem using the Kolmogorov-Smirnov ambiguity set (see Appendix E.3 for formal definition). This set enjoys strong large-sample statistical guarantees (Bertsimas et al. 2018).

The next three policies, **JS-Fixed**, **S-SAA-Fixed** and **Oracle-Fixed**, each shrink towards the uniform distribution, i.e., a fixed anchor. They differ in the amount of shrinkage. JS-Fixed, i.e., $\boldsymbol{x}(\alpha^{\mathsf{JS}}_{\boldsymbol{p}_0}, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$, pools according to Theorem 2.1; S-SAA-Fixed, i.e., $\boldsymbol{x}(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0}, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$, is our Shrunken-SAA algorithm; and Oracle-Fixed, i.e., $\boldsymbol{x}(\alpha^{\mathsf{OR}}_{\boldsymbol{p}_0}, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ is the oracle shrinkage.

The next two policies, **S-SAA-Beta** and **Oracle-Beta**, each shrink towards a data-driven choice of anchor in $\mathcal{P}$, where $\mathcal{P}$ consists of scaled beta-distributions (cf. Appendix E.3). S-SAA, i.e., $\boldsymbol{x}(\alpha_{\mathcal{P}}^{\mathsf{S\text{-}SAA}}, h_{\mathcal{P}}(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}})$, uses $h_{\mathcal{P}}$, while Oracle-Beta, i.e., $\boldsymbol{x}(\alpha_{\mathcal{P}}^{\mathsf{OR}}, \boldsymbol{q}_{\mathcal{P}}^{\mathsf{OR}}, \hat{\boldsymbol{m}})$, uses the oracle anchor.

Finally, the last set of policies, **JS-GM**, **S-SAA-GM** and **Oracle-GM** each shrink towards the grand-mean distribution, $\hat{\boldsymbol{p}}^{\mathsf{GM}}$. They differ in the amount of shrinkage. JS-GM, pools according to Theorem 2.1, S-SAA-GM is our Shrunken-SAA Algorithm, and Oracle-GM is the oracle pooling.

Intuitively, the difference between the JS policies and the decoupled policies illustrates the value of data-pooling in a "generic" fashion that does not account for the shape of the cost functions. By contrast, the difference between the Shrunken-SAA policies and the JS policies quantifies the additional benefit of tailoring the amount of pooling to the specific newsvendor cost function. Similarly, the difference between the "Beta" anchor versions and the Fixed versions help quantify the value of a good choice of anchor, and, as we will see, the GM variants highlight that the grand-mean is often a good heuristic choice of anchor.

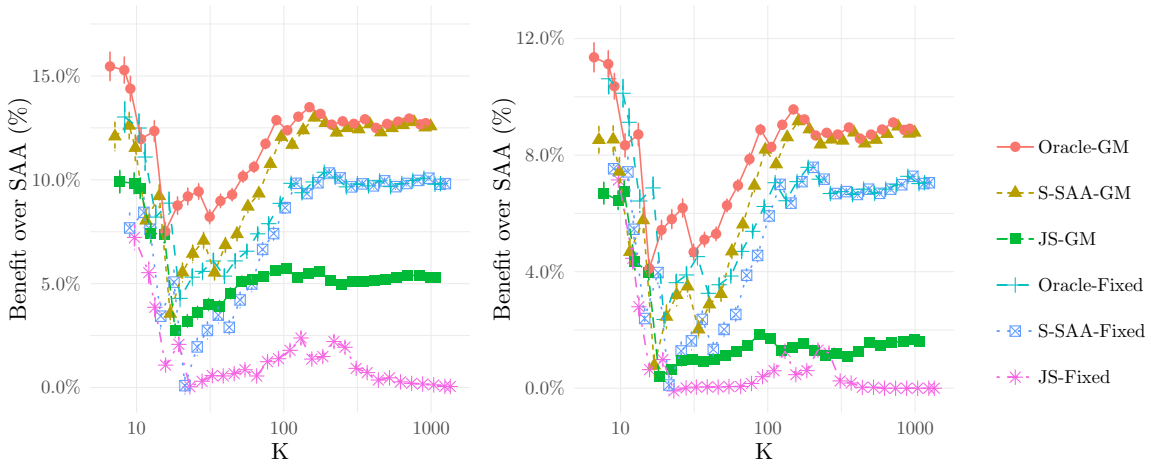
Before presenting the details, we summarize our main findings. When N is moderate to large, all methods (including Shrunken-SAA) perform comparably to the full-information solution. When N is small to moderate, however, our Shrunken-SAA policies provide a significant benefit over SAA and a substantial benefit over JS variants that do not leverage the optimization structure. This is true even for moderate K ($K \leq 100$) and even when $\hat{N}_k$ are fixed (violating Assumption 3.1). The value of d has little effect on the performance of Shrunken-SAA; it strongly outperforms decoupling even as $d \to \infty$. Finally, our GM heuristic has very strong performance, comparable to the Beta variants which optimize the choice of anchor, at a much smaller computational cost.

For ease of comparison in what follows, we present all results as "% Benefit over SAA," i.e., bigger values are better. In many cases, to aid readability, we only present a subset of benchmark policies on a graph. In these cases, larger tables with all benchmarks are available in Appendix E.

6.1. Data Description

Our dataset consists of daily sales at the store level for a European pharmacy chain with locations across 7 countries. We treat these aggregated store sales as if they were the realized daily demand of a single product. Although this is clearly a simplification of the underlying inventory management problem, we do not believe it significantly impacts the study of our key questions outlined above. Additionally, aggregating over products makes demand censoring insignificant.

The original dataset contains 942 days of data across 1115 stores. After some preliminary data-cleaning (see Appendix E.3), we are left with 629 days. Due to local holidays, individual stores may still be closed on these 629 days. Almost all (1105) stores have at least one missing day, and 16% of stores have 20% of days missing.



(a) $\hat{N}_k \sim \text{Poisson}(10)$ (b) $\hat{N}_k = 10$ (non-random)

Figure 6 **Robustness to Assumption 3.1.** Performance of policies on simulated data. In the first panel, the amount of data per store follows Assumption 3.1 with $N_k = 10$. In the second panel, the amount of data is fixed at $\hat{N}_k = 10$ for all runs. Error bars show ± 1 standard error.

Stores vary in size, available assortment of products, promotional activities and prices, creating significant heterogeneity in demand. The average daily demand ranges from 3,183 to 23,400. The first panel of Fig. EC.3 in Appendix E plots the average daily demand by store. The second panel provides a more fine-grained perspective, showing the distribution of daily demand for a few representative stores. The distributions are quite distinct, at least partially because the overall scale of daily sales differs wildly between stores.

Finally, with the exception of Appendix E.7, we discretize demand by dividing the range of observations into d equally-spaced bins to form the true distributions $\boldsymbol{p}_k$. Figure EC.2 plots $\boldsymbol{p}_k$ for some representative stores when $d = 20$. We consider these distributions to be quite diverse and far from the uniform distribution (our fixed anchor). We also plot the distribution of the 95% quantile with respect to this discretization in the second panel of Fig. EC.2. Note that it is not the case that 95% quantile occurs in the same (discretized) bin for each $\boldsymbol{p}_k$, i.e., the quantile itself displays some heterogeneity, unlike Example 2.3.

6.2. An Idealized Synthetic Dataset

We first consider an ideal setting for Shrunken-SAA. Specifically, after discretizing demand for each store into $d = 20$ buckets, we set $\boldsymbol{p}_k$ to be the empirical distribution of demand over the *entire* dataset with respect to these buckets. We then simulate synthetic data according to Eq. (2.1) under Assumption 3.1. We train each of our methods using this data, and then evaluate their true performance using the $\boldsymbol{p}_k$. We repeat this process 200 times. The left panel of Fig. 6 shows the average results for a subset of the policies. Table EC.1 in the appendix includes all policies.

As suggested by Theorems 4.5 and 4.6, Shrunken-SAA significantly outperforms decoupling even for K as small as 10. For large K, the benefit is as large as $10-15\%$. Both of our Shrunken-SAA policies converge quickly to their oracle benchmarks. We note the JS policies also outperform the decoupled solutions, but by a smaller amount (5-10%). For both sets of policies, shrinking to the grand mean outperforms shrinking to the uniform distribution, since, as observed earlier, the true distributions are far from uniform and have quantiles far from the uniform quantile. Indeed, the grand-mean policies perform comparably to our Beta policies (cf. Table EC.1).

We also illustrate the standard deviation of the performance for each of these methods in Fig. EC.4 in Appendix E. For all approaches, the standard deviation tends to zero as $K \to \infty$, because the true performance concentrates at its expectation for each method. For small K, our Shrunken-SAA approaches exhibit significantly smaller standard deviation than SAA, and, for larger K, the standard deviation is comparable to the oracle values, and much less than JS variants. The reduction in variability compared to SAA follows intuitively since pooling increases stability.

Finally, we plot the average amount of shrinkage across runs as a function of K for each method in Fig. EC.5 in Appendix E. We observe that the shrinkage amount converges quickly as $K \to \infty$, and that our Shrunken-SAA methods pool much more than the JS variants. In particular, when shrinking to the grand-mean or to an optimized Beta distribution, our Shrunken-SAA methods use a value of $\alpha \geq 30$ for large K, i.e., placing 3 times more weight on the anchor than the data, itself. By contrast, JS variants eventually engage in almost no pooling.

6.3. Relaxing Assumption 3.1

We next consider robustness to Assumption 3.1. Specifically, we repeat the experiment of the previous section but now simulate data with $\hat{N}_k = 10$ for all k and all runs. Results are shown in the second panel of Fig. 6, and Figs. EC.4 and EC.5 and Table EC.2 in Appendix E. We see the same qualitative features. Specifically, our Shrunken-SAA methods converge to oracle performance, and, even for moderate K, they significantly outperform decoupling. The JS methods offer a much smaller improvement over SAA. Many of the other features with respect to convergence in α and standard deviation of the performance are also qualitatively similar.

6.4. Historical Backtest

For our remaining tests we consider a more realistic setting for Shrunken-SAA. Specifically, we employ repeated random subsampling validation with our data to assess each method: for each store we select $\hat{N}_k = 10$ days randomly from the dataset, then train each method with these points, and finally evaluate their out-of-sample performance on $N_{\text{test}} = 10$ data points, again chosen randomly from the dataset. Note that unlike the previous experiment, it is possible that some of sampled

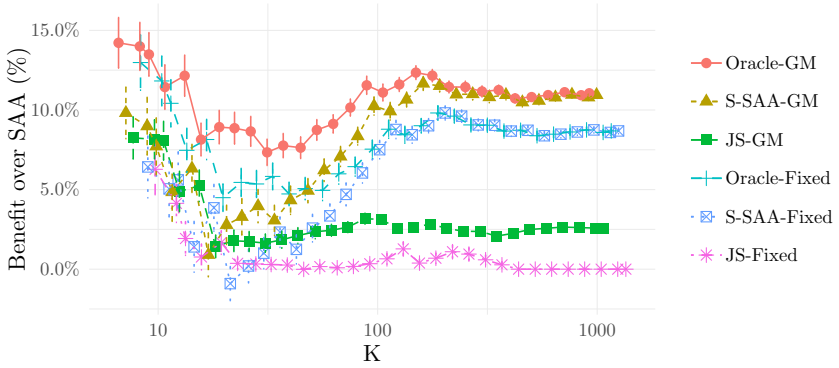


Figure 7 **Historical Backtest.** We evaluate our policies on historical data using $d = 20$. Error bars show ± 1 standard error.

training days have missing data for store k. In this cases, we will have fewer than $\hat{N}_k$ points when training store k. Similar missing data occur for the N_{test} testing points. We prefer repeated, random subsampling validation to more traditional 5-fold cross-validation when evaluating our methods, in order to finely control the number of data points $\hat{N}_k$ used in each subproblem.

We evaluate each of our policies using our historical backtest set-up with $d = 20$ in Fig. 7. For readability, the figure shows a subset of policies. Table EC.3 in the appendix shows all policies. Importantly, we see the same features as in our synthetic data experiment: our Shrunken-SAA methods converge to oracle optimality and offer a substantive improvement over SAA for large enough K. They also outperform JS variants that do not leverage the optimization structure.

6.5. Other Experiments with Synthetic and Real Data

Appendices E.7 to E.9 in the appendix study the robustness of Shrunken-SAA to the number of support points d, its performance as $N \to \infty$, and compares computationally cheaper variants of the algorithm that substitute 2–fold or 5–fold cross-validation for the LOO validation step. We omit details for space. Generally, we find that: i) Shrunken-SAA is quite robust to d. ii) As N increases Shrunken-SAA retains many of SAA's strong large-sample properties. Namely, both methods approach full-information optimum, so there is less "room" to improve upon decoupling, but Shrunken-SAA offers some marginal benefit for large K. iii) Other forms of cross-validation perform quite well and are viable alternatives in computationally limited settings.

7. Conclusion and Future Directions

In this paper, we introduce and study the data-pooling phenomenon for stochastic optimization problems, i.e., that when solving many separate data-driven stochastic optimization subproblems, there exist algorithms which pool data across subproblems that outperform decoupling, even when 1) the underlying subproblems are distinct and unrelated, and 2) data for each subproblem are independent. We propose a simple algorithm Shrunken-SAA that exploits this phenomenon by pooling data in a particular fashion motivated by a Bayes model. We prove that under frequentist

assumptions, in the limit as the number of subproblems grows large, Shrunken-SAA identifies whether pooling in this way can improve upon decoupling, and, if so, the ideal amount to pool, even if the amount of data per subproblem is fixed and small. In other words, Shrunken-SAA identifies an optimal level of pooling in the so-called small-data, large-scale regime. In particular, we prove explicit high-probability bounds on the performance of Shrunken-SAA relative to an oracle benchmark that decay like $\tilde{\mathcal{O}}(1/\sqrt{K})$ where K is the number of subproblems.

Shrunken-SAA need not offer a strict benefit over decoupling in all instances. Hence, we also introduce the Sub-Optimality-Instability tradeoff, a decomposition of the benefits of data-pooling that provides strong intuition into the kinds of problems for which data-pooling offers a benefit. Overall, this intuition and empirical evidence with real data suggest Shrunken-SAA offers significant benefits in the small-data, large-scale regime for a variety of problems.

We hope our work inspires fellow researchers to think of data-pooling as an "additional knob" that might be leveraged to improve performance when designing algorithms for data-driven decision-making under uncertainty.

Acknowledgments

The authors would like to thank the editorial team including 3 anonymous reviewers for the constructive comments on an earlier draft. Grant Funding: V.G. is partially supported by the National Science Foundation under Grant No. 1661732. N.K. is partially supported by the National Science Foundation under Grant No. 1656996.

References

Beran, R. 1996. Stein estimation in high dimensions: A retrospective. *Madan Puri Festschrift* 91–110.

Bertsimas, D., V. Gupta, N. Kallus. 2018. Robust sample average approximation. *Mathematical Programming* **171**(1-2) 217–282.

Bousquet, O., A. Elisseeff. 2002. Stability and generalization. *Journal of Machine Learning Research* **2**(March) 499–526.

Brown, L.D. 1971. Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics* **42**(3) 855–903.

Brown, L.D., L.H. Zhao, et al. 2012. A geometrical explanation of Stein shrinkage. *Statistical Science* **27**(1) 24–30.

Chen, L.H.Y. 1975. Poisson approximation for dependent trials. *The Annals of Probability* 534–545.

Davarnia, D., G. Cornuéjols. 2017. From estimation to optimization via shrinkage. *Operations Research Letters* **45**(6) 642–646.

Deheuvels, P., D. Pfeifer. 1988. Poisson approximations of multinomial distributions and point processes. *Journal of Multivariate Analysis* **25**(1) 65–89.

DeMiguel, V., A. Martin-Utrera, F.J. Nogales. 2013. Size matters: Optimal calibration of shrinkage estimators for portfolio selection. *Journal of Banking & Finance* **37**(8) 3018–3034.

Efron, B., T. Hastie. 2016. *Computer Age Statistical Inference*, vol. 5. Cambridge University Press.

Efron, B., C. Morris. 1977. Stein's paradox in statistics. *Scientific American* **236**(5) 119–127.

Esfahani, P.M., D. Kuhn. 2018. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* **171**(1-2) 115–166.

Friedman, J., T. Hastie, R. Tibshirani. 2001. *The Elements of Statistical Learning*. 10, Springer series in statistics New York.

Gupta, V., P. Rusmevichientong. 2017. Small-data, large-scale linear optimization with uncertain objectives. URL `https://ssrn.com/abstract=3065655`. To Appear in Management Science.

Jorion, P. 1986. Bayes-Stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis* **21**(3) 279–292.

Kleywegt, A.J., A. Shapiro, T. Homem-de Mello. 2002. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* **12**(2) 479–502.

Levi, R., G. Perakis, J. Uichanco. 2015. The data-driven newsvendor problem: New bounds and insights. *Operations Research* **63**(6) 1294–1306.

McDonald, D.R. 1980. On the Poisson approximation to the multinomial distribution. *Canadian Journal of Statistics* **8**(1) 115–118.

Mukherjee, G., L.D. Brown, P. Rusmevichientong. 2015. Efficient empirical Bayes prediction under check loss using asymptotic risk estimates. *arXiv preprint arXiv:1511.00028* .

Munkres, J.R. 1974. *Topology: A First Course*. Prentice-Hall.

Pollard, D. 1990. Empirical processes: Theory and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*. JSTOR, i–86.

Shalev-Shwartz, S., O. Shamir, N. Srebro, K. Sridharan. 2010. Learnability, stability and uniform convergence. *Journal of Machine Learning Research* **11**(Oct) 2635–2670.

Shapiro, A., D. Dentcheva, A. Ruszczyński. 2009. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM.

Stanley, R.P. 2004. An introduction to hyperplane arrangements. *IAS/Park City Mathematics Series* **14**.

Stein, C. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of 3rd Berkeley Symposium on Mathematical Statistics and Probability* **I** 197–206.

Stigler, S.M. 1990. The 1988 Neyman Memorial Lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science* **5**(1) 147–155.

Van der Vaart, A.W. 2000. *Asymptotic statistics*, vol. 3. Cambridge University Press.

Van der Vaart, A.W., J. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer.

Yu, B. 2013. Stability. *Bernoulli* **19**(4) 1484–1500.

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Online Appendix: Data-Pooling for Stochastic Optimization

Appendix A: Proof of Theorem 2.1: Data-Pooling for MSE

Proof of Theorem 2.1. First note that

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^K \boldsymbol{p}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k^{\mathsf{SAA}}) - \frac{1}{K}\sum_{k=1}^K \boldsymbol{p}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha^{\mathsf{JS}}_{\boldsymbol{p}_0}, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)) \;-\; \frac{\left(\frac{1}{K}\sum_{k=1}^K \sigma_k^2/\hat{N}\right)^2}{\frac{1}{K}\sum_{k=1}^K \sigma_k^2/\hat{N} + \frac{1}{K}\sum_{k=1}^K (\mu_k-\mu_{k0})^2} \\
&\quad= \left(\frac{1}{K}\sum_{k=1}^K \left(\sigma_k^2 + (\mu_k - \hat{\mu}_k(0))^2\right) - \frac{1}{K}\sum_{k=1}^K \left(\sigma_k^2 + (\mu_k - \hat{\mu}_k(\alpha^{\mathsf{JS}}))^2\right)\right) \\
&\qquad - \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \left(\sigma_k^2 + (\mu_k - \hat{\mu}_k(0))^2\right) - \frac{1}{K}\sum_{k=1}^K \left(\sigma_k^2 + (\mu_k - \hat{\mu}_k(\alpha^{\mathsf{AP}}))^2\right) \middle| \hat{N}\right] \\
&\quad\le \left|\frac{1}{K}\sum_{k=1}^K \left((\mu_k - \hat{\mu}_k(0))^2 - \mathbb{E}\left[(\mu_k - \hat{\mu}_k(0))^2 \mid \hat{N}\right]\right)\right| + \left|\frac{1}{K}\sum_{k=1}^K \left((\mu_k - \hat{\mu}_k(\alpha^{\mathsf{JS}}))^2 - \mathbb{E}\left[(\mu_k - \hat{\mu}_k(\alpha^{\mathsf{AP}}))^2 \mid \hat{N}\right]\right)\right| \\
&\quad\le 2\sup_{\alpha\ge 0}\left|\frac{1}{K}\sum_{k=1}^K \left((\mu_k - \hat{\mu}_k(\alpha))^2 - \mathbb{E}\left[(\mu_k - \hat{\mu}_k(\alpha))^2 \mid \hat{N}\right]\right)\right| \qquad \text{(EC.A.1)} \\
&\qquad + \left|\frac{1}{K}\sum_{k=1}^K \left(\mathbb{E}\left[(\mu_k - \hat{\mu}_k(\alpha^{\mathsf{JS}}))^2 \mid \hat{N}\right] - \mathbb{E}\left[(\mu_k - \hat{\mu}_k(\alpha^{\mathsf{AP}}))^2 \mid \hat{N}\right]\right)\right|. \qquad \text{(EC.A.2)}
\end{aligned}$$

We begin by showing Eq. (EC.A.1) converges to zero in probability. Notice Eq. (EC.A.1) is the maximal deviation of a stochastic process (indexed by α) composed of averages of independent, but not identically distributed, random variables. Such processes are discussed in Theorem 4.1, and we follow that approach to establish convergence here.

We first claim that the constants $F_k = 4a_{\max}^2$ yield an envelope. Specifically,

$$|\mu_k - \hat{\mu}_k(\alpha)| \;\le\; \left|\boldsymbol{p}^\top \boldsymbol{a}_k\right| + \left|\hat{\boldsymbol{p}}(\alpha)^\top \boldsymbol{a}_k\right| \;\le\; 2\|\boldsymbol{a}_k\|_\infty.$$

which is at most $2a_{\max}$. Hence $(\mu_k - \hat{\mu}_k(\alpha))^2 \le F_k$.

We next show that the set $\left\{\left((\mu_k - \hat{\mu}_k(\alpha))^2\right)_{k=1}^K : \alpha \ge 0\right\} \subseteq \mathbb{R}^K$ has pseudo-dimension at most 3. Indeed, this set is contained within the set

$$\left\{\left(\left(\theta(\mu_k - \mu_{k0}) + (1-\theta)(\mu_k - \hat{\mu}_k)\right)^2\right)_{k=1}^K : \theta \in \mathbb{R}\right\} \;\subseteq\; \mathbb{R}^K$$

This set is the range of a quadratic function of θ, and is hence contained within a linear subspace of dimension at most 3. Thus, it has pseudo-dimension at most 3.

Since this set has pseudo-dimension at most 3, there exists a constant A_1 (not depending on K or other problem parameters) such that the corresponding Dudley integral can be bounded as

$J \le \mathrm{A}_1 \|\boldsymbol{F}\|_2$ (Pollard 1990, pg. 37). Theorem 4.1 with $p=1$ thus implies there exists a constant A_2 (not depending on K or other problem parameters) such that

$$\mathbb{E}\left[\sup_{\alpha\ge 0}\left|\frac{1}{K}\sum_{k=1}^{K}\left((\mu_k-\hat{\mu}_k(\alpha))^2-\mathbb{E}\left[(\mu_k-\hat{\mu}_k(\alpha))^2\right]\right)\right|\right] \;\le\; \mathrm{A}_2\cdot a_{\max}^2/\sqrt{K}.$$

Markov's inequality then yields the convergence of Eq. (EC.A.1) to 0.

We will next show that Eq. (EC.A.2) converges to 0. Let $\theta^{\mathsf{JS}}=\frac{\alpha^{\mathsf{JS}}}{\alpha^{\mathsf{JS}}+\hat{N}}$ and $\theta^{\mathsf{AP}}=\frac{\alpha^{\mathsf{AP}}}{\alpha^{\mathsf{AP}}+\hat{N}}$ and note $\theta^{\mathsf{JS}},\theta^{\mathsf{AP}}\in[0,1]$ almost surely. Write,

$$\begin{aligned}
&\left|\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}\left[(\mu_k-\hat{\mu}_k(\alpha^{\mathsf{JS}}))^2-(\mu_k-\hat{\mu}_k(\alpha^{\mathsf{AP}}))^2\mid\hat{N}\right]\right)\right| \\
&\quad\le\; \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left|(\mu_k-\hat{\mu}_k(\alpha^{\mathsf{JS}}))^2-(\mu_k-\hat{\mu}_k(\alpha^{\mathsf{AP}}))^2\right|\mid\hat{N}\right] \\
&\quad=\; \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left|(\mu_k-\hat{\mu}_k+\theta^{\mathsf{JS}}(\hat{\mu}_k-\mu_{k0}))^2-(\mu_k-\hat{\mu}_k+\theta^{\mathsf{AP}}(\hat{\mu}_k-\mu_{k0}))^2\right|\mid\hat{N}\right].
\end{aligned}$$

Consider the function $\theta\mapsto(\mu_k-\hat{\mu}_k+\theta(\hat{\mu}_k-\mu_{k0}))^2$. For $|\theta|\le 1$, its derivative is bounded in magnitude by

$$\begin{aligned}
2\left|\mu_k-\hat{\mu}_k+\theta(\hat{\mu}_k-\mu_{k0})\right|\left|\hat{\mu}_k-\mu_{k0}\right| \;&\le\; 2\Big(\left|\mu_k-\hat{\mu}_k\right|+\left|\hat{\mu}_k-\mu_{k0}\right|\Big)\left|\hat{\mu}_k-\mu_{k0}\right| \\
&\le\; 2\Big(2a_{\max}+2a_{\max}\Big)2a_{\max} \;=\; 16a_{\max}^2.
\end{aligned}$$

Hence, by the mean-value theorem,

$$\mathbb{E}\left[\left|(\mu_k-\hat{\mu}_k+\theta^{\mathsf{JS}}(\hat{\mu}_k-\mu_{k0}))^2-(\mu_k-\hat{\mu}_k+\theta^{\mathsf{AP}}(\hat{\mu}_k-\mu_{k0}))^2\right|\mid\hat{N}\right] \;\le\; 16a_{\max}^2\mathbb{E}\left[\left|\theta^{\mathsf{JS}}-\theta^{\mathsf{AP}}\right|\mid\hat{N}\right].$$

We will next show that, conditional on $\hat{N}$, $\theta^{JS}-\theta^{\mathsf{AP}}\to_p 0$ as $K\to\infty$. Since $\left|\theta^{\mathsf{JS}}-\theta^{\mathsf{AP}}\right|\le 2$ almost surely, this will imply that $\mathbb{E}\left[\left|\theta^{\mathsf{JS}}-\theta^{\mathsf{AP}}\right|\mid\hat{N}\right]\to 0$ as $K\to\infty$, completing the proof.

Since $\hat{N}\ge 1,\alpha^{\mathsf{JS}}\ge 0,\alpha^{\mathsf{AP}}\ge 0$, we have $\left|\theta^{\mathsf{JS}}-\theta^{\mathsf{AP}}\right|=\frac{\hat{N}}{(\alpha^{\mathsf{JS}}+\hat{N})(\alpha^{\mathsf{AP}}+\hat{N})}\left|\alpha^{\mathsf{JS}}-\alpha^{\mathsf{AP}}\right|\le\left|\alpha^{\mathsf{JS}}-\alpha^{\mathsf{AP}}\right|$. We proceed to show $\alpha^{\mathsf{JS}}\to_p\alpha^{\mathsf{AP}}$. We show this second convergence by showing that both the numerator and denominator converge in probability. For the numerator,

$$0\le\frac{1}{\hat{N}-1}\sum_{i=1}^{\hat{N}}(\hat{\xi}_{ki}-\hat{\mu}_k)^2 \;\le\; \frac{\hat{N}}{\hat{N}-1}4a_{\max}^2 \;\le\; 8a_{\max}^2,$$

since $\hat{N}\ge 2\implies\frac{\hat{N}}{\hat{N}-1}\le 2$. By Hoeffding's inequality, for any $t>0$,

$$\mathbb{P}\left(\left|\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\hat{N}-1}\sum_{i=1}^{\hat{N}}(\hat{\xi}_{ki}-\hat{\mu}_k)^2-\frac{1}{K}\sum_{k=1}^{K}\sigma_k^2\right|>t\;\mid\;\hat{N}\right) \;\le\; 2\exp\left(-\frac{Kt^2}{32a_{\max}^4}\right)\;\to\;0,$$

as $K\to\infty$. Thus, $\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\hat{N}-1}\sum_{i=1}^{\hat{N}}(\hat{\xi}_{ki}-\hat{\mu}_k)^2\to_p\frac{1}{K}\sum_{k=1}^{K}\sigma_k^2$.

Entirely analogously, $0 \le (\hat{\mu}_k - \mu_{k0})^2 = ((\hat{\boldsymbol{p}}_k - \boldsymbol{p}_0)^\top \boldsymbol{a}_k)^2 \le 4a_{\max}^2$. Hence, by Hoeffding's inequality,

$$\mathbb{P}\left(\left|\frac{1}{K}\sum_{k=1}^{K}\left((\mu_{k0}-\hat{\mu}_k)^2 - \mathbb{E}\left[(\mu_{k0}-\hat{\mu}_k)^2\right]\right)\right| > t \;\middle|\; \hat{N}\right) \;\le\; 2\exp\left(-\frac{Kt^2}{8a_{\max}^2}\right) \;\to\; 0,$$

as $K \to \infty$. Recall $\mathbb{E}\left[(\mu_{k0}-\hat{\mu}_k)^2 \mid \hat{N}\right] = \sigma_k^2/\hat{N} + (\mu_{k0}-\mu_k)^2$ by the bias-variance decomposition. Combining the numerator and denominator, we have by Slutsky's Theorem that $\alpha^{\mathsf{JS}} \to \alpha^{\mathsf{AP}}$. □

Appendix B: Auxiliary Lemmas

In this section,, we first prove some auxiliary lemmas that we will need when proving our performance guarantees. These results are largely elementary or well-known facts about tails of random variables.

LEMMA B.1 **(Bounding a Gaussian Integral)**. *Suppose $t \ge 1$. Then*

$$\int_0^1 \sqrt{\log(t/\epsilon)}d\epsilon \le \sqrt{\log t} + \sqrt{\pi}/2.$$

Proof. Make the substitution $u = \sqrt{2\log(t/\epsilon)}$. Then,

$$\begin{aligned}
\int_0^1 \sqrt{\log(t/\epsilon)}d\epsilon &= \frac{t}{\sqrt{2}}\int_{\sqrt{2\log t}}^{\infty} u^2 e^{-u^2/2}du \\
&= \frac{tu}{\sqrt{2}}e^{-u^2/2}\Big|_{\infty}^{\sqrt{2\log t}} + \frac{t}{\sqrt{2}}\int_{\sqrt{2\log t}}^{\infty} e^{-u^2/2}du \qquad \text{(integration by parts)} \\
&= \sqrt{\log t} + \frac{t}{\sqrt{2}}\int_{\sqrt{2\log t}}^{\infty} e^{-u^2/2}du.
\end{aligned}$$

Consider $t \mapsto \frac{t}{\sqrt{2}}\int_{\sqrt{2\log t}}^{\infty} e^{-u^2/2}du$. Its derivative with respect to t is

$$\frac{1}{\sqrt{2}}\int_{\sqrt{2\log t}}^{\infty} e^{-u^2/2}du - \frac{1}{2t\sqrt{\log t}} \;\le\; \frac{1}{t\sqrt{2}\cdot\sqrt{2\log t}} - \frac{1}{2t\sqrt{\log t}} \;=\; 0,$$

where we have a standard inequality for the tail CDF of the normal distribution: $\int_x^\infty e^{-u^2/2}du \le x^{-1}\cdot e^{-\frac{x^2}{2}}$. Since the derivative is always non-positive, the integral is non-increasing in t. Thus,

$$t\int_{\sqrt{\log t}}^{\infty} e^{-u^2}du \;\le\; 1\int_{\sqrt{\log 1}}^{\infty} e^{-u^2}du = \frac{\sqrt{\pi}}{2}.$$

Substituting above completes the proof. □

LEMMA B.2 **(L_p-norms of Products)**. *For any $p \ge 1$ and random variables X, Y. Then, $\|XY\|_p \le \|X\|_{2p}\|Y\|_{2p}$.*

Proof. By Hölder's inequality, $\mathbb{E}[|XY|^p] \le \sqrt{\mathbb{E}[X^{2p}]}\cdot\sqrt{\mathbb{E}[Y^{2p}]}$. Taking the p^{th} root of both side yields the result. □

The following lemma is a specific case of Lemma 2.2.2 of Van der Vaart and Wellner (1996) with explicit constants:

LEMMA B.3 **(Tails of the Maximum)**. *Suppose the random variables $Y_1,\dots,Y_K$ satisfy $\mathbb{E}\exp(\beta_0 Y_k)\le 2$ for all $k=1,\dots,K$, $K\ge 2$. Let $Y_{\max}=\max_{k=1,\dots,K} Y_k$, and define $\beta=\frac{\beta_0}{1+\log K}$. Then, $\mathbb{E}\exp(\beta Y_{\max})\le 6$.*

Proof. By definition of β,

$$t\le \beta Y_{\max} \iff 1\le e^{\beta_0 Y_{\max}-t(1+\log K)}. \tag{EC.B.1}$$

Then, writing $\exp(\cdot)$ as an integral,

$$\begin{aligned}
\exp(\beta Y_{\max}) &= e+\int_1^{\beta Y_{\max}} e^t dt \\
&\le e+\int_1^{\beta Y_{\max}} e^{\beta_0 Y_{\max}}\cdot e^{-t(1+\log K)}\cdot e^t dt && \text{(Eq. (EC.B.1))}\\
&\le e+\int_1^{\beta Y_{\max}} e^{\beta_0 Y_{\max}}\cdot e^{-t\log K} dt \\
&\le e+\sum_{k=1}^K \int_1^{\infty} e^{\beta_0 Y_k}\cdot e^{-t\log K} dt,
\end{aligned}$$

where in the last step we have bounded the maximum by a sum and extended the limits of integration because the integrand is positive. Now take expectations of both sides and evaluate the integral, yielding

$$\mathbb{E}\left[\exp(\beta Y_{\max})\right] \;\le\; e+2K\int_1^{\infty} e^{-t\log K} dt \;=\; e+\frac{2}{\log K}\le 6,$$

since $K\ge 2$. $\square$

Recall, for any random variable Y and function $\Psi(\cdot)$, $\|Y\|_\Psi \equiv \inf\{\beta>0 : \mathbb{E}\left[\Psi\left(|Y|\,\beta^{-1}\right)\right]\le 1\}$ is the Orlicz norm of Y with respect to $\Psi(\cdot)$.

LEMMA B.4 **(Relating Ψ-norm and L_p-norm)**. *Fix $p\ge 1$. Let $\Psi(t)=\frac{1}{5}\exp(t^2)$, and $\|\cdot\|_\Psi$ be the corresponding Orlicz norm. Then,*

i) For any $t\ge 0$, $t^p\le \left(\frac{p}{e}\right)^p e^t$.

ii) For any $t\ge 0$, $t^p\le \left(\frac{p}{2}\right)^{\frac{p}{2}} e^{-\frac{p}{2}} e^{t^2}$.

iii) Let $C_p=5^{1/p}\left(\frac{p}{2}\right)^{1/2} e^{-1/2}$. For any random variable Y, $\|Y\|_p\le C_p\|Y\|_\Psi$.

iv) For any random variable $Y\ge 1$, $\|\sqrt{\log Y}\|_p\le 5^{1/p}\left(\frac{p}{2e}\right)^{1/2}\max(1,\sqrt{\mathbb{E}[Y]}/2)$.

Proof. Consider the optimization $\max_{t\ge 0} t^p e^{-t}$. Taking derivatives shows the optimal solution is $t^*=p$, and the optimal value is $p^p e^{-p}$. Hence, $t^p e^{-t}\le p^p e^{-p}$ for all t. Rearranging proves the first statement. The second follows from the first since, $t^p=(t^2)^{\left(\frac{p}{2}\right)}\le \left(\frac{p}{2}\right)^{p/2} e^{-\frac{p}{2}} e^{t^2}$.

For the third, statement, let $\beta = \|Y\|_\Psi$, i.e., $\mathbb{E}\left[\exp\left(\frac{Y^2}{\beta^2}\right)\right] \le 5$. Then,

$$\mathbb{E}\left[\left(\frac{|Y|}{C_p\beta}\right)^p\right] \;=\; \frac{1}{C_p^p}\mathbb{E}\left[\left(\frac{|Y|}{\beta}\right)^p\right] \;\le\; \frac{1}{C_p^p}\left(\frac{p}{2}\right)^{p/2} e^{-\frac{p}{2}}\mathbb{E}\left[e^{\frac{Y^2}{\beta^2}}\right] \;\le\; 1.$$

Rearranging and taking the p^{th} root of both sides proves the third statement.

Finally, for the last statement, we will first bound $\|\sqrt{\log Y}\|_\Psi$ where $\Psi(t) = \frac{1}{5}\exp(t^2)$. To this end, it suffices to find a $B > 0$ such that

$$\frac{1}{5}\mathbb{E}\left[\exp(\log(Y)/B^2)\right] \le 1 \;\text{ or, equivalently, }\; \mathbb{E}\left[Y^{1/B^2}\right] \le 5.$$

We have two possibilities: Suppose first $\mathbb{E}[Y] \le 5$. Then $B = 1$ is feasible above, and so $\|\sqrt{\log(Y)}\|_\Psi \le 1$.

On the other hand, suppose $\mathbb{E}[Y] > 5$. Consider $\theta = \frac{4}{\mathbb{E}[Y]-1} \in (0,1)$. Then, from convexity of the function $t \mapsto \mathbb{E}[Y^t]$,

$$\mathbb{E}[Y^\theta] \;\le\; \theta\mathbb{E}[Y^1] + (1-\theta)\mathbb{E}[Y^0] \;=\; \theta\mathbb{E}[Y] + (1-\theta) \;=\; 5.$$

Thus, if we let $B = \sqrt{\mathbb{E}[Y]}/2$, we have

$$\mathbb{E}\left[Y^{1/B^2}\right] \;=\; \mathbb{E}\left[Y^{4/\mathbb{E}[Y]}\right] \;\le\; \mathbb{E}\left[Y^{4/(\mathbb{E}[Y]-1)}\right] \;=\; \mathbb{E}[Y^\theta] \le 5.$$

Hence, $\|\sqrt{\log(Y)}\|_\Psi \le \sqrt{\mathbb{E}[Y]}/2$. Combining both cases proves $\|\sqrt{\log(Y)}\|_\Psi \le \max(1, \sqrt{\mathbb{E}[Y]}/2)$.

Apply Part iii) to complete the proof. □

LEMMA B.5 **(Properties of Poisson Random Variables)**. *Suppose* $\hat{N}_k \sim \text{Poisson}(N_k)$, *for* $k = 1, \ldots, K$, *where* $N_k \ge 1$ *for all* k, *and* $K \ge 2$. *Let* $\hat{N}_{\max} \equiv \max_k \hat{N}_k$, $N_{\max} \equiv \max_k N_k$, $\hat{N}_{\min} \equiv \min_k \hat{N}_k$ *and* $N_{\min} \equiv \min_k N_k$. *Then for any* $p \ge 1$*:*

i) $\mathbb{E}\left[\exp\left(\frac{\hat{N}_k}{2N_k}\right)\right] \;\le\; 2,$

ii) $\mathbb{E}\left[\exp\left(\frac{N_k}{2(\hat{N}_k+1)}\right)\right] \;\le\; 2,$

iii) $\mathbb{E}\left[\exp\left(\frac{\hat{N}_{\max}}{2(1+\log K)N_{\max}}\right)\right] \;\le\; 6,$

iv) $\mathbb{E}\left[\exp\left(\frac{N_{\min}}{2(1+\log(K))(\hat{N}_{\min}+1)}\right)\right] \;\le\; 6,$

v) $\|\hat{N}_{\max}\|_p \;\le\; 6^{1/p}\left(\frac{2p}{e}\right) N_{\max}(1+\log(K)) \;\le\; 6^{1/p}\left(\frac{6p}{e}\right) N_{\max}\log(K),$

vi) $\left\|\sqrt{\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}}\right\|_p \;\le\; 6^{1/p}\left(\frac{6p}{e}\right)\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\cdot\log(K).$

Proof.

Part i) Let $\beta_0 \equiv \log\left(1 + \frac{\log 2}{N_k}\right)$. From the Poisson moment generating function,

$$\mathbb{E}\left[\exp(\beta_0\hat{N}_k)\right] \;=\; \exp(N_k(e_0^\beta - 1)) = 2.$$

Thus, to prove i), it suffices to show that $\beta_0 = \log\left(1+\frac{\log 2}{N_k}\right) \geq \frac{1}{2N_k}$. The function $N \mapsto \log\left(1+\frac{\log 2}{N}\right) - \frac{1}{2N}$ is positive at $N=1$ and tends to zero as $N \to \infty$. By differentiating, we see it has one critical point at $N = \frac{\log 2}{2\log 2 - 1}$ which by inspection is a maximum. Hence, it is always non-negative, proving the claim and the first statement.

Part ii) Use the Poisson probability mass function to write

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{N_k}{2(\hat{N}_k+1)}\right)\right] &= e^{-N_k}\sum_{n=0}^{\infty}\frac{N_k^n}{n!}\cdot\exp\left(\frac{N_k}{2(n+1)}\right) \\
&= e^{-N_k}\sum_{n=0}^{\infty}\frac{N_k^n}{n!}\cdot\sum_{j=0}^{\infty}\left(\frac{N_k}{2(n+1)}\right)^j\frac{1}{j!} \\
&= e^{-N_k}\sum_{j=0}^{\infty}\frac{1}{j!}\left(\frac{N_k}{2}\right)^j\cdot\sum_{n=0}^{\infty}\frac{N_k^n}{n!}\left(\frac{1}{n+1}\right)^j,
\end{aligned}$$

where the first equality uses the Taylor expansion of $\exp(\cdot)$ and the second from reversing the summations. Since $\frac{1}{n+1} \leq \frac{i}{n+i}$ for all $n, i \geq 1$, we obtain that

$$\left(\frac{1}{n+1}\right)^j \leq \frac{1}{n+1}\cdot\frac{2}{n+2}\cdots\frac{j}{n+j} = \frac{n!j!}{(n+j)!}.$$

Substituting above yields

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\frac{N_k}{2(\hat{N}_k+1)}\right)\right] &\leq e^{-N_k}\sum_{j=0}^{\infty}\frac{1}{j!}\left(\frac{N_k}{2}\right)^j\cdot\sum_{n=0}^{\infty}\frac{N_k^n}{n!}\frac{n!j!}{(n+j)!} \\
&= e^{-N_k}\sum_{j=0}^{\infty}\frac{1}{2^j}\cdot\sum_{n=0}^{\infty}\frac{N_k^{n+j}}{(n+j)!} \\
&= e^{-N_k}\sum_{j=0}^{\infty}\frac{1}{2^j}\cdot\sum_{n=j}^{\infty}\frac{N_k^n}{n!} \\
&\leq e^{-N_k}\sum_{j=0}^{\infty}\frac{1}{2^j}\cdot\sum_{n=0}^{\infty}\frac{N_k^n}{n!} \\
&= 2.
\end{aligned}$$

Parts iii) and iv) These results follow by combining Lemma B.3 with parts i) and ii) respectively.

Part v) Let $\beta = \frac{1}{2(1+\log K)N_{\max}}$. Then, from Lemma B.4 Part i),

$$\mathbb{E}[\hat{N}_{\max}^p] = \beta^{-p}\mathbb{E}[(\beta\hat{N}_{\max})^p] \leq \beta^{-p}\left(\frac{p}{e}\right)^p\mathbb{E}[\exp(\beta\hat{N}_{\max})] \leq 6\left(\frac{2p}{e}\right)^p N_{\max}^p(1+\log K)^p,$$

where the second inequality uses Part iii). Taking the p^{th} root of both sides proves the first statement. The second follows because $K \geq 2$ implies that $1+\log K \leq 3\log K$.

Part vi) Applying an identical argument to the previous part but with Part iv) , we have

$$\mathbb{E}\left[(\hat{N}_{\min}+1)^{-p}\right] \;\le\; 6\left(\frac{2p}{eN_{\min}}\right)^{p}(1+\log K)^{p}.$$

Therefore, we have

$$\begin{aligned}
\left\|\sqrt{\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}}\right\|_p^p &= \mathbb{E}\left[\left(\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}\right)^{p/2}\right] \\
&\le \sqrt{\mathbb{E}\hat{N}_{\max}^{p}}\cdot\sqrt{\mathbb{E}\left[(\hat{N}_{\min}+1)^{-p}\right]} \qquad \text{(Cauchy-Schwarz Inequality)} \\
&\le 6\cdot 2^{p}\left(\log(K)+1\right))^{p}e^{-p}p^{p}\left(\frac{N_{\max}}{N_{\min}}\right)^{p/2} \\
&= 6\cdot\left(\frac{2p}{e}\right)^{p}(\log(K)+1)^{p}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{p/2}.
\end{aligned}$$

Finally, since $K\ge 2$, we have $1+\log K\le 3\log K$. Making this substitution and simplifying completes the proof. □

Appendix C: Deferred Proofs for Sub-Optimality Guarantees from Section 4

In this section, we provide the complete proofs for the high-probability sub-optimality bounds presented in Section 4.

C.1. Proof of Theorem 4.2: Shrunken-SAA with Fixed Anchors for Strongly Convex Problems

We first prove the results summarized in Section 4.2.

C.1.1. Proof of continuity lemma and packing number bounds As mentioned in the main text, the key idea is to establish continuity of the solutions $\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)$ in the parameters.

LEMMA C.1 **(Continuity properties of $\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)$).** *Under the assumptions of Theorem 4.2,*

i) (Continuity in anchor) For any $\alpha\ge 0$, and any $\boldsymbol{p},\overline{\boldsymbol{p}}\in\Delta_d$,

$$\|\boldsymbol{x}_k(\alpha,\boldsymbol{p},\hat{\boldsymbol{m}}_k)-\boldsymbol{x}_k(\alpha,\overline{\boldsymbol{p}},\hat{\boldsymbol{m}}_k)\|_2 \;\le\; \frac{L}{\gamma}\cdot\|\boldsymbol{p}-\overline{\boldsymbol{p}}\|_1.$$

ii) (Continuity in $\hat{\boldsymbol{m}}_k$) For any $\hat{\boldsymbol{m}}_k$ such that $\hat{N}_k\ge 1$ we have

$$\|\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)-\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i)\|_2 \;\le\; \frac{4L}{\gamma\hat{N}_k}.$$

iii) (Continuity in α) For any $\alpha,\overline{\alpha}\ge 0$, and $\boldsymbol{p}_0\in\Delta_d$,

$$\|\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)-\boldsymbol{x}_k(\overline{\alpha},\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)\|_2 \;\le\; \frac{4L}{\gamma}\cdot\frac{|\alpha-\overline{\alpha}|}{\hat{N}_k+1}.$$

iv) (Continuity at $\alpha=\infty$) For any $\alpha\geq 0$ and $\boldsymbol{p}_0\in\Delta_d$ such that $\max(\alpha,\hat{N}_k)>0$,

$$\|\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)-\boldsymbol{x}_k(\infty,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)\|_2 \;\leq\; \frac{2L}{\gamma}\frac{\hat{N}_k}{\hat{N}_k+\alpha}.$$

Proof. Fix k. For any $\boldsymbol{q}\in\Delta_d$, define

$$f_{\boldsymbol{q}}(\boldsymbol{x})\equiv \boldsymbol{q}^\top \boldsymbol{c}_k(\boldsymbol{x}), \qquad \boldsymbol{x}(\boldsymbol{q})\in\arg\min_{\boldsymbol{x}\in\mathcal{X}_k} f_{\boldsymbol{q}}(\boldsymbol{x}).$$

We first prove the general inequality for any $\boldsymbol{q},\overline{\boldsymbol{q}}\in\Delta_d$,

$$\|\boldsymbol{x}(\boldsymbol{q})-\boldsymbol{x}(\overline{\boldsymbol{q}})\|_2\leq \frac{L}{\gamma}\cdot\|\boldsymbol{q}-\overline{\boldsymbol{q}}\|_1. \tag{EC.C.1}$$

We will then use this general purpose inequality to prove the various parts of the lemma by choosing particular values for $\boldsymbol{q}$ and $\overline{\boldsymbol{q}}$.

Note that since each $c_{ki}(\boldsymbol{x})$ is γ-strongly convex for each i, $f_{\boldsymbol{q}}(\boldsymbol{x})$ is also γ-strongly convex. From the first-order optimality conditions, $\nabla f_{\boldsymbol{q}}(\boldsymbol{x}(\boldsymbol{q}))^\top(\boldsymbol{x}(\overline{\boldsymbol{q}})-\boldsymbol{x}(\boldsymbol{q}))\geq 0$. Then, from strong-convexity,

$$\begin{aligned} f_{\boldsymbol{q}}(\boldsymbol{x}(\overline{\boldsymbol{q}}))-f_{\boldsymbol{q}}(\boldsymbol{x}(\boldsymbol{q})) \;&\geq\; \nabla f_{\boldsymbol{q}}(\boldsymbol{x}(\boldsymbol{q}))^\top(\boldsymbol{x}(\overline{\boldsymbol{q}})-\boldsymbol{x}(\boldsymbol{q}))+\frac{\gamma}{2}\|\boldsymbol{x}(\boldsymbol{q})-\boldsymbol{x}(\overline{\boldsymbol{q}})\|_2^2\\ &\geq\; \frac{\gamma}{2}\|\boldsymbol{x}(\boldsymbol{q})-\boldsymbol{x}(\overline{\boldsymbol{q}})\|_2^2. \end{aligned}$$

A symmetric argument holds switching $\boldsymbol{q}$ and $\overline{\boldsymbol{q}}$ yielding

$$f_{\overline{\boldsymbol{q}}}(\boldsymbol{x}(\boldsymbol{q}))-f_{\overline{\boldsymbol{q}}}(\boldsymbol{x}(\overline{\boldsymbol{q}}))\geq\frac{\gamma}{2}\|\boldsymbol{x}(\boldsymbol{q})-\boldsymbol{x}(\overline{\boldsymbol{q}})\|_2^2.$$

Adding yields,

$$\begin{aligned} \gamma\|\boldsymbol{x}(\boldsymbol{q})-\boldsymbol{x}(\overline{\boldsymbol{q}})\|_2^2 \;&\leq\; \Big(f_{\overline{\boldsymbol{q}}}(\boldsymbol{x}(\boldsymbol{q}))-f_{\boldsymbol{q}}(\boldsymbol{x}(\boldsymbol{q}))\Big)+\Big(f_{\boldsymbol{q}}(\boldsymbol{x}(\overline{\boldsymbol{q}}))-f_{\overline{\boldsymbol{q}}}(\boldsymbol{x}(\overline{\boldsymbol{q}}))\Big)\\ &=\; (\overline{\boldsymbol{q}}-\boldsymbol{q})^\top\big(\boldsymbol{c}_k(\boldsymbol{x}(\boldsymbol{q}))-\boldsymbol{c}_k(\boldsymbol{x}(\overline{\boldsymbol{q}}))\big)\\ &\leq\; \big\|\boldsymbol{c}_k(\boldsymbol{x}(\boldsymbol{q}))-\boldsymbol{c}_k(\boldsymbol{x}(\overline{\boldsymbol{q}}))\big\|_\infty\|\boldsymbol{q}-\overline{\boldsymbol{q}}\|_1\\ &\leq\; L\|\boldsymbol{x}(\boldsymbol{q})-\boldsymbol{x}(\overline{\boldsymbol{q}})\|_2\|\boldsymbol{q}-\overline{\boldsymbol{q}}\|_1, \end{aligned}$$

by the Hölder inequality and assumed Lipschitz constant. Rearranging proves Eq. (EC.C.1).

We can now prove each part of the lemma.

Part i) First suppose $\alpha+\hat{N}_k>0$. Take

$$\boldsymbol{q}=\frac{\alpha}{\alpha+\hat{N}_k}\boldsymbol{p}+\frac{\hat{N}_k}{\hat{N}_k+\alpha}\hat{\boldsymbol{p}}_k, \quad \text{and} \quad \overline{\boldsymbol{q}}=\frac{\alpha}{\alpha+\hat{N}_k}\overline{\boldsymbol{p}}+\frac{\hat{N}_k}{\hat{N}_k+\alpha}\hat{\boldsymbol{p}}_k.$$

Then, $\|\boldsymbol{q}-\overline{\boldsymbol{q}}\|_1 \;=\; \frac{\alpha}{\hat{N}_k+\alpha}\|\boldsymbol{p}-\overline{\boldsymbol{p}}\|_1 \;\leq\; \|\boldsymbol{p}-\overline{\boldsymbol{p}}\|_1$. Substituting into Eq. (EC.C.1) proves the result in this case. Next, suppose $\alpha+\hat{N}_k=0$. Then, applying Eq. (EC.C.1) with $\boldsymbol{q}=\boldsymbol{p}$ and $\overline{\boldsymbol{q}}=\overline{\boldsymbol{p}}$ yields the result.

Part ii) First suppose $\hat{N}_k \geq 2$. Take

$$\boldsymbol{q} = \frac{\alpha}{\hat{N}_k + \alpha}\boldsymbol{p}_0 + \frac{1}{\hat{N}_k + \alpha}\hat{\boldsymbol{m}}_k \quad \text{and} \quad \overline{\boldsymbol{q}} = \frac{\alpha}{\hat{N}_k + \alpha - 1}\boldsymbol{p}_0 + \frac{1}{\hat{N}_k + \alpha - 1}(\hat{\boldsymbol{m}}_k - \boldsymbol{e}_i).$$

Then,

$$\begin{aligned}
\|\boldsymbol{q} - \overline{\boldsymbol{q}}\|_1 &\leq \left|\frac{\alpha}{\hat{N}_k + \alpha} - \frac{\alpha}{\hat{N}_k + \alpha - 1}\right| \|\boldsymbol{p}_0\|_1 + \left|\frac{1}{\hat{N}_k + \alpha} - \frac{1}{\hat{N}_k + \alpha - 1}\right| \|\hat{\boldsymbol{m}}_k\|_1 + \frac{1}{\hat{N}_k + \alpha - 1} \\
&= \frac{2}{\hat{N}_k - 1 + \alpha} \\
&\leq \frac{4}{\hat{N}_k}.
\end{aligned}$$

Substituting into Eq. (EC.C.1) proves the result when $\hat{N}_k \geq 2$.

Next, when $\hat{N}_k = 1$, let $\boldsymbol{q}$ be as above and $\overline{\boldsymbol{q}} = \boldsymbol{p}_0$. Then $\|\boldsymbol{q} - \overline{\boldsymbol{q}}\|_1 \leq 2 \leq \frac{4}{\hat{N}_k}$. Again, substituting into Eq. (EC.C.1) proves the result.

Part iii) Notice if $\hat{N}_k = 0$, then $\|\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) - \boldsymbol{x}_k(\overline{\alpha}, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)\|_2 = 0$ and the bounds holds trivially. Hence, suppose $\hat{N}_k \geq 1$. Consider taking $\boldsymbol{q} = \hat{\boldsymbol{p}}_k(\alpha)$ and $\overline{\boldsymbol{q}} = \hat{\boldsymbol{p}}_k(\overline{\alpha})$. Then

$$\begin{aligned}
\|\boldsymbol{q} - \overline{\boldsymbol{q}}\|_1 &= \left\| \left(\left(\frac{\alpha}{\hat{N}_k + \alpha} - \frac{\overline{\alpha}}{\hat{N}_k + \overline{\alpha}} \right) \boldsymbol{p}_0 + \left(\frac{\hat{N}_k}{\hat{N}_k + \alpha} - \frac{\hat{N}_k}{\hat{N}_k + \overline{\alpha}} \right) \hat{\boldsymbol{p}}_k \right) \right\|_1 \\
&\leq \left(\left| \frac{\alpha}{\hat{N}_k + \alpha} - \frac{\overline{\alpha}}{\hat{N}_k + \overline{\alpha}} \right| + \left| \frac{\hat{N}_k}{\hat{N}_k + \alpha} - \frac{\hat{N}_k}{\hat{N}_k + \overline{\alpha}} \right| \right) \\
&= 2 \left| \frac{\hat{N}_k}{\hat{N}_k + \alpha} - \frac{\hat{N}_k}{\hat{N}_k + \overline{\alpha}} \right|, \\
&= \frac{2\hat{N}_k \, |\alpha - \overline{\alpha}|}{(\hat{N}_k + \alpha)(\hat{N}_k + \overline{\alpha})},
\end{aligned}$$

where second equality follows because $\left| \frac{\alpha}{\hat{N}_k + \alpha} - \frac{\overline{\alpha}}{\hat{N}_k + \overline{\alpha}} \right| = \left| \frac{\hat{N}_k}{\hat{N}_k + \alpha} - \frac{\hat{N}_k}{\hat{N}_k + \overline{\alpha}} \right|$. Next write,

$$\frac{2\hat{N}_k \, |\alpha - \overline{\alpha}|}{(\hat{N}_k + \alpha)(\hat{N}_k + \overline{\alpha})} \leq \frac{2 \, |\alpha - \overline{\alpha}|}{(\hat{N}_k + \overline{\alpha})} \leq \frac{2 \, |\alpha - \overline{\alpha}|}{\hat{N}_k} \leq \frac{4 \, |\alpha - \overline{\alpha}|}{(\hat{N}_k + 1)},$$

where the last inequality follows because $\frac{1}{N} \leq \frac{2}{N+1}$ for $N \geq 1$.

Substituting into Eq. (EC.C.1) completes the proof of part iii).

Part iv) Take $\boldsymbol{q} = \boldsymbol{p}_0$ and $\overline{\boldsymbol{q}} = \hat{\boldsymbol{p}}_k(\alpha)$. Then,

$$\begin{aligned}
\|\boldsymbol{q} - \overline{\boldsymbol{q}}\|_1 &= \left\| \left(1 - \frac{\alpha}{\hat{N}_k + \alpha} \right) \boldsymbol{p}_0 + \left(0 - \frac{\hat{N}_k}{\hat{N}_k + \alpha} \right) \hat{\boldsymbol{p}}_k \right\|_1 \\
&\leq \left| 1 - \frac{\alpha}{\hat{N}_k + \alpha} \right| + \left| 0 - \frac{\hat{N}_k}{\hat{N}_k + \alpha} \right| \\
&= 2 \frac{\hat{N}_k}{\hat{N}_k + \alpha}.
\end{aligned}$$

Again, substituting into Eq. (EC.C.1) proves the inequality. □

LEMMA C.2 **(Packing Numbers for Strongly-Convex Problems)**. *Under the assumptions of Theorem 4.2, we have for any* $0 < \epsilon \leq 1$,

$$D\left(\epsilon\|\mathbf{F}^{\mathsf{Perf}}\|_2,\ \{\mathbf{Z}(\alpha,\boldsymbol{p}_0):\alpha\geq 0\}\right) \ \leq\ 2+\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}\frac{32L^4}{C^2\gamma^2\epsilon^2}, \tag{EC.C.2}$$

$$D\left(\epsilon\|\mathbf{F}^{\mathsf{LOO}}\|_2,\ \left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0):\alpha\geq 0\right\}\right) \ \leq\ 2+\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}\frac{32L^4}{C^2\gamma^2\epsilon^2}. \tag{EC.C.3}$$

Proof. We first prove Eq. (EC.C.2). We proceed by constructing an $\frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{Perf}}\|_2$-covering. The desired packing number is at most the size of this covering. Recall, by Lemma 4.2, $\|\mathbf{F}^{\mathsf{Perf}}\|_2^2 = \frac{C^2}{\lambda_{\text{avg}}^2}\|\boldsymbol{\lambda}\|_2^2$, and let $Z_k(\infty,\boldsymbol{p}_0) = \frac{1}{\lambda_{\text{avg}}}\sum_{i=1}^d \lambda_k p_{ki} c_{ki}(\boldsymbol{x}_k(\infty,\boldsymbol{p}_0))$.

First, suppose $\hat{N}_{\max}=0$, which implies $\hat{N}_k=0$ for all $k=1,\ldots,K$. In this case, $\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k) = \boldsymbol{x}_k(\infty,\boldsymbol{p}_0)$ for all k, whereby $\{\mathbf{Z}(\alpha,\boldsymbol{p}_0):\alpha\geq 0\} = \{\mathbf{Z}(\infty,\boldsymbol{p}_0)\}$, and the covering number is 1, so the above bound is valid.

Now suppose $\hat{N}_{\max}>0$. Let $\alpha_{\max} = \frac{4L^2\hat{N}_{\max}}{C\gamma\epsilon}$. For any $\alpha\geq\alpha_{\max}>0$,

$$\begin{aligned}
|Z_k(\alpha,\boldsymbol{p}_0)-Z_k(\infty,\boldsymbol{p}_0)| \ &\leq\ \frac{\lambda_k}{\lambda_{\text{avg}}}\sum_{i=1}^d p_{ki}\,|c_{ki}(\boldsymbol{x}_k(\alpha,\hat{\boldsymbol{m}}))-c_{ki}(\boldsymbol{x}_k(\infty))| \\
&\leq\ \frac{\lambda_k}{\lambda_{\text{avg}}}\sum_{i=1}^d p_{ki}L\|\boldsymbol{x}_k(\alpha,\hat{\boldsymbol{m}}))-\boldsymbol{x}_k(\infty)\|_2 && \text{(Lipschitz continuity)} \\
&\leq\ \frac{\lambda_k}{\lambda_{\text{avg}}}\frac{2L^2}{\gamma}\cdot\frac{\hat{N}_k}{\hat{N}_k+\alpha} && \text{(Lemma C.1, part iv) since } \alpha>0).
\end{aligned}$$

It follows that for all $\alpha\geq\alpha_{\max}$ we have

$$\begin{aligned}
\|\mathbf{Z}(\alpha,\boldsymbol{p}_0)-\mathbf{Z}(\infty,\boldsymbol{p}_0)\|_2 \ &\leq\ \left(\frac{4L^4}{\lambda_{\text{avg}}^2\gamma^2}\sum_{k=1}^K\lambda_k^2\left(\frac{\hat{N}_k}{\hat{N}_k+\alpha}\right)^2\right)^{1/2} \\
&\leq\ \frac{2L^2\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}\gamma}\left(\frac{\hat{N}_{\max}}{\hat{N}_{\max}+\alpha}\right) \\
&\leq\ \frac{2L^2\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}\gamma}\left(\frac{\hat{N}_{\max}}{\hat{N}_{\max}+\alpha_{\max}}\right) \\
&\leq\ \frac{2L^2\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}\gamma}\left(\frac{1}{1+\frac{4L^2}{C\gamma\epsilon}}\right) \\
&\leq\ \frac{2L^2\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}\gamma}\cdot\frac{C\gamma\epsilon}{4L^2} \\
&=\ \frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{Perf}}\|.
\end{aligned}$$

Thus, in our covering, we place one point at $\mathbf{Z}(\infty,\boldsymbol{p}_0)$ to cover all points $\mathbf{Z}(\alpha,\boldsymbol{p}_0)$ with $\alpha\geq\alpha_{\max}$.

Next let $\{\alpha_1,\dots,\alpha_M\}$ be a $\frac{\gamma(\hat{N}_{\min}+1)C\epsilon}{8L^2}$ covering of $[0,\alpha_{\max}]$. Note, $M \le 1+\frac{8L^2\alpha_{\max}}{\gamma(\hat{N}_{\min}+1)C\epsilon}$. We claim $\{\mathbf{Z}(\alpha,\boldsymbol{p}_0),\dots,\mathbf{Z}(\alpha_M,\boldsymbol{p}_0)\}$ is an $\frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{Perf}}\|$-covering of $\{\mathbf{Z}(\alpha,\boldsymbol{p}_0):\alpha\in[0,\alpha_{\max}]\}$. Indeed, for any $\alpha\in[0,\alpha_{\max}]$, let α_j be the nearest element of the α-covering. Then,

$$\begin{aligned}
|Z_k(\alpha,\boldsymbol{p}_0)-Z_k(\alpha_j,\boldsymbol{p}_0)| &\le \frac{\lambda_k}{\lambda_{\text{avg}}}\sum_{i=1}^d p_{ki}\,|c_{ki}(\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_j))-c_{ki}(\boldsymbol{x}_k(\alpha_j,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k))|\\
&\le \frac{\lambda_k}{\lambda_{\text{avg}}}\sum_{i=1}^d p_{ki}L\|\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_j)-\boldsymbol{x}_k(\alpha_j,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)\|_2\\
&\le \frac{\lambda_k}{\lambda_{\text{avg}}}\frac{4L^2}{\gamma(\hat{N}_{\min}+1)}|\alpha-\alpha_j| && \text{(Lemma C.1, part iii))}\\
&\le \frac{\lambda_k}{\lambda_{\text{avg}}}\frac{4L^2}{\gamma(\hat{N}_{\min}+1)}\cdot\frac{\gamma(\hat{N}_{\min}+1)C\epsilon}{8L^2}\\
&= \frac{C\epsilon\lambda_k}{2\lambda_{\text{avg}}}
\end{aligned}$$

Thus, $\|\mathbf{Z}(\alpha,\boldsymbol{p}_0)-\mathbf{Z}(\alpha_j,\boldsymbol{p}_0)\|_2 \le \frac{C\epsilon\|\boldsymbol{\lambda}\|_2}{2\lambda_{\text{avg}}} = \frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{Perf}}\|$ as was to be shown.

The total size of the covering is thus

$$1+M \;\le\; 2+\frac{8L^2\alpha_{\max}}{\gamma(1+\hat{N}_{\min})C\epsilon} \;=\; 2+\frac{\hat{N}_{\max}}{1+\hat{N}_{\min}}\frac{32L^4}{C^2\gamma^2\epsilon^2}.$$

We next prove Eq. (EC.C.3). We again proceed by constructing an $\frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{LOO}}\|$-covering, since the desired packing is at most the size of this covering. Recall by Lemma 4.2, $\|\mathbf{F}^{\mathsf{LOO}}\|_2^2 = \frac{C^2}{N^2\lambda_{\text{avg}}^2}\|\hat{\boldsymbol{N}}\|_2^2$.

If $\hat{N}_{\max}=0$, then $\hat{N}_k=0$ for all k, and $\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0):\alpha\ge 0\}=\{\mathbf{0}\}$, so this covering number is 1.

Otherwise, $\hat{N}_{\max}>0$. Let $\alpha_{\max}=\frac{4\hat{N}_{\max}L^2}{C\gamma\epsilon}$. Then, for any $\alpha\ge\alpha_{\max}>0$,

$$\begin{aligned}
\left|Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)-Z_k^{\mathsf{LOO}}(\infty,\boldsymbol{p}_0)\right| &\le \frac{1}{N\lambda_{\text{avg}}}\sum_{i=1}^d \hat{m}_{ki}\,|c_{ki}(\boldsymbol{x}_k(\alpha,\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i))-c_{ki}(\boldsymbol{x}_k(\infty))|\\
&\le \frac{L}{N\lambda_{\text{avg}}}\sum_{i=1}^d \hat{m}_{ki}\|\boldsymbol{x}_k(\alpha,\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i)-\boldsymbol{x}_k(\infty)\|_2 && \text{(Lipschitz-Continuity)}\\
&\le \frac{L}{N\lambda_{\text{avg}}}\sum_{i=1}^d \hat{m}_{ki}\frac{2L}{\gamma}\frac{\hat{N}_k-1}{\hat{N}_k-1+\alpha} && \text{(Lemma C.1, part iv))}\\
&\le \frac{2L^2\hat{N}_k}{\gamma N\lambda_{\text{avg}}}\frac{\hat{N}_k}{\hat{N}_k+\alpha},
\end{aligned}$$

because $x\mapsto\frac{x}{x+\alpha}$ is an increasing function. Thus, for any $\alpha\ge\alpha_{\max}$,

$$\begin{aligned}
\|\mathbf{Z}_k^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)-\mathbf{Z}_k^{\mathsf{LOO}}(\infty,\boldsymbol{p}_0)\|_2 &\le \frac{2L^2}{\gamma}\left(\sum_{k=1}^K\frac{\hat{N}_k^2}{N^2\lambda_{\text{avg}}^2}\cdot\left(\frac{\hat{N}_k}{\hat{N}_k+\alpha}\right)^2\right)^{1/2}\\
&\le \frac{2L^2}{\gamma C}\left(\sum_{k=1}^K\frac{C^2\hat{N}_k^2}{N^2\lambda_{\text{avg}}^2}\right)^{1/2}\cdot\frac{\hat{N}_{\max}}{\hat{N}_{\max}+\alpha}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{2L^2}{\gamma C}\|\mathbf{F}^{\mathsf{LOO}}\|_2 \ \frac{\hat{N}_{\max}}{\hat{N}_{\max}+\alpha} \\
&\leq \frac{2L^2}{\gamma C}\|\mathbf{F}^{\mathsf{LOO}}\|_2 \ \frac{\hat{N}_{\max}}{\hat{N}_{\max}+\alpha_{\max}} \\
&= \frac{2L^2}{\gamma C}\|\mathbf{F}^{\mathsf{LOO}}\|_2 \ \frac{1}{1+\frac{4L^2}{C\gamma\epsilon}} \\
&\leq \frac{2L^2}{\gamma C}\|\mathbf{F}^{\mathsf{LOO}}\|_2 \ \frac{C\gamma\epsilon}{4L^2} \\
&= \frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{LOO}}\|_2
\end{aligned}
$$

Thus, in our covering, we place one point at $\mathbf{Z}^{\mathsf{LOO}}(\infty, \boldsymbol{p}_0)$ to cover all points $\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)$ for $\alpha \geq \alpha_{\max}$.

Next let $\{\alpha_1, \ldots, \alpha_M\}$ be a $\frac{\gamma(\hat{N}_{\min}+1)C\epsilon}{8L^2}$-covering of $[0, \alpha_{\max}]$. Note, $M \leq 1 + \frac{8L^2\alpha_{\max}}{\gamma(\hat{N}_{\min}+1)C\epsilon}$. We claim this covering induces an $\frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{LOO}}\|_2$-covering of $\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \in [0, \alpha_{\max}]\}$. Indeed, for any $\alpha \in [0, \alpha_{\max}]$, let α_j be the nearest element of the α-covering. Then, for any k such that $\hat{N}_k \geq 1$,

$$
\begin{aligned}
&\left|Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) - Z_k^{\mathsf{LOO}}(\alpha_j, \boldsymbol{p}_0)\right| \\
&\qquad \leq \frac{1}{N\lambda_{\text{avg}}} \sum_{i=1}^{d} \hat{m}_{ki} \left| c_{ki}(\boldsymbol{x}_k(\alpha, \hat{\boldsymbol{m}}_{ki} - \boldsymbol{e}_i)) - c_{ki}(\boldsymbol{x}_k(\alpha_j, \hat{\boldsymbol{m}}_{ki} - \boldsymbol{e}_i))\right| \\
&\qquad \leq \frac{L}{N\lambda_{\text{avg}}} \sum_{i=1}^{d} \hat{m}_{ki} \|\boldsymbol{x}_k(\alpha, \hat{\boldsymbol{m}}_{ki} - \boldsymbol{e}_i)) - \boldsymbol{x}_k(\alpha_j, \hat{\boldsymbol{m}}_{ki} - \boldsymbol{e}_i)\|_2 \qquad \text{(Lipschitz Continuity)} \\
&\qquad \leq \frac{\hat{N}_k}{N\lambda_{\text{avg}}} \cdot \frac{4L^2}{\gamma(\hat{N}_{\min}+1)} \cdot |\alpha - \alpha_j| \qquad \text{(Lemma C.1, part iii))} \\
&\qquad \leq \frac{\hat{N}_k}{N\lambda_{\text{avg}}} \cdot \frac{4L^2}{\gamma(\hat{N}_{\min}+1)} \cdot \frac{\gamma(\hat{N}_{\min}+1)C\epsilon}{8L^2} \\
&\qquad = C\frac{\hat{N}_k}{N\lambda_{\text{avg}}}\frac{\epsilon}{2}.
\end{aligned}
$$

On the other hand, for any k such that $\hat{N}_k = 0$, $\left|Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) - Z_k^{\mathsf{LOO}}(\alpha_j, \boldsymbol{p}_0)\right| = 0$. In total, this implies $\|\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) - \mathbf{Z}^{\mathsf{LOO}}(\alpha_j, \boldsymbol{p}_0)\|_2^2 \leq \frac{\epsilon^2}{4}\frac{C^2}{N^2\lambda_{\text{avg}}^2}\|\hat{\mathbf{N}}\|_2^2$, which implies $\|\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) - \mathbf{Z}^{\mathsf{LOO}}(\alpha_j, \boldsymbol{p}_0)\| \leq \frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{LOO}}\|_2$, as was to be proven.

Thus, the total size of the covering is at most

$$
1 + M \ \leq \ 2 + \frac{8L^2\alpha_{\max}}{\gamma(1+\hat{N}_{\min})C\epsilon} \ = \ 2 + \frac{\hat{N}_{\max}}{1+\hat{N}_{\min}}\frac{32L^4}{C^2\gamma^2\epsilon^2}.
$$

This completes the proof. □

C.1.2. Maximal deviation bounds. We next use the above lemmas to bound the maximal deviations of interest via Theorem 4.1.

LEMMA C.3 **(Bounding the Maximal Deviations)**. *Suppose* $\frac{4L^2}{C\gamma} \geq 1$. *Then, under the assumptions of Theorem 4.2, there exists a universal constant* A *such that for any* $0 < \delta < 1/2$, *the following two statements each hold (separately) with probability at least* $1-\delta$*:*

$$\sup_{\alpha \geq 0} \left| \frac{1}{K} \sum_{k=1}^{K} (Z_k(\alpha, \boldsymbol{p}_0) - \mathbb{E}[Z_k(\alpha, \boldsymbol{p}_0)]) \right| \leq \mathrm{A} \cdot L \sqrt{\frac{C}{\gamma}} \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{5/4} \cdot \frac{\log(1/\delta) \cdot \sqrt{\log(K)}}{\sqrt{K}},$$

$$\sup_{\alpha \geq 0} \left| \frac{1}{K} \sum_{k=1}^{K} \left(Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) - \mathbb{E}[Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)] \right) \right| \leq \mathrm{A} \cdot L \sqrt{\frac{C}{\gamma}} \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^{5/4} \cdot \frac{\log^2(1/\delta) \cdot \log^{3/2}(K)}{\sqrt{K}}.$$

Proof. To prove the first inequality, our strategy will be to apply Theorem 4.1 to the process $\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}$. To that end, we first bound the variable J in Eq. (4.3). Recall by Lemma 4.2, the size of the envelope is at most $C\frac{\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}}$. Using the bound on the packing numbers from Lemma C.2,

$$J \;\leq\; 9C \frac{\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}} \int_0^1 \sqrt{\log\left(2 + \frac{\hat{N}_{\max}}{\hat{N}_{\min}+1} \frac{32L^4}{C^2\gamma^2\epsilon^2}\right)} d\epsilon \;\leq\; 9C \frac{\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}} \int_0^1 \sqrt{\log\left(\frac{t}{\epsilon^2}\right)} d\epsilon$$

where the second inequality uses $2 \leq 2/\epsilon^2$ and $t = 2 + \frac{\hat{N}_{\max}}{\hat{N}_{\min}+1} \frac{32L^4}{C^2\gamma^2}$ Substitute $\log(t/\epsilon^2) = 2\log(\sqrt{t}/\epsilon)$ in the integral above, and then apply Lemma B.1, yielding

$$\begin{aligned} J \;&\leq\; 9\sqrt{2} \cdot C \frac{\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}} \left(\sqrt{\pi}/2 + \sqrt{\log\left(\sqrt{2 + \frac{\hat{N}_{\max}}{\hat{N}_{\min}+1} \frac{32L^4}{C^2\gamma^2}} \right)} \right) \\ &\leq\; 9\sqrt{2} \cdot C \frac{\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}} (\sqrt{\pi}+1) \sqrt{\log\left(\sqrt{2 + \frac{\hat{N}_{\max}}{\hat{N}_{\min}+1} \frac{32L^4}{C^2\gamma^2}} \right)}, \end{aligned}$$

where in the second inequality we have used $\sqrt{\pi \log\left(\sqrt{2 + \frac{\hat{N}_{\max}}{\hat{N}_{\min}+1} \frac{32L^4}{C^2\gamma^2}}\right)} \geq \sqrt{\pi \log(\sqrt{2})} > \sqrt{\pi}/2$. Thus, taking the p-norm of both sides and rounding up the leading constant shows that there exists a universal constant A_1 such that

$$\|J\|_p \;\leq\; \mathrm{A}_1 \cdot C \frac{\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}} \left\| \sqrt{\log\left(\sqrt{2 + \frac{\hat{N}_{\max}}{\hat{N}_{\min}+1} \frac{32L^4}{C^2\gamma^2}} \right)} \right\|_p. \tag{EC.C.4}$$

We next bound the p-norm on the right. Invoke Lemma B.4 Part iv) with $Y = \sqrt{2 + \frac{\hat{N}_{\max}}{\hat{N}_{\min}+1} \frac{32L^4}{C^2\gamma^2}} \geq \sqrt{2} \geq 1$. Notice $\sqrt{2} \cdot \sqrt{\mathbb{E}[Y]} \geq 1$, which implies

$$\max(1, \sqrt{\mathbb{E}[Y]}/2) \;\leq\; 1 + \sqrt{\mathbb{E}[Y]}/2 \;\leq\; (\sqrt{2}+1)\sqrt{\mathbb{E}[Y]}/2 \leq \sqrt{\mathbb{E}[Y]}.$$

Hence, the norm on the right-hand side of Eq. (EC.C.4) is at most

$$5^{1/p} \left(\frac{p}{2e}\right)^{1/2} \sqrt{\mathbb{E}\left[\sqrt{2 + \frac{\hat{N}_{\max}}{\hat{N}_{\min}+1} \frac{32L^4}{C^2\gamma^2}} \right]}$$

$$\leq\ 5^{1/p}\left(\frac{p}{2e}\right)^{1/2}\sqrt[4]{2+\mathbb{E}\left[\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}\right]\cdot\frac{32L^4}{C^2\gamma^2}}, \qquad \text{(Jensen's Inequality)}$$

$$\leq\ 5^{1/p}\left(\frac{p}{2e}\right)^{1/2}\sqrt[4]{2+\frac{32L^4}{C^2\gamma^2}\cdot 6\left(\frac{12}{e}\right)^2\frac{\lambda_{\max}}{\lambda_{\min}}\log^2 K}, \qquad \text{(Lemma B.5 Part vi))}$$

We next use the assumptions on the parameters to rewrite this bound more simply. By the assumption that $\frac{4L^2}{C\gamma}\geq 1$, we have $\frac{32L^4}{C^2\gamma^2}\geq 2$. Moreover, since $K\geq 2$, $(\frac{12}{e}\log K)^2\geq 1$. Hence, the term under the square root is at most $\frac{64L^4}{C^2\gamma^2}\cdot 6\left(\frac{12}{e}\right)^2\frac{\lambda_{\max}}{\lambda_{\min}}\log^2 K$.

Substituting and simplifying thus shows there exists a universal constant A_2 such that

$$\left\|\sqrt{\log\left(\sqrt{2+\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}\frac{32L^4}{C^2\gamma^2}}\right)}\right\|_p\ \leq\ \mathrm{A}_2\cdot 5^{1/p}\sqrt{p}\cdot\frac{L}{\sqrt{C\gamma}}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/4}\sqrt{\log K}$$

Hence, substituting above into Eq. (EC.C.4) shows there exists a universal constant A_3 such that

$$\|J\|_p\ \leq\ \mathrm{A}_3\cdot L\sqrt{\frac{C}{\gamma}}\cdot\frac{\lambda_{\max}}{\lambda_{\min}}^{5/4}\cdot 5^{\frac{1}{p}}\sqrt{p}\cdot\sqrt{K\log K}.$$

Finally, applying Theorem 4.1 yields

$$\sup_{\alpha\geq 0}\left|\frac{1}{K}\sum_{k=1}^{K}\left(Z_k(\alpha,\boldsymbol{p}_0)-\mathbb{E}[Z_k(\alpha,\boldsymbol{p}_0)]\right)\right|\ \leq\ \mathrm{A}_3\cdot\left(\frac{25}{\delta}\right)^{1/p}p\cdot L\sqrt{\frac{C}{\gamma}}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4}\frac{\sqrt{\log K}}{\sqrt{K}}.$$

This expression is minimized to first order by taking $p=2\log(1/\delta)\geq 1$ and observing $\left(\frac{25}{\delta}\right)^{\frac{1}{2\log(1/\delta)}}$ is at most a constant for $0<\delta<\frac{1}{2}$. Substituting and simplifying proves the first result.

The proof of the second result is very similar, applying Theorem 4.1 to the process $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0):\alpha\geq 0\right\}$. The only key difference is the envelope of this process is now $\frac{C}{N\lambda_{\text{avg}}}\|\hat{\boldsymbol{N}}\|_2\leq\frac{C\sqrt{K}}{N\lambda_{\text{avg}}}\hat{N}_{\max}$ (cf. Lemma 4.2). Thus, following the same steps that lead to Eq. (EC.C.4) but with this envelope shows that J for this process satisfies

$$\begin{aligned}\|J\|_p\ &\leq\ \mathrm{A}_4\cdot\frac{C\sqrt{K}}{N\lambda_{\text{avg}}}\left\|\hat{N}_{\max}\cdot\sqrt{\log\left(\sqrt{2+\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}\frac{32L^4}{C^2\gamma^2}}\right)}\right\|_p\\ &\leq\ \mathrm{A}_4\cdot\frac{C\sqrt{K}}{N\lambda_{\text{avg}}}\left\|\hat{N}_{\max}\right\|_{2p}\cdot\left\|\sqrt{\log\left(\sqrt{2+\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}\frac{32L^4}{C^2\gamma^2}}\right)}\right\|_{2p},\end{aligned}$$

for some constant A_4, where the second inequality follows from Hölder's Inequality (cf. Lemma B.2).

Following an argument entirely analogous to the one that followed Eq. (EC.C.4) but with p replaced by $2p$ shows

$$\left\|\sqrt{\log\left(\sqrt{2+\frac{\hat{N}_{\max}}{\hat{N}_{\min}+1}\frac{32L^4}{C^2\gamma^2}}\right)}\right\|_{2p}\ \leq\ \mathrm{A}_5\cdot 5^{\frac{1}{2p}}\cdot\sqrt{p}\cdot\frac{L}{\sqrt{C\gamma}}\cdot\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{1/4}\cdot\sqrt{\log K}$$

We bound $\|\hat{N}_{\max}\|_{2p}$ using Lemma B.5 Part v).

Then combining these bounds proves

$$\|J\|_p \;\le\; \mathrm{A}_6 \cdot L\sqrt{\frac{C}{\gamma}} \cdot \frac{\lambda_{\max}}{\lambda_{\min}}^{5/4} 6^{1/p} p^{3/2} \cdot \sqrt{K}\log^{3/2}(K).$$

Applying Theorem 4.1, substituting $p = 2\log(1/\delta) > 1$ and simplifying yields the result. □

C.1.3. Proof of Theorem 4.2 We now can prove our main result:

Proof of Theorem 4.2. Combining Lemmas C.3 and 4.1 shows if $\frac{4L^2}{C\gamma} \ge 1$, then there exists a universal constant A such that

$$\mathsf{SubOpt}_{\boldsymbol{p}_0,K}(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0}) \;\le\; \mathrm{A}\cdot L\sqrt{\frac{C}{\gamma}} \cdot \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \cdot \frac{\log^2(1/\delta)\cdot\log^{3/2}(K)}{\sqrt{K}}.$$

If $\frac{4L^2}{C\gamma} < 1$, we can always increase L until $\frac{4L^2}{C\gamma} = 1$ as the larger L remains a valid Lipschitz constant. Increasing the leading constant in this case proves the theorem. □

C.2. Deferred Proofs from Section 4.3: Shrunken-SAA with Data-Driven Anchors for Strongly-Convex Problems

Our strategy to proving Theorems 4.3 and 4.4 is similar to proving to Theorem 4.2 except that our process is now indexed by both $\alpha \ge 0$ and $\boldsymbol{q} \in \mathcal{P}$.

C.2.1. Maximal deviation bounds. Our first step is to use Lemma C.1, part i) to reduce bounding the maximal deviations of $\overline{Z}_K(\cdot,\cdot)$, $\overline{Z}^{\mathsf{LOO}}_K(\cdot,\cdot)$ to bounding the maximal deviations of $\overline{Z}_K(\cdot,\boldsymbol{q})$, $\overline{Z}^{\mathsf{LOO}}_K(\cdot,\boldsymbol{q})$ for a finite number of fixed anchors $\boldsymbol{q} \in \mathcal{P}$.

Lemma C.4 **(Reduction to Maximal Deviations with Fixed Anchor)**. *Under the assumptions of Theorem 4.3, if $\{\boldsymbol{q}^1,\ldots,\boldsymbol{q}^M\}$ is an ϵ_0-covering of $\mathcal{P}$ with respect to ℓ_1, then*

$$\sup_{\alpha\ge 0,\boldsymbol{q}\in\mathrm{Im}(h)} \left|\overline{Z}(\alpha,\boldsymbol{q}) - \mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q})]\right| \;\le\; \frac{2L^2}{\gamma}\epsilon_0 + \max_{j=1,\ldots,M}\sup_{\alpha\ge 0}\left|\overline{Z}(\alpha,\boldsymbol{q}^j) - \mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q}^j)]\right|, \quad \text{(EC.C.5)}$$

$$\begin{aligned}\sup_{\alpha\ge 0,\boldsymbol{q}\in\mathrm{Im}(h)} \left|\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) - \mathbb{E}[\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q})]\right| \;&\le\; \frac{2L^2}{\gamma}\frac{\hat{N}_{avg}}{N\lambda_{\mathrm{avg}}}\epsilon_0 \\ &\quad + \max_{j=1,\ldots,M}\sup_{\alpha\ge 0}\left|\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j) - \mathbb{E}[\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)]\right|.\end{aligned} \quad \text{(EC.C.6)}$$

Proof. Consider the first inequality. Fix some $\boldsymbol{q} \in \mathcal{P}$, and suppose $\boldsymbol{q}^j$ is the closest member of the covering. Then,

$$\begin{aligned}\left|Z_k(\alpha,\boldsymbol{q}) - Z_k(\alpha,\boldsymbol{q}^j)\right| \;&\le\; \frac{\lambda_k}{\lambda_{\mathrm{avg}}}\left|\boldsymbol{p}_k{}^\top\left(\boldsymbol{c}_k(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k)) - \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha,\boldsymbol{q}^j,\hat{\boldsymbol{m}}_k))\right)\right| \\ &\le\; L\cdot\frac{\lambda_k}{\lambda_{\mathrm{avg}}}\left\|\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k) - \boldsymbol{x}_k(\alpha,\boldsymbol{q}^j,\hat{\boldsymbol{m}}_k)\right\|_2 && \text{(Lipschitz Continuity)} \\ &\le\; \frac{L^2}{\gamma}\left\|\boldsymbol{q}-\boldsymbol{q}^j\right\|_1\frac{\lambda_k}{\lambda_{\mathrm{avg}}} && \text{(Lemma C.1, part i))} \\ &\le\; \frac{L^2}{\gamma}\epsilon_0\frac{\lambda_k}{\lambda_{\mathrm{avg}}}\end{aligned}$$

Averaging over k shows $\left|\overline{Z}(\alpha,\boldsymbol{q})-\overline{Z}(\alpha,\boldsymbol{q}^j)\right| \leq \frac{L^2}{\gamma}\epsilon_0$. By Jensen's inequality, this bound also implies that $\left|\mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q})]-\mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q}^j)]\right| \leq \mathbb{E}\left[\left|\overline{Z}(\alpha,\boldsymbol{q})-\overline{Z}(\alpha,\boldsymbol{q}^j)\right|\right] \leq \frac{L^2}{\gamma}\epsilon_0$. Hence, by the triangle inequality,

$$\begin{aligned}\left|\overline{Z}(\alpha,\boldsymbol{q})-\mathbb{E}\left[\overline{Z}(\alpha,\boldsymbol{q})\right]\right| &\leq \left|\overline{Z}(\alpha,\boldsymbol{q})-\overline{Z}(\alpha,\boldsymbol{q}^j)\right|+\left|\mathbb{E}\left[\overline{Z}(\alpha,\boldsymbol{q})-\overline{Z}(\alpha,\boldsymbol{q}^j)\right]\right|+\left|\overline{Z}(\alpha,\boldsymbol{q}^j)-\mathbb{E}\left[\overline{Z}(\alpha,\boldsymbol{q}^j)\right]\right|.\\ &\leq \frac{2L^2}{\gamma}\epsilon_0+\left|\overline{Z}(\alpha,\boldsymbol{q}^j)-\mathbb{E}\left[\overline{Z}(\alpha,\boldsymbol{q}^j)\right]\right|.\end{aligned}$$

Substituting yields the first inequality in the result.

We next prove the second inequality. Fix some $\boldsymbol{q}\in\mathcal{P}$, and suppose $\boldsymbol{q}^j$ is the closest member of the covering. Then,

$$\begin{aligned}&\left|\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q})-\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)\right|\\ &\leq \frac{1}{KN\lambda_{\text{avg}}}\sum_{k=1}^{K}\sum_{i=1}^{d}\hat{m}_{ki}\left|c_{ki}(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i))-c_{ki}(\boldsymbol{x}_k(\alpha,\boldsymbol{q}^j,\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i))\right|\\ &\leq \frac{L}{KN\lambda_{\text{avg}}}\sum_{k=1}^{K}\sum_{i=1}^{d}\hat{m}_{ki}\left\|\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i)-\boldsymbol{x}_k(\alpha,\boldsymbol{q}^j,\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i)\right\|_2 && \text{(Lipschitz Continuity)}\\ &\leq \frac{L^2}{N\lambda_{\text{avg}}\gamma}\left\|\boldsymbol{q}-\boldsymbol{q}^j\right\|_1\frac{1}{K}\sum_{k=1}^{K}\hat{N}_k && \text{(Lemma C.1, part i))}\\ &\leq \frac{L^2}{\gamma}\frac{\hat{N}_{\text{avg}}}{N\lambda_{\text{avg}}}\epsilon_0\end{aligned}$$

By Jensen's inequality, this further implies that $\left|\mathbb{E}[\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q})]-\mathbb{E}[\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)]\right| \leq \mathbb{E}\left[\left|\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q})-\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)\right|\right] \leq \frac{L^2}{\gamma}\frac{\hat{N}_{\text{avg}}}{N\lambda_{\text{avg}}}\epsilon_0$. Using the triangle inequality as before and applying the two bounds above yields our second inequality in the result. □

We next use the above lemmas to bound the maximal deviations of interest via Theorem 4.1:

LEMMA C.5 **(Bounding Maximal Deviations General Anchors)**. *Under the assumptions of Theorem 4.3, there exists a universal constant* A *such that for any* $0<\delta<\frac{1}{2}$*, the following two statements each hold (separately) with probability at least* $1-\delta$*:*

$$\begin{aligned}\sup_{\alpha\geq 0,\ \boldsymbol{q}\in\mathcal{P}}\left|\frac{1}{K}\sum_{k=1}^{K}Z_k(\alpha,\boldsymbol{q})-\mathbb{E}[Z_k(\alpha,\boldsymbol{q})]\right| &\leq \mathrm{A}\cdot\max\left(C,\ \frac{L^2}{\gamma}+L\sqrt{\frac{C}{\gamma}}\right)\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4}\frac{d_0\log^{3/2}(K)\log(1/\delta)}{\sqrt{K}},\\ \sup_{\alpha\geq 0,\ \boldsymbol{q}\in\mathcal{P}}\left|\frac{1}{K}\sum_{k=1}^{K}Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{q})-\mathbb{E}[Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{q})]\right| &\leq \mathrm{A}\cdot\max\left(C,\ \frac{L^2}{\gamma}\ +\ L\sqrt{\frac{C}{\gamma}}\right)\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4}\frac{d_0^2\log^{7/2}(K)\log^2(1/\delta)}{\sqrt{K}}.\end{aligned}$$

Proof. First consider the case $\frac{4L^2}{C\gamma}\geq 1$. Fix some $0<\epsilon_0<\frac{1}{2}$ and consider a minimal ϵ_0-covering of $\mathcal{P}$ with respect to ℓ_1. Denote its size by M. Necessarily, $M\leq D_1(\epsilon_0,\mathcal{P})$ (cf. Pollard 1990, pg. 10). Apply Lemma C.4 with this covering, and then apply the first part of Lemma C.3 with

$\delta \leftarrow \delta/M$ to bound the remaining suprema. This shows that there exists a constant A_1 such that with probability at least $1-\delta$,

$$\sup_{\alpha\geq 0,\ \boldsymbol{q}\in\mathcal{P}} \left| \frac{1}{K}\sum_{k=1}^{K} Z_k(\alpha,\boldsymbol{q}) - \mathbb{E}[Z_k(\alpha,\boldsymbol{q})] \right| \ \leq\ \mathrm{A}_1 \cdot \frac{L^2}{\gamma}\epsilon_0 \ +\ \mathrm{A}_1 L\sqrt{\frac{C}{\gamma}} \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \frac{\log^{1/2}(K)}{\sqrt{K}} \cdot \log\left(\frac{D_1(\epsilon_0,\mathcal{P})}{\delta}\right).$$

Directly optimizing the choice of ϵ_0 appears difficult. We instead take the (suboptimal) choice $\epsilon_0 = \frac{1}{2\sqrt{K}}$ and note $\epsilon_0 < \frac{1}{2}$ since $K \geq 2$. Furthermore, by assumptions on the parameters, $d_0 \geq 1$, $2\log K \geq 1$ and $2\log(1/\delta) \geq 1$. Hence,

$$\begin{aligned}
\log(D_1(\epsilon_0,\mathcal{P})/\delta) \ &\leq\ \log(1/\delta) + d_0\log(1/\epsilon_0) \\
&=\ \log(1/\delta) + d_0\log 2 + \frac{d_0}{2}\log K \\
&\leq\ 2d_0\log K\log(1/\delta) + 2d_0\log K\log(1/\delta) + d_0\log K\log(1/\delta) \\
&=\ 5d_0\log K\log(1/\delta).
\end{aligned}$$

Substituting above shows there exists a constant A_2 such that

$$\begin{aligned}
\sup_{\alpha\geq 0,\ \boldsymbol{q}\in\mathcal{P}} \left| \frac{1}{K}\sum_{k=1}^{K} Z_k(\alpha,\boldsymbol{q}) - \mathbb{E}[Z_k(\alpha,\boldsymbol{q})] \right| \ &\leq\ \mathrm{A}_2 \cdot \frac{L^2}{\gamma\sqrt{K}} \ +\ \mathrm{A}_2 L\sqrt{\frac{C}{\gamma}} \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \frac{d_0\log^{3/2}(K)\log(1/\delta)}{\sqrt{K}}, \\
&\leq\ \mathrm{A}_3 \cdot \left(\frac{L^2}{\gamma} \ +\ L\sqrt{\frac{C}{\gamma}}\right) \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \frac{d_0\log^{3/2}(K)\log(1/\delta)}{\sqrt{K}},
\end{aligned}$$

by collecting constants.

In the case $\frac{4L^2}{C\gamma} < 1$, we can always increase L until $\frac{4L^2}{C\gamma} = 1$ as the larger L remains a valid Lipschitz constant. Substituting this increased L yields the leading term $3C/4$ and proves the first inequality.

The proof of the second inequality is very similar. Assume $\frac{4L^2}{C\gamma} \geq 1$. Again, applying Lemma C.4 over an ϵ_0-covering and using Lemma C.3 with $\delta \leftarrow \frac{\delta}{2M}$ to bound the remaining suprema shows that with probability at least $1-\delta/2$,

$$\begin{aligned}
&\sup_{\alpha\geq 0,\ \boldsymbol{q}\in\mathcal{P}} \left| \frac{1}{K}\sum_{k=1}^{K} Z_k(\alpha,\boldsymbol{q}) - \mathbb{E}[Z_k(\alpha,\boldsymbol{q})] \right| \\
&\quad\leq\ \mathrm{A}_4 \cdot \frac{L^2}{\gamma}\frac{\hat{N}_{\mathrm{avg}}}{N\lambda_{\mathrm{avg}}}\epsilon_0 \ +\ \mathrm{A}_4 L\sqrt{\frac{C}{\gamma}} \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \frac{\log^{3/2}(K)}{\sqrt{K}} \cdot \log^2\left(\frac{2D_1(\epsilon_0,\mathcal{P})}{\delta}\right).
\end{aligned}$$

Take the (suboptimal) choice $\epsilon_0 = \frac{1}{2\sqrt{K}}$. The same simplifications from above show that

$$\log(2D_1(\epsilon_0,\mathcal{P})/\delta) \ \leq\ \log 2 + 5d_0\log K\log(1/\delta) \ \leq\ 7d_0\log K\log(1/\delta),$$

whereby with probability at least $1-\delta/2$,

$$\sup_{\alpha\geq 0,\ \boldsymbol{q}\in\mathcal{P}}\left|\frac{1}{K}\sum_{k=1}^{K}Z_k(\alpha,\boldsymbol{q})-\mathbb{E}[Z_k(\alpha,\boldsymbol{q})]\right| \leq \mathrm{A}_5\cdot\frac{L^2}{\gamma\sqrt{K}}\frac{\hat{N}_{\mathrm{avg}}}{N\lambda_{\mathrm{avg}}} + \mathrm{A}_5 L\sqrt{\frac{C}{\gamma}}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4}\frac{d_0^2\log^{7/2}(K)\log^2(1/\delta)}{\sqrt{K}}.$$

It remains to bound the fraction $\frac{\hat{N}_{\mathrm{avg}}}{N\lambda_{\mathrm{avg}}}=\frac{K\hat{N}_{\mathrm{avg}}}{KN\lambda_{\mathrm{avg}}}$. Notice $K\hat{N}_{\mathrm{avg}}\sim\mathrm{Poisson}(KN\lambda_{\mathrm{avg}})$. From Lemma B.5 Part i) applied to $K\hat{N}_{\mathrm{avg}}$ and Markov's inequality, we have that with probability at least $1-\delta/2$, $\frac{\hat{N}_{\mathrm{avg}}}{N\lambda_{\mathrm{avg}}}\leq\log(4/\delta)$.

Substitute this bound above, apply the union bound and collect constants to show that with probability at least $1-\delta$

$$\sup_{\alpha\geq 0,\ \boldsymbol{q}\in\mathcal{P}}\left|\frac{1}{K}\sum_{k=1}^{K}Z_k(\alpha,\boldsymbol{q})-\mathbb{E}[Z_k(\alpha,\boldsymbol{q})]\right| \leq \mathrm{A}_6\cdot\left(\frac{L^2}{\gamma}+L\sqrt{\frac{C}{\gamma}}\right)\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4}\frac{d_0^2\log^{7/2}(K)\log^2(1/\delta)}{\sqrt{K}}.$$

In the case $\frac{4L^2}{C\gamma}<1$, we can again increase L until $\frac{4L^2}{C\gamma}=1$ since the larger L is still a valid Lipschitz constant. Substituting this increased L yields the leading term $3C/4$ and proves the second claim. □

C.2.2. Proofs of Theorems 4.3 and 4.4. We can now prove the main results of the section via our previously outlined strategy.

Proof of Theorems 4.3 and 4.4. The proofs of both theorems are identical. For both theorems, by Lemma 4.1, the quantity to be bounded is bounded by the sum of the same two maximal deviations. These are in turn bounded by Lemma C.5. Instantiating each bound for $\delta\leftarrow\delta/2$, adding the right hand sides and applying the union bound yields a bound on the sub-optimality. Collecting dominant terms yields the result. □

C.3. Proof of Theorem 4.5: Shrunken-SAA with Fixed Anchors for Discrete Problems

We first use Corollary 4.1 proven in Section 4.4 to prove the following bounds on the maximal deviations of interest via Theorem 4.1.

Lemma C.6 **(Bounding Maximal Deviations for Discrete Problems)**. *Under the assumptions of Theorem 4.5, there exists a constant* A *such that for any* $0<\delta<1/2$*, the following two statements hold (separately) each with probability at least* $1-\delta$*:*

$$\sup_{\alpha\geq 0}\left|\frac{1}{K}\sum_{k=1}^{K}Z_k(\alpha,\boldsymbol{p}_0)-\mathbb{E}[Z_k(\alpha,\boldsymbol{p}_0)]\right| \leq \mathrm{A}\cdot C\frac{\lambda_{\max}}{\lambda_{\min}}\cdot\sqrt{\log\left(\sum_{k=1}^{K}|\mathcal{X}_k|\right)}\cdot\frac{\sqrt{\log\left(\frac{1}{\delta}\right)}}{\sqrt{K}},$$

$$\sup_{\alpha\geq 0}\left|\frac{1}{K}\sum_{k=1}^{K}Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)-\mathbb{E}[Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)]\right| \leq \mathrm{A}\cdot C\frac{\lambda_{\max}}{\lambda_{\min}}\cdot\sqrt{\log\left(N_{\max}\sum_{k=1}^{K}|\mathcal{X}_k|\right)}\cdot\frac{\log^{3/2}(K)\cdot\log^{3/2}(1/\delta)}{\sqrt{K}}.$$

Proof. Consider the first inequality. We first bound the variable J in Eq. (4.3) corresponding to the process $\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}$ with the envelope given by Lemma 4.2. By Corollary 4.1,

$$J \;\leq\; 9C \cdot \frac{\lambda_{\max}}{\lambda_{\min}} \cdot \sqrt{K} \sqrt{\log\left(2\sum_{k=1}^{K} |\mathcal{X}_k|\right)},$$

where we have upper bounded $\|\boldsymbol{\lambda}\|_2 \leq \lambda_{\max}\sqrt{K}$. From Theorem 4.1, there exists a constant A_1 such that with probability at least $1-\delta$,

$$\sup_{\alpha \geq 0} \left| \frac{1}{K} \sum_{k=1}^{K} Z_k(\alpha, \boldsymbol{p}_0) - \mathbb{E}[Z_k(\alpha, \boldsymbol{p}_0)] \right| \leq \mathrm{A}_1 \cdot \left(\frac{5}{\delta}\right)^{1/p} p^{1/2} \cdot C \frac{\lambda_{\max}}{\lambda_{\min}} \cdot \sqrt{\frac{\log\left(2\sum_{k=1}^{K} |\mathcal{X}_k|\right)}{K}}.$$

Let $p = 2\log(1/\delta) > 1$, and collect constants to complete the proof.

The proof of the second inequality is similar but uses different envelopes (cf. Lemma 4.2) and the larger packing numbers of Corollary 4.1. Specifically, we note that $\min(d, \hat{N}_k) \leq \hat{N}_{\max}$ and $\|\hat{\boldsymbol{N}}\|_2 \leq \hat{N}_{\max}\sqrt{K}$, and bound J as

$$J \;\leq\; 9\frac{C\sqrt{K}}{N\lambda_{\text{avg}}} \hat{N}_{\max} \sqrt{\log\left(1 + 2\hat{N}_{\max}\sum_{k=1}^{K} |\mathcal{X}_k|\right)}.$$

Recall $N_{\max} \equiv N\lambda_{\max} \geq N\lambda_{\min} \geq 1$. Thus, we can upper bound the logarithm as

$$\begin{aligned}
\log\left(1 + 2\hat{N}_{\max}\sum_{k=1}^{K} |\mathcal{X}_k|\right) \;&\leq\; \log\left(6N_{\max}\sum_{k=1}^{K} |\mathcal{X}_k| + 2\hat{N}_{\max}\sum_{k=1}^{K} |\mathcal{X}_k|\right) \\
&=\; \underbrace{\log\left(2N_{\max}\sum_{k=1}^{K} |\mathcal{X}_k|\right)}_{\geq \log 4} + \underbrace{\log\left(3 + \frac{\hat{N}_{\max}}{N_{\max}}\right)}_{\geq \log 3} \\
&\leq\; 2\log\left(2N_{\max}\sum_{k=1}^{K} |\mathcal{X}_k|\right) \cdot \log\left(3 + \frac{\hat{N}_{\max}}{N_{\max}}\right),
\end{aligned}$$

where the last inequality follows because $a + b \leq 2ab$ when $a, b \geq 1$

Substituting above and taking the p-norm shows there exists a constant A_2 such that

$$\begin{aligned}
\|J\|_p \;&\leq\; \mathrm{A}_2 \cdot \frac{C\sqrt{K}}{N\lambda_{\text{avg}}} \sqrt{\log\left(2N_{\max}\sum_{k=1}^{K} |\mathcal{X}_k|\right)} \cdot \left\| \hat{N}_{\max} \sqrt{\log\left(3 + \frac{\hat{N}_{\max}}{N_{\max}}\right)} \right\|_p \\
&\leq\; \mathrm{A}_2 \cdot \frac{C\sqrt{K}}{N\lambda_{\text{avg}}} \sqrt{\log\left(2N_{\max}\sum_{k=1}^{K} |\mathcal{X}_k|\right)} \cdot \left\| \hat{N}_{\max} \right\|_{2p} \cdot \left\| \sqrt{\log\left(3 + \frac{\hat{N}_{\max}}{N_{\max}}\right)} \right\|_{2p},
\end{aligned}$$

where the second inequality follows from Hölder's Inequality (cf. Lemma B.2) We next bound these two $2p$-norms.

We bound the second $2p$-norm using Lemma B.4 Part iv) with $Y = 3 + \frac{\hat{N}_{\max}}{N_{\max}} > 3$, yielding

$$\begin{aligned}\left\|\sqrt{\log\left(3+\frac{\hat{N}_{\max}}{N_{\max}}\right)}\right\|_{2p} &\le 5^{\frac{1}{2p}}\sqrt{\frac{p}{e}}\max\left(1,\frac{1}{2}\sqrt{3+\mathbb{E}\left[\frac{\hat{N}_{\max}}{N_{\max}}\right]}\right)\\ &\le 5^{\frac{1}{2p}}\sqrt{\frac{p}{e}}\max\left(1,\frac{1}{2}\sqrt{3+\frac{36}{e}\log K}\right) \qquad \text{(Lemma B.5 Part v))}\\ &\le 2\cdot 5^{\frac{1}{2p}}\sqrt{p}\sqrt{\log K},\end{aligned}$$

since $K \ge 2$.

Similarly, bound $\left\|\hat{N}_{\max}\right\|_{2p}$ using Lemma B.5 Part v).

Combining shows

$$\|J\|_p \;\le\; \mathrm{A}_3\cdot\frac{CN_{\max}\sqrt{K}}{N\lambda_{\text{avg}}}\sqrt{\log\left(2N_{\max}\sum_{k=1}^{K}|\mathcal{X}_k|\right)}\cdot 6^{\frac{1}{p}}p^{3/2}\cdot\log^{3/2}(K).$$

Applying Theorem 4.1 and substituting $p = 2\log(1/\delta) > 1$ proves the second inequality. □

We can now prove the main result of the section.

Proof of Theorem 4.5. Lemma C.6 bound the maximal deviations in Lemma 4.1. Instantiating them for $\delta \leftarrow \delta/2$, adding their righthand sides and applying the union bound bounds the sub-optimality. Collecting dominant terms proves the result. □

C.4. Deferred Proofs from Section 4.5: Shrunken-SAA with Data-Driven Anchors for Discrete Problems.

As a first step towards our proof, we prove Lemma 4.4. Recall the $m \equiv \sum_{k=1}^{K}\binom{|\mathcal{X}_k|}{2}$ hyperplanes defined in Section 4.5:

$$H_{kij} = \left\{\boldsymbol{\theta}\in\mathbb{R}^{d_0} : (V\boldsymbol{\theta}+\hat{\boldsymbol{m}}_k)^\top(\boldsymbol{c}_k(\boldsymbol{x}_{ki})-\boldsymbol{c}_k(\boldsymbol{x}_{kj})) = \mathbf{0}\right\}, \quad \forall\, k=1,\ldots,K,\, i\neq j = 1,\ldots,|\mathcal{X}_k|\,.$$

In words, for $\boldsymbol{\theta}$ on H_{kij} we are indifferent between $\boldsymbol{x}_{ki}$ and $\boldsymbol{x}_{kj}$ when using $\boldsymbol{\theta}$ in Eq. (4.5). On either side, we strictly prefer one solution.

For any fixed $\boldsymbol{\theta}\in\mathbb{R}^{d_0}$, we considered the polyhedron induced by the equality constraints of those hyperplanes containing $\boldsymbol{\theta}$, and the inequality constraints defined by the side on which $\boldsymbol{\theta}$ lies for the remaining hyperplanes. We call such polyhedra *fully-specified* because they are defined by their relationship to *all* m hyperplanes in the arrangement. Because this polyhedron lives in $\mathbb{R}^{d_0}$, it necessarily has dimension $j \le d_0$. For example the shaded region in Fig. 4 is a fully-specified polyhedron with $j=2$, the bold line segment has $j=1$ and the bold point has $j=0$. As argued in the main text, to bound $|\{\mathbf{Z}(\alpha,\boldsymbol{q}) : \alpha\ge 0, \boldsymbol{q}\in\mathcal{P}\}|$ it suffices to count the number of j-dimensional fully-specified polyhedron in the arrangement of the above m hyperplanes for all $0\le j\le d_0$.

Counting the polyhedra induced by hyperplane arrangements is a classical problem in geometry. For example, it is well-known that the number of d_0-dimensional, fully-specified polyhedra in a hyperplane arrangement with m hyperplanes in $\mathbb{R}^{d_0}$ is at most $\sum_{i=0}^{d_0}\binom{m}{i}$ (Stanley 2004, Prop. 2.4). We first use this result to bound the total number of polyhedra in an arbitrary arrangement with m hyperplanes in $\mathbb{R}^{d_0}$.

LEMMA C.7 **(Number of Fully-Specified Polyhedra)**. *In a hyperplane arrangement with m hyperplanes in $\mathbb{R}^{d_0}$, the number of fully-specified polyhedra is at most*

$$\sum_{j=0}^{d_0}\binom{m}{d_0-j}\sum_{i=0}^{j}\binom{m-d_0+j}{i} \;\le\; (1+2m)^{d_0}.$$

Proof of Lemma C.7 Each fully-specified polyhedron has some dimension, $0\le j\le d_0$. We will count the number of such fully-specified polyhedra by counting for each dimension j.

Fix some $0\le j\le d_0$. Notice that each j-dimensional polyhedron lives in a j-dimensional subspace defined by d_0-j linearly independent hyperplanes from the arrangement. There are at most $\binom{m}{d_0-j}$ ways to choose these linearly independent d_0-j hyperplanes. Next project the remaining hyperplanes onto this subspace which yields at most $m-d_0+j$ non-trivial hyperplanes in the subspace, i.e., hyperplanes that are neither the whole subspace nor the empty set. These non-trivial hyperplanes "cut up" the subspace into various polyhedra, including j-dimensional, fully-specified polyhedra. By (Stanley 2004, Prop. 2.4), the number of j-dimensional, fully-specified polyhedra in this hyerplane arrangement of at most $m-d_0+j$ hyperplanes in j-dimensional space is at most $\sum_{i=0}^{j}\binom{m-d_0+j}{i}$. In summary, it follows that there are at most $\binom{m}{d_0-j}\sum_{i=0}^{j}\binom{m-d_0+j}{i}$ j-dimensional, fully-specified polyhedra in the arrangement.

Summing over j gives the lefthand side of the bound in the lemma.

For the righthand side, recall that

$$\sum_{i=0}^{j}\binom{m-d_0+j}{i} \;\le\; \sum_{i=0}^{j}(m-d_0+j)^i\cdot 1^{m-d_0+j-i} \;\le\; (1+m-d_0+j)^j \;\le\; (1+m)^j,$$

where the penultimate inequality is the binomial expansion and the last follow because $j\le d_0$. Next,

$$\begin{aligned}\sum_{j=0}^{d_0}\binom{m}{d_0-j}\sum_{i=0}^{j}\binom{m-d_0+j}{i} &\le \sum_{j=0}^{d_0}\binom{m}{d_0-j}(1+m)^j\\ &\le \sum_{j=0}^{d_0} m^{d_0-j}(1+m)^j\\ &= (1+2m)^{d_0},\end{aligned}$$

where the last equality is again the binomial expansion. □

We can now bound the cardinality of the relevant solution sets.

Proof of Lemma 4.4. Recall there are $m = \sum_{k=1}^K \binom{|\mathcal{X}_k|}{2}$ hyperplanes in the arrangement Eq. (4.6) in $\mathbb{R}^{d_0}$, and the number of fully-specified polyhedra in this arrangement upper-bounds $|\{\mathbf{Z}(\alpha,\boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \mathcal{P}\}|$. Noting $1 + 2m = 1 + \sum_{k=1}^K |\mathcal{X}_k| (|\mathcal{X}_k| - 1) \leq \sum_{k=1}^K |\mathcal{X}_k|^2$ yields the first bound.

A similar argument can be used to bound $\left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \mathcal{P}\right\}\right|$. Suppose first $\hat{N}_{\max} = 0$. Then this set has size 1. On the other hand, if $\hat{N}_{\max} > 0$, let $\mathcal{I}_k = \{i = 1, \ldots, d : \hat{m}_{ki} > 0\}$, so that

$$\begin{aligned}\left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \mathcal{P}\right\}\right| &\leq \left|\left\{(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k - \boldsymbol{e}_i))_{k=1,\ldots,K, i\in\mathcal{I}_k} \;:\; \boldsymbol{q} \in \mathcal{P}, \alpha \geq 0\right\}\right| \\ &\leq \left|\left\{(\boldsymbol{x}_k(\|V\boldsymbol{\theta}\|_1, V\boldsymbol{\theta}/\|V\boldsymbol{\theta}\|_1, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i))_{k=1,\ldots,K, i\in\mathcal{I}_k} \;:\; \boldsymbol{\theta} \in \mathbb{R}^{d_0}, V\boldsymbol{\theta} \in \mathbb{R}^d_+\right\}\right|. \qquad \text{(EC.C.7)}\end{aligned}$$

We then consider the arrangement generated by

$$H_{kijl} = \left\{\boldsymbol{\theta} \in \mathbb{R}^{d_0} : (V\boldsymbol{\theta} + \hat{\boldsymbol{m}}_k - \boldsymbol{e}_l)^\top (\boldsymbol{c}_k(\boldsymbol{x}_{ki}) - \boldsymbol{c}_k(\boldsymbol{x}_{kj})) = \boldsymbol{0}\right\},$$

for all $k = 1, \ldots, K$, $i, j = 1, \ldots, |\mathcal{X}_k|$ with $i \neq j$, and $l \in \mathcal{I}_k$. Notice that since $|\mathcal{I}_k| \leq \hat{N}_k$ there are at most $\hat{N}_{\max} \sum_{k=1}^K \binom{|\mathcal{X}_k|}{2}$ such hyperplanes. Moreover, $\left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \mathcal{P}\right\}\right|$ is upper-bounded by the number of fully-specified polyhedra in this arrangement. Note that $1 + 2\hat{N}_{\max} \sum_{k=1}^K \binom{|\mathcal{X}_k|}{2} = 1 + \hat{N}_{\max} \sum_{k=1}^K |\mathcal{X}_k| (|\mathcal{X}_k| - 1) \leq \hat{N}_{\max} \sum_{k=1}^K |\mathcal{X}_k|^2$. Adding 1 covers the case $\hat{N}_{\max} = 0$. Plugging in this value into Lemma C.7 yields the second bound above. □

C.4.1. Maximal Deviation Bounds. We next use Lemma 4.4 to bound the maximal deviations of interest via Theorem 4.1.

LEMMA C.8 **(Bounding Maximal Deviations, Discrete Case, General Anchors)**. *Under the assumptions of Theorem 4.6, there exists a constant* A *such that for any* $0 < \delta < \frac{1}{2}$*, both of the following statements hold (separately) with probability at least* $1 - \delta$*:*

$$\sup_{\alpha\geq 0,\ \boldsymbol{q}\in\mathcal{P}} \left|\frac{1}{K}\sum_{k=1}^K Z_k(\alpha,\boldsymbol{q}) - \mathbb{E}[Z_k(\alpha,\boldsymbol{q})]\right| \;\leq\; \mathrm{A}\cdot C\frac{\lambda_{\max}}{\lambda_{\min}} \cdot \sqrt{d_0 \log\left(\sum_{k=1}^K |\mathcal{X}_k|\right)} \cdot \frac{\sqrt{\log(1/\delta)}}{\sqrt{K}}.$$

$$\sup_{\alpha\geq 0,\ \boldsymbol{q}\in\mathcal{P}} \left|\frac{1}{K}\sum_{k=1}^K Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) - \mathbb{E}[Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{q})]\right| \;\leq\; \mathrm{A}\cdot C\frac{\lambda_{\max}}{\lambda_{\min}} \sqrt{d_0 \log\left(N_{\max}\sum_{k=1}^K |\mathcal{X}_k|\right)} \cdot \frac{\log^{3/2}(K)\log^2(1/\delta)}{\sqrt{K}}.$$

Proof. Using Lemmas 4.2 and 4.4 to bound the variable J in Eq. (4.3) and since $\left(\sum_{k=1}^K |\mathcal{X}_k|^2\right)^{d_0} \leq \left(\sum_{k=1}^K |\mathcal{X}_k|\right)^{2d_0}$, proves

$$\|J\|_p \;\leq\; 9C\frac{\lambda_{\max}}{\lambda_{\min}}\sqrt{K}\sqrt{2d_0 \log\left(\sum_{k=1}^K |\mathcal{X}_k|\right)}.$$

Next apply Theorem 4.1 and let $p = 2\log(1/\delta)$ to prove the first statement.

For the second inequality, we follow a similar strategy with the appropriate envelope and packing number (cf. Lemmas 4.2 and 4.4). In this case,

$$J \;\le\; \frac{9C\hat{N}_{\max}\sqrt{K}}{N\lambda_{\text{avg}}} \cdot \sqrt{\log\left(1+\hat{N}_{\max}^{d_0}\left(\sum_{k=1}^{K}|\mathcal{X}_k|\right)^{2d_0}\right)}.$$

Consider the inner logarithm, and let $\mathcal{X}_{\text{tot}} \equiv \sum_{k=1}^{K}|\mathcal{X}_k| \ge K \ge 2$. Then,

$$\begin{aligned}
\log\left(1+\hat{N}_{\max}^{d_0}\mathcal{X}_{\text{tot}}^{2d_0}\right) \;&\le\; d_0\log\left(1+\hat{N}_{\max}\mathcal{X}_{\text{tot}}^{2}\right) && \text{(since } \hat{N}_{\max}\mathcal{X}_{\text{tot}}^2>0\text{)}\\
&\le d_0\log\left(3N_{\max}\mathcal{X}_{\text{tot}}^{2}+\hat{N}_{\max}\mathcal{X}_{\text{tot}}^{2}\right)\\
&\le\; d_0\left(\underbrace{\log\left(N_{\max}\mathcal{X}_{\text{tot}}^{2}\right)}_{\ge\log(4)}+\underbrace{\log\left(3+\frac{\hat{N}_{\max}}{N_{\max}}\right)}_{\ge\log(3)}\right)\\
&\le\; 2d_0\log\left(N_{\max}\mathcal{X}_{\text{tot}}^{2}\right)\cdot\log\left(3+\frac{\hat{N}_{\max}}{N_{\max}}\right),
\end{aligned}$$

where the last inequality follows because $a+b\le 2ab$ for $a,b\ge 1$.

Substituting above shows

$$\begin{aligned}
\|J\|_p \;&\le\; \frac{9C\sqrt{K}}{N\lambda_{\text{avg}}}\cdot\sqrt{2d_0\log\left(N_{\max}\mathcal{X}_{\text{tot}}^{2}\right)}\left\|\hat{N}_{\max}\cdot\sqrt{\log\left(3+\frac{\hat{N}_{\max}}{N_{\max}}\right)}\right\|_p.\\
&\le\; \frac{9C\sqrt{K}}{N\lambda_{\text{avg}}}\cdot\sqrt{2d_0\log\left(N_{\max}\mathcal{X}_{\text{tot}}^{2}\right)}\cdot\left\|\hat{N}_{\max}\right\|_{2p}\cdot\left\|\sqrt{\log\left(3+\frac{\hat{N}_{\max}}{N_{\max}}\right)}\right\|_{2p},
\end{aligned}$$

We next bound these norms. The first is bounded by Lemma B.5 Part v). The second was bounded in the proof of Lemma C.6 as

$$\left\|\sqrt{\log\left(3+\frac{\hat{N}_{\max}}{N_{\max}}\right)}\right\|_{2p} \le 2\cdot\; 5^{\frac{1}{2p}}\sqrt{p}\sqrt{\log K}.$$

Combining proves

$$\|J\|_p \;\le\; \mathrm{A}_3\cdot C\cdot\frac{\lambda_{\max}}{\lambda_{\min}}\sqrt{d_0\log\left(N_{\max}\mathcal{X}_{\text{tot}}^{2}\right)}\cdot 6^{\frac{1}{p}}p^{3/2}\cdot\log^{3/2}(K)\sqrt{K},$$

for some constant A_3. Now apply Theorem 4.1 and substitute $p=2\log(1/\delta)$ to prove the second inequality. $\square$

C.4.2. Proofs of Theorems 4.6 and 4.7. We can now prove the main results of the section via our usual strategy.

Proof of Theorems 4.6 and 4.7. The proofs of both theorems are identical. For both theorems, by Lemma 4.1, the quantity to be bounded is bounded by the sum of the same two maximal deviations. These are in turn bounded by Lemma C.8. Instantiating each bound for $\delta \leftarrow \delta/2$, adding the right hand sides and applying the union bound yields a bound on the sub-optimality. Collecting dominant terms yields the result. $\square$

Appendix D: Contrasting the Sub-Optimality-Stability Bias-Variance Tradeoffs

We here expand on the discussion from Section 5 comparing the Sub-Optimality-Stability tradeoff to the classic bias-variance tradeoff. As mentioned in Section 5, one important distinction is that the former applies to general optimization problems. In the following we will show that they are different even when we restrict to the case of MSE (cf. Example 2.1).

To be more precise, fix the cost functions $c_k(x,\xi) = (x-\xi)^2$, let μ_k and σ_k^2 denote the mean and variance of $\xi_k \in \mathbb{R}$ and assume $\lambda_k = 1$ for all k for simplicity. There are at least two ways to interpret the classical bias-variance tradeoff in context of Assumption 3.1. First, we can decompose conditionally on $\hat{\boldsymbol{N}}$, yielding

$$\mathbb{E}\left[\overline{Z}_K(\alpha,\boldsymbol{p}_0) \mid \hat{\boldsymbol{N}}\right] = \frac{1}{K}\sum_{k=1}^K \underbrace{\left(\frac{\alpha}{\hat{N}_k+\alpha}\right)^2 (\mu_k-\mu_{k0})^2}_{\text{Conditional Bias Squared}} + \underbrace{\left(\frac{\hat{N}_k}{\hat{N}_k+\alpha}\right)^2 \frac{\sigma_k^2}{\hat{N}_k}}_{\text{Conditional Variance}},$$

where $\mu_{k0} = \boldsymbol{p}_0^\top \boldsymbol{a}_k$. Taking expectations of both sides yields the identity for $\alpha > 0$

$$\mathbb{E}\left[\overline{Z}_K(\alpha,\boldsymbol{p}_0)\right] = \underbrace{\frac{1}{K}\sum_{k=1}^K \mathbb{E}\left[\left(\frac{\alpha}{\hat{N}_k+\alpha}\right)^2\right](\mu_k-\mu_{k0})^2}_{\text{Expected Conditional Bias Squared}} + \underbrace{\mathbb{E}\left[\frac{\hat{N}_k}{(\hat{N}_k+\alpha)^2}\right]\sigma_k^2}_{\text{Expected Conditional Variance}}. \quad \text{(EC.D.1)}$$

This perspective is perhaps most appropriate if view Assumption 3.1 as a smoothing that randomizes over instances.

Alternatively, we can apply the bias-variance decomposition unconditionally, yielding for $\alpha > 0$,

$$\begin{aligned}\mathbb{E}\left[\overline{Z}_K(\alpha,\boldsymbol{p}_0)\right] &= \frac{1}{K}\sum_{k=1}^K \left(\mathbb{E}\left[x_k(\alpha,\boldsymbol{p}_0,\hat{\mu}_k)-\mu_k\right]\right)^2 + \mathrm{Var}(x_k(\alpha,\boldsymbol{p}_0,\hat{\mu}_k)),\\ &= \frac{1}{K}\sum_{k=1}^K \underbrace{\left(\mathbb{E}\left[\frac{\alpha}{\hat{N}_k+\alpha}\right]\right)^2 (\mu_{0k}-\mu_k)^2}_{\text{Bias Squared}} + \underbrace{\mathrm{Var}(x_k(\alpha,\boldsymbol{p}_0,\hat{\mu}_k))}_{\text{Variance}}, \quad \text{(EC.D.2)}\end{aligned}$$

(We can, if desired, evaluate the second term using the law of total variance after conditioning on $\hat{N}_k$, but this expression will not be needed in what follows.) This perspective is perhaps most appropriate if we view the randomization of $\hat{N}_k$ as intrinsic to the data-generating process.

Finally, from Lemma 3.1 and our previous comments, we have that

$$\mathbb{E}\left[\overline{Z}_K(\alpha, \boldsymbol{p}_0)\right] = \frac{1}{N\lambda_{\text{avg}}}\left(\mathbb{E}\left[\text{SAA-SubOptimality}(\alpha)\right] + \mathbb{E}\left[\text{Instability}(\alpha)\right] + \mathbb{E}\left[\text{SAA}(0)\right]\right),$$

where, again, SAA(0) does not depend on α. A straightforward calculation yields,

LEMMA D.1 **(SAA-Sub-Optimality for MSE)**. *For $\alpha > 0$, we have*

$$\text{SAA-SubOpt}(\alpha) = \frac{1}{K}\sum_{k=1}^{K} \hat{N}_k \left(\frac{\alpha}{\hat{N}_k + \alpha}\right)^2 (\hat{\mu}_k - \mu_{k0})^2$$

$$\mathbb{E}\left[\text{SAA-SubOpt}(\alpha)\right] = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\hat{N}_k \left(\frac{\alpha}{\hat{N}_k + \alpha}\right)^2\right] (\mu_k - \mu_{k0})^2 + \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\left[\left(\frac{\alpha}{\hat{N}_k + \alpha}\right)^2\right] \sigma_k^2,$$

where $\hat{\mu}_k$ is the sample mean for the k^{th} subproblem.

Proof of Lemma D.1 By definition, the k^{th} term of SAA-SubOpt(α) is

$$\begin{aligned}\sum_{i=1}^{d} \hat{m}_{ki}\left(c_{ki}(x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)) - c_{ki}(x_k(0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))\right) &= \hat{N}_k \sum_{i=1}^{d} \hat{\boldsymbol{p}}_{ki}\left(c_{ki}(x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)) - c_{ki}(x_k(0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))\right) \\ &= \hat{N}_k\left(\mathbb{E}\left[(\hat{\xi}_k - \hat{\mu}_k(\alpha))^2 \mid \hat{\boldsymbol{m}}_k\right] + \mathbb{E}\left[(\hat{\xi}_k - \hat{\mu}_k)^2 \mid \hat{\boldsymbol{m}}_k\right]\right)\end{aligned}$$

where $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) = \hat{\mu}_k(\alpha) \equiv \frac{\alpha}{\hat{N}_k+\alpha}\mu_{k0} + \frac{\hat{N}_k}{\hat{N}_k+\alpha}\hat{\mu}_k$, and $\hat{\xi}_k \sim \hat{\boldsymbol{p}}_k$.

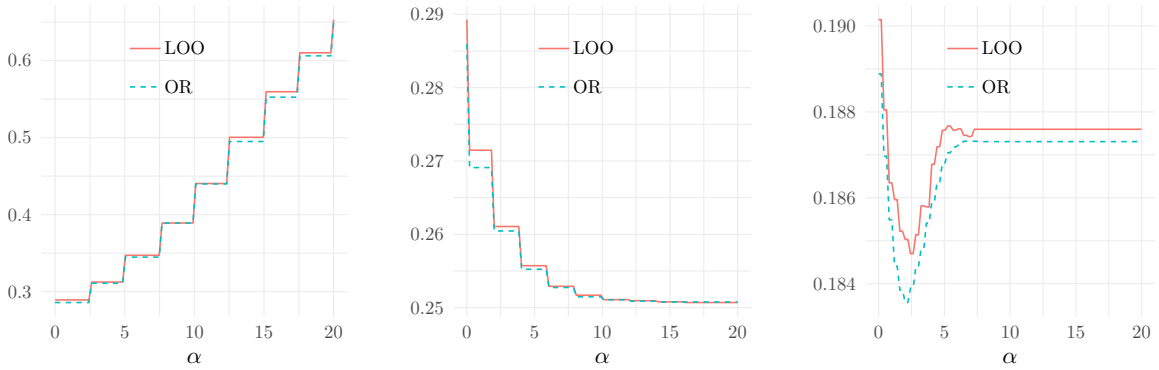
Note $\mathbb{E}\left[(\hat{\xi}_k - \hat{\mu}_k(\alpha))^2 \mid \hat{\boldsymbol{m}}_k\right] = (\hat{\mu}_k - \hat{\mu}_k(\alpha))^2 + \hat{\sigma}_k^2$, where $\hat{\sigma}_k^2$ is the variance of $\hat{\xi}_k \mid \hat{\boldsymbol{m}}_k$. Similarly, $\mathbb{E}\left[(\hat{\xi}_k - \hat{\mu}_k)^2 \mid \hat{\boldsymbol{m}}_k\right] = \hat{\sigma}_k^2$. Hence from above, the k^{th} term of SAA-SubOpt(α) is $\hat{N}_k(\hat{\mu}_k - \hat{\mu}_k(\alpha))^2$. Using the definition of $\hat{\mu}_k(\alpha)$ we have $(\hat{\mu}_k - \hat{\mu}_k(\alpha))^2 = \left(\frac{\alpha}{\hat{N}_k+\alpha}\right)^2 (\mu_0 - \hat{\mu}_k)^2$. Summing across the k terms yields the expression for SAA-SubOpt(α) in the lemma.

Now consider taking the conditional expectation of the k^{th} term of SAA-SubOpt(α) where we condition on $\hat{\boldsymbol{N}}$. From our previous expression, this is simply

$$\begin{aligned}\hat{N}_k \left(\frac{\alpha}{\hat{N}_k + \alpha}\right)^2 \mathbb{E}\left[(\mu_0 - \hat{\mu}_k)^2 \mid \hat{\boldsymbol{N}}\right] &= \hat{N}_k \left(\frac{\alpha}{\hat{N}_k + \alpha}\right)^2 \left((\mu_0 - \mu_k)^2 + \frac{\sigma_k^2}{\hat{N}_k}\right). \\ &= \hat{N}_k \left(\frac{\alpha}{\hat{N}_k + \alpha}\right)^2 (\mu_0 - \mu_k)^2 + \left(\frac{\alpha}{\hat{N}_k + \alpha}\right)^2 \sigma_k^2.\end{aligned}$$

Taking expectations and then averaging over k yields the expression for $\mathbb{E}\left[\text{SAA-SubOpt}(\alpha)\right]$, completing the lemma. □

By inspection, $\frac{1}{N\lambda_{\text{avg}}}\mathbb{E}[\text{SAA-SubOpt}(\alpha)]$ involves a non-zero term that depends on both σ_k^2 and α. Consequently, it must differ from the bias-squared term in Eq. (EC.D.2) and the expected conditional bias-squared term in Eq. (EC.D.1). In particular, since the difference depends on α and SAA(0) does not depend on α, the difference is not solely due to the treatment of this constant. Finally, since each of the identities decomposes the same quantity $\mathbb{E}\left[\overline{Z}_K \alpha, \boldsymbol{p}_0\right]$, it follows that the bias-variance tradeoff and the Sub-Optimality-Instability Tradeoff are fundamentally different for this example.



(a) $\alpha^{\text{S-SAA}}_{\boldsymbol{p}_0} = \alpha^{\text{OR}}_{\boldsymbol{p}_0} = 0$ (b) $\alpha^{\text{S-SAA}}_{\boldsymbol{p}_0} \approx \alpha^{\text{OR}}_{\boldsymbol{p}_0} \approx 16.16$ (c) $\alpha^{\text{S-SAA}}_{\boldsymbol{p}_0} \approx 2.42$, $\alpha^{\text{OR}}_{\boldsymbol{p}_0} \approx 2.22$

Figure EC.1 **LOO and Oracle Curves**. We consider $K = 10,000$ newsvendors where $p_{k1} \sim \text{Uniform}[.6, .9]$, $\hat{N}_k \sim \text{Poisson}(10)$. We consider a single data draw. The values of p_{01} and the critical fractile s are $(p_{01}, s) = (.3, .5)$, $(p_{01}, s) = (.75, .5)$, and $(p_{01}, s) = (.3, .2)$, respectively. In the first panel, instability initially increases, and there is no benefit to pooling. In the second and third, instability is decreasing and there is a benefit to pooling.

Appendix E: Computational Details and Additional Numerical Experiments

E.1. Simulation Set-up for Fig. 1

For $d = 10$, we generate $5,000$ distributions $\boldsymbol{p}_k$ according to a uniform distribution on the simplex and additional $5,000$ distributions $\boldsymbol{p}_k$ according to the Dirichlet distribution with parameter $(3, \ldots, 3)$, for a total of $K = 10,000$ subproblems. We take $\lambda_k = 1$ for all k. Across all runs, these $\boldsymbol{p}_k$ and λ_k are fixed. Then, for each run, for each k, we then generate $\hat{N}_k = 20$ data points independently according to Eq. (2.1). We train each of our policies on these data, and evaluate against the true $\boldsymbol{p}_k$. Results are averaged across $10,000$ runs.

E.2. Additional Figures from Example 5.1.

Figure EC.1 shows the companion figures for Example 5.1 from Section 5.

E.3. Implementation Details for Computational Experiments from Section 6

On average, less than 2.5% of stores are open on weekends, and hence we drop all weekends from our dataset. Similarly, the data exhibits a mild upward linear trend at a rate of 215 units a year (approximately 3.7% increase per year), with a p-value $< .001$. This trend is likely due to inflation and growing GDP over the time frame. We remove this trend using simple ordinary least squares. Finally, many stores engage in promotional activities periodically throughout the month of December leading up to Christmas. These promotions distort sales in the surrounding period. Hence we drop data for the month of December from our dataset.

Throughout, $\alpha^{\text{OR}}_{\boldsymbol{p}_0}, \alpha^{\text{S-SAA}}_{\boldsymbol{p}_0}$ are obtained by exhaustively searching a grid of length 120 points from 0 to 180. The grand-mean and Beta variants are obtained similarly. Notice when $\hat{N}_k = 10$, a value of

$\alpha = 180$ amounts to having 18 times more weight on the anchor point than the data, itself. Unless otherwise specified in an experiment, $d = 20$ and $\hat{N}_k = 10$ (fixed, non-random for all k).

The "KS" policy described in the main-text corresponds to solving a data-driven distributionally robust version of the newsvendor problem, namely,

$$\boldsymbol{x}_k^{\mathsf{KS}}(\rho_k, \mathcal{S}_k) \in \min_x \sup_{\mathbb{P} \in \mathcal{P}^{\mathsf{KS}}(\rho_k, \mathcal{S}_k)} \mathbb{E}_{\xi \sim \mathbb{P}} \left[\max \left\{ \frac{s}{1-s}(\xi - x), (x - \xi) \right\} \right],$$

where the ambiguity set $\mathcal{P}^{\mathsf{KS}}(\rho_k, \mathcal{S}_k)$ is the Kolmogorov-Smirnov ball around the empirical distribution, i.e.,

$$\mathcal{P}^{\mathsf{KS}}(\rho_k, \mathcal{S}_k) \equiv \left\{ \mathbb{P} \ : \ \sup_{t \in \mathbb{R}} \left| \mathbb{P}(\xi \le t) - \frac{1}{\hat{N}_k} \sum_{j=1}^{\hat{N}_k} \mathbb{I}\left[\hat{\xi}_{jk} \le t \right] \right| \le \rho_k \right\}.$$

This ambiguity set enjoys strong statistical guarantees in the large-sample setting, and, for the special case of the newsvendor problem, $\boldsymbol{x}_k^{\mathsf{KS}}(\rho_k, \mathcal{S}_k)$ can be evaluated in closed-form (Bertsimas et al. 2018). For these reasons, we employ it in our experiments as a strong, distributionally robust benchmark. Throughout, we select the parameters ρ_k in a decoupled fashion, using 5-fold cross-validation on $\mathcal{S}_k$ to select ρ_k for each k.

As mentioned, our "Beta" policies use data-driven anchors selected from $\mathcal{P}$, the class of all (scaled) Beta-distributions. More specifically, this class consists of all Beta $\left(\frac{\mu}{1-\mu}\theta_2, \theta_2\right)$ distributions with mean $\mu \in \{1e-6, .05, .1, \ldots, 1\}$ and shape parameter $\theta_2 \in \{0, .05, .1, .15, \ldots, 3\}$. (In cases where $d < \infty$, we discretize this distribution into d equal sized bins on $[0, 1]$.) This beta-distribution should be interpreted as the distribution of the *normalized* demand at the k^{th} store. Said differently, when shrinking the k^{th} problem, we shrink to the un-normalized demand, i.e., towards the distribution of $\hat{\xi}_{k,\min} + (\hat{\xi}_{k,\min} - \hat{\xi}_{k,\max}) \cdot \text{Beta}\left(\frac{\mu}{1-\mu}\theta_2, \theta_2\right)$.

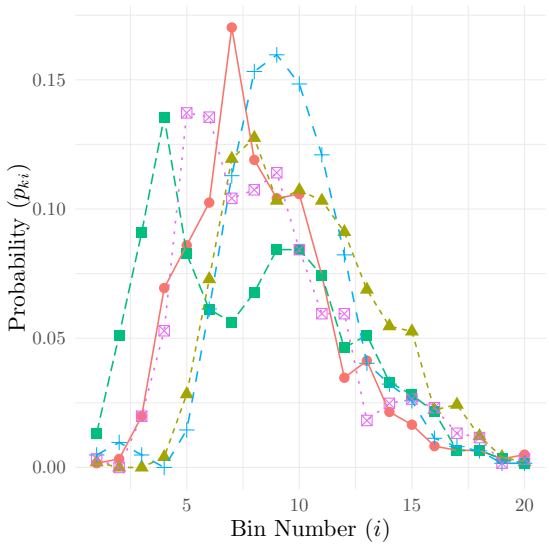
E.4. Summary of Historical Dataset

Figure EC.2 illustrates typical demand distributions at our stores as described in Section 6. The stores display significant heterogeneity.

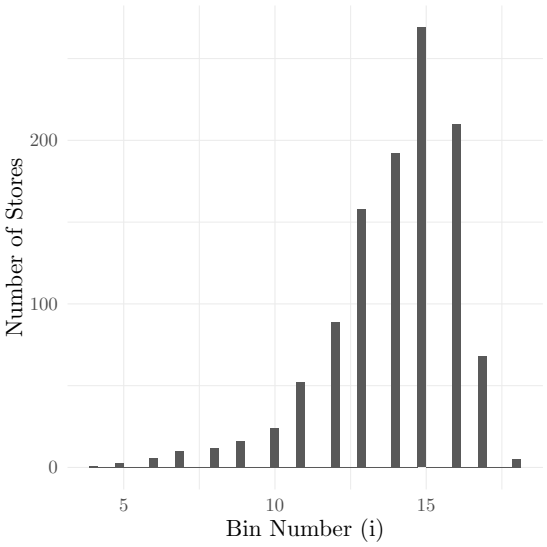
The first panel of Fig. EC.3 shows the average daily demand by store for each of the 1,115 stores in our dataset. The second panel shows estimates of the demand distributions at a few stores. We stress that the individual demand distributions exhibit markedly different means, variances and skewness.

E.5. Additional Figures from Sections 6.2 and 6.3.

The relative performance improvement over all SAA for all of our policies from the experiment in Section 6.2 is displayed in Tables EC.1 and EC.2 for the case where $\hat{N}_k$ is random and non-random, respectively. To ease comparison, policies that shrink to the same type of anchor are grouped together. Notice qualitative features are similar in both tables.

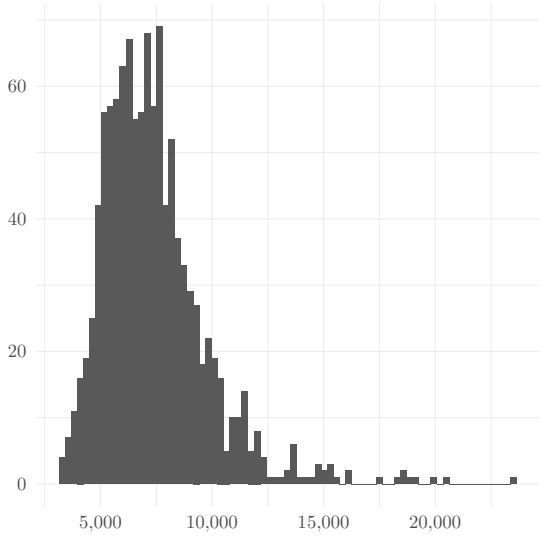


(a) Representative $\boldsymbol{p}_k$

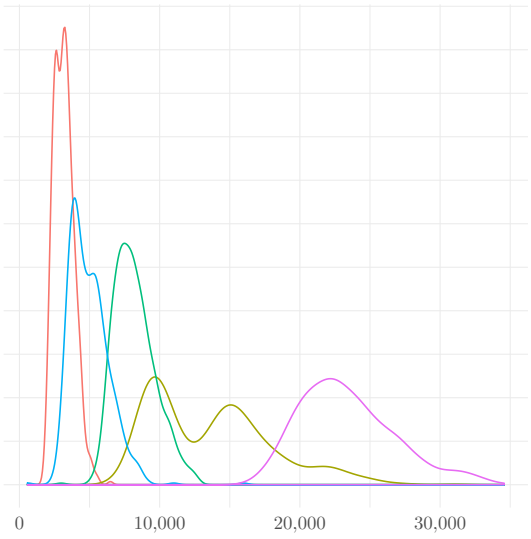


(b) Distribution of Critical Quantile across Bins

Figure EC.2 **Heterogeneity in $\boldsymbol{p}_k$ across stores.** The left panel shows some representative (discretized) distributions $\boldsymbol{p}_k$ when $d = 20$ for several stores. The right panel shows a histogram of the number of stores whose critical quantile occurs in each bin.



(a) **Distribution of Average Daily Demand**



(b) **Demand Distributions by Store**

Figure EC.3 **Heterogeneity in Store Demand**. The first panel shows a histogram of average daily demand by store across 1,115 stores in a European drugstore chain. The second panel shows estimates of the demand distribution at a few representative stores.

Table EC.1 Relative Performance Improvement over SAA (%), $\hat{N}_k \sim \text{Poisson}(10)$.
Performance using simulated data as described in Section 6.2.

	Beta		Grand-Mean			Fixed (Uniform)			Decoupled	
K	Oracle	S-SAA	Oracle	S-SAA	JS	Oracle	S-SAA	JS	SAA	KS
10	17.20	10.19	15.28	12.61	9.83	12.49	8.43	5.53	0	-8.71
32	11.02	6.42	9.44	7.07	4.00	6.09	3.48	0.56	0	-12.05
64	11.34	8.57	10.17	8.71	5.20	7.40	6.65	1.24	0	-11.57
128	13.04	11.75	12.38	11.68	5.27	9.38	9.34	1.37	0	-11.49
256	13.10	12.37	12.66	12.27	4.94	9.66	9.66	0.92	0	-10.71
362	13.08	12.57	12.69	12.43	5.13	9.71	9.71	0.36	0	-10.43
431	13.26	12.80	12.91	12.68	5.13	9.95	9.95	0.46	0	-10.25
512	12.95	12.48	12.50	12.29	5.21	9.67	9.67	0.27	0	-10.64
609	13.12	12.72	12.69	12.49	5.32	9.82	9.82	0.20	0	-10.57
724	13.21	12.85	12.80	12.63	5.39	9.97	9.97	0.17	0	-10.43
861	13.35	13.04	12.95	12.78	5.40	10.08	10.08	0.13	0	-10.46
1024	13.07	12.79	12.67	12.52	5.29	9.78	9.78	0.05	0	-10.62
1115	13.12	12.86	12.73	12.58	5.27	9.82	9.82	0.05	0	-10.68

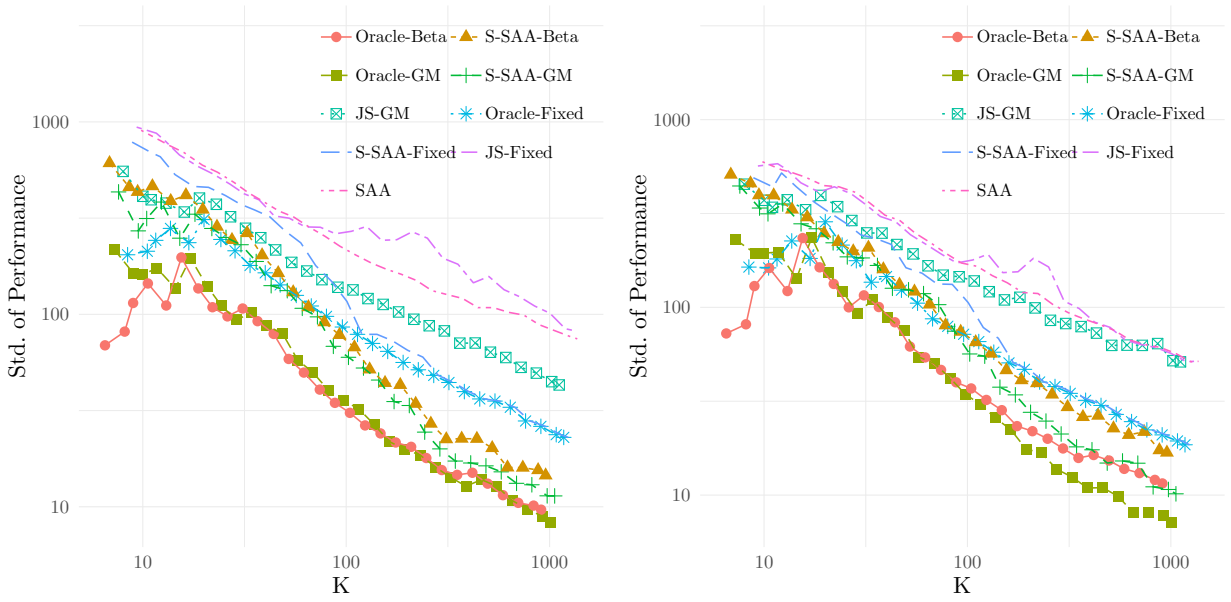
Table EC.2 Relative Performance Improvement over SAA (%), $\hat{N}_k = 10$ **(non-random).**
Performance using simulated data as described in Section 6.2.

	Beta		Grand-Mean			Fixed (Uniform)			Decoupled	
K	Oracle	S-SAA	Oracle	S-SAA	JS	Oracle	S-SAA	JS	SAA	KS
10	13.07	6.89	11.13	8.54	6.46	10.12	7.42	4.46	0	-13.89
32	7.37	3.19	6.19	3.49	1.00	4.52	2.36	0.05	0	-17.05
64	7.09	4.75	6.27	4.70	1.27	4.70	3.88	0.16	0	-16.93
128	8.71	7.62	8.28	7.69	1.28	6.43	6.35	0.47	0	-17.20
256	8.92	8.25	8.67	8.37	1.13	6.68	6.68	0.25	0	-16.26
362	8.93	8.47	8.71	8.50	1.08	6.65	6.65	0.03	0	-16.06
431	9.11	8.75	8.95	8.78	1.26	6.83	6.83	0.03	0	-15.92
512	8.87	8.52	8.57	8.40	1.55	6.69	6.69	0.00	0	-16.22
609	9.03	8.70	8.70	8.53	1.47	6.83	6.83	0.00	0	-16.33
724	9.16	8.88	8.88	8.73	1.57	6.98	6.98	0.00	0	-16.21
861	9.42	9.15	9.12	8.98	1.61	7.27	7.27	0.00	0	-16.26
1024	9.19	8.96	8.86	8.74	1.67	7.02	7.02	0.00	0	-16.45
1115	9.22	8.98	8.90	8.77	1.62	7.05	7.05	0.00	0	-16.49

Figure EC.4 shows the standard deviation of each of our methods on simulated data from Section 6.2 as a function of K, both when Assumption 3.1 holds and when it is violated and the amount of data is fixed. Performance is again quite similar in both cases.

Figure EC.5 shows the average amount of pooling by method by K on our simulated data set from Section 6.2, both when Assumption 3.1 holds and when the amount of data is fixed. Again, in both cases the performance is quite similar, and we see that both Shrunken-SAA and the oracle method when using $\hat{\boldsymbol{p}}^{\text{GM}}$ shrink more than the other methods.

E.6. Additional Figures from Section 6.4: Historical Backtest



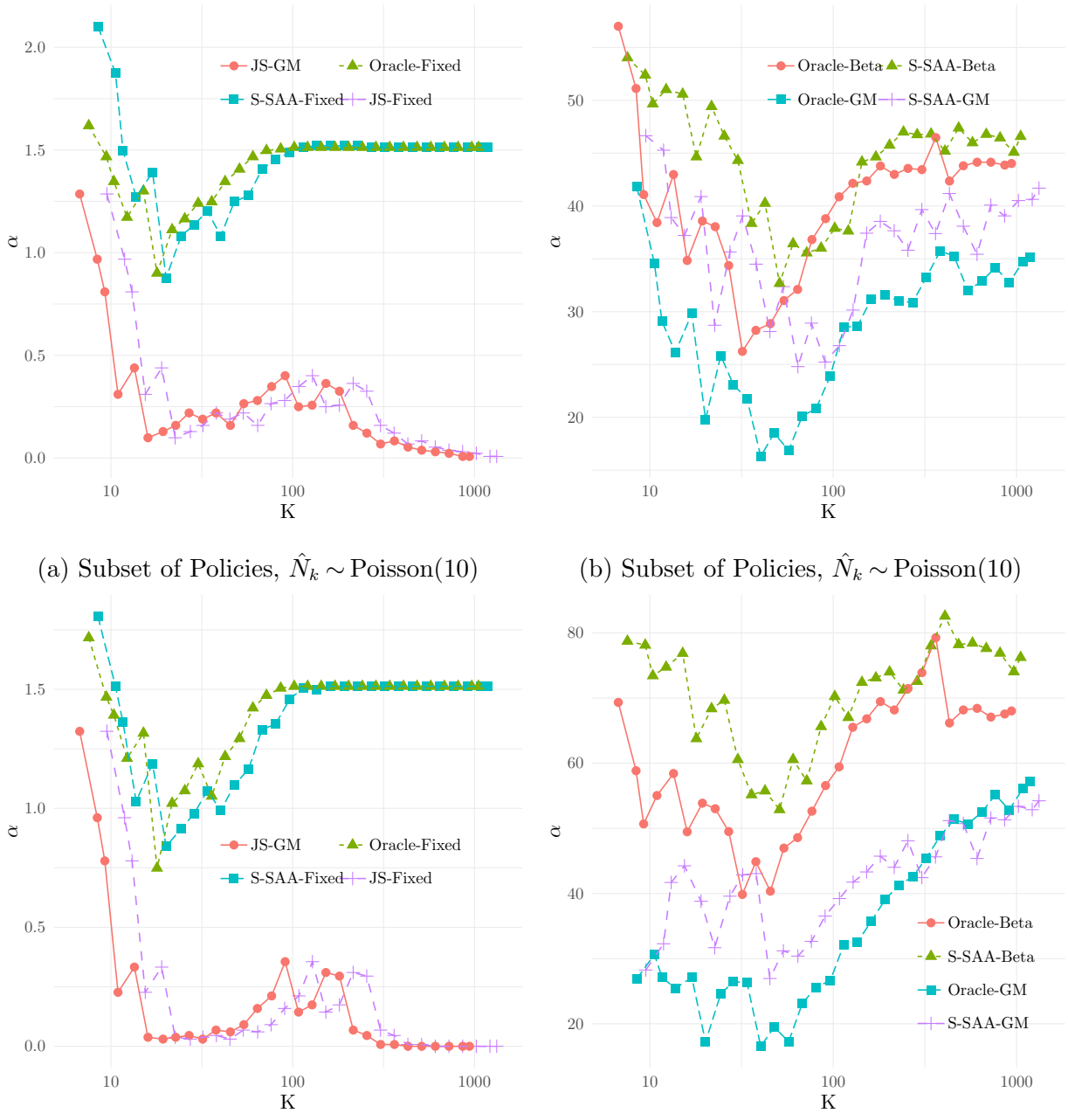
(a) $\hat{N}_k \sim \text{Poisson}(10)$ (b) $\hat{N}_k = 10$ (non-random)

Figure EC.4 **Standard Deviation of Performance** For each method, the standard deviation of converges to zero because performance concentrates at its expectation as $K \to \infty$. Notice that our Shrunken-SAA methods are less variable than the decoupled SAA solution because pooling increases stability.

Table EC.3 shows the relative performance improvement over SAA for all of our policies in the historical data experiment described in Section 6.4 with $d = 20$. For convenience, policies with the same type of anchor are grouped together for comparison.

Table EC.3 **Relative Performance Improvement over SAA (%), Historical Data.** Performance using historical data as described in Section 6.4, $d = 20$.

	Beta		Grand-Mean			Fixed (Uniform)			Decoupled	
K	Oracle	S-SAA	Oracle	S-SAA	JS	Oracle	S-SAA	JS	SAA	KS
10	18.96	4.72	13.99	8.98	8.16	11.82	5.04	4.13	0	-12.56
32	11.34	4.17	8.65	3.96	1.63	5.83	2.32	0.26	0	-14.62
64	10.47	6.25	8.74	6.22	2.44	5.99	4.70	0.17	0	-14.02
128	11.88	9.92	11.10	9.92	2.55	8.44	8.44	0.38	0	-13.25
256	11.92	10.89	11.44	10.98	2.38	9.06	9.06	0.59	0	-12.60
362	11.49	10.78	11.16	10.81	2.08	8.67	8.67	0.00	0	-12.44
431	11.55	10.89	11.25	10.95	2.25	8.72	8.72	0.00	0	-12.28
512	11.12	10.43	10.73	10.48	2.49	8.38	8.38	0.00	0	-12.50
609	11.19	10.57	10.81	10.58	2.57	8.48	8.48	0.00	0	-12.42
724	11.25	10.77	10.94	10.79	2.65	8.62	8.62	0.00	0	-12.31
861	11.40	11.01	11.12	10.96	2.61	8.75	8.75	0.00	0	-12.47
1024	11.20	10.85	10.93	10.80	2.58	8.59	8.59	0.00	0	-12.56
1115	11.30	10.95	11.05	10.94	2.55	8.68	8.68	0.00	0	-12.61



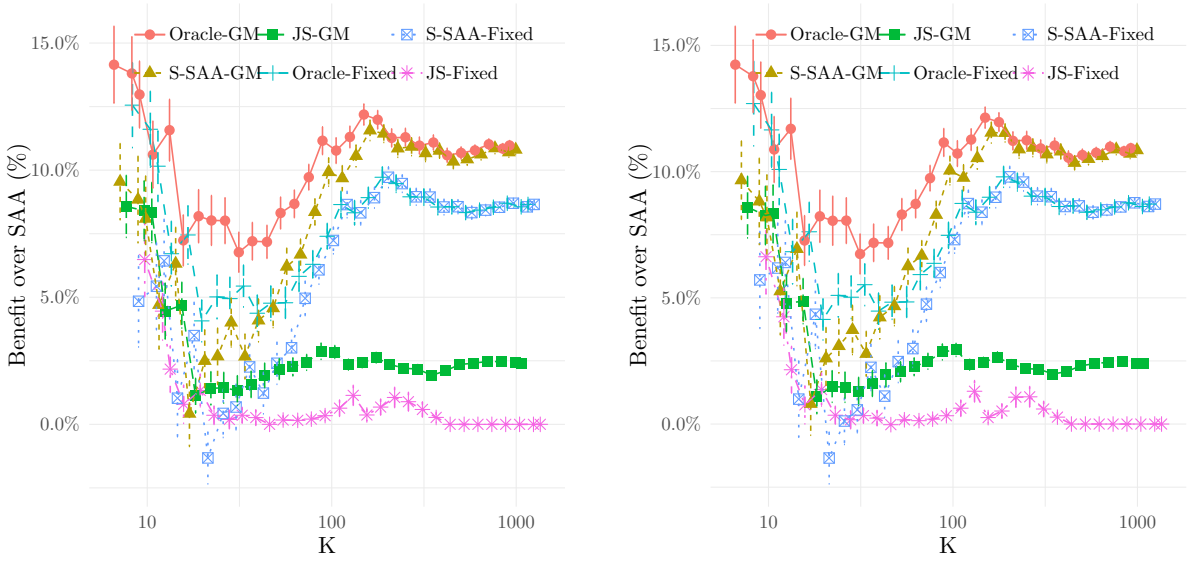
(a) Subset of Policies, $\hat{N}_k \sim \text{Poisson}(10)$

(b) Subset of Policies, $\hat{N}_k \sim \text{Poisson}(10)$

(c) Subset of Policies, $\hat{N}_k = 10$ (non-random)

(d) Subset of Policies, $\hat{N}_k = 10$ (non-random)

Figure EC.5 **Amount of Pooling by Method** We plot the amount of data-pooling (α) for each of the above methods (plotted separately for clarity). In panels a) and b), the amount of data follows Assumption 3.1. In the remainder, it is fixed. In general, optimization-aware methods shrink much more aggressively in both instances.



(a) $d = 50$ (b) $d = \infty$

Figure EC.6 **Robustness to choice of d.** Performance of policies on our historical data. In the first panel, $d = 50$. In the second panel, the distributions $\mathbb{P}_k$ are treated as continuous in the Shrunken-SAA algorithm, i.e., $d = \infty$. Error bars show ± 1 standard error. The differences between the plots are essentially indiscernible.

E.7. Performance as $d \to \infty$

Recall that the Shrunken-SAA algorithm, does not require that the random variables $\boldsymbol{\xi}_k$ have discrete support (cf. Remark 3.1). Consequently, we next study the robustness of Shrunken-SAA to d, the number of support points of $\boldsymbol{\xi}_k$.

To this end, we increase d from our base case. Figure EC.6 below shows results for $d = 50$ and $d = \infty$, i.e., not performing any discretization. The complete set of policies can be seen in Tables EC.4 and EC.5 below.

The performance is nearly identical to the case of $d = 20$. To make this clearer, in the second panel of Fig. EC.7 we plot the performance of our Shrunken-SAA methods for varying d. Again, the differences are quite small. In our opinion, these results suggest that the performance of Shrunken-SAA is quite robust to size of the support of $\boldsymbol{\xi}_k$, and is still effective if $\boldsymbol{\xi}_k$ may be continuous.

E.8. Performance as $N \to \infty$

We next study the performance of our methods as we increase $\hat{N}_k$. Recall in the experiment above, $\hat{N}_k = 10$, with some instances having fewer training points due to missing values. In Fig. EC.8 we consider $\hat{N}_k = 20$ days and $\hat{N}_k = 40$ days for training (again with some instances having fewer data points), and let $d = \infty$. (See also Tables EC.6 and EC.7 for all benchmarks.) As $\hat{N}_k$ increases for

Table EC.4 Relative Performance Improvement over SAA (%).

Performance using historical data as described in Appendix E.7, $d = 50$.

	Beta		Grand-Mean			Fixed (Uniform)			Decoupled	
K	Oracle	S-SAA	Oracle	S-SAA	JS	Oracle	S-SAA	JS	SAA	KS
10	18.39	6.38	13.81	8.84	8.41	11.61	5.43	4.45	0	-17.50
32	10.73	3.44	8.01	4.01	1.33	5.44	2.26	0.27	0	-18.56
64	10.01	6.51	8.31	6.21	2.26	5.82	4.96	0.21	0	-18.02
128	11.60	9.96	10.77	9.68	2.35	8.32	8.32	0.37	0	-16.03
256	11.73	10.86	11.27	10.86	2.21	8.95	8.95	0.58	0	-15.86
362	11.36	10.80	10.97	10.67	1.93	8.55	8.55	0.00	0	-16.01
431	11.44	10.92	11.09	10.77	2.11	8.56	8.56	0.00	0	-15.69
512	11.02	10.57	10.59	10.34	2.35	8.33	8.33	0.00	0	-16.14
609	11.08	10.70	10.68	10.43	2.40	8.43	8.43	0.00	0	-16.03
724	11.15	10.82	10.78	10.62	2.47	8.54	8.54	0.00	0	-15.76
861	11.36	11.09	11.02	10.86	2.47	8.70	8.70	0.00	0	-15.72
1024	11.20	10.93	10.85	10.70	2.44	8.56	8.56	0.00	0	-15.98
1115	11.32	11.08	10.97	10.81	2.40	8.65	8.65	0.00	0	-15.85

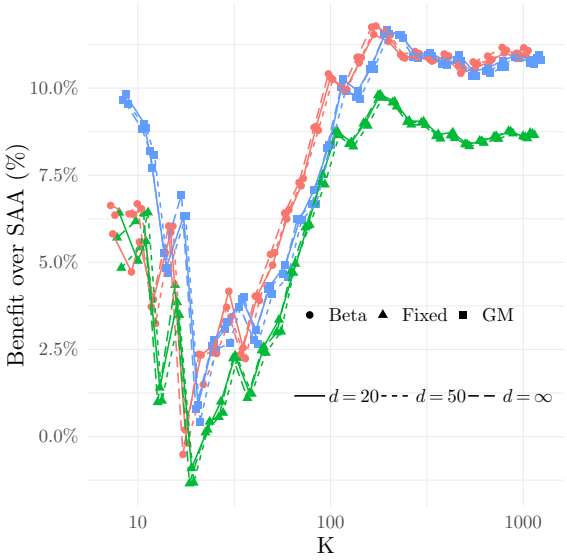


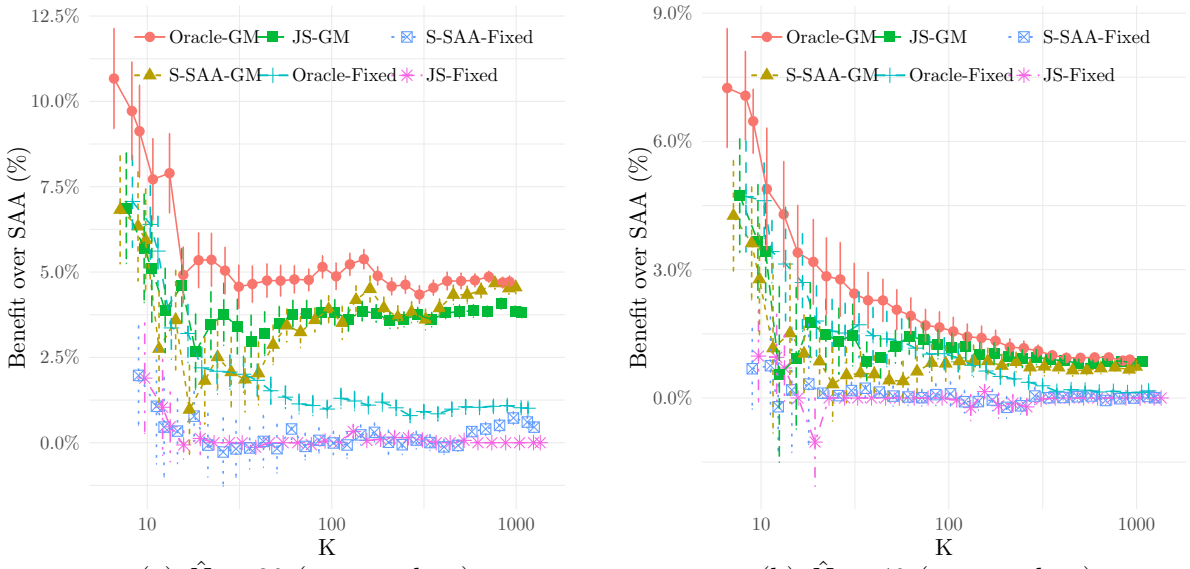
Figure EC.7 Robustness to d on Historical Data. We limit attention to the Shrunken-SAA policies and compare them on the same historical datasets for $d = 20, 50, \infty$. The performance of each variant is insensitive to d.

all k, SAA, itself, converges in performance to the full-information optimum. Consequently, there is "less-room" to improve upon SAA, and we see that for $\hat{N}_k = 40$, our methods still improve upon decoupling, but by a smaller amount. We also note that the JS-GM variant performs relatively better than for small $\hat{N}_k$. We intuit this is because as $\hat{N}_k \to \infty$, the empirical distribution $\hat{p}_k$ converges in probability to the true distribution $\boldsymbol{p}_k$, i.e., the variance of $\hat{\boldsymbol{p}}_k$ around $\boldsymbol{p}_k$ decreases. For large enough $\hat{N}_k$, this variance is a "second order" concern, and hence accounting for discrepancy in the mean (which is how $\alpha^{\mathsf{JS}}_{\boldsymbol{p}_0}$ is chosen) captures most of the benefits. This viewpoint accords more generally with intuition that estimate-then-optimize procedures work well in environments with high signal-to-noise ratios.

Table EC.5 Relative Performance Improvement over SAA (%).

Performance using historical data as described in Appendix E.7, $d = \infty$.

	Beta		Grand-Mean			Fixed (Uniform)			Decoupled	
K	Oracle	S-SAA	Oracle	S-SAA	JS	Oracle	S-SAA	JS	SAA	KS
10	18.13	6.39	13.78	8.82	8.26	11.65	6.18	4.25	0	-23.75
32	10.60	3.71	8.06	3.73	1.29	5.52	2.25	0.24	0	-26.99
64	9.94	6.41	8.30	6.25	2.27	5.92	4.75	0.19	0	-25.18
128	11.52	9.98	10.72	9.76	2.35	8.39	8.39	0.26	0	-23.38
256	11.66	10.96	11.22	10.88	2.19	9.03	9.03	0.60	0	-21.83
362	11.31	10.83	10.92	10.68	1.98	8.62	8.62	0.00	0	-21.49
431	11.38	10.98	11.03	10.80	2.08	8.64	8.64	0.00	0	-20.74
512	10.99	10.65	10.55	10.35	2.31	8.39	8.39	0.00	0	-21.51
609	11.06	10.75	10.66	10.50	2.39	8.48	8.48	0.00	0	-21.43
724	11.13	10.92	10.75	10.62	2.45	8.60	8.60	0.00	0	-21.09
861	11.34	11.17	10.99	10.88	2.48	8.77	8.77	0.00	0	-20.76
1024	11.17	11.00	10.81	10.71	2.40	8.62	8.62	0.00	0	-20.88
1115	11.29	11.16	10.93	10.85	2.38	8.71	8.71	0.00	0	-20.83



(a) $\hat{N}_k = 20$ (non-random) (b) $\hat{N}_k = 40$ (non-random)

Figure EC.8 Dependence on N. Evaluated on historical data with $d = \infty$. Error bars show ± 1 standard error.

In summary, we believe these preliminary studies support the idea that Shrunken-SAA retains many of SAA's strong large-sample properties, but still offers a marginal benefit for large K.

E.9. Other Forms of Cross-Validation

Our theoretical development of Shrunken-SAA naturally motivated our Modified-LOO procedure in Algorithm 1. When $K\hat{N}_{\text{avg}}$ is very large, however, LOO may be computationally demanding, and simpler 5-fold or 10-fold cross-validation methods might be preferred. We next study the

Table EC.6 Relative Performance Improvement over SAA (%) when $N = 20$.
Evaluated on historical data with $d = \infty$.

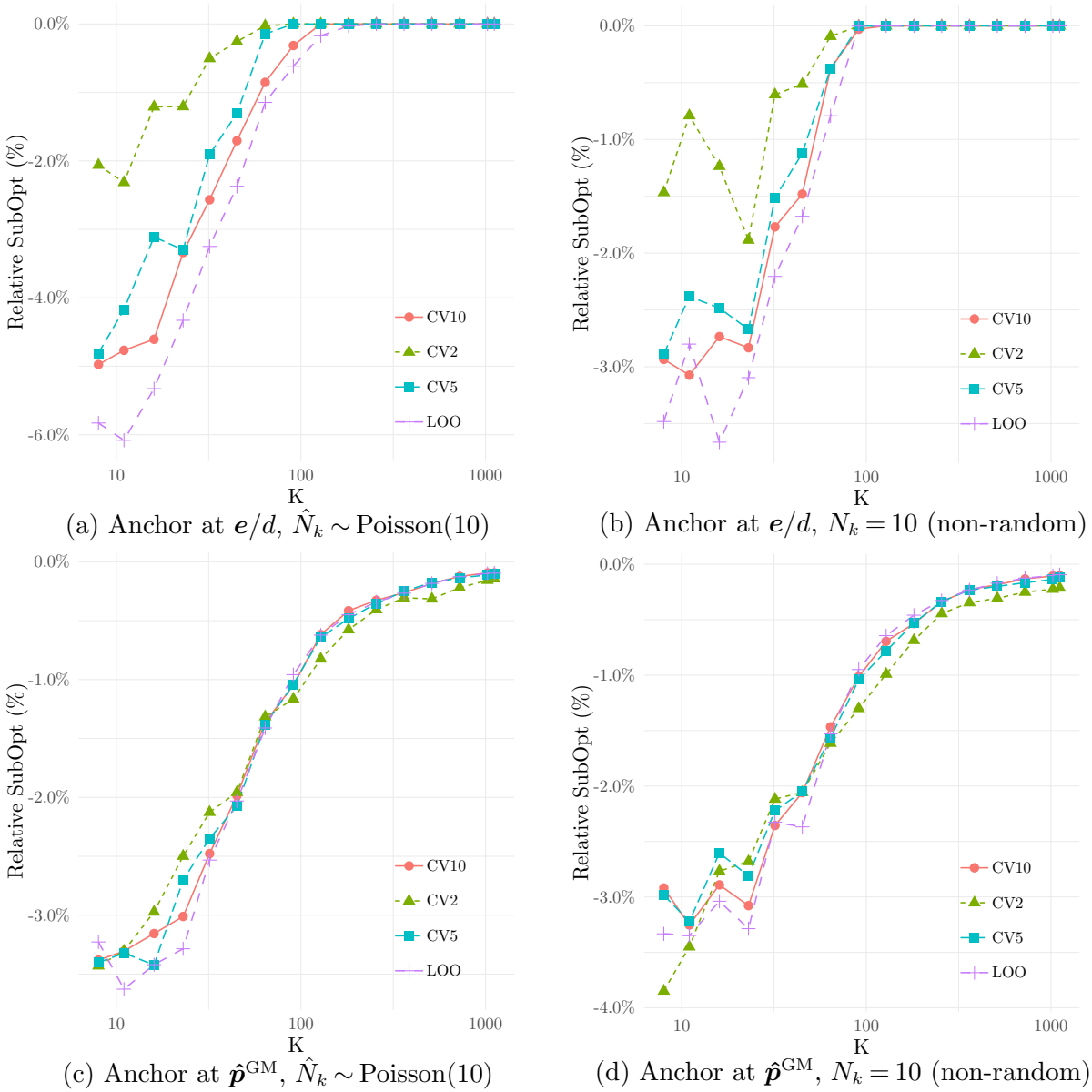
	Beta		Grand-Mean			Fixed (Uniform)			Decoupled	
K	Oracle	S-SAA	Oracle	S-SAA	JS	Oracle	S-SAA	JS	SAA	KS
10	15.09	5.58	10.72	7.57	7.29	8.26	2.90	1.03	0	-7.79
32	8.74	2.76	5.85	2.96	3.81	2.81	0.29	-0.14	0	-16.33
64	6.73	3.62	5.27	3.87	3.98	1.74	-0.17	-0.06	0	-16.46
128	6.32	4.53	5.21	3.93	3.89	1.51	0.18	0.02	0	-16.94
256	5.87	4.32	4.93	3.92	3.80	1.23	0.02	-0.13	0	-15.75
362	5.68	4.56	4.75	4.04	3.89	1.28	0.05	0.00	0	-15.72
431	5.65	4.59	4.81	4.23	3.96	1.32	0.23	0.07	0	-15.69
512	5.72	4.79	4.98	4.55	4.00	1.35	0.50	0.00	0	-16.09
609	5.58	4.88	4.89	4.56	3.94	1.28	0.58	0.00	0	-16.33
724	5.54	4.85	4.87	4.56	3.97	1.29	0.62	0.00	0	-16.28
861	5.54	5.09	5.04	4.82	4.09	1.32	0.74	0.00	0	-16.34
1024	5.50	4.97	4.93	4.75	3.95	1.26	0.45	0.00	0	-16.58
1115	5.44	4.95	4.90	4.75	3.90	1.19	0.30	0.00	0	-16.50

Table EC.7 Relative Performance Improvement over SAA (%) when $N = 40$.
Evaluated on historical data with $d = \infty$.

	Beta		Grand-Mean			Fixed (Uniform)			Decoupled	
K	Oracle	S-SAA	Oracle	S-SAA	JS	Oracle	S-SAA	JS	SAA	KS
10	10.09	1.21	7.07	3.62	3.66	4.61	0.76	0.95	0	-3.58
32	5.44	0.63	2.77	0.53	1.47	1.71	0.23	0.00	0	-9.09
64	3.71	0.80	2.06	0.39	1.45	1.17	0.01	0.00	0	-6.20
128	2.55	1.44	1.56	0.86	1.20	0.77	-0.08	0.15	0	-5.58
256	2.08	1.51	1.19	0.84	0.92	0.37	0.05	-0.03	0	-5.40
362	1.92	1.16	1.10	0.74	0.89	0.16	0.01	0.00	0	-5.35
431	1.85	1.32	1.00	0.72	0.88	0.19	0.03	0.00	0	-5.23
512	1.67	1.22	0.94	0.65	0.80	0.18	0.05	0.00	0	-5.06
609	1.49	1.11	0.94	0.65	0.80	0.14	-0.06	0.00	0	-5.20
724	1.57	1.22	0.95	0.69	0.84	0.16	-0.02	0.00	0	-5.04
861	1.49	1.16	0.95	0.71	0.86	0.13	-0.00	0.00	0	-4.99
1024	1.41	1.13	0.87	0.67	0.81	0.13	0.04	0.00	0	-5.05
1115	1.44	1.19	0.90	0.73	0.84	0.17	0.01	0.00	0	-4.99

performance of Algorithm 1 when we replace the Modified-LOO Cross-Validation step by a simpler Modified κ-fold cross-validation step, where $\kappa \in \{2, 5, 10\}$. Here the qualifier "Modified" indicates that, as in Algorithm 1, we do not update the anchor (even if it depends on the data) for each fold.

Figure EC.9 shows some indicative results under the synthetic data setting of Section 6.2 in the case $d = \infty$ (continuous data). We consider both $\hat{N}_k \sim \text{Poisson}(10)$ (left panels) or $\hat{N}_k = 10$ (right panels). In both settings, each form of cross-validation converges to oracle performance qualitatively similarly to the LOO performance. We have repeated this test for other values of N and d and with our historical data setting of Section 6.4, and largely observe similar results. In summary,



(a) Anchor at $\boldsymbol{e}/d$, $\hat{N}_k \sim \text{Poisson}(10)$

(b) Anchor at $\boldsymbol{e}/d$, $N_k = 10$ (non-random)

(c) Anchor at $\hat{\boldsymbol{p}}^{\text{GM}}$, $\hat{N}_k \sim \text{Poisson}(10)$

(d) Anchor at $\hat{\boldsymbol{p}}^{\text{GM}}$, $N_k = 10$ (non-random)

Figure EC.9 **Other Types of Cross-Validation** We compare variants of the Shrunken-SAA procedure that leverage (modified) 2, 5 or 10-fold cross-validation instead of LOO cross-validation in Algorithm 1, both when the amount of data is random (Assumption 3.1 hold) and when it is fixed. In each case, all methods of cross-validation seem to converge to oracle optimality. Plots show relative suboptimality to oracle performance.

this suggests empirically that when computational budgets require it, Shrunken-SAA can safely be implemented with other forms of cross-validation.

Appendix F: Extension to Continuous Distributions

In this section, we extend our results from Sections 4.2 and 4.3 to the case where the random variables $\boldsymbol{\xi}_k$ may have continuous support and discuss the challenges of similar extensions for

discrete problems. Specifically, we no longer assume $\boldsymbol{\xi}_k \in \{\boldsymbol{a}_{k1}, \ldots, \boldsymbol{a}_{kd}\}$, i.e., that $\boldsymbol{\xi}_k$ is supported on a finite set. Instead we allow any compact support.

ASSUMPTION F.1 **(Compact Support for $\mathbb{P}_k$ and $h(\mathcal{S})$)**. *There exists a compact set $\Xi \subseteq \mathbb{R}^\ell$ such that, for each $k = 1, \ldots, K$, $\boldsymbol{\xi}_k \sim \mathbb{P}_k$ is an ℓ-dimensional real random vector whose support is contained in Ξ, and, with probability 1 with respect $\mathcal{S}$, $h(\mathcal{S}) \in \mathcal{P}$ and has support contained in Ξ.*

As mentioned in Section 4.6, our proof technique will be to consider a discretized system whose performance is arbitrarily close to the true, continuous system and invoke our results for this discretized system. In order to construct an arbitrarily close discretized system, we will require some additional continuity on the cost functions.

ASSUMPTION F.2 **(Equicontinuity)**. *For each $k = 1, \ldots, K$, $\{c_k(\boldsymbol{x}, \boldsymbol{\xi}_k) \ : \ \boldsymbol{x} \in \mathcal{X}_k\}$ is equicontinuous in $\boldsymbol{\xi}$ for all $\boldsymbol{\xi} \in \Xi$. Namely, for every $\epsilon > 0, \boldsymbol{\xi} \in \Xi$ there exists $\delta > 0$ such that $|c_k(\boldsymbol{x}, \boldsymbol{\xi}) - c_k(\boldsymbol{x}, \boldsymbol{\xi}')| \leq \epsilon$ for all $\boldsymbol{x} \in \mathcal{X}$, $\|\boldsymbol{\xi} - \boldsymbol{\xi}'\| \leq \delta$.*

REMARK F.1. Notice that in principle, $c_k(\boldsymbol{x}, \boldsymbol{\xi}_k)$ need only be defined for $\boldsymbol{\xi}_k$ in the support of $\mathbb{P}_k$. Assuming that it is defined and equicontinuous on the larger Ξ is without loss of generality via the Tietze continuous extension theorem (Munkres 1974, Theorem 3.2). $\square$

Finally, we assume the same assumptions on the cost functions as in Sections 4.2 and 4.3. We restate these below in terms of $\boldsymbol{\xi}$ that may not be finitely supported.

ASSUMPTION F.3 **(Bounded, Lipschitz, Strongly-Convex Optimization)**. *There exists L, γ such that $c_k(\boldsymbol{x}, \boldsymbol{\xi})$ are γ-strongly convex and L-Lipschitz over $\mathcal{X}_k$, and, moreover, $\mathcal{X}_k$ is non-empty and convex, for all $k = 1, \ldots, K$, and $\boldsymbol{\xi} \in \Xi$.*

For clarity, we repeat the definitions of some of our primitives, but now in terms of general distributions and data sets $\mathcal{S}_k$ and $\mathcal{S}$:

$$\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k) \in \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \sum_{j=1}^{\hat{N}_k} c_k(\boldsymbol{x}_k, \hat{\boldsymbol{\xi}}_{kj}) + \alpha \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{Q}}\left[c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k)\right],$$

$$\overline{Z}_K(\alpha, \mathbb{Q}) = \frac{1}{K}\sum_{k=1}^K Z_k(\alpha, \mathbb{Q}), \quad \text{where} \quad Z_k(\alpha, \mathbb{Q}) = \frac{\lambda_k}{\lambda_{\text{avg}}} \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{P}_k}[c_k(\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k), \boldsymbol{\xi}_k)],$$

$$\overline{Z}_K^{\mathsf{LOO}}(\alpha, \mathbb{Q}) = \frac{1}{K}\sum_{k=1}^K Z_k^{\mathsf{LOO}}(\alpha, \mathbb{Q}), \quad \text{where} \quad Z_k^{\mathsf{LOO}}(\alpha, \mathbb{Q}) = \frac{1}{N\lambda_{\text{avg}}} \sum_{j=1}^{\hat{N}_k} c_k(\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k \backslash \{\hat{\boldsymbol{\xi}}_{kj}\}), \hat{\boldsymbol{\xi}}_{kj}).$$

Notice $\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k)$ is precisely as in Algorithm 1.

The oracle pooling amount for a specified $h(\cdot)$ is given by

$$\alpha_h^{\mathsf{OR}} \in \arg\min_{\alpha \geq 0} \overline{Z}_K(\alpha, h(\mathcal{S})),$$

and the simultaneous oracle pooling amount and oracle anchor within a class $\mathcal{P}$ is given by

$$(\alpha_{\mathcal{P}}^{\mathsf{OR}}, \mathbb{Q}_{\mathcal{P}}^{\mathsf{OR}}) \in \operatorname*{arg\,min}_{\alpha\geq 0,\, \mathbb{Q}\in\mathcal{P}} \overline{Z}_K(\alpha, \mathbb{Q}).$$

Again, we will measure performance of a policy relative to these oracle benchmarks:

$$\mathsf{SubOpt}_{h,K}(\alpha) = \overline{Z}_K(\alpha, h(\mathcal{S})) - \overline{Z}_K(\alpha_h^{\mathsf{OR}}, h(\mathcal{S})),$$
$$\mathsf{SubOpt}_{\mathcal{P},K}(\alpha, \mathbb{Q}) = \overline{Z}_K(\alpha, \mathbb{Q}) - \overline{Z}_K(\alpha_{\mathcal{P}}^{\mathsf{OR}}, \mathbb{Q}_{\mathcal{P}}^{\mathsf{OR}}).$$

For convenience, we again often refer to the constant function $\mathcal{S}\mapsto\mathbb{Q}$ as just $\mathbb{Q}$. Notice that in the special case that $\boldsymbol{\xi}_k$ has finite, discrete support, each of these above definitions is equivalent to our original definitions in Section 2.

We can now prove an extension of Theorem 4.2 to the case of continuous random variables.

THEOREM F.1. **(Shrunken-SAA with Fixed Anchors for Strongly-Convex Problems and Continuous Distributions)** *Fix any $\mathbb{P}_0$. Suppose Assumptions 3.1 and F.1 to F.3 hold, $K\geq 2$ and $N\lambda_{\min}\geq 1$. Then, there exists a universal constant* A *such that for any $0<\delta<1/2$, with probability at least $1-\delta$, we have that*

$$\mathsf{SubOpt}_{\mathbb{P}_0,K}(\alpha_{\mathbb{P}_0}^{\mathsf{S\text{-}SAA}}) \;\leq\; \mathrm{A}\cdot\max\left(C, L\sqrt{\frac{C}{\gamma}}\right)\cdot\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4}\cdot\frac{\log^2(1/\delta)\cdot\log^{3/2}(K)}{\sqrt{K}}.$$

The first step in the proof of Theorem F.1 is to construct our approximate discrete system:

LEMMA F.1 **(A Discrete Approximate System)**. *Suppose Assumptions F.1 to F.3 hold. Then, for any $\epsilon>0$, there exists a finite partition $B_1',\ldots,B_d'$ of Ξ and, for each $k=1,\ldots,K$, random variables $\boldsymbol{\xi}_k^{\mathsf{disc}}$ supported on $\{\boldsymbol{a}_{k1},\ldots,\boldsymbol{a}_{kd}\}$ such that*

i) $\boldsymbol{a}_{ki}\in B_i'$,

ii) $\mathbb{P}(\boldsymbol{\xi}_k^{\mathsf{disc}}=\boldsymbol{a}_{ki}) = \mathbb{P}(\boldsymbol{\xi}_k\in B_i')$, *and*

iii) $|c_k(\boldsymbol{x}_k,\boldsymbol{\xi}_k) - c_k(\boldsymbol{x},\boldsymbol{a}_{ki})|\leq\epsilon$ *for all* $k=1,\ldots,K$, $i=1,\ldots,d$, $\boldsymbol{\xi}_k\in B_i'$ *and* $\boldsymbol{x}\in\mathcal{X}_k$.

Proof. Since K is finite, Assumption F.2 implies that the larger set $\{c_k(\boldsymbol{x};\boldsymbol{\xi}) : \boldsymbol{x}\in\mathcal{X}_k,\, k=1,\ldots,K\}$ is equicontinuous in $\boldsymbol{\xi}$ for all $\boldsymbol{\xi}\in\Xi$. In other words, for every $\boldsymbol{\xi}\in\Xi$, there exists $\delta(\boldsymbol{\xi})>0$ such that $|c_k(\boldsymbol{x};\boldsymbol{\xi}) - c_k(\boldsymbol{x};\boldsymbol{\xi}')|\leq\epsilon$ for all $k=1,\ldots,K$ and $\boldsymbol{x}\in\mathcal{X}_k$ whenever $\|\boldsymbol{\xi}-\boldsymbol{\xi}'\|\leq\delta(\boldsymbol{\xi})$ and $\boldsymbol{\xi}'\in\Xi$. Let $B(\boldsymbol{\xi}) = \{\boldsymbol{\xi}'\in\Xi : \|\boldsymbol{\xi}-\boldsymbol{\xi}'\|\leq\delta(\boldsymbol{\xi})\}$. Then $\bigcup_{\boldsymbol{\xi}\in\Xi} B(\boldsymbol{\xi})$ necessarily covers Ξ. Since Ξ is compact, there exists a finite subcover, $B(\boldsymbol{\xi}_1),\ldots,B(\boldsymbol{\xi}_d)$. We construct a partition from this finite subcover, namely,

$$B_i' = B(\boldsymbol{\xi}_i)\setminus\bigcup_{1\leq j\leq i-1} B(\boldsymbol{\xi}_j)\cap B(\boldsymbol{\xi}_i).$$

In words, B_i' is the same as $B(\boldsymbol{\xi}_i)$ but omits any point that was already covered by a previous set. Let $\chi:\Xi\to\{1,\ldots,d\}$ be the indicator of this partition, i.e., $\boldsymbol{\xi}\in B_{\chi(\boldsymbol{\xi})}'$ for all $\boldsymbol{\xi}\in\Xi$.

Now let

$$\boldsymbol{a}_{ki} \equiv \boldsymbol{\xi}_i, \quad \text{for } i=1,\dots,d, \text{ and } k=1,\dots,K$$

and define the discrete random variable $\boldsymbol{\xi}_k^{\mathsf{disc}}$ such that $\mathbb{P}(\boldsymbol{\xi}_k^{\mathsf{disc}} = \boldsymbol{a}_{ki}) = \mathbb{P}(\boldsymbol{\xi}_k \in B_i')$.

Then the first two claims in the lemma are immediate. For the last, notice by construction of the partition, $\|\boldsymbol{\xi}_k - \boldsymbol{a}_{ki}\| \le \delta(\boldsymbol{a}_{ki})$ so that the third claim holds by equicontinuity. □

We will now apply our existing analysis to the discretized system. For clarity, given $B_i', \hat{\boldsymbol{\xi}}_k^{\mathsf{disc}}$ as in Lemma F.1, we define

$$\begin{aligned}
c_{ki}(\boldsymbol{x}) &\equiv c_k(\boldsymbol{x}, \boldsymbol{a}_{ki}), && \forall i=1,\dots,d,\ k=1,\dots,K,\\
p_{ki} &\equiv \mathbb{P}_k(B_i'), && \forall i=1,\dots,d,\ k=1,\dots,K,\\
\hat{m}_{ki} &\equiv \sum_{j=1}^{\hat{N}_k} \mathbb{I}\left[\hat{\boldsymbol{\xi}}_{kj} \in B_i'\right], && \forall i=1,\dots,d,\ k=1,\dots,K\\
\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k) &\in \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} (\hat{\boldsymbol{m}}_k + \alpha \boldsymbol{q})^T \boldsymbol{c}_k(\boldsymbol{x}_k), && \forall k=1,\dots,K.
\end{aligned}$$

One can confirm directly that, under Assumption F.3, $c_{ki}(\boldsymbol{x})$ are each C-bounded, L-Lipschitz, and γ-strongly convex for every k, i. Finally, for any distribution $\mathbb{Q}$ on $\mathbb{R}^\ell$, define its discretization $\mathsf{disc}(\mathbb{Q}) = (\mathbb{Q}(B_1'), \dots, \mathbb{Q}(B_d')) \in \Delta_d$.

The next step of the proof establishes that the policies $\boldsymbol{x}_k^{\mathsf{disc}}(\cdot,\cdot,\cdot)$ of the discretized system are suitably close to the policies $\boldsymbol{x}_k(\cdot\cdot,\cdot)$ of the original, continuous system.

Lemma F.2 **(Bounding Differences in Policies)**. *Suppose Assumptions F.1 to F.3 hold. For given $\epsilon > 0$, consider the discretization given by Lemma F.1. Then for any $\mathbb{Q} \in \mathcal{P}$ and data set $\mathcal{S}_k$,*

$$\left\|\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k) - \boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k)\right\|_2 \le \sqrt{\frac{2\epsilon}{\gamma}}.$$

Proof. Define

$$\begin{aligned}
f_k^{\mathsf{disc}}(\boldsymbol{x}_k) &\equiv \left(\frac{\hat{\boldsymbol{m}}_k + \alpha \mathsf{disc}(\mathbb{Q})}{\hat{N}_k + \alpha}\right)^T \boldsymbol{c}_k(\boldsymbol{x}_k) = \frac{1}{\hat{N}_k + \alpha} \sum_{j=1}^{\hat{N}_k} c_k(\boldsymbol{x}_k, \xi_{\chi(\hat{\boldsymbol{\xi}}_{kj})}) + \frac{\alpha}{\hat{N}_k + \alpha} \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{Q}}\left[c_k(\boldsymbol{x}_k, \boldsymbol{a}_{k,\chi(\boldsymbol{\xi}_k)})\right],\\
f_k^{\mathsf{cts}}(\boldsymbol{x}_k) &\equiv \frac{1}{\hat{N}_k + \alpha} \sum_{j=1}^{\hat{N}_k} c_k(\boldsymbol{x}_k, \hat{\boldsymbol{\xi}}_{kj}) + \frac{\alpha}{\hat{N}_k + \alpha} \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{Q}}\left[c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k)\right].
\end{aligned}$$

Using Lemma F.1 part iii) and the triangle inequality, we have that $\left|f_k^{\mathsf{disc}}(\boldsymbol{x}_k) - f_k^{\mathsf{disc}}(\boldsymbol{x}_k)\right| \le \epsilon$ for all $\boldsymbol{x}_k \in \mathcal{X}_k$, and all k.

By construction f_k^{disc} and f_k^{cts} are both γ-strongly convex, and $\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k)$ and $\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k)$ are their respective optimizers. Hence, we can use an argument similar to Lemma C.1 to show that $\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k)$ and $\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k)$ are close. More specifically, by strong-convexity

$$f_k^{\mathsf{cts}}(\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k)) - f_k^{\mathsf{cts}}(\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k)) \geq \frac{\gamma}{2} \left\| \boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k) - \boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k) \right\|_2^2$$
$$f_k^{\mathsf{disc}}(\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k)) - f_k^{\mathsf{disc}}(\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k)) \geq \frac{\gamma}{2} \left\| \boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k) - \boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k) \right\|_2^2.$$

Combining, we obtain

$$\begin{aligned}
\gamma \left\| \boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k) - \boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k) \right\|_2^2 &\leq \left| f_k^{\mathsf{cts}}(\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k)) - f_k^{\mathsf{disc}}(\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k)) \right| \\
&\quad + \left| f_k^{\mathsf{disc}}(\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k)) - f_k^{\mathsf{cts}}(\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k)) \right| \\
&\leq 2\epsilon.
\end{aligned}$$

Rearranging proves the result. □

Finally, we introduce discrete analogues of our usual stochastic processes

$$\begin{aligned}
\overline{Z}_K^{\mathsf{disc}}(\alpha, \boldsymbol{q}) &= \frac{1}{K} \sum_{k=1}^K Z_k^{\mathsf{disc}}(\alpha, \boldsymbol{q}) \quad \text{where} \quad Z_k^{\mathsf{disc}}(\alpha, \boldsymbol{q}) = \frac{\lambda_k}{\lambda_{\text{avg}}} \boldsymbol{p}_k^T \boldsymbol{c}_k(\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \boldsymbol{q})), \\
\overline{Z}_K^{\mathsf{LOO,disc}}(\alpha, \boldsymbol{q}) &= \frac{1}{K} \sum_{k=1}^K Z_k^{\mathsf{disc}}(\alpha, \boldsymbol{q}) \quad \text{where} \quad Z_k^{\mathsf{LOO,disc}}(\alpha, \boldsymbol{q}) = \frac{1}{N\lambda_{\text{avg}}} \sum_{i=1}^d \hat{m}_{ki} c_{ki}(\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)).
\end{aligned}$$

We can now prove our first main result.

Proof of Theorem F.1. Fix any k, and $\boldsymbol{x}_k, \boldsymbol{y}_k \in \mathcal{X}_k$. Then,

$$\begin{aligned}
\left| \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{P}_k} \left[c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k) \right] - \boldsymbol{p}_k^\top \boldsymbol{c}_k(\boldsymbol{y}_k) \right| &= \left| \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{P}_k} \left[c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k) \right] - \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{P}_k} \left[\sum_{i=1}^d \mathbb{I}\left[\boldsymbol{\xi}_k \in B_i' \right] c_k(\boldsymbol{y}_k, \boldsymbol{a}_{ki}) \right] \right| \\
&= \left| \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{P}_k} \left[\sum_{i=1}^d c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k) \mathbb{I}\left[\boldsymbol{\xi}_k \in B_i' \right] \right] - \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{P}_k} \left[\sum_{i=1}^d \mathbb{I}\left[\boldsymbol{\xi}_k \in B_i' \right] c_k(\boldsymbol{y}_k, \boldsymbol{a}_{ki}) \right] \right| \\
&\leq \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{P}_k} \left[\sum_{i=1}^d \mathbb{I}\left[\boldsymbol{\xi}_k \in B_i' \right] \left| c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k) - c_k(\boldsymbol{y}_k, \boldsymbol{a}_{ki}) \right| \right],
\end{aligned}$$

where the first equality uses the definition of p_{ki}, the second equality uses that B_i' form a partition, the last inequality uses the triangle inequality. Now, whenever $\boldsymbol{\xi}_k \in B_i'$,

$$\left| c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k) - c_k(\boldsymbol{y}_k, \boldsymbol{a}_{ki}) \right| \leq \left| c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k) - c_k(\boldsymbol{x}_k, \boldsymbol{a}_{ki}) \right| + \left| c_k(\boldsymbol{x}_k, \boldsymbol{a}_{ki}) - c_k(\boldsymbol{y}_k, \boldsymbol{a}_{ki}) \right| \leq \epsilon + L \|\boldsymbol{x}_k - \boldsymbol{y}_k\|_2.$$

Substituting above shows

$$\left| \mathbb{E}_{\boldsymbol{\xi}_k \sim \mathbb{P}_k} \left[c_k(\boldsymbol{x}_k, \boldsymbol{\xi}_k) \right] - \boldsymbol{p}_k^\top \boldsymbol{c}_k(\boldsymbol{y}_k) \right| \leq \epsilon + L \|\boldsymbol{x}_k - \boldsymbol{y}_k\|_2,$$

by construction of B_i' and the Assumption F.3.

Now for any α, $\mathbb{Q}$, we can instantiate this inequality with $\boldsymbol{x}_k \leftarrow \boldsymbol{x}(\alpha, \mathbb{Q}, \mathcal{S}_k)$ an $\boldsymbol{y}_k \leftarrow \boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k)$ to see that

$$\left|Z_k(\alpha,\mathbb{Q}) - \overline{Z}_k^{\mathsf{disc}}(\alpha,\mathsf{disc}(\mathbb{Q}))\right| \;\le\; \frac{\lambda_k}{\lambda_{\text{avg}}}\left(\epsilon + L\|\boldsymbol{x}(\alpha,\mathbb{Q},\mathcal{S}_k) - \boldsymbol{x}_k^{\mathsf{disc}}(\alpha,\mathsf{disc}(\mathbb{Q}),\hat{\boldsymbol{m}}_k)\|\right) \;\le\; \frac{\lambda_k}{\lambda_{\text{avg}}}\left(\epsilon + L\sqrt{\frac{2\epsilon}{\gamma}}\right),$$

by Lemma F.2. Averaging over k proves

$$\left|\overline{Z}_K(\alpha,\mathbb{Q}) - \overline{Z}_K^{\mathsf{disc}}(\alpha,\mathsf{disc}(\mathbb{Q}))\right| \le L\sqrt{\frac{2\epsilon}{\gamma}} + \epsilon.$$

An entirely analogous argument yields

$$\left|\overline{Z}_K^{\mathsf{LOO}}(\alpha,\mathbb{Q}) - \overline{Z}_K^{\mathsf{LOO,disc}}(\alpha,\mathsf{disc}(\mathbb{Q}))\right| \le \left(L\sqrt{\frac{2\epsilon}{\gamma}} + \epsilon\right)\frac{\hat{N}_{\text{avg}}}{N\lambda_{\text{avg}}}.$$

Notice $K\hat{N}_{\text{avg}} \sim \text{Poisson}(KN\lambda_{\text{avg}})$. From Lemma B.5 Part i) applied to $K\hat{N}_{\text{avg}}$ and Markov's inequality, we have that with probability at least $1-\delta/2$, $\frac{\hat{N}_{\text{avg}}}{N\lambda_{\text{avg}}} \le \log(4/\delta)$.

Now suppose $\frac{4L^2}{C\gamma} \ge 1$. Then, applying Lemma C.3 and Lemma 3.1 to $\overline{Z}_K^{\mathsf{disc}}$ and $\overline{Z}_K^{\mathsf{LOO,disc}}$ with $\boldsymbol{p}_0 \leftarrow \mathsf{disc}(\mathbb{P}_0)$ and $\delta \leftarrow \delta/4$ shows that there exists a universal constant A_1 such that with probability at least $1-\delta/2$,

$$\sup_{\alpha\ge 0}\left|\overline{Z}_K^{\mathsf{disc}}(\alpha,\mathsf{disc}(\mathbb{P}_0)) - \overline{Z}_K^{\mathsf{LOO,disc}}(\alpha,\mathsf{disc}(\mathbb{P}_0))\right| \le \mathrm{A}_1 \cdot L\sqrt{\frac{C}{\gamma}}\cdot\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4}\cdot\frac{\log^2(1/\delta)\cdot\log^{3/2}(K)}{\sqrt{K}},$$

Therefore, with probability at least $1-\delta$,

$$\begin{aligned}
&\sup_{\alpha\ge 0}\left|\overline{Z}_K(\alpha,\mathbb{P}_0) - \overline{Z}_K^{\mathsf{LOO}}(\alpha,\mathbb{P}_0)\right| \\
&\;\le\; \mathrm{A}_1 \cdot L\sqrt{\frac{C}{\gamma}}\cdot\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4}\cdot\frac{\log^2(1/\delta)\cdot\log^{3/2}(K)}{\sqrt{K}} + \left(L\sqrt{\frac{2\epsilon}{\gamma}} + \epsilon\right)(1+\log(4/\delta)), \qquad \text{(EC.F.1)}
\end{aligned}$$

Similar to Lemma 4.1, $\overline{Z}_K^{\mathsf{LOO}}(\alpha_{\mathbb{P}_0}^{\mathsf{S\text{-}SAA}},\mathbb{P}_0) \le \overline{Z}_K^{\mathsf{LOO}}(\alpha_{\mathbb{P}_0}^{\mathsf{OR}},\mathbb{P}_0)$ implies that

$$\begin{aligned}
\mathsf{SubOpt}_{\mathbb{P}_0,K}(\alpha_{\mathbb{P}_0}^{\mathsf{S\text{-}SAA}}) \;&\le\; \overline{Z}_k(\alpha_{\mathbb{P}_0}^{\mathsf{S\text{-}SAA}},\mathbb{P}_0) - \overline{Z}_k^{\mathsf{LOO}}(\alpha_{\mathbb{P}_0}^{\mathsf{S\text{-}SAA}},\mathbb{P}_0) \\
&\quad+\; \overline{Z}_k^{\mathsf{LOO}}(\alpha^{\mathsf{OR}}\mathbb{P}_0,\mathbb{P}_0) - \overline{Z}_k(\alpha_{\mathbb{P}_0}^{\mathsf{OR}},\mathbb{P}_0) \\
&\le\; 2\sup_{\alpha\ge 0}\left|\overline{Z}_K(\alpha,\mathbb{P}_0) - \overline{Z}_K^{\mathsf{LOO}}(\alpha,\mathbb{P}_0)\right|,
\end{aligned}$$

and, hence, $\mathsf{SubOpt}_{\mathbb{P}_0,K}(\alpha_{\mathbb{P}_0}^{\mathsf{S\text{-}SAA}})$ is at most twice Eq. (EC.F.1).

Finally, recall the choice of $\epsilon > 0$ was arbitrary. Thus, taking a limit $\epsilon \to 0$, shows that there exists a constant A_2 such that

$$\mathsf{SubOpt}_{\mathbb{P}_0,K}(\alpha_{\mathbb{P}_0}^{\mathsf{S\text{-}SAA}}) \le \mathrm{A}_2 \cdot L\sqrt{\frac{C}{\gamma}}\cdot\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4}\cdot\frac{\log^2(1/\delta)\cdot\log^{3/2}(K)}{\sqrt{K}}.,$$

In the case that that $\frac{4L^2}{C\gamma} < 1$, we can always increase L until $\frac{4L^2}{C\gamma} = 1$, since the larger L is still a valid Lipschitz constant. Substituting this larger L above and collecting constants proves the theorem. $\square$

The same key idea can also be used to prove analogues of Theorems 4.3 and 4.4. In the case of continuous distributions, we measure the complexity of $\mathcal{P}$ by its packing number with respect to total variation distance. Specifically, let $D_{\mathsf{TV}}(\epsilon, \mathcal{P})$ be the largest number of elements of $\mathcal{P}$ that are each at least ϵ separated in total-variation distance.

THEOREM F.2. **(Shrunken-SAA with Data-Driven Anchors for Strongly-Convex Problems and Continuous Distributions)** *Fix any $h(\cdot)$. Suppose Assumptions 3.1 and F.1 to F.3 hold, $K \geq 2$ and $N\lambda_{\min} \geq 1$. Suppose moreover that there exists d_0 such that for any $0 < \epsilon < 1/2$, $\log D_{\mathsf{TV}}(\epsilon, \mathcal{P}) \leq d_0 \log(1/\epsilon)$. Then, there exists a universal constant* A *such that for any $0 < \delta < 1/2$, with probability at least $1-\delta$, we have that*

$$\mathsf{SubOpt}_{h,K}(\alpha_h^{\mathsf{S\text{-}SAA}}) \;\leq\; \mathrm{A} \cdot \max\left(C,\; \frac{L^2}{\gamma} \;+\; L\sqrt{\frac{C}{\gamma}}\right) \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \frac{d_0^2 \log^{7/2}(K) \log^2(1/\delta)}{\sqrt{K}}.$$

Proof. Fix an $\epsilon > 0$. We apply the same discretization as in the proof of Theorem F.1. Let $\mathcal{P}^{\mathsf{disc}} = \{\mathsf{disc}(\mathbb{Q}) \,:\, \mathbb{Q} \in \mathcal{P}\}$. Since $\|\mathsf{disc}(\mathbb{Q}) - \mathsf{disc}(\mathbb{Q}')\|_1 \leq 2\,\|\mathbb{Q} - \mathbb{Q}'\|_{\mathsf{TV}}$, we have that $\log D_1(\epsilon, \mathcal{P}^{\mathsf{disc}}) \leq 2d_0 \log(1/\epsilon)$. Thus, the assumptions of Theorem 4.3 hold for $\mathcal{P} \leftarrow \mathcal{P}^{\mathsf{disc}}$ and $d_0 \leftarrow 2d_0$, and we can apply Lemma C.5 to bound the maximal deviations in the discrete system.

The remainder of the proof follows the proof of Theorem F.1 closely. Specifically, we bound the difference between the discrete system and the original continuous system, and then bound $\mathsf{SubOpt}_{h,K}(\alpha_h^{\mathsf{S\text{-}SAA}})$ by twice the maximal deviations and take a limit as $\epsilon \to 0$ to yield the result. □

THEOREM F.3. **(Shrunken-SAA with $h_{\mathcal{P}}$ for Strongly-Convex Problems and Continuous Distributions)** *Under the assumptions of Theorem F.2, there exists a universal constant* A *such that for any $0 < \delta < 1/2$, with probability at least $1-\delta$, we have that*

$$\mathsf{SubOpt}_{\mathcal{P},K}(\alpha_{h_{\mathcal{P}}}^{\mathsf{S\text{-}SAA}}, h_{\mathcal{P}}(\hat{\boldsymbol{m}})) \;\leq\; \mathrm{A} \cdot \max\left(C,\; \frac{L^2}{\gamma} \;+\; L\sqrt{\frac{C}{\gamma}}\right) \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{5/4} \frac{d_0^2 \log^{7/2}(K) \log^2(1/\delta)}{\sqrt{K}}.$$

Proof. The proof is the same as Theorem F.2. □

REMARK F.2 (CHALLENGES WITH DISCRETE PROBLEMS). Proving similar extensions for continuous distributions and discrete problems poses some technical challenges . The key issue appears to be establishing an analogue of Lemma F.2. Indeed, without further assumptions, it is not clear that the set of policies $\left\{(\boldsymbol{x}_k(\alpha, \mathbb{Q}, \mathcal{S}_k))_{k=1}^K : \alpha \geq 0,\; \mathbb{Q} \in \mathcal{P}\right\}$ as indexed by α and $\mathbb{Q}$ will be identical to $\left\{(\boldsymbol{x}_k^{\mathsf{disc}}(\alpha, \mathsf{disc}(\mathbb{Q}), \hat{\boldsymbol{m}}_k))_{k=1}^K : \alpha \geq 0,\; \mathbb{Q} \in \mathcal{P}\right\}$, and, it also not clear in what sense these two sets might be "approximately equal" and under what conditions. Thus, we leave establishing the suitable additional assumptions to analyze this case to future work.