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.

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 $\{\hat{\boldsymbol{\xi}}_{k,1},\ldots,\hat{\boldsymbol{\xi}}_{k,\hat{N}_k}\}$ drawn i.i.d. from $\mathbb{P}_k$ independently across k, intuition *still* strongly suggests decoupling is without loss of generality and that data-driven procedures should 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 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

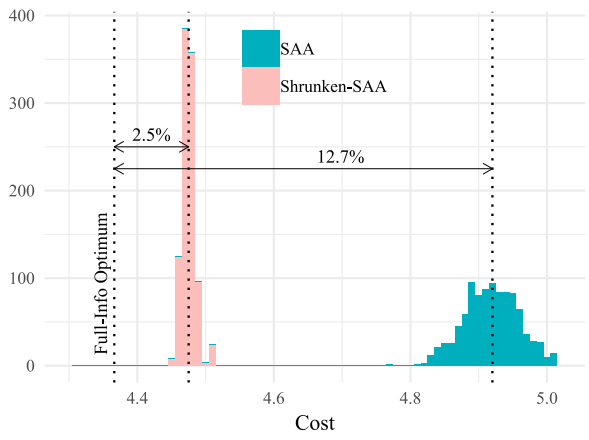


Figure 1 **The Data-Pooling Phenomenon** We consider $K = 10,000$ data-driven newsvendor problems each with critical fractile 90% and 20 data points per subproblem, drawn independently across subproblems. SAA decouples the problems and orders the 90^{th}-sample quantile in each. Shrunken-SAA (cf. Algorithm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

in Section 3, but, loosely speaking Shrunken-SAA proceeds by replacing the k^{th} dataset 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 it reduces by over 80% the loss to full-information optimum 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 only assume that each $\mathbb{P}_k$ has known, finite, discrete support (potentially differing across k). 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 to 4.5). These theoretical results study Problem Eq. (1.2) when the amount of data available for the k^{th} subproblem is, itself, random (see Assumption 3.1 and surrounding discussion), but, numerical experiments suggest this assumption is not crucial.

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) already converge to the full-information optimum in the large-sample regime, the small-data, large-scale regime is in some ways the more interesting regime in which to study the potential 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 data so strongly. Shrunken-SAA works by navigating this tradeoff seeking a "sweet spot" to improve performance. (See Section 5 for a fuller 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. Lemma D.1 and subsequent remark). 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 at the same time 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. Moreover, Shrunken-SAA recovers SAA when no pooling is optimal. For these reasons, and its aforementioned strong theoretical and practical performance, we use SAA throughout as the natural, "apples-to-apples" decoupled benchmark to which to compare our data-pooling procedure.

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 does not require strong 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 essentially any optimization in which traditional 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.

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. 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 $\boldsymbol{\xi}_{kj} \sim \mathbb{P}_k$, $j = 1, \ldots, \hat{N}_k$, denote the $\hat{N}_k$ i.i.d. data draws. Since $\mathbb{P}_k$ is discrete, we summarize these data via counts, $\hat{\boldsymbol{m}}_k = (\hat{m}_{k1}, \ldots, \hat{m}_{kd})$, where $\hat{m}_{ki}$ denotes the number of times that $\boldsymbol{a}_{ki}$ was observed in subproblem k, and $\boldsymbol{e}^\top \hat{\boldsymbol{m}}_k = \hat{N}_k$. More precisely, we have

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

Let $\hat{\boldsymbol{m}} = (\hat{\boldsymbol{m}}_1, \ldots, \hat{\boldsymbol{m}}_K)$ 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. Again, for convenience, we let $\hat{N}_{\max} = \max_k \hat{N}_k$. Finally, let $\hat{\boldsymbol{p}}_k \equiv \hat{\boldsymbol{m}}_k / \hat{N}_k$ denote the empirical distribution of data 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. 3.1 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:

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

Our goal is to identify a data-driven policy, i.e., a function $\boldsymbol{x}(\hat{\boldsymbol{m}}) = (\boldsymbol{x}_1(\hat{\boldsymbol{m}}), \ldots, \boldsymbol{x}_K(\hat{\boldsymbol{m}}))$ that maps the data $\hat{\boldsymbol{m}}$ to $\mathcal{X}_1 \times \cdots \times \mathcal{X}_K$ which has good performance in Problem (2.2), i.e., 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, 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 $\hat{\boldsymbol{p}}_k$:

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

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.

2.1. A Bayesian Perspective of Data-Pooling

To begin to motivate the Shrunken-SAA algorithm, 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 $\hat{\boldsymbol{p}}_k(\alpha)$ is a convex combination between $\hat{\boldsymbol{p}}_k$, 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 completeness in what follows, we also define $\boldsymbol{x}_k(0,\boldsymbol{p}_0,\mathbf{0}) = \lim_{\alpha\to 0}\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\mathbf{0})$, so that $\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\cdot)$ is continuous in α. Observe $\boldsymbol{x}_k(0,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)$ is *not* continuous in $\hat{\boldsymbol{m}}_k$ at $\hat{N}_k = 0$. Furthermore, for $\alpha > 0$, $\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\mathbf{0})$ is the (non-random) solution to the fully shrunken k^{th} subproblem. To emphasize this non-randomness, 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 $\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\mathbf{0}) = \boldsymbol{x}_k(\infty,\boldsymbol{p}_0)$ for all $\alpha > 0$.

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 $\sigma_k^2 = \boldsymbol{p}_k^\top(\boldsymbol{a}_k - \mu_k)^2$.

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

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

where $\hat{\mu}_k = \frac{1}{\hat{N}}\sum_{i=1}^{\hat{N}} \xi_i^k$ 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 as $K \to \infty$,

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K} \left(\sigma_k^2 + (\mu_k - \hat{\mu}_k(\alpha))^2\right) &\to_p \frac{1}{K}\left(\sum_{k=1}^{K} \sigma_k^2 + \mathbb{E}\left[(\mu_k - \hat{\mu}_k(\alpha))^2\right]\right) \\ &= \frac{1}{K}\sum_{k=1}^{K}\left(\sigma_k^2 + \left(\frac{\alpha}{\hat{N}+\alpha}\right)^2 (\mu_k - \mu_{k0})^2 + \left(\frac{\hat{N}}{\hat{N}+\alpha}\right)^2 \frac{\sigma_k^2}{\hat{N}}\right), \qquad (2.5)\end{aligned}$$

again using the bias-variance decomposition of MSE. We can minimize the righthand side of Eq. (2.5) 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, comparing the value of Eq. (2.5) at $\alpha = 0$ and at $\alpha = \alpha_{\boldsymbol{p}_0}^{\mathsf{AP}}$ suggests that for large K, data-pooling can (in principle) decrease the MSE by approximately

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

Notice this value 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, 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

$$\theta \;=\; \frac{1}{\hat{N}} \frac{\frac{1}{K}\sum_{k=1}^{K} \frac{1}{\hat{N}-1}\sum_{i=1}^{\hat{N}} (\hat{\xi}_{ki} - \hat{\mu}_k)^2}{\frac{1}{K}\sum_{k=1}^{K}(\mu_{k0} - \hat{\mu}_k)^2}.$$

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 that 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, $\lambda_k = \lambda_{\text{avg}}$, $\boldsymbol{p}_k$ 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 symmetric holding and back-ordering costs 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 $x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) = \mathbb{I}\left[\hat{p}_{k1} \geq \frac{1}{2} + \frac{\alpha}{\hat{N}_k}(\frac{1}{2} - p_{01})\right]$.[1] Further, for any α (possibly depending on $\hat{\boldsymbol{m}}$),

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

[1] 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.

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.

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 for every data realization, shrinkage by any $\alpha > 0$ cannot improve upon the decoupled SAA solution. Moreover, if shrinking by α performs comparably to the SAA solution, $\boldsymbol{x}(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$ is, itself, a solution to the decoupled problem. In other words, the shrunken solution either performs strictly worse than decoupling, or is, itself, an SAA solution.*

On the other hand, there exist examples where the traditional James-Stein estimator might suggest the benefits of pooling are marginal, but, by data-pooling in way that exploits the optimization structure, we can achieve significant benefits. Indeed, Theorem 2.1 and Efron and Morris (1977) both suggest that data-pooling is most useful when the $\boldsymbol{p}_k$ are clustered near each other, i.e., all close to a single anchor in ℓ_2-norm. When they are dispersed, the benefits of pooling over decoupling might appear marginal (cf. Eq. (2.6)). However, for general optimization problems, this is not always true. Consider the following example.

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.

The James-Stein estimator does not shrink very much in this example. A straightforward computation shows that for K sufficiently large, $\alpha_{\boldsymbol{p}_0}^{\mathsf{JS}} \leq \frac{(1-s)d^2}{s}$ 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.

Algorithm 1 The Shrunken-SAA Algorithm. Input: data $\hat{\boldsymbol{m}}$, anchor $h(\hat{\boldsymbol{m}})$

Fix a grid of $\alpha \in \mathcal{A} \subseteq [0,\infty)$

for all $\alpha \in \mathcal{A}, \quad k=1,\ldots,K, \quad i=1,\ldots,d$ **do**

$\boldsymbol{x}_k(\alpha, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i) \leftarrow \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \ (\hat{\boldsymbol{m}}_k - \boldsymbol{e}_i + \alpha h(\hat{\boldsymbol{m}}))^\top \boldsymbol{c}_k(\boldsymbol{x}_k)$ // Compute LOO solutions

end for

$\alpha_h^{\mathsf{S\text{-}SAA}} \leftarrow \arg\min_{\alpha \in \mathcal{A}} \sum_{i=1}^{\hat{N}_k} \sum_{k=1}^{K} \hat{m}_{ki} c_{ki}(\boldsymbol{x}_k(\alpha, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i))$ // Modified LOO-Cross-Validation

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

$\boldsymbol{x}_k(\alpha^{\mathsf{S\text{-}SAA}}, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k) \leftarrow \arg\min_{\boldsymbol{x}_k \in \mathcal{X}_k} \ (\hat{\boldsymbol{m}}_k + \alpha_h^{\mathsf{S\text{-}SAA}} h(\hat{\boldsymbol{m}}))^\top \boldsymbol{c}_k(\boldsymbol{x}_k)$ // Compute solutions with $\alpha_h^{\mathsf{S\text{-}SAA}}$

end for

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

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.

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 will show that Shrunken-SAA always the best-possible shrinkage in an optimization-aware fashion for many types of problems.

3. Motivating the Shrunken SAA Algorithm

Algorithm 1 formally defines Shrunken-SAA. For clarity, $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ and $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)$ agree with the definition in Eq. (2.4); we have simply scaled by $\hat{N}_k + \alpha$ and $\hat{N}_k - 1 + \alpha$. Consequently, Shrunken-SAA retains the particular pooling structure suggested by our previous Bayesian argument, but allows for an arbitrary user-defined anchor and chooses the amount of pooling via a particular cross-validation scheme. The user-defined anchor may depend on the data. To emphasize this dependence, we denote the anchor by $h(\hat{\boldsymbol{m}}) \in \Delta_d$ in what follows. Two choices we will often use are a fixed (non-data-driven) anchor, i.e., $h(\hat{\boldsymbol{m}}) = \boldsymbol{p}_0$, and the grand-mean of the empirical distributions, i.e., $h(\hat{\boldsymbol{m}}) = \hat{\boldsymbol{p}}^{\mathsf{GM}} \equiv \frac{1}{K}\sum_{k=1}^{K} \hat{\boldsymbol{m}}_k / \hat{N}_k$. Moreover, note that we presented Algorithm 1 as it may be implemented in practice, using a grid of $\alpha \in \mathcal{A}$, but our theory will study the idealized Shrunken-SAA algorithm with $\mathcal{A} = [0,\infty)$.

REMARK 3.1 (COMPUTATIONAL COMPLEXITY OF SHRUNKEN-SAA). The computational bottleneck in Algorithm 1 is computing $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)$. This step is computationally equivalent to solving the k^{th} subproblem by SAA. In this sense, we consider Shrunken-SAA to be *roughly* as tractable as SAA. We say "roughly" because, in the worst-case, one must solve $|\mathcal{A}| \sum_{k=1}^{K} \min(\hat{N}_k, d)$ such problems in the LOO-cross-validation step. Fortunately, we can parallelize these problems distributed computing environments and use previous iterations to "warm-start" solvers.

To motivate the choice of $\alpha_h^{\mathsf{S\text{-}SAA}}$ in Algorithm 1, we first define an "ideal" amount of pooling. From Theorem 2.2, it need not be the case (for $h(\hat{\boldsymbol{m}})$) that data-pooling improves upon decoupling. Hence, to establish an appropriate benchmark, we define the *oracle* pooling amount for $h(\hat{\boldsymbol{m}})$, i.e.,

$$\alpha_h^{\mathsf{OR}} \in \arg\min_{\alpha \geq 0} \overline{Z}_K(\alpha, h(\hat{\boldsymbol{m}})), \quad \text{where} \quad \overline{Z}_K(\alpha, \boldsymbol{q}) = \frac{1}{K} \sum_{k=1}^{K} Z_k(\alpha, \boldsymbol{q}), \tag{3.1}$$
$$Z_k(\alpha, \boldsymbol{q}) = \frac{\lambda_k}{\lambda_{\text{avg}}} \boldsymbol{p}_k{}^{\top} \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k)).$$

Notice α_h^{OR} is random, depending on the entire data-sequence. By construction, $\overline{Z}_K(\alpha_h^{\mathsf{OR}}, h(\hat{\boldsymbol{m}}))$ upper 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_h^{OR} 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 on a particular data-realization by

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

Good pooling procedures will have small sub-optimality with high-probability with respect to the data. Note we allow for the possibility that $\alpha_h^{\mathsf{OR}} = 0$, 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 $\alpha_h^{\mathsf{OR}} > 0$ gives intuition into when and why data-pooling is helpful, a task we take up in Section 5.

In motivating $\alpha^{\mathsf{S\text{-}SAA}}$, it will simplify the exposition considerably to first consider the special case of non-data-driven anchors, i.e. when $h(\hat{\boldsymbol{m}}) = \boldsymbol{p}_0$ is a constant function. This special case is interesting in its own right. It generalizes Theorem 2.1 and might be appropriate in applications where one can identify a canonical measure a priori, e.g., $p_{0i} = 1/d$. In this special case, we abuse notation slightly, replacing the map $\hat{\boldsymbol{m}} \mapsto \boldsymbol{p}_0$ with the constant $\boldsymbol{p}_0$ when it is clear from context. In particular, we define

$$\alpha_{\boldsymbol{p}_0}^{\mathsf{OR}} \in \arg\min_{\alpha \geq 0} \overline{Z}_K(\alpha, \boldsymbol{p}_0). \tag{3.2}$$

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

We first consider the case of a non-data-driven anchor $h(\hat{\boldsymbol{m}}) = \boldsymbol{p}_0$. 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. (3.2) 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. $\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \frac{\lambda_k}{\lambda_{\text{avg}}} \hat{\boldsymbol{p}}_k^\top \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha, \boldsymbol{q}, \hat{\boldsymbol{m}}_k))\right] \neq \mathbb{E}\left[\overline{Z}_K(\alpha, \boldsymbol{p}_0)\right]$, 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 the 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 often termed termed the "optimizer's curse" – in-sample costs are always 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 $\hat{\boldsymbol{m}}_k$. For convenience, define $\hat{N}_{\max} \equiv \max_k \hat{N}_k$.

From an analytical point of view, the benefit of Assumption 3.1 is that it allows us to employ a Poisson-splitting argument to break the dependence across i in $\hat{\boldsymbol{m}}_k$. More specifically, by the poisson-splitting property, under Assumption 3.1,

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

and, furthermore, the $\hat{m}_{ki}$ are independent across i and k.

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 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,

$$W \sim \text{Poisson}(\lambda) \implies \lambda \mathbb{E}[f(W+1)] = \mathbb{E}[W f(W)]. \tag{3.3}$$

The proof of the identity is immediate from the Poisson probability mass function.[2]

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

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

We then have

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, $\mathbb{E}\left[\overline{Z}_K^{\mathsf{LOO}}(\alpha, \boldsymbol{q})\right] = \mathbb{E}\left[\overline{Z}_K(\alpha, \boldsymbol{q})\right]$.*

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 $i = 1, \dots, d$. Let $\hat{m}_{k,-i}$ denote $(\hat{m}_{k,j})_{j \neq i}$. Then, by Eq. (3.3),

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

Taking expectations of both sides, summing over $i = 1, \dots, d$ 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. □

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

We 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. Thus, we let

$$\alpha_{\boldsymbol{p}_0}^{\mathsf{S\text{-}SAA}} \in \arg\min_{\alpha \geq 0} \sum_{k=1}^{K} \sum_{i=1}^{d} \hat{m}_{ki} c_{ki}(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)). \tag{3.4}$$

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

Although we motivated Eq. (3.4) 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

$$\mathbb{E}\left[C\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K, \kappa, \boldsymbol{\xi}\right)\right].$$

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 $\mathbb{E}\left[C\left(\boldsymbol{x}_1(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}), \ldots, \boldsymbol{x}_K(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}), \kappa, \boldsymbol{\xi}\right)\right]$ 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. (As a computational matter, only the policy corresponding to the left-out realization of κ needs to be trained.) We repeat this procedure for each point in the grand data set and then 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. (3.4) 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).

In the case of a data-driven anchor $h(\hat{\boldsymbol{m}})$, however, we stress that unlike traditional leave-one-out validation, we do *not* use one fewer points when computing the anchor point in Algorithm 1; we use $h(\hat{\boldsymbol{m}})$ for all iterations. In this sense, 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, i.e.,

$$\mathsf{SubOpt}_K(\alpha_h^{\mathsf{S\text{-}SAA}}, h) \to 0 \quad \text{as } K \to \infty, \qquad \text{almost surely,} \tag{4.1}$$

even if the expected amount of data per subproblem remains fixed. More precisely, we prove a stronger result for finite K, i.e., for any $K \geq 2$ and any $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\mathsf{SubOpt}_K(\alpha_h^{\mathsf{S\text{-}SAA}}, h) \leq \tilde{\mathcal{O}}\left(\frac{\log^{\beta}(2/\delta)}{\sqrt{K}}\right), \tag{4.2}$$

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. (See Theorems 4.2 to 4.5 below for precise statements.) High-probability, finite K results like Eq. (4.2) are stronger than asymptotic results like Eq. (4.1) in the sense that once the precise asymptotic framework and probability spaces are defined, proving Eq. (4.1) from Eq. (4.2) is straightforward using an invocation of the Borel-Cantelli lemma.

We will prove separate results below for the case of fixed, non-data-driven anchors ($h(\hat{\boldsymbol{m}}) = \boldsymbol{p}_0$) and data-driven anchors. Importantly, our bounds for data-driven anchors will hold for *any* measurable function $h(\hat{\boldsymbol{m}})$. Thus, they apply to several interesting variants of Shrunken-SAA beyond shrinking to the grand mean ($h(\hat{\boldsymbol{m}}) = \hat{\boldsymbol{p}}^{\mathsf{GM}}$). For example, in cases where one has domain knowledge suggesting a particular parametric model (normal, lognormal, etc.), one can fit that parametric model to the data using *any* statistical technique and set $h(\hat{\boldsymbol{m}})$ to the fitted value. Our bounds will still apply. In fact, one could in principle *simultaneously* optimize over α and $\boldsymbol{p}_0$ in Eq. (3.2), even approximately, and set $h(\hat{\boldsymbol{m}})$ to the optimizer since this procedure is still a function of the data. In this sense, our performance guarantees are fairly general purpose.

4.1. Overview of Proof Technique

To prove performance guarantees like Eq. (4.2), we first bound the sub-optimality of $\alpha_h^{\mathsf{S\text{-}SAA}}$ by bounding 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)**. *For a non-data-driven anchor $h(\hat{\boldsymbol{m}}) = \boldsymbol{p}_0$,*

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

Similarly, for a general data-driven anchor,

$$\mathsf{SubOpt}_K(\alpha^{\mathsf{S\text{-}SAA}}_h, h) \le 2 \underbrace{\sup_{\alpha\ge 0,\ \boldsymbol{q}\in\Delta_d} \left|\overline{Z}_K(\alpha, \boldsymbol{q}) - \mathbb{E}\left[\overline{Z}_K(\alpha, \boldsymbol{q})\right]\right|}_{\textit{Maximal Stochastic Deviation in } \overline{Z}_K(\cdot,\cdot)} + 2 \underbrace{\sup_{\alpha\ge 0,\ \boldsymbol{q}\in\Delta_d} \left|\overline{Z}^{\mathsf{LOO}}_K(\alpha, \boldsymbol{q}) - \mathbb{E}\left[\overline{Z}^{\mathsf{LOO}}_K(\alpha, \boldsymbol{q})\right]\right|}_{\textit{Maximal Stochastic Deviation in } \overline{Z}^{\mathsf{LOO}}_K(\cdot,\cdot)}.$$

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

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

By Lemma 3.1, the last term is zero. The proof of the second statement is nearly identical, but in the second inequality, we take an additional supremum over $\boldsymbol{q} \in \Delta^d$ in place of $h(\hat{\boldsymbol{m}})$. □

Proving a performance guarantee for $\alpha^{\mathsf{S\text{-}SAA}}_{h,\boldsymbol{p}_0}$ 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 sample (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 $D(S, \epsilon)$, 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 $|W_k(t)| \le F_k$ 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 $p \geq 1$,[3]

$$\mathbb{E}\left[\sup_{t\in\mathcal{T}}\left|\overline{W}_K(t)-\mathbb{E}[\overline{W}_K(t)]\right|^p\right] \;\leq\; 5\,(2p)^{p/2}\,e^{-p/2}\mathbb{E}[J^p]K^{-p}.$$

In particular, by Markov's Inequality, for any $0<\delta<1$*, with probability at least* $1-\delta$*,*

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

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* in the empirical process literature. Note that if $\mathcal{T}$ is finite, a simple union bound bounds the maximal deviation. Theorem 4.1 extends beyond this simple case by characterizing the complexity of $\mathbf{W}(t)$ over $t\in\mathcal{T}$. Namely, while packing numbers describe the size of a set at scale ϵ, the Dudley integral roughly describes "the complexity" 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:

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

To apply Theorem 4.1 we need to, first, identify envelopes for each process and, second, compute the Dudley integral 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* $k=1\ldots,K$*,* $\sup_{\boldsymbol{q}\in\Delta^d}|c_{ki}(\boldsymbol{x}_k(\infty,\boldsymbol{q}))|\leq C$.

Notice that $\sup_{\alpha\geq 0,\ \boldsymbol{q}\in\Delta_d}|c_{ki}(\boldsymbol{x}_k(\alpha,\boldsymbol{q}))|=\sup_{\boldsymbol{q}\in\Delta_d}|c_k(\boldsymbol{x}_k(\infty,\boldsymbol{q})|$, 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 a valid envelope for* $\mathbf{Z}(\alpha,\boldsymbol{q})$ *with*

$$\|\mathbf{F}^{\mathsf{Perf}}\|_2 \;=\; \frac{C}{\lambda_{\text{avg}}}\|\boldsymbol{\lambda}\|_2 \;\leq\; \frac{C\lambda_{\max}}{\lambda_{\min}}\sqrt{K}.$$

2. *The random vector* $\mathbf{F}^{\mathsf{LOO}}$ *such that* $F_k^{\mathsf{LOO}}=C\frac{\hat{N}_k}{N\lambda_{\text{avg}}}$ *is a valid envelope for* $\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q})$ *with*

$$\|\mathbf{F}^{\mathsf{LOO}}\|_2 \;\leq\; \frac{C}{N\lambda_{\text{avg}}}\|\hat{\boldsymbol{N}}\|_2 \;\leq\; \frac{C}{N\lambda_{\min}}\hat{N}_{\max}\sqrt{K}.$$

[3] Strictly speaking, eq. (7.5) of Pollard (1990) shows that $\mathbb{E}\left[\left|\sup_{t\in\mathcal{T}}\left|\overline{W}_K(t)-\mathbb{E}[\overline{W}_K(t)]\right|\right|^p\right]\leq 2^pC_p^p\mathbb{E}\left[J^p\right]K^{-p}$, for some constant C_p that relates the ℓ_p norm of a random variable and a particular Orlicz norm. In Lemma B.1, we prove that it suffices to take $C_p=5^{1/p}\left(\frac{p}{2e}\right)^{1/2}$.

We next seek to bound the packing number (and Dudley integrals) for the sets

$$\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\} \subseteq \mathbb{R}^K, \qquad \left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\right\} \subseteq \mathbb{R}^K,$$

for the case of fixed anchors and the sets

$$\{\mathbf{Z}(\alpha, \boldsymbol{q}) : \alpha \geq 0,\ \boldsymbol{q} \in \Delta_d\} \subseteq \mathbb{R}^K, \qquad \left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{q}) : \alpha \geq 0,\ \boldsymbol{q} \in \Delta_d\right\} \subseteq \mathbb{R}^K,$$

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). In the remainder of the section we focus on two general classes of optimization problems – smooth, convex optimization problems and discrete optimization problems. Although we focus on these classes, we expect a similar proof strategy and technique might be employed to attack other classes of optimization problems.

In summary, in the next sections we prove performance guarantees for the above two classes of optimization problems by the following strategy: 1) Compute the packing numbers and Dudley integrals for relevant sets above 2) Apply Theorem 4.1 to bound the relevant maximal deviations and 3) Use these bounds in Lemma 4.1 to bound the sub-optimality.

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 techniques very similar to those above, i.e., reducing the problem to bounding an appropriate 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 many of the results below and somewhat standard arguments for empirical processes. Moreover, the result is perhaps unsurprising, as all reasonable 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 a consistent estimator of $\boldsymbol{p}_k$ for all k in this regime. Consequently, we here focus on the small-data, large-scale regime.

4.2. Performance Guarantees for Smooth, Convex Optimization Problems

In this section, we treat the case where the K subproblems are smooth enough so that $\boldsymbol{x}_k(\alpha, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k)$ is smooth in α and $\boldsymbol{p}_0$ for each k. Specifically, in this section we assume

Assumption 4.2 **(Smooth, 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 $k = 1, \ldots, K$, $i = 1, \ldots, d$,.*

We first treat the case of fixed anchors and prove:

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

$$\mathsf{SubOpt}_K(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0}, \boldsymbol{p}_0) \;\leq\; \mathrm{A} \cdot \max\left(L\sqrt{\frac{C}{\gamma}}, \frac{C}{4}\right) \cdot N^{1/4} \frac{\lambda_{\max}^{5/4}}{\lambda_{\min}} \cdot \frac{\log^{5/4}(K)\log^{7/4}(2/\delta)}{\sqrt{K}}.$$

REMARK 4.2. Note that the above bound does not explicitly depend on d, the size of the support for the $\boldsymbol{p}_k$. This implies that a similar performance guarantee holds for a modified Shrunken-SAA with fixed anchors when the $\mathbb{P}_k$ in Problem (1.2) have continuous support and cost functions are sufficiently smooth. Specifically, one can first discretize each subproblem very finely and then apply Shrunken-SAA to the discretized subproblems. If the functions $\boldsymbol{c}_k(\boldsymbol{x}, \boldsymbol{\xi})$ are uniformly smooth in $\boldsymbol{x}$ over $\boldsymbol{\xi}$, then Theorem 4.2 provides the same bound for arbitrarily small discretizations, so we derive a performance guarantee on the original problem. Algorithmically, Shrunken-SAA remains the same no matter how fine the discretization, so can be applied when $\mathbb{P}_k$ are continuous. We demonstrate this in Section 6.5.

To prove the theorem, we follow the approach outlined in Section 4.1 and first seek to bound the ϵ-packing numbers of $\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}$ and $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\right\}$. A key observation is that since the subproblems are smooth and strongly-convex, the optimal solutions $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$ are also smooth as functions of α and $\boldsymbol{p}_0$ for each k. Specifically,

LEMMA 4.3 **(Continuity properties of $\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)$)**. *Under the assumptions of Theorem 4.2,*

i) (Continuity in α) For any $0 \leq \alpha_0 \leq \alpha$ and $\hat{\boldsymbol{m}}_k$ such that $\max(\alpha, \hat{N}_k) > 0$, we have

$$\|\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) - \boldsymbol{x}_k(\alpha_0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k)\|_2 \;\leq\; \sqrt{\frac{4C}{\gamma}}\sqrt{\frac{\hat{N}_k}{(\hat{N}_k+\alpha_0)^2}}\sqrt{\alpha-\alpha_0} \;\leq\; \sqrt{\frac{4C}{\gamma}}\sqrt{\alpha-\alpha_0}\;.$$

ii) (Limit as $\alpha \to \infty$) For any $0 \leq \alpha_0$ and $\hat{\boldsymbol{m}}_k$ such that $\max(\alpha_0, \hat{N}_k) > 0$, we have

$$\|\boldsymbol{x}_k(\alpha_0, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k) - \boldsymbol{x}_k(\infty, \boldsymbol{p}_0)\|_2 \;\leq\; \sqrt{\frac{4C}{\gamma}}\sqrt{\frac{\hat{N}_k}{\hat{N}_k+\alpha_0}}.$$

iii) (Continuity in anchor) For any $\alpha \geq 0$, and any $\boldsymbol{p}, \overline{\boldsymbol{p}} \in \Delta^d$,

$$\|\boldsymbol{x}_k(\alpha, \boldsymbol{p}, \hat{\boldsymbol{m}}_k) - \boldsymbol{x}_k(\alpha, \overline{\boldsymbol{p}}, \hat{\boldsymbol{m}}_k)\|_2 \;\leq\; \sqrt{\frac{2C}{\gamma}}\sqrt{\frac{\alpha}{\hat{N}_k+\alpha}}\sqrt{\|\boldsymbol{p}-\overline{\boldsymbol{p}}\|_1}.$$

Using this continuity, we can now bound the requisite packing numbers. First consider $\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}$. Continuity 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 the set $\left\{(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))_{k=1}^K : \alpha \geq 0\right\}$, which in turn yields a covering of the set $\{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}$. By carefully choosing the initial grid of α's, we can ensure

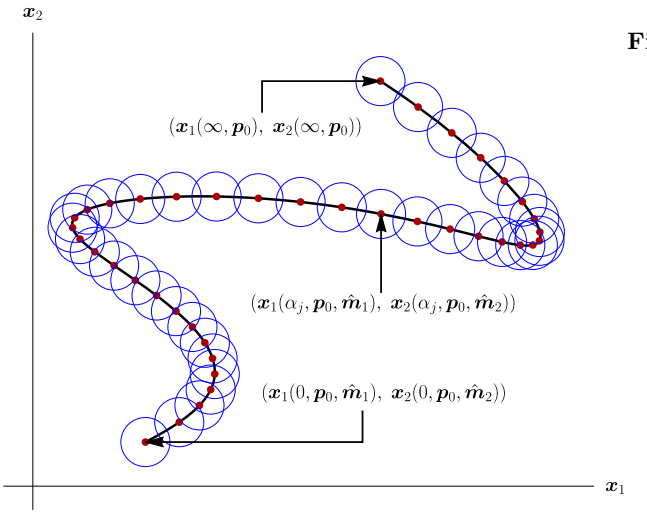


Figure 2 **Covering a continuous process.** The set $\{(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k))_{k=1}^K : \alpha \geq 0\}$ can be thought of as a parametric curve indexed by α in the space $\prod_{k=1}^K \mathcal{X}_k$. Because of the square-root continuity in α (Lemma 4.3i)), to cover this curve for any compact set $\alpha \in [0, \alpha_{\max}]$ requires $\mathcal{O}(1/\epsilon^2)$ balls of size $\epsilon/2$. Because of the continuity at $\alpha = \infty$ (Lemma 4.3ii)), it suffices to take $\alpha_{\max} = \mathcal{O}(1/\epsilon^2)$. This yields a packing number bound of $\mathcal{O}(1/\epsilon^4)$ (c.f. Lemma 4.4).

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 $\mathcal{O}(1/\epsilon^4)$. 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 $\{(x_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i) : i = 1, \ldots, d,\ k = 1, \ldots, K) : \alpha \geq 0\}$.

We state the formal result below which is proven in full in the appendix.

LEMMA 4.4 **(Packing Numbers for Smooth, Convex Problems)**. *Under the assumptions of Theorem 4.2, if* $\frac{16L^2}{C\gamma} \geq 1$, *then for any* $0 < \epsilon \leq 1$,

$$D\left(\epsilon \|\mathbf{F}^{\mathsf{Perf}}\|_2,\ \{\mathbf{Z}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}\right) \leq 1 + \hat{N}_{avg} \frac{2^8 L^4}{\gamma^2 C^2 \epsilon^4}, \quad \textit{where} \quad \hat{N}_{avg} = \frac{1}{\|\boldsymbol{\lambda}\|_2^2} \sum_{k=1}^K \lambda_k^2 \hat{N}_k, \tag{4.4}$$

$$D\left(\epsilon \|\mathbf{F}^{\mathsf{LOO}}\|_2,\ \{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\}\right) \leq 1 + \hat{N}_{\max} \frac{2^8 L^4}{\gamma^2 C^2 \epsilon^4}, \quad \textit{where} \quad \hat{N}_{\max} = \max_k \hat{N}_k. \tag{4.5}$$

These packing number bounds lead to a bound on the Dudley integral and, by leveraging Theorem 4.1 and Lemma 4.2, to a bound on the maximal deviations of $\overline{Z}_K(\cdot, \boldsymbol{p}_0), \overline{Z}_K^{\mathsf{LOO}}(\cdot, \boldsymbol{p}_0)$. The details of this are given in Appendix B.2. Finally, combining these results with Lemma 4.1 proves Theorem 4.2 above (cf. Appendix B.2).

We next consider the case of a data-driven anchor $h(\hat{\boldsymbol{m}})$. By covering α and $\boldsymbol{p}_0 \in \Delta_d$ simultaneously, we can extend the above argument to prove:

THEOREM 4.3 **(Shrunken-SAA with Data-Driven Anchors for Smooth, Convex Problems)**. *Suppose Assumptions 4.1 and 4.2 hold. Assume* $K \geq 2$, $N\lambda_{\min} \geq 1$. *Then, there exists a universal constant* A *such that for any* $0 < \delta < 1/2$, *with probability at least* $1 - \delta$, *we have that*

$$\mathsf{SubOpt}_K(\alpha_h^{\mathsf{S\text{-}SAA}}, h) \leq \mathrm{A} \cdot \max\left(\frac{L\sqrt{C}}{\sqrt{\gamma}}, \frac{C}{4}\right) \cdot N^{1/4} \frac{\lambda_{\max}^{5/4}}{\lambda_{\min}} d^{7/4} \cdot \log^{7/4}\left(\frac{1}{\delta}\right) \cdot \frac{\log^3(K)}{\sqrt{K}}.$$

Notice that unlike Theorem 4.2, the bound in Theorem 4.3 depends polynomially on d. This dependence arises because we impose no assumptions on $h(\hat{\boldsymbol{m}}) \in \Delta_d$, and hence must control behavior across the entire d-dimensional simplex. From a purely theoretical point of view, with stronger assumptions on the anchor $h(\hat{\boldsymbol{m}})$, one might be able to remove this dependence. For $h(\hat{\boldsymbol{m}}) = \boldsymbol{p}^{\mathsf{GM}}$, we find Shrunken-SAA performs well numerically even if d is large as seen in Section 6.5. A full proof of Theorem 4.3 can be found in Appendix B.3.

4.3. Performance Guarantees for Fixed Anchors for 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 nonlinear 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 $k = 1, \ldots, K$, there exists $\sigma_k : 2^{\mathcal{X}_k} \to \mathcal{X}_k$ such that*

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

The main result that we will prove in this section follows:

Theorem 4.4 **(Shrunken-SAA with Fixed Anchors for Discrete Problems)**. *Suppose that $|\mathcal{X}_k| < \infty$ for each k and that Assumptions 4.1 and 4.3 hold. Then, there exists a universal constant* A *such that with probability at least $1 - \delta$,*

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

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 $\log(d)$. This mild dependence on d stems from the fact that we have made *no assumptions of continuity* on the functions $\boldsymbol{c}_k(\boldsymbol{x}, \boldsymbol{\xi})$

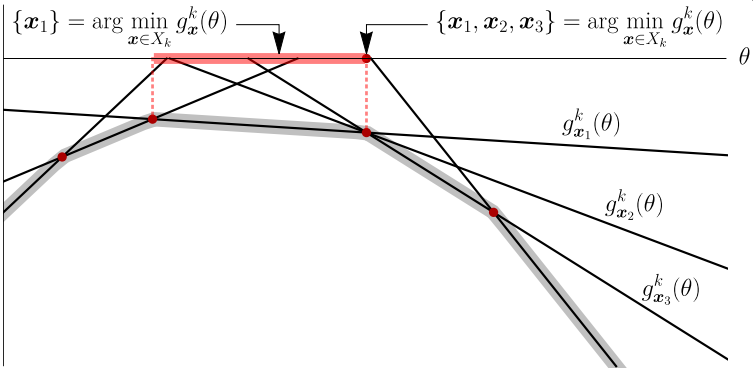


Figure 3 **Counting Discrete Solutions.** A convex piecewise-linear function consisting of $|\mathcal{X}_k|$ lines has at most $|\mathcal{X}_k| - 1$ breakpoints, between which the set of active supporting lines is constant. Any function of this set of active supporting lines is piecewise constant with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

in $\boldsymbol{x}$ or $\boldsymbol{\xi}$. Since these functions could be arbitrarily non-smooth, we need to control their behavior across all i, which introduces a d dependence. However, we argue that this is largely a theoretical issue, not a practical one. In Section 6.5, we show that the empirical performance of Shrunken-SAA for these types of discrete problems is fairly insensitive to the value of d.

To prove Theorem 4.4, 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, $2\sum_{k=1}^K |\mathcal{X}_k|$, 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 re-parameterizing 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

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

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 convex, 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 $\left\{(\boldsymbol{p}^0 - \hat{\boldsymbol{p}}_k)^\top \boldsymbol{c}_k(\boldsymbol{x}) \ : \ \boldsymbol{x} \in \arg\min_{\boldsymbol{x}_k\in\mathcal{X}_k} g^k_{\boldsymbol{x}}(\theta)\right\}$.

Notice that the set of active supporting hyperplanes is constant between breakpoints, so that 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. 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. We illustrate this argument in Fig. 3 and summarize the conclusion as follows:

LEMMA 4.5. *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.*

By taking the union of all these points of discontinuity over $k = 1, \ldots, K$, we get 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 $(\boldsymbol{x}_k(\alpha, \boldsymbol{p}_0, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i))_{k=1,i=1}^{K,d}$ gives a corresponding cardinality bound on $\left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0\right\}\right|$. We summarize this result as follows:

COROLLARY 4.1 **(Bounding Cardinality of Solution Sets for Discrete Problems)**. *Suppose Assumption 4.3 holds. Then,*

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

Although these bounds may appear large, the important feature is that they are only linear in K as long as $\left|\mathcal{X}^k\right|$ 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 4.4. The details are in Appendix B.4.

4.4. Performance Guarantees for Data-Driven Anchors for Discrete Optimization Problems

We next extend the results of Section 4.3 to the case of a data-driven anchor, $h(\hat{\boldsymbol{m}})$. We prove that

THEOREM 4.5 **(Shrunken-SAA with Data-Driven Anchors for Discrete Problems)**. *Suppose that $|\mathcal{X}_k| < \infty$ for each k and that Assumptions 4.1 and 4.3 hold. Then, there exists a universal constant* A *such that with probability at least $1 - \delta$,*

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

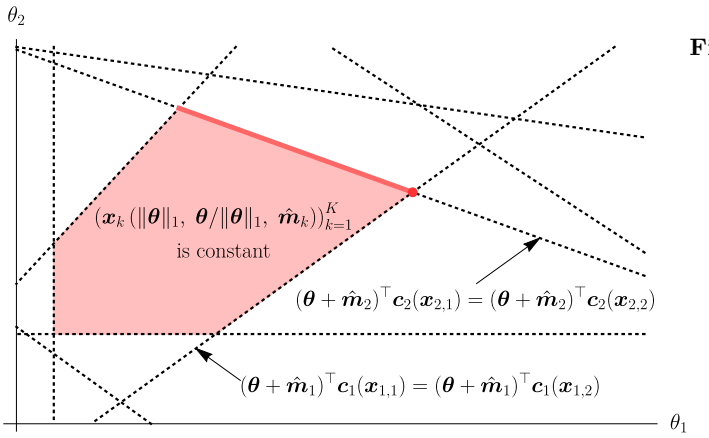


Figure 4 **Solution Induced Hyperplane Arrangement.** The hyperplanes H_{kij} (cf. Eq. (4.8)) in $\mathbb{R}^d$ are indifference curves between solutions i and j in subproblem k. The total ordering on each set $\mathcal{X}_k$ induced by $\boldsymbol{x} \mapsto (\alpha \boldsymbol{q} + \hat{\boldsymbol{m}}_k)^\top \boldsymbol{c}_k(\boldsymbol{x})$ is thus constant on the interior of the fully-specified polyhedra defined by the hyperplanes.

Our strategy is again to bound the cardinality of $\left\{\mathbf{Z}(\alpha,\boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \Delta^d\right\}$, $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \Delta^d\right\}$. The key to obtaining a tight bound is realizing that ranging $d+1$ parameters, $\alpha \geq 0$ and $\boldsymbol{q} \in \Delta^d$, does not achieve all $\prod_{k=1}^K |\mathcal{X}_k|$ possible combinations of solutions.

We first reparameterize our policies. Let $\boldsymbol{\theta} \in \mathbb{R}^d_+$ and define $\alpha(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$ and $q(\boldsymbol{\theta}) = \boldsymbol{\theta}/\|\boldsymbol{\theta}\|_1$ for $\boldsymbol{\theta} \neq \mathbf{0}$ and $q_i(\mathbf{0}) = 1/d$. Notice,

$$\begin{aligned}\left|\left\{\mathbf{Z}(\alpha,\boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \Delta^d\right\}\right| &\leq \left|\left\{(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k))_{k=1}^K \ :\ \boldsymbol{q} \in \Delta_d, \alpha \geq 0\right\}\right| \\ &\leq \left|\left\{(\boldsymbol{x}_k(\alpha(\boldsymbol{\theta}),\boldsymbol{q}(\boldsymbol{\theta}),\hat{\boldsymbol{m}}_k))_{k=1}^K \ :\ \boldsymbol{\theta} \in \mathbb{R}^d_+\right\}\right|. \end{aligned} \tag{4.6}$$

Hence, it suffices to bound the right most sides of Eq. (4.6). An advantage of this $\boldsymbol{\theta}$-parameterization over the original $(\alpha,\boldsymbol{q})$-parameterization is that (by scaling)

$$\boldsymbol{x}_k(\alpha(\boldsymbol{\theta}),\boldsymbol{q}(\boldsymbol{\theta}),\hat{\boldsymbol{m}}_k) \in \arg\min_{\boldsymbol{x}\in\mathcal{X}_k} (\boldsymbol{\theta}+\hat{\boldsymbol{m}}_k)^\top \boldsymbol{c}_k(\boldsymbol{x}), \tag{4.7}$$

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

Next index the set $\mathcal{X}_k = \left\{\boldsymbol{x}_{k,1},\ldots,\boldsymbol{x}_{k,|\mathcal{X}^k|}\right\}$ in some order and define the hyperplanes

$$H_{kij} = \left\{\boldsymbol{\theta} \in \mathbb{R}^d : (\boldsymbol{\theta}+\hat{\boldsymbol{m}}_k)^\top (\boldsymbol{c}_k(\boldsymbol{x}_{ki}) - \boldsymbol{c}_k(\boldsymbol{x}_{kj})) = \mathbf{0}\right\}, \quad \forall\, k = 1,\ldots,K,\ i \neq j = 1,\ldots,\left|\mathcal{X}^k\right|. \tag{4.8}$$

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. (4.7). On either side, we strictly prefer one solution. These hyperplanes induce the hyperplane arrangement seen in Fig. 4, consisting of $m \equiv \sum_{k=1}^K \binom{|\mathcal{X}_k|}{2}$ hyperplanes in $\mathbb{R}^d$.

Now fix any $\boldsymbol{\theta} \in \mathbb{R}^d$ and consider 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 in the arrangement. 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 $\mathbb{R}^d$, it necessarily has dimension $j \leq d$. 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 $j = 0$.

The key to bounding Eq. (4.6) is recognizing that under Assumption 4.3, $(\boldsymbol{x}_k(\alpha(\boldsymbol{\theta}),\boldsymbol{q}(\boldsymbol{\theta}),\hat{\boldsymbol{m}}_k))_{k=1}^K$ is constant on the relative interior of any j-dimensional fully-specified polyhedron with $j \geq 1$. This is because for any $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ both in the relative interior of the same fully-specified polyhedron, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are on the same "side" of all the hyperplanes, and hence induce the same solutions to Eq. (4.7). When $j=0$, the relevant fully-specified polyhedron is simply a point, and hence, solutions are still constant over the polyhedron. Since the fully-specified polyhedra with dimensions $0 \leq j \leq d$ fully partition $\mathbb{R}^d$, it follows that Eq. (4.6) is at most the number of such polyhedra in the arrangement. Notice this argument generalizes our previous argument from counting breakpoints in a univariate piecewise affine function to counting the pieces in a multivariate piecewise affine function.

Appendix B.5 provides a geometric argument to count such polyhedra and bound the cardinality and packing number. A similar argument (with a different hyperplane arrangement) can be used to bound the cardinality of $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) : \alpha \geq 0, \boldsymbol{q} \in \Delta^d\right\}$. Equipped with both, we follow our usual strategy via Theorem 4.1 and Lemma 4.1 to prove Theorem 4.5. The details are in Appendix B.5.

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, 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 4.4 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

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

where

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

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 can 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 the original data $\hat{\boldsymbol{m}}_k$. 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 for the above approximations to hold. 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 4.5 show that the 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 ensure 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.

At a high-level, the Sub-Optimality-Instability Tradeoff resembles the classical bias-variance tradeoff for MSE. Loosely speaking, 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 (c.f. Example 2.1), however, the two tradeoffs still differ and are two different ways to split the "whole" into "pieces." We discuss this in detail in Appendix D.

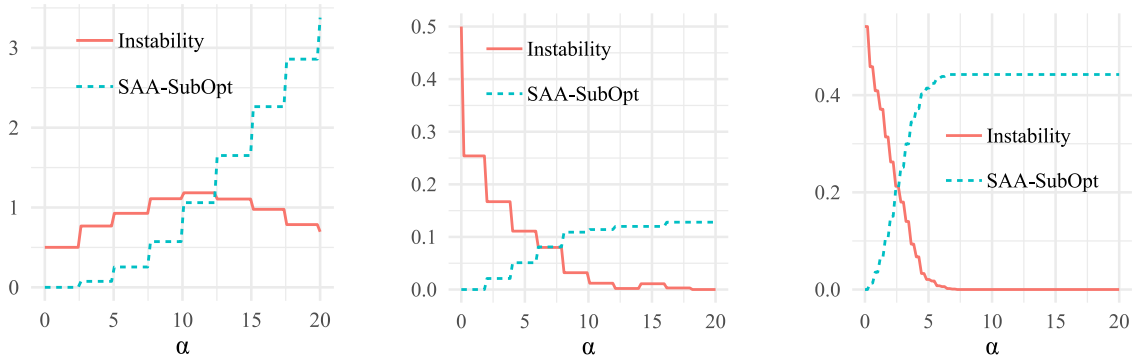
5.1. Sub-Optimality-Instability Tradeoff as a Diagnostic Tool

Our comments above are qualitative, focusing on developing intuition. However, the Sub-Optimality-Instability Tradeoff also provides a quantitative diagnostic tool for studying the effects of 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 often sheds insight into how data-pooling improves (or does not improve) solution quality.

We illustrate this idea using a simple optimization problem, i.e., Example 2.2, to facilitate comparison to analytical results:

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, but ultimately decreases as α



(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 $K = 1000$ 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.

becomes sufficiently large. Because of this initial increasing behavior, the "gains" in stability never outweigh the costs of sub-optimality, and hence decoupling is best. Indeed, as seen in the first panel of Fig. EC.1 in the appendix, $\alpha^{\text{S-SAA}}_{\boldsymbol{p}_0} = \alpha^{\text{OR}}_{\boldsymbol{p}_0} = 0.0$ for this example.

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. EC.1 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. Indeed, the second panel Fig. EC.1 shows $\alpha^{\text{S-SAA}}_{\boldsymbol{p}_0} \approx 12.12$ and $\alpha^{\text{OR}}_{\boldsymbol{p}_0} \approx 16.16$ for this example.

Finally, as mentioned previously, the potential benefits of data-pooling also depends on the problem structure, and the Sub-Optimality-Instability tradeoff again allows us to study this dependence. Consider letting $p_{01} = .3$ again, but now change the objective to a newsvendor cost function 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 also shows $\alpha^{\text{S-SAA}}_{\boldsymbol{p}_0} \approx 2.02$ and $\alpha^{\text{OR}}_{\boldsymbol{p}_0} \approx 2.42$.

In summary, while $\alpha^{\text{S-SAA}}_h$ 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 $\boldsymbol{x}_k(\alpha, h(\hat{\boldsymbol{m}}), \hat{\boldsymbol{m}}_k)$.

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 *BLINDED FOR REVIEW*. We

focus on assessing the degree to which Shrunken-SAA is robust to violations of the assumptions underlying Theorems 4.2 to 4.5. Specifically, we ask how does Shrunken-SAA perform when

- K is small to moderate, and not growing to infinity;
- Assumption 3.1 is violated, i.e., each $\hat{N}_k$ is fixed and non-random;
- d is large / potentially infinite, i.e., the true $\mathbb{P}_k$ do not have finite, discrete support; or
- N grows large.

For simplicity, we take each subproblem to be a newsvendor problem with a critical quantile of $s = 95\%$. 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:

i) **SAA**: the decoupled-benchmark, $\boldsymbol{x}(0, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$,
ii) **JS-Fixed**: a policy inspired by James-Stein estimation and a fixed anchor, $\boldsymbol{x}(\alpha^{\mathsf{JS}}_{\boldsymbol{p}_0}, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$,
iii) **S-SAA-Fixed**: the Shrunken-SAA policy with a fixed anchor, $\boldsymbol{x}(\alpha^{\mathsf{S\text{-}SAA}}_{\boldsymbol{p}_0}, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$
iv) **Oracle-Fixed**: the oracle policy with a fixed anchor, $\boldsymbol{x}(\alpha^{\mathsf{OR}}_{\boldsymbol{p}_0}, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$,
v) **JS-GM**:, a policy inspired by James-Stein estimation and the grand-mean anchor $\boldsymbol{x}(\alpha^{\mathsf{JS}}_{\hat{\boldsymbol{p}}^{\mathsf{GM}}}, \hat{\boldsymbol{p}}^{\mathsf{GM}}, \hat{\boldsymbol{m}})$,
vi) **S-SAA-GM**: the Shrunken-SAA policy with the grand-mean anchor, $\boldsymbol{x}(\alpha^{\mathsf{S\text{-}SAA}}_{\hat{\boldsymbol{p}}^{\mathsf{GM}}}, \hat{\boldsymbol{p}}^{\mathsf{GM}}, \hat{\boldsymbol{m}})$ and
vii) **Oracle-GM**: the oracle policy with the grand-mean anchor, $\boldsymbol{x}(\alpha^{\mathsf{OR}}_{\hat{\boldsymbol{p}}^{\mathsf{GM}}}, \boldsymbol{p}_0, \hat{\boldsymbol{m}})$.

For each policy we take the fixed anchor to be the uniform distribution $\boldsymbol{p}_{0i} = \frac{1}{d}$, and recall that $\hat{\boldsymbol{p}}^{\mathsf{GM}} \equiv \frac{1}{K}\sum_{k=1}^{K} \hat{\boldsymbol{p}}_k$. In each case, the requisite α is computed by exhaustively searching a finite grid. Unless otherwise specified, $d = 20$, and $N_k = 10$. For ease of comparison, we present all results as "% Benefit over SAA," i.e., bigger values are better.

Intuitively, the difference between the JS policies and the SAA policy 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.

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 $d \to \infty$.

6.1. Data Description

Our dataset consists of daily sales at the store level for a European pharmacy chain with locations across 7 countries. For the purpose of our experiments, 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 the many products makes demand censoring an insignificant effect.

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

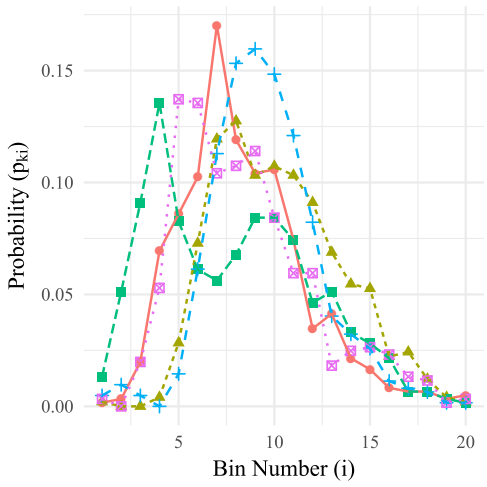
Stores vary in size, available assortment of products, promotional activities and prices, creating significant heterogeneity in demand. The average daily demand ranges from 3,183 to 23,400. The first panel of Fig. EC.2 in Appendix E plots the average daily demand by store. The second panel provides a 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 Section 6.5, we discretize demand by dividing the range of observations into d equally-spaced bins to form the true distributions $\boldsymbol{p}_k$. Figure 6 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. 6. 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.

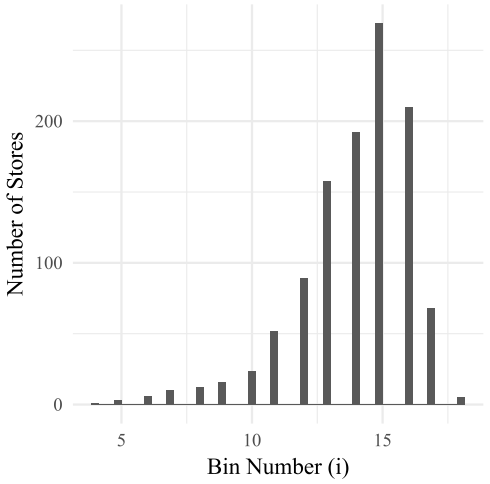
6.2. An Idealized Synthetic Dataset

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. 7 shows the average results.

As suggested by Theorems 4.4 and 4.5, Shrunken-SAA significantly outperforms decoupling even for K as small as 32. 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.

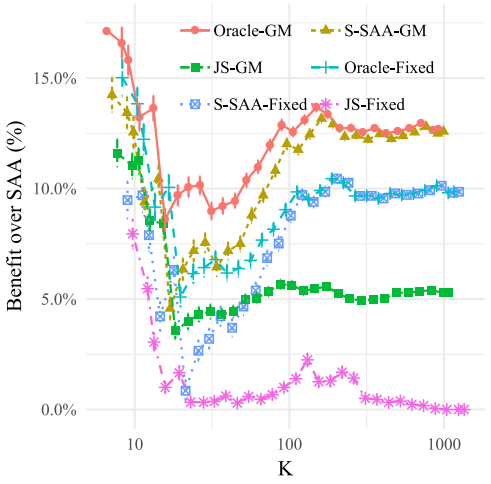


(a) Representative $\boldsymbol{p}_k$

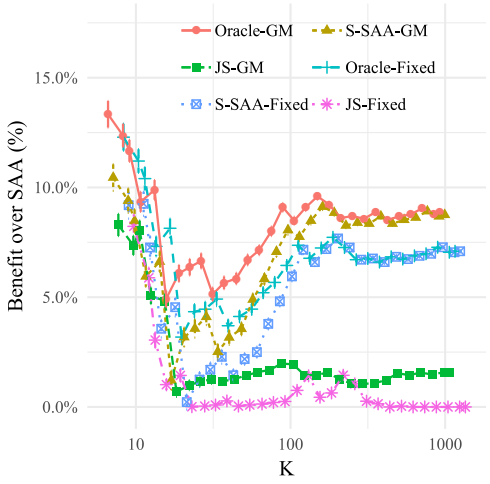


(b) Distribution of Critical Quantile across Bins

Figure 6 **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) $\hat{N}_k \sim \text{Poisson}(N_k)$



(b) $\hat{N}_k = 10$ fixed

Figure 7 **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.

We also illustrate the standard deviation of the performance for each of these methods in Fig. EC.3 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.4 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, our Shrunken-SAA methods use a value of $\alpha \approx 35$, i.e., placing 3.5 times more weight on the grand-mean measure 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. 7. Although the magnitude of the benefit is somewhat reduced, we largely 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. Overall, we attribute these similarities to the fact that when $d = 20$, a Multinomial$(10, \boldsymbol{p})$ random variable is very well-approximated by independent poisson random variables, provided $\boldsymbol{p}$ is not too close to a unit vector. Hence, Assumption 3.1, while not strictly true, is approximately satisfied.

6.4. Historical Backtest

For our remaining tests we consider a less ideal, but more realistic setting for Shrunken-SAA. Specifically, we repeated random subsampling validation 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. We repeat this process 200 times. Note that unlike the previous experiment, it is possible that some of sampled 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 issues with missing data occur for the N_{test} testing points. Thus, missing data poses an additional challenge in this setting. We prefer repeated, random subsampling validation to say, more traditional 5-fold cross-validation, because we would like to be able finely control the number of data points $\hat{N}_k$ used in each subproblem.

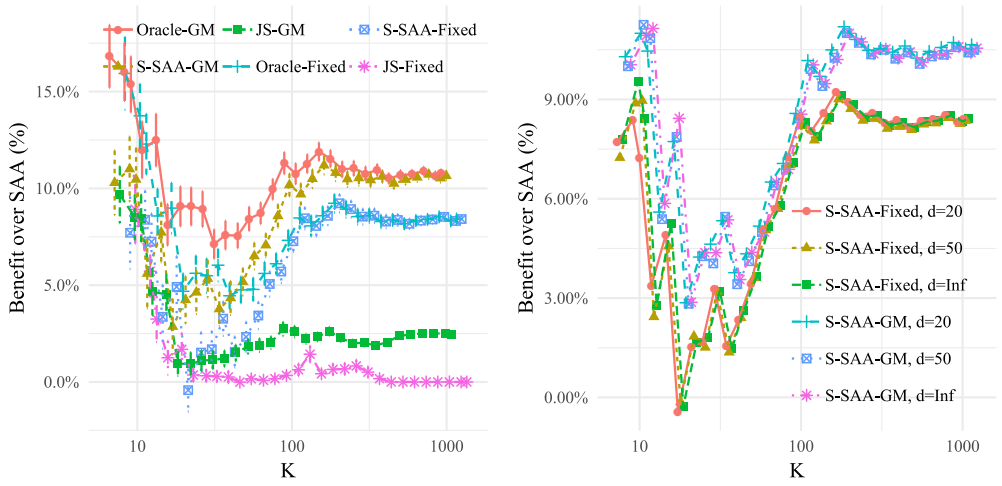


Figure 8 **Robustness to choice of d.** In the first panel, we evaluate our policies on historical data using $d = 20$. Error bars show ± 1 standard error. In the second panel, we limit attention to the Shrunken-SAA policies and compare them on the same historical datasets for $d = 20, 50, \infty$.

6.5. Performance as $d \to \infty$

Recall that the Shrunken-SAA algorithm, itself, only relies on the value d through the distributions $\hat{\boldsymbol{p}}_k$ and their support. Like traditional SAA, however, Shrunken-SAA can be applied to distributions with continuous support by simply using the empirical distributions $\hat{\mathbb{P}}_k$ without any discretization. This amounts to treating each observation as if it were in its own bin and is equivalent to setting $d = \infty$. In this sense, the choice to discretize the data into d bins is a modeling choice more than an algorithmic requirement.

Consequently, we next study the robustness of Shrunken-SAA to this choice of d. As a base case, we first evaluate each of our policies using the our historical backtest set-up for $d = 20$ in the first panel of Fig. 8. Importantly, we see the same qualitative features as in our synthetic data experiment: our Shruken-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.

We next increase d. Figure EC.5 in Appendix E shows results for $d = 50$ and $d = \infty$, i.e., not performing any discretization. The performance is nearly identical to the case of $d = 20$. To make this clearer, in the second panel of Fig. 8 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 focus on finite d is primarily a mathematical convenience to facilitate a simpler proof, but not intrinsic to the algorithm or required for good performance.

6.6. Performance as $N \to \infty$

As a final test, we 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.6 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 $d = \infty$. As $\hat{N}_k$ increases for 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.

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.

7. Conclusion and Future Directions

In this paper, we introduce and study the phenomenon of data-pooling 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 1) when the underlying subproblems are distinct and unrelated, and 2) data for each subproblem are independent. We propose a simple, novel algorithm, Shrunken-SAA, that exploits this phenomenon by pooling data in a particular fashion motivated by the empirical Bayes literature. 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 for all optimization instances. Consequently, 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, and data-pooling more generally, offer significant benefits in the small-data, large-scale regime for a variety of problems.

Finally, Shrunken-SAA is merely one possible algorithm to exploit data-pooling. Others certainly exist. Our Sub-Optimality-Instability tradeoff provides a general intuition for analyzing methods for simultaneously solving many data-driven stochastic optimization problems. It suggests that any algorithm that can be tuned to trade off between in-sample optimality and stability might be adapted to exploit data-pooling. Natural candidates include data-driven distributionally robust procedures and regularization approaches. A formal study of these techniques in a data-pooling context and of the conditions under which they might outperform Shrunken-SAA is an interesting area of future study.

Overall, 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

BLINDED FOR REVIEW. 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.

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. *SSRN: Preprint* URL `https://ssrn.com/abstract=3065655`.

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* .

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.

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

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

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, processes (indexed by α). Such processes are discussed in Theorem 4.1 above, 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 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 in probability.

We will next show that Eq. (EC.A.2) converges to 0 in probability. Recall that for any α

$$\mathbb{E}\left[(\mu_k-\hat{\mu}_k(\alpha))^2\right] = \left(\frac{\alpha}{\hat{N}+\alpha}\right)^2(\mu_k-\mu_{k0})^2+\left(\frac{\hat{N}}{\hat{N}+\alpha}\right)^2\frac{\sigma_k^2}{\hat{N}}.$$

Differentiating the right hand side by α and taking absolute values yields

$$\left|\frac{2\alpha\hat{N}}{(\hat{N}+\alpha)^3}\cdot(\mu_k-\mu_{k0})^2-\frac{2\hat{N}^2}{(\hat{N}+\alpha)^3}\frac{\sigma_k^2}{\hat{N}}\right| \;\le\; 2(\mu_k-\mu_{k0})^2+2\sigma_k^2 \;\le\; 10a_{\max}^2.$$

Hence by the mean-value theorem, $\left|\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}\left[(\mu_k-\hat{\mu}_k(\alpha^{\mathsf{JS}}))^2\right]-\mathbb{E}\left[(\mu_k-\hat{\mu}_k(\alpha^{\mathsf{AP}}))^2\right]\right)\right|$ is bounded above by $10a_{\max}^2\left|\alpha^{\mathsf{JS}}-\alpha^{\mathsf{AP}}\right|$.

We will now complete the proof by showing that $\alpha^{\mathsf{JS}} \to_p \alpha^{\mathsf{AP}}$. We do this 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 $t>0$,

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

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,

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

as $K\to\infty$. Recall $\mathbb{E}\left[(\mu_{k0}-\hat{\mu}_k)^2\right]=\sigma_k^2/\hat{N}+(\mu_{k0}-\mu_k)^2$ 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: 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 3.

B.1. Additional Lemmas

We first prove two additional lemmas that we need in what follows.

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

i) For any t, $t^p \leq p^p e^{-p} e^t$.

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

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

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

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

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

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

LEMMA B.2 **(Bounding Tails of $\hat{N}_{\max}$)**. *Define the constant $N_{\max} = N\lambda_{\max}$, the random variable $\hat{N}_{\max} \equiv \max_k \hat{N}_k$ and assume $N_{\max} \geq 1$ and $K \geq 2$. Let $\beta = \frac{\log\left(1+\frac{\log 2}{N_{\max}}\right)}{1+\log K}$. Then, under Assumption 3.1*

i) $\mathbb{E}[\exp(\beta \hat{N}_{\max})] \leq 6$,

ii) $\beta \geq \frac{1}{6N_{\max}\log K}$,

iii) For any $p > 0$, $\mathbb{E}[\hat{N}_{\max}^p] \leq 6\left(\frac{6p}{e}\right)^p N_{\max}^p \log^p K$.

Proof. We first observe that for $\beta_0 \equiv \log\left(1+\frac{\log 2}{N_{\max}}\right)$, $\mathbb{E}[\exp(\beta_0 \hat{N}_k)] \leq 2$. Indeed, this is immediate from the poisson moment generating function,

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

We now prove the first claim. Note $\beta = \frac{\beta_0}{1+\log K}$. Writing $\exp(\cdot)$ as an integral,

$$\begin{aligned}
\exp(\beta \hat{N}_{\max}) &= e + \int_1^{\beta \hat{N}_{\max}} e^t dt \\
&\leq e + \int_1^{\beta \hat{N}_{\max}} e^{\beta_0 \hat{N}_{\max}} \cdot e^{-t(1+\log K)} \cdot e^t dt \quad (\text{since } t \leq \beta \hat{N}_{\max} \iff 1 \leq e^{\beta_0 \hat{N}_{\max} - t(1+\log K)}) \\
&\leq e + \int_1^{\beta \hat{N}_{\max}} e^{\beta_0 \hat{N}_{\max}} \cdot e^{-t\log K} dt \\
&\leq e + \sum_{k=1}^K \int_1^\infty e^{\beta_0 \hat{N}_k} \cdot e^{-t\log K} dt,
\end{aligned}$$

[4] Namely, for any random variable Y, $\|Y\|_\Psi \equiv \inf\{C > 0 : \mathbb{E}[\Psi(|Y|/C)] \leq 1\}$.

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

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

since $K \ge 2$.

To prove the second claim, observe that $K \ge 2$ implies that $1 + \log K \le 3 \log K$. Furthermore, we claim that $\log(1 + \frac{\log 2}{N_{\max}}) \ge \frac{1}{2N_{\max}}$ for $N_{\max} \ge 1$. Indeed, the function $\log(1 + \frac{\log 2}{N_{\max}}) - \frac{1}{2N_{\max}}$ is positive at $N_{\max} = 1$, and tends to 0 as $N_{\max} \to \infty$. By differentiating, we see it only admits one critical point at $N_{\max} = \frac{\log 2}{2 \log 2 - 1}$, which by inspection is a maximum. This proves that $\log(1 + \frac{\log 2}{N_{\max}}) \ge \frac{1}{2N_{\max}}$ for $N_{\max} \ge 1$. Substituting into the definition of β proves the second claim.

For the third claim, notice from Lemma B.1 that

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

where we have used the second claim to simplify. This concludes the proof. □

B.2. Proof of Theorem 4.2: Shrunken-SAA with Fixed Anchors for Smooth, Convex Problems

We first prove the results summarized in Section 4.2.

B.2.1. Proof of continuity lemma and packing number bounds

Proof of Lemma 4.3. 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 $\boldsymbol{q}, \overline{\boldsymbol{q}} \in \Delta_d$,

$$\|\boldsymbol{x}(\boldsymbol{q}) - \boldsymbol{x}(\overline{\boldsymbol{q}})\|_2 \le \sqrt{\frac{2C}{\gamma}} \cdot \sqrt{\|\boldsymbol{q} - \overline{\boldsymbol{q}}\|_1}. \tag{EC.B.1}$$

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, $\nabla f_{\boldsymbol{q}}(\boldsymbol{x}(\boldsymbol{q}))^\top (\boldsymbol{x}(\overline{\boldsymbol{q}}) - \boldsymbol{x}(\boldsymbol{q})) \ge 0$. Then, from strong-convexity,

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

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

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

Adding yields,

$$\begin{aligned}\gamma\|\boldsymbol{x}(\boldsymbol{q})-\boldsymbol{x}(\overline{\boldsymbol{q}})\|_2^2 &\le \left(f_{\overline{\boldsymbol{q}}}(\boldsymbol{x}(\boldsymbol{q}))-f_{\boldsymbol{q}}(\boldsymbol{x}(\boldsymbol{q}))\right)+\left(f_{\boldsymbol{q}}(\boldsymbol{x}(\overline{\boldsymbol{q}}))-f_{\overline{\boldsymbol{q}}}(\boldsymbol{x}(\overline{\boldsymbol{q}}))\right)\\ &= (\overline{\boldsymbol{q}}-\boldsymbol{q})^\top\left(\boldsymbol{c}_k(\boldsymbol{x}(\boldsymbol{q}))-\boldsymbol{c}_k(\boldsymbol{x}(\overline{\boldsymbol{q}}))\right)\\ &\le 2C\|\boldsymbol{q}-\overline{\boldsymbol{q}}\|_1,\end{aligned}$$

by the Cauchy-Schwarz inequality. Rearranging proves Eq. (EC.B.1).

We can now prove each part of the lemma.

i) Consider taking $\boldsymbol{q}=\hat{\boldsymbol{p}}_k(\alpha)$ and $\overline{\boldsymbol{q}}=\hat{\boldsymbol{p}}_k(\alpha_0)$. Then

$$\begin{aligned}\|\boldsymbol{q}-\overline{\boldsymbol{q}}\|_1 &= \left\|\left(\left(\frac{\alpha}{\hat{N}_k+\alpha}-\frac{\alpha_0}{\hat{N}_k+\alpha_0}\right)\boldsymbol{p}_0+\left(\frac{\hat{N}_k}{\hat{N}_k+\alpha}-\frac{\hat{N}_k}{\hat{N}_k+\alpha_0}\right)\hat{\boldsymbol{p}}_k\right)\right\|_1\\ &\le \left(\left|\frac{\alpha}{\hat{N}_k+\alpha}-\frac{\alpha_0}{\hat{N}_k+\alpha_0}\right|+\left|\frac{\hat{N}_k}{\hat{N}_k+\alpha}-\frac{\hat{N}_k}{\hat{N}_k+\alpha_0}\right|\right),\\ &= 2\left|\frac{\hat{N}_k}{\hat{N}_k+\alpha}-\frac{\hat{N}_k}{\hat{N}_k+\alpha_0}\right|,\end{aligned}$$

since $\left|\frac{\alpha}{\hat{N}_k+\alpha}-\frac{\alpha_0}{\hat{N}_k+\alpha_0}\right|=\left|\frac{\hat{N}_k}{\hat{N}_k+\alpha}-\frac{\hat{N}_k}{\hat{N}_k+\alpha_0}\right|$.

Finally,

$$\left|\frac{\hat{N}_k}{\hat{N}_k+\alpha}-\frac{\hat{N}_k}{\hat{N}_k+\alpha_0}\right| = \frac{(\alpha-\alpha_0)\hat{N}_k}{(\hat{N}_k+\alpha)(\hat{N}_k+\alpha_0)} \le \frac{(\alpha-\alpha_0)\hat{N}_k}{(\hat{N}_k+\alpha_0)^2},$$

because $\alpha_0\le\alpha$. Substituting into Eq. (EC.B.1) proves the first inequality. The second follows because $\alpha,\alpha_0\ge 0$ and $\hat{N}_k\ge 1$.

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

$$\begin{aligned}\|\boldsymbol{q}-\overline{\boldsymbol{q}}\|_1 &= \left\|\left(1-\frac{\alpha_0}{\hat{N}_k+\alpha_0}\right)\boldsymbol{p}_0+\left(0-\frac{\hat{N}_k}{\hat{N}_k+\alpha_0}\right)\hat{\boldsymbol{p}}_k\right\|_1\\ &\le \left|1-\frac{\alpha_0}{\hat{N}_k+\alpha_0}\right|+\left|0-\frac{\hat{N}_k}{\hat{N}_k+\alpha_0}\right|\\ &= 2\frac{\hat{N}_k}{\hat{N}_k+\alpha_0}.\end{aligned}$$

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

iii) Finally take

$$\begin{aligned}\boldsymbol{q} &= \frac{\alpha}{\alpha+\hat{N}_k}\boldsymbol{p}+\frac{\hat{N}_k}{\hat{N}_k+\alpha}\hat{\boldsymbol{p}}_k\\ \overline{\boldsymbol{q}} &= \frac{\alpha}{\alpha+\hat{N}_k}\overline{\boldsymbol{p}}+\frac{\hat{N}_k}{\hat{N}_k+\alpha}\hat{\boldsymbol{p}}_k\end{aligned}$$

Then,

$$\|\boldsymbol{q}-\overline{\boldsymbol{q}}\|_1 \;\le\; \frac{\alpha}{\hat{N}_k+\alpha}\|\boldsymbol{p}-\overline{\boldsymbol{p}}\|_1.$$

Substituting into Eq. (EC.B.1) proves the result. □

Proof of Lemma 4.4. We first prove Eq. (4.4). 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. 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 $\hat{N}_{\text{avg}}=0$, which implies $\hat{N}_k=0$ for all $k=1,\dots,K$. 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 $\{\mathbf{Z}(\alpha,\boldsymbol{p}_0):\alpha\ge 0\}=\{\mathbf{Z}(\infty,\boldsymbol{p}_0)\}$, and the covering number is 1, so the above bound is valid.

Now suppose $\hat{N}_{\text{avg}}>0$. Let $\alpha_{\max}=\hat{N}_{\text{avg}}\left(\frac{16L^2}{C\gamma\epsilon^2}-1\right)$. By assumption on the parameters, $\alpha_{\max}>0$. Then, for any $\alpha\ge\alpha_{\max}$,

$$\begin{aligned}
|Z_k(\alpha,\boldsymbol{p}_0)-Z_k(\infty,\boldsymbol{p}_0)| \;&\le\; \frac{\lambda_k}{\lambda_{\text{avg}}}\sum_{i=1}^d p_{ki}\,|c_{ki}(\boldsymbol{x}_k(\alpha,\hat{\boldsymbol{m}}))-c_{ki}(\boldsymbol{x}_k(\infty))| \\
&\le\; \frac{\lambda_k}{\lambda_{\text{avg}}}\sum_{i=1}^d p_{ki}L\|\boldsymbol{x}_k(\alpha,\hat{\boldsymbol{m}}))-\boldsymbol{x}_k(\infty)\|_2 \qquad \text{(Lipschitz continuity)} \\
&\le\; \frac{\lambda_k}{\lambda_{\text{avg}}}L\sqrt{\frac{4C}{\gamma}}\cdot\sqrt{\frac{\hat{N}_k}{\hat{N}_k+\alpha}} \qquad \text{(Lemma 4.3, part ii) since } \alpha>0).
\end{aligned}$$

It follows that

$$\|\mathbf{Z}(\alpha,\boldsymbol{p}_0)-\mathbf{Z}(\infty,\boldsymbol{p}_0)\|_2^2 \;\le\; \frac{L^2 4C}{\lambda_{\text{avg}}^2\gamma}\sum_{k=1}^K\lambda_k^2\frac{\hat{N}_k}{\hat{N}_k+\alpha} \;\le\; \frac{4CL^2\|\boldsymbol{\lambda}\|_2^2}{\gamma\lambda_{\text{avg}}^2}\frac{\hat{N}_{\text{avg}}}{\hat{N}_{\text{avg}}+\alpha}$$

where the last inequaity follows from Jensen's after noting that $x\mapsto\frac{x}{x+\alpha}$ is concave. Using $\alpha\ge\alpha_{\max}$, we conclude the last line is at most $\frac{\epsilon^2C^2\|\boldsymbol{\lambda}\|_2^2}{\lambda_{\text{avg}}^2}/4$, i.e., $\|\mathbf{Z}(\alpha,\boldsymbol{p}_0)-\mathbf{Z}(\infty,\boldsymbol{p}_0)\|_2\le\frac{\epsilon C\|\boldsymbol{\lambda}\|_2}{2\lambda_{\text{avg}}}=\frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{Perf}}\|$ for all $\alpha\ge\alpha_{\max}$. Thus, to construct our covering, we place one point at $\mathbf{Z}(\infty,\boldsymbol{p}_0)$ to cover all points $\mathbf{Z}(\alpha,\boldsymbol{p}_0)$ with $\alpha\ge\alpha_{\max}$.

Next let $\{\alpha_1,\dots,\alpha_M\}$ be a $\frac{\gamma C\epsilon^2}{16L^2}$ covering of $[0,\alpha_{\max}]$. Note, $M\le\frac{16L^2\alpha_{\max}}{\gamma C\epsilon^2}$. We claim $\{\mathbf{Z}(\alpha,\boldsymbol{p}_0),\dots,\mathbf{Z}(\alpha_M,\boldsymbol{p}_0)\}$ is a $\frac{C\epsilon\|\boldsymbol{\lambda}\|_2}{2\lambda_{\text{avg}}}$-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, and then

$$\begin{aligned}
|Z_k(\alpha,\boldsymbol{p}_0)-Z_k(\alpha_j,\boldsymbol{p}_0)| \;&\le\; \frac{\lambda_k}{\lambda_{\text{avg}}}\sum_{i=1}^d p_{ki}\,|c_{ki}(\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_j))-c_{ki}(\boldsymbol{x}_k(\alpha_j,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k))| \\
&\le\; \frac{\lambda_k}{\lambda_{\text{avg}}}\sum_{i=1}^d p_{ki}L\|\boldsymbol{x}_k(\alpha,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_j)-\boldsymbol{x}_k(\alpha_j,\boldsymbol{p}_0,\hat{\boldsymbol{m}}_k)\|_2 \\
&\le\; \frac{\lambda_k}{\lambda_{\text{avg}}}L\sqrt{\frac{4C}{\gamma}}\sqrt{|\alpha-\alpha_j|} \qquad \text{(Lemma 4.3, part i))}
\end{aligned}$$

$$\leq \frac{\lambda_k}{\lambda_{\text{avg}}} L \sqrt{\frac{4C}{\gamma}} \sqrt{\frac{\gamma C \epsilon^2}{16L^2}}$$
$$= \frac{C\epsilon\lambda_k}{2\lambda_{\text{avg}}}$$

It follows that $\|\mathbf{Z}(\alpha, \boldsymbol{p}_0) - \mathbf{Z}(\alpha_j, \boldsymbol{p}_0)\|_2 \leq \frac{C\epsilon\|\boldsymbol{\lambda}\|_2}{2\lambda_{\text{avg}}}$ as was to be shown.

The total size of the covering is thus

$$1 + M \;\leq\; 1 + \frac{16L^2\alpha_{\max}}{\gamma C\epsilon^2} \;\leq\; 1 + \frac{16L^2}{\gamma C\epsilon^2}\hat{N}_{\text{avg}}\left(\frac{16L^2}{C\gamma\epsilon^2} - 1\right) \;\leq\; 1 + \hat{N}_{\text{avg}}\frac{16^2L^4}{\gamma^2C^2\epsilon^4}\ .$$

We next prove Eq. (4.5). We again proceed by constructing an $\frac{\epsilon\|\mathbf{F}^{\mathsf{LOO}}\|_2}{2}$-covering, since the desired packing is at most the size of this covering. By Lemma 4.2, $\|\mathbf{F}^{\mathsf{LOO}}\|_2^2 = \frac{C^2}{N^2\lambda_{\text{avg}}^2}\sum_{k=1}^K \hat{N}_k^2$.

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

Otherwise, $\hat{N}_{\max} > 0$. Let $\alpha_{\max} = \hat{N}_{\max}\left(\frac{16L^2}{\epsilon^2 C\gamma} - 1\right)$. By assumption on the parameters, $\alpha_{\max} > 0$. Then, for any $\alpha \geq \alpha_{\max}$,

$$\begin{aligned}
\left|Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) - Z_k^{\mathsf{LOO}}(\infty, \boldsymbol{p}_0)\right| &\leq \frac{1}{N\lambda_{\text{avg}}}\sum_{i=1}^d \hat{m}_{ki}\left|c_{ki}(\boldsymbol{x}_k(\alpha, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i)) - c_{ki}(\boldsymbol{x}_k(\infty))\right| \\
&\leq \frac{L}{N\lambda_{\text{avg}}}\sum_{i=1}^d \hat{m}_{ki}\|\boldsymbol{x}_k(\alpha, \hat{\boldsymbol{m}}_k - \boldsymbol{e}_i) - \boldsymbol{x}_k(\infty)\|_2 && \text{(Lipschitz-Continuity)} \\
&\leq \frac{L}{N\lambda_{\text{avg}}}\sum_{i=1}^d \hat{m}_{ki}\sqrt{\frac{4C}{\gamma}}\sqrt{\frac{\hat{N}_k - 1}{\hat{N}_k - 1 + \alpha}} && \text{(Lemma 4.3, part ii))} \\
&\leq L\sqrt{\frac{4C}{\gamma}} \cdot \frac{\hat{N}_k}{N\lambda_{\text{avg}}} \cdot \sqrt{\frac{\hat{N}_k}{\hat{N}_k + \alpha}},
\end{aligned}$$

because $x \mapsto \frac{x}{x+\alpha}$ is an increasing function. Thus,

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

where the last inequality again follows because $x \mapsto x/(x+\alpha)$ is increasing. Since $\alpha \geq \alpha_{\max}$, this last term is at most

$$\frac{4L^2}{C\gamma}\|\mathbf{F}^{\mathsf{LOO}}\|_2^2 \cdot \frac{\epsilon^2 C\gamma}{16L^2} \;=\; \frac{\|\mathbf{F}^{\mathsf{LOO}}\|_2^2\epsilon^2}{4},$$

which implies $\|\mathbf{Z}_k^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)-\mathbf{Z}_k^{\mathsf{LOO}}(\infty,\boldsymbol{p}_0)\|_2 \le \frac{\|\mathbf{F}^{\mathsf{LOO}}\|_2\epsilon}{2}$. Thus, to construct 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 $\alpha \ge \alpha_{\max}$.

Next let $\{\alpha_1,\dots,\alpha_M\}$ be an $\frac{\epsilon^2\gamma C}{16L^2}$-covering of $[0,\alpha_{\max}]$. Note $M \le \frac{16L^2\alpha_{\max}}{\epsilon^2\gamma C}$. 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 $\hat{N}_k>0$,

$$\begin{aligned}
&\left|Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)-Z_k^{\mathsf{LOO}}(\alpha_j,\boldsymbol{p}_0)\right| \\
&\qquad\le \frac{1}{N\lambda_{\text{avg}}}\sum_{i=1}^{d}\hat{m}_{ki}\left|c_{ki}(\boldsymbol{x}_k(\alpha,\hat{\boldsymbol{m}}_{ki}-\boldsymbol{e}_i))-c_{ki}(\boldsymbol{x}_k(\alpha_j,\hat{\boldsymbol{m}}_{ki}-\boldsymbol{e}_i))\right| \\
&\qquad\le \frac{L}{N\lambda_{\text{avg}}}\sum_{i=1}^{d}\hat{m}_{ki}\|\boldsymbol{x}_k(\alpha,\hat{\boldsymbol{m}}_{ki}-\boldsymbol{e}_i))-\boldsymbol{x}_k(\alpha_j,\hat{\boldsymbol{m}}_{ki}-\boldsymbol{e}_i)\|_2 && \text{(Lipschitz Continuity)} \\
&\qquad\le \frac{L\hat{N}_k}{N\lambda_{\text{avg}}}\cdot\sqrt{\frac{4C}{\gamma}}\cdot\sqrt{|\alpha-\alpha_j|} && \text{(Lemma 4.3, part i))} \\
&\qquad\le \frac{L\hat{N}_k}{N\lambda_{\text{avg}}}\cdot\sqrt{\frac{4C}{\gamma}}\cdot\frac{\epsilon}{4L}\sqrt{\gamma C} \\
&\qquad= C\frac{\hat{N}_k}{N\lambda_{\text{avg}}}\frac{\epsilon}{2}.
\end{aligned}$$

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 $\|\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)-\mathbf{Z}^{\mathsf{LOO}}(\alpha_j,\boldsymbol{p}_0)\|_2^2 \le \frac{\epsilon^2}{4}\frac{C^2}{N^2\lambda_{\text{avg}}^2}\sum_{k=1}^{K}\hat{N}_k^2$, which implies $\|\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)-\mathbf{Z}^{\mathsf{LOO}}(\alpha_j,\boldsymbol{p}_0)\| \le \frac{\epsilon}{2}\|\mathbf{F}^{\mathsf{LOO}}\|_2$, as was to be proven.

Thus, the total size of the covering is at most

$$1+M \le 1+\frac{16L^2\alpha_{\max}}{\epsilon^2\gamma C} \;\le\; 1+\frac{16L^2}{\epsilon^2\gamma C}\cdot\hat{N}_{\max}\left(\frac{16L^2}{\epsilon^2 C\gamma}-1\right) \;\le\; 1+\hat{N}_{\max}\frac{16^2L^4}{\epsilon^4\gamma^2C^2}.$$

This completes the proof. □

B.2.2. Maximal deviation bounds and performance guarantee. We next use the above lemmas to bound the maximal deviations of interest via Theorem 4.1.

LEMMA B.3 **(Uniform Convergence of True Performance)**. *Under the assumptions of Theorem 4.2 and* $\frac{16L^2}{C\gamma}\ge 1$*, there exists a universal constant* A *such that with probability at least* $1-\delta$

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

Proof. We first bound the variable J in Eq. (4.3) corresponding to the process $\{\mathbf{Z}(\alpha,\boldsymbol{p}_0):\alpha\ge 0\}$ with the envelope given by Lemma 4.2. Notice, if $\hat{N}_{\text{avg}}=0$, then by Eq. (4.4) the integrand in Eq. (4.3) is zero, and hence $J=0$. Thus, we focus on the case $\hat{N}_{\text{avg}}\ge 1$.

By assumption $\frac{16L^2}{\gamma C} \geq 1$, hence $\hat{N}_{\text{avg}} \frac{16^2 L^4}{\gamma^2 C^2 x^4} \geq 1$ for all $x \in [0,1]$. Thus, using Eq. (4.4),

$$J \;\leq\; 9C \frac{\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}} \int_0^1 \sqrt{\log\left(1 + \hat{N}_{\text{avg}} \frac{16^2 L^4}{\gamma^2 C^2 x^4}\right)} dx \;\leq\; 9C \frac{\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}} \int_0^1 \sqrt{\log\left(\hat{N}_{\text{avg}} \cdot \frac{2 \cdot 16^2 L^4}{\gamma^2 C^2 x^4}\right)} dx.$$

For convenience, let $t = \hat{N}_{\text{avg}} \frac{2 \cdot 16^2 L^4}{\gamma^2 C^2}$. Now make the substitution $u = \sqrt{\log(t/x^4)} \iff x = t^{1/4} e^{-u^2/4}$ in the integrand to yield,

$$\frac{t^{1/4}}{2} \int_{\log t}^{\infty} u^2 e^{\frac{-u^2}{4}} du \;\leq\; \frac{t^{1/4}}{2} \int_{-\infty}^{\infty} u^2 e^{\frac{-u^2}{4}} du \;=\; \frac{t^{1/4}\sqrt{4\pi}}{2} \cdot \frac{1}{\sqrt{4\pi}} \int_{-\infty}^{\infty} u^2 e^{\frac{-u^2}{4}} du \;=\; t^{1/4}\sqrt{4\pi},$$

where we recognize the last integral as the second moment of a mean-zero gaussian with variance 2. Substituting above shows there exists a universal constant A_J such that

$$J \;\;\leq\;\; \mathrm{A}_J \frac{\|\boldsymbol{\lambda}\|_2}{\lambda_{\text{avg}}} \cdot L \sqrt{\frac{C}{\gamma}} \hat{N}_{\text{avg}}^{1/4} \;\;\leq\;\; \mathrm{A}_J \frac{\lambda_{\max}}{\lambda_{\min}} \cdot L \sqrt{\frac{C}{\gamma}} \hat{N}_{\text{avg}}^{1/4} \cdot \sqrt{K}$$

Notice this expression also holds when $\hat{N}_{\text{avg}} = 0$.

We next bound the p^{th} norm of J. Note by assumption $N\lambda_{\min} \geq 1$, which implies $N_{\max} \geq 1$. By Lemma B.2,

$$\mathbb{E}[\hat{N}_{\text{avg}}^{p/4}] \;\leq\; \mathbb{E}[\hat{N}_{\max}^{p/4}] \;\leq\; 6 \left(\frac{6p}{4e}\right)^{p/4} N_{\max}^{p/4} \log^{p/4} K.$$

Hence, there exists a constant A_0 such that

$$\begin{aligned}
\sqrt[p]{\mathbb{E}[J^p]} \;&\leq\; \mathrm{A}_J \frac{\lambda_{\max}}{\lambda_{\min}} \cdot L \sqrt{\frac{C}{\gamma}} \cdot \sqrt{K} \cdot \sqrt[p]{\mathbb{E}[\hat{N}_{\text{avg}]}^{p/4}} \\
&\leq\; \mathrm{A}_0 \frac{\lambda_{\max}}{\lambda_{\min}} \cdot L \sqrt{\frac{C}{\gamma}} \cdot N_{\max}^{1/4} \cdot p^{1/4} 6^{1/p} \cdot \sqrt{K} \log^{1/4} K
\end{aligned}$$

Hence, from Theorem 4.1, there exists a universal constant A_1 such that with probability at least $1 - \delta$,

$$\sup_{\alpha \geq 0} \left| \frac{1}{K} \sum_{k=1}^{K} Z_k(\alpha, \boldsymbol{p}_0) - \mathbb{E}[Z_k(\alpha, \boldsymbol{p}_0)] \right| \;\leq\; \mathrm{A}_1 \left(\frac{6 \cdot 5}{\delta}\right)^{1/p} p^{3/4} \cdot L \sqrt{\frac{C}{\gamma}} \cdot N_{\max}^{1/4} \frac{\lambda_{\max}}{\lambda_{\min}} \cdot \frac{\log^{1/4}(K)}{\sqrt{K}}.$$

This bound is minimized to first order in δ by choosing any $p^* = \mathrm{A}_2 \log(1/\delta)$. Substituting this value, collecting constants, and simplifying shows there exists a universal constant A_3 such that with probability at least $1 - \delta$

$$\sup_{\alpha \geq 0} \left| \frac{1}{K} \sum_{k=1}^{K} Z_k(\alpha, \boldsymbol{p}_0) - \mathbb{E}[Z_k(\alpha, \boldsymbol{p}_0)] \right| \;\leq\; \mathrm{A}_3 \log^{3/4}(1/\delta) \cdot L \sqrt{\frac{C}{\gamma}} \cdot N_{\max}^{1/4} \frac{\lambda_{\max}}{\lambda_{\min}} \cdot \frac{\log^{1/4}(K)}{\sqrt{K}}.$$

This completes the proof after noting that $N_{\max} = N\lambda_{\max}$. $\square$

LEMMA B.4 **(Uniform Convergence of LOO Performance)**. *Under the assumptions of Theorem 4.2 and* $\frac{16L^2}{C\gamma} \geq 1$*, there exists a universal constant* A *such that*

$$\sup_{\alpha \geq 0} \left| \frac{1}{K} \sum_{k=1}^{K} \left(Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) - \mathbb{E}[Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)] \right) \right| \;\leq\; \mathrm{A} \cdot L \sqrt{\frac{C}{\gamma}} \cdot N^{1/4} \frac{\lambda_{\max}^{5/4}}{\lambda_{\min}} \cdot \frac{\log^{5/4}(K)}{\sqrt{K}} \cdot \log^{7/4}\left(\frac{1}{\delta}\right).$$

Proof. The proof follows a similar structure to Lemma B.3. We first bound the variable J in Eq. (4.3) corresponding to the process $\left\{ \mathbf{Z}^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) : \alpha \geq 0 \right\}$ with the envelope given by Lemma 4.2. Notice, if $\hat{N}_{\max} = 0$, then by Eq. (4.5) the integrand in Eq. (4.3) is zero, and hence $J = 0$. Thus, we focus on the case $\hat{N}_{\text{avg}} \geq 1$.

Since $\frac{16L^2}{\gamma C} \geq 1$, we have that $\hat{N}_{\max} \frac{16^2 L^4}{x^4 \gamma^2 C^2} \geq 1$ for all $0 \leq x \leq 1$. Using Eq. (4.5), we then upper bound

$$J \;\leq\; 9\|\mathbf{F}^{\mathsf{LOO}}\|_2 \int_0^1 \sqrt{\log\left(1 + \hat{N}_{\max} \frac{16^2 L^4}{x^4 \gamma^2 C^2}\right)} dx \;\leq\; 9\|\mathbf{F}^{\mathsf{LOO}}\|_2 \int_0^1 \sqrt{\log\left(\hat{N}_{\max} \frac{2 \cdot 16^2 L^4}{x^4 \gamma^2 C^2}\right)} dx.$$

Let $t = \hat{N}_{\max} \frac{2 \cdot 16^2 L^4}{\gamma^2 C^2}$. The same transformation as in Lemma B.3 allows us to upperbound the integral. We conclude that there exists a universal constant A_J such that

$$J(\hat{\boldsymbol{m}}) \leq \mathrm{A}_J \cdot \frac{L}{\sqrt{\gamma C}} \cdot \|\mathbf{F}^{\mathsf{LOO}}\|_2 \hat{N}_{\max}^{1/4}.$$

We next bound $\mathbb{E}[J^p]$. Recall by assumption $N\lambda_{\min} \geq 1$, which implies $N_{\max} \geq 1$. Then,

$$\begin{aligned}
\mathbb{E}\left[J^p\right] &\leq \mathrm{A}_J^p \left(\frac{L}{\sqrt{\gamma C}}\right)^p \frac{C^p K^{p/2}}{N^p \lambda_{\min}^p} \mathbb{E}[\hat{N}_{\max}^{5p/4}] && \text{(Lemma 4.2)} \\
&\leq 6\mathrm{A}_J^p \left(\frac{L}{\sqrt{\gamma C}}\right)^p \frac{C^p K^{p/2}}{N^p \lambda_{\min}^p} \left(\frac{30p}{4e}\right)^{5p/4} N_{\max}^{5p/4} \log^{5p/4}(K) && \text{(Lemma B.2).}
\end{aligned}$$

Using Theorem 4.1 shows that there exists a universal constant A_1 such that with probability at least $1 - \delta$,

$$\sup_{\alpha \geq 0} \left| \frac{1}{K} \sum_{k=1}^{K} Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) - \mathbb{E}[Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)] \right| \leq \mathrm{A}_1 \left(\frac{6 \cdot 5}{\delta}\right)^{1/p} p^{7/4} \cdot L \sqrt{\frac{C}{\gamma}} \cdot \frac{N_{\max}^{5/4}}{N\lambda_{\min}} \frac{\log^{5/4}(K)}{\sqrt{K}}$$

This bound is minimized to first order in δ for by any $p^* = \mathrm{A}_2 \log(1/\delta)$. S ubstituting this value, collecting constants, and simplifying shows there exists a universal constant A_3 such that with probability at least $1 - \delta$

$$\sup_{\alpha \geq 0} \left| \frac{1}{K} \sum_{k=1}^{K} Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0) - \mathbb{E}[Z_k^{\mathsf{LOO}}(\alpha, \boldsymbol{p}_0)] \right| \leq \mathrm{A}_3 L \sqrt{\frac{C}{\gamma}} \cdot \frac{N_{\max}^{5/4}}{N\lambda_{\min}} \cdot \frac{\log^{5/4}(K)}{\sqrt{K}} \log^{7/4}\left(\frac{1}{\delta}\right).$$

Note that $N_{\max} = N\lambda_{\max}$ and simplify to complete the proof. □

We can now prove the main result of the section.

Proof of Theorem 4.2. We first consider the case that $\frac{16L^2}{C\gamma} \geq 1$. Then, Lemmas B.3, B.4 and 4.4 bound the maximal deviations in Lemma 4.1. Instantiate the lemmas with $\delta \to \delta/2$, adding their right hand sides and applying the union bound thus bounds the sub-optimality. Collecting dominant terms yields the bound

$$\mathrm{A} \cdot L\sqrt{\frac{C}{\gamma}} \cdot N^{1/4} \frac{\lambda_{\max}^{5/4}}{\lambda_{\min}} \cdot \frac{\log^{5/4}(K)}{\sqrt{K}} \cdot \log^{7/4}\left(\frac{2}{\delta}\right).$$

where A is some universal constant not depending on problem parameters.

In the case that $\frac{16L^2}{C\gamma} < 1$, we can always increase L to $\sqrt{C\gamma}/4$ and apply the case above with this larger Lipschitz constant. The resulting bound is of the same form except the leading term is now $C/4$. Taking the maximum of these two cases proves the result. $\square$

B.3. Proof of Theorem 4.3: Shrunken-SAA with Data-Driven Anchors for Smooth, Convex Problems

Our strategy to proving Theorem 4.3 is similar to proving to Theorem 4.2 except that our process is now indexed by both $\alpha \geq 0$ and $\boldsymbol{q} \in \Delta^d$. Using Lemma 4.3, part iii), we can reduce bounding the maximal deviations of $\overline{Z}_K(\cdot,\cdot)$, $\overline{Z}_K^{\mathsf{LOO}}(\cdot,\cdot)$ to bounding the maximal deviations of $\overline{Z}_K(\cdot,\boldsymbol{q})$, $\overline{Z}_K^{\mathsf{LOO}}(\cdot,\boldsymbol{q})$ for a finite number of fixed anchors $\boldsymbol{q}$.

LEMMA B.5 **(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 Δ^d with respect to ℓ_1, then*

$$\sup_{\alpha\geq 0,\boldsymbol{q}\in\Delta^d} \left|\overline{Z}(\alpha,\boldsymbol{q}) - \mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q})]\right| \;\leq\; L\sqrt{\epsilon_0}\sqrt{\frac{8C}{\gamma}} + \max_{j=1,\ldots,M} \sup_{\alpha\geq 0} \left|\overline{Z}(\alpha,\boldsymbol{q}^j) - \mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q}^j)]\right|, \tag{EC.B.2}$$

$$\begin{aligned}\sup_{\alpha\geq 0,\boldsymbol{q}\in\Delta^d} \left|\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}) - \mathbb{E}[\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q})]\right| \;\leq\; & L\sqrt{\epsilon_0}\sqrt{\frac{2C}{\gamma}}\left(\frac{\hat{N}_{avg}}{N\lambda_{\mathrm{avg}}}+1\right) \\ & + \max_{j=1,\ldots,M} \sup_{\alpha\geq 0} \left|\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j) - \mathbb{E}[\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)]\right|.\end{aligned} \tag{EC.B.3}$$

Proof. Let $\{\boldsymbol{q}^1,\ldots,\boldsymbol{q}^M\}$ be an ϵ_0 covering of Δ^d with respect to ℓ_1.

We first prove the first inequality. Consider some $\boldsymbol{q}\in\Delta^d$, and suppose $\boldsymbol{q}^j$ is the closest member of the covering. Then,

$$\begin{aligned}\left|\overline{Z}(\alpha,\boldsymbol{q}) - \overline{Z}(\alpha,\boldsymbol{q}^j)\right| &\leq \frac{1}{K}\sum_{k=1}^K \frac{\lambda_k}{\lambda_{\mathrm{avg}}} \left|\boldsymbol{p}_k{}^\top\left(\boldsymbol{c}_k(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k)) - \boldsymbol{c}_k(\boldsymbol{x}_k(\alpha,\boldsymbol{q}^j,\hat{\boldsymbol{m}}_k))\right)\right| \\ &\leq \frac{L}{K}\sum_{k=1}^K \frac{\lambda_k}{\lambda_{\mathrm{avg}}} \left\|\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k) - \boldsymbol{x}_k(\alpha,\boldsymbol{q}^j,\hat{\boldsymbol{m}}_k)\right\|_2 && \text{(Lipschitz Continuity)} \\ &\leq L\sqrt{\frac{2C}{\gamma}}\sqrt{\|\boldsymbol{q}-\boldsymbol{q}^j\|_1}\cdot\frac{1}{K}\sum_{k=1}^K\sqrt{\frac{\alpha}{\hat{N}_k+\alpha}} && \text{(Lemma 4.3, part iii))} \\ &\leq L\sqrt{\frac{2C}{\gamma}}\sqrt{\epsilon_0}.\end{aligned}$$

By Jensen's inequality, this further implies that $\left|\mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q})]-\mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q}^j)]\right| \le \mathbb{E}\left[\left|\overline{Z}(\alpha,\boldsymbol{q})-\overline{Z}(\alpha,\boldsymbol{q}^j)\right|\right] \le L\sqrt{\frac{2C}{\gamma}}\sqrt{\epsilon_0}$. By the triangle inequality,

$$\left|\overline{Z}(\alpha,\boldsymbol{q})-\mathbb{E}\left[\overline{Z}(\alpha,\boldsymbol{q})\right]\right| \le \left|\overline{Z}(\alpha,\boldsymbol{q})-\overline{Z}(\alpha,\boldsymbol{q}^j)\right| + \left|\mathbb{E}\left[\overline{Z}(\alpha,\boldsymbol{q})-\overline{Z}(\alpha,\boldsymbol{q}^j)\right]\right| + \left|\overline{Z}(\alpha,\boldsymbol{q}^j)-\mathbb{E}\left[\overline{Z}(\alpha,\boldsymbol{q}^j)\right]\right|.$$

Applying the above bounds we arrive at the first inequality in the result.

We next prove the second inequality. Now consider some $\boldsymbol{q}\in\Delta^d$, and suppose $\boldsymbol{q}^j$ is the closest member of the covering. Then,

$$\begin{aligned}
&\left|\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q})-\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)\right| \\
&\le \frac{1}{KN\lambda_{\text{avg}}}\sum_{k=1}^{K}\sum_{i=1}^{d}\hat{m}_{ki}\left|c_{ki}(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i))-c_{ki}(\boldsymbol{x}_k(\alpha,\boldsymbol{q}^j,\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i))\right| \\
&\le \frac{L}{KN\lambda_{\text{avg}}}\sum_{k=1}^{K}\sum_{i=1}^{d}\hat{m}_{ki}\left\|\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i)-\boldsymbol{x}_k(\alpha,\boldsymbol{q}^j,\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i)\right\|_2 && \text{(Lipschitz Continuity)} \\
&\le \frac{L\sqrt{\epsilon_0}}{KN\lambda_{\text{avg}}}\sqrt{\frac{2C}{\gamma}}\sum_{k=1}^{K}\hat{N}_k\sqrt{\frac{\alpha}{\hat{N}_k-1+\alpha}}\mathbb{I}\left[\hat{N}_k>0\right] && \text{(Lemma 4.3, part iii))} \\
&\le \frac{L\sqrt{\epsilon_0}}{KN\lambda_{\text{avg}}}\sqrt{\frac{2C}{\gamma}}\sum_{k=1}^{K}\hat{N}_k \\
&= L\sqrt{\epsilon_0}\sqrt{\frac{2C}{\gamma}}\frac{\hat{N}_{\text{avg}}}{N\lambda_{\text{avg}}}
\end{aligned}$$

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

B.3.1. Maximal deviation bounds and performance guarantee. We next use the above lemmas to bound the maximal deviations of interest via Theorem 4.1

LEMMA B.6 **(Bound on Maximal Deviation of True Performance for General Anchors).** *Suppose* $\frac{16L^2}{C\gamma}\ge 1$. *Then, under the assumptions of Theorem 4.3, there exists a constant* A *such that for any* $0<\delta<\frac{1}{2}$, *with probability at least* $1-\delta$,

$$\sup_{\alpha\ge 0,\ \boldsymbol{q}\in\Delta^d}\left|\frac{1}{K}\sum_{k=1}^{K}Z_k(\alpha,\boldsymbol{q})-\mathbb{E}[Z_k(\alpha,\boldsymbol{q})]\right| \ \le\ \mathrm{A}\cdot L\sqrt{\frac{C}{\gamma}}\cdot N^{1/4}\frac{\lambda_{\max}^{5/4}}{\lambda_{\min}}\cdot d^{3/4}\frac{\log(K)}{\sqrt{K}}\cdot\log^{3/4}\left(\frac{1}{\delta}\right).$$

Proof. Let $\{\boldsymbol{q}^1,\ldots,\boldsymbol{q}^M\}$ be an ϵ_0 covering of Δ^d with respect to ℓ_1. Note that $M\le\frac{3^d}{\epsilon_0^d}$ (cf. (Pollard 1990, Lemma 4.1)). Combining a union bound with Lemmas B.3 and B.5 shows that there exists a constant A_2 such that with probability at least $1-\delta$,

$$\sup_{\alpha\ge 0,\boldsymbol{q}\in\Delta^d}\left|\overline{Z}(\alpha,\boldsymbol{q})-\mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q})]\right| \le \mathrm{A}_2 L\sqrt{\frac{C}{\gamma}}\left(\sqrt{\epsilon_0}+N^{1/4}\frac{\lambda_{\max}^{5/4}}{\lambda_{\min}}\cdot\frac{\log^{1/4}(K)}{\sqrt{K}}\log^{3/4}\left(\frac{M}{\delta}\right)\right)$$

$$\leq \mathrm{A}_2 L\sqrt{\frac{C}{\gamma}}\left(\sqrt{\epsilon_0}+N^{1/4}\frac{\lambda_{\max}^{5/4}}{\lambda_{\min}}\cdot\frac{\log^{1/4}(K)}{\sqrt{K}}\left(\log\left(\frac{1}{\delta}\right)+d\log\left(\frac{3}{\epsilon_0}\right)\right)^{3/4}\right),$$

where we have used the aforementioned bound on M.

Directly optimizing the choice of ϵ_0 appears difficult. We instead take (the suboptimal choice) $\epsilon_0=\frac{3}{K}$. Notice then, $d\log(3/\epsilon_0)=d\log K>1$, because $d\geq 2$ and $K\geq 2$. Since $\delta<\frac{1}{2}$, $2\log(1/\delta)>1$. Hence,

$$\log\left(\frac{1}{\delta}\right)+d\log\left(\frac{3}{\epsilon_0}\right) \;\leq\; d\log(K)\log\left(\frac{1}{\delta}\right)+2d\log(K)\log\left(\frac{1}{\delta}\right) \;=\; 3d\log(K)\log\left(\frac{1}{\delta}\right).$$

Substituting above shows there exists a constant A_3 such that with probability at least $1-\delta$,

$$\sup_{\alpha\geq 0,\boldsymbol{q}\in\Delta^d}\left|\overline{Z}(\alpha,\boldsymbol{q})-\mathbb{E}[\overline{Z}(\alpha,\boldsymbol{q})]\right|$$
$$\leq\; \mathrm{A}_3\cdot L\sqrt{\frac{C}{\gamma}}\left(\frac{1}{\sqrt{K}}+N^{1/4}\frac{\lambda_{\max}^{5/4}}{\lambda_{\min}}\cdot d^{3/4}\frac{\log(K)}{\sqrt{K}}\log^{3/4}\left(\frac{1}{\delta}\right)\right)$$

We now "clean-up" the bound. Recall $N\lambda_{\min}\geq 1$ by assumption, which implies $N\lambda_{\max}\geq 1$. Hence, $N^{1/4}\frac{\lambda_{\max}^{5/4}}{\lambda_{\min}}=(N\lambda_{\max})^{1/4}\frac{\lambda_{\max}}{\lambda_{\min}}\geq 1$. Moreover, $2d^{3/4}\log(K)\log^{3/4}(1/\delta)\geq 1$. Hence we can increase to

$$\frac{1}{\sqrt{K}}\to 2N^{1/4}\frac{\lambda_{\max}^{5/4}}{\lambda_{\min}}\cdot d^{3/4}\log(K)\log^{3/4}(1/\delta).$$

Substituting above, simplifying and collecting constants completes the proof. □

LEMMA B.7 **(Bound on Maximal Deviation of LOO Performance for General Anchors)**. *Suppose* $\frac{16L^2}{C\gamma}\geq 1$. *Then, under the assumptions of Theorem 4.3, there exists a constant* A *such that for any* $0<\delta<1$ *with probability at least* $1-\delta$

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

Proof. Let $\{\boldsymbol{q}^1,\ldots,\boldsymbol{q}^M\}$ be an ϵ_0 covering of Δ^d with respect to ℓ_1. Note that $M\leq\frac{3^d}{\epsilon_0^d}$ (cf. (Pollard 1990, Lemma 4.1)). We proceed by applying Lemma B.5 and working term by term in Eq. (EC.B.3). To analyze the first term of Eq. (EC.B.3), we note from Lemma B.2 that for $\beta=\frac{\log\left(1+\frac{\log 2}{N_{\max}}\right)}{1+\log K}$, $\mathbb{E}[\exp(\beta N_{\max})]\leq 6$. Hence, by Markov's Inequality, with probability at least $1-\delta/2$,

$$\hat{N}_{\max} \;\leq\; \frac{\log\left(\frac{12}{\delta}\right)}{\beta} \;\leq\; 6N_{\max}\log\left(\frac{12}{\delta}\right)\log K,$$

using the second part of Lemma B.2 to simplify. This inequality, in turn, implies that with probability at least $1-\delta/2$,

$$L\sqrt{\epsilon_0}\sqrt{\frac{2C}{\gamma}}\frac{\hat{N}_{\text{avg}}}{N\lambda_{\text{avg}}} \;\leq\; 6L\sqrt{\epsilon_0}\sqrt{\frac{2C}{\gamma}}\frac{\lambda_{\max}}{\lambda_{\text{avg}}}\log\left(\frac{12}{\delta}\right)\log K,$$

To analyze the second term in Eq. (EC.B.3), we first combine a union bound with Lemma B.4 to argue that there exists a constant A_2 such that with probability at least $1-\delta/2$,

$$\sup_{\substack{\alpha\geq 0\\ 1\leq j\leq M}} \left|\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)-\mathbb{E}[\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)]\right| \leq \mathrm{A}_2 L\sqrt{\frac{C}{\gamma}}\cdot\frac{\lambda_{\max}}{\lambda_{\min}}N_{\max}^{1/4}\cdot\frac{\log^{5/4}(K)}{\sqrt{K}}\log^{7/4}\left(\frac{2M}{\delta}\right)$$
$$= \mathrm{A}_2 L\sqrt{\frac{C}{\gamma}}\cdot\frac{\lambda_{\max}}{\lambda_{\min}}N_{\max}^{1/4}\cdot\frac{\log^{5/4}(K)}{\sqrt{K}}\left(\log\left(\frac{2}{\delta}\right)+d\log\left(\frac{3}{\epsilon_0}\right)\right)^{7/4}$$

where we have used the aforementioned bound on M.

Again, optimizing ϵ_0 appears difficult so we instead choose $\epsilon_0 = 3/K$. Then, $d\log(3/\epsilon_0) = d\log K > 1$, because $d\geq 2$ and $K\geq 2$. Similarly, $\delta<1$ implies $2\log(2/\delta)>1$. Therefore,

$$\log(2/\delta)+d\log(3/\epsilon_0) \;\leq\; d\log(K)\log(2/\delta)+2d\log(K)\log(2/\delta) \;=\; 3d\log K\log(2/\delta).$$

Substituting above shows there exists a constant A_3 such that with probability at least $1-\delta/2$,

$$\sup_{\substack{\alpha\geq 0\\ 1\leq j\leq M}} \left|\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)-\mathbb{E}[\overline{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}^j)]\right| \;\leq\; \mathrm{A}_3\cdot L\sqrt{\frac{C}{\gamma}}\cdot\frac{\lambda_{\max}}{\lambda_{\min}}N_{\max}^{1/4}\cdot d^{7/4}\cdot\frac{\log^3 K}{\sqrt{K}}\cdot\log^{7/4}\left(\frac{2}{\delta}\right). \quad \text{(EC.B.4)}$$

Moreover, substituting ϵ_0 into our earlier bound on the first term in Eq. (EC.B.3) shows that with probability at least $1-\delta/2$,

$$L\sqrt{\epsilon_0}\sqrt{\frac{2C}{\gamma}}\frac{\hat{N}_{\mathrm{avg}}}{N\lambda_{\mathrm{avg}}} \;\leq\; 6L\sqrt{\frac{2C}{\gamma}}\frac{\lambda_{\max}}{\lambda_{\mathrm{avg}}}\log\left(\frac{12}{\delta}\right)\frac{\log K}{\sqrt{K}},$$

Combining these two terms bounds the relevant maximal deviation. We next "clean-up" the bound slightly. Since $\delta<1$, $\log(12/\delta)\leq 5\log^{7/4}(2/\delta)$. Recall $N_{\max}\geq 1$ by assumption which implies $N_{\max}^{1/4}\geq 1$. Finally, $d\geq 2$ and $K\geq 2$ implies that $d^{7/4}\log^2(K)\geq 1$. Hence, replacing our bound on the first term with

$$30L\sqrt{\frac{2C}{\gamma}}N_{\max}^{1/4}\frac{\lambda_{\max}}{\lambda_{\mathrm{avg}}}\cdot d^{7/4}\cdot\frac{\log^3 K}{\sqrt{K}}\cdot\log^{7/4}\left(\frac{2}{\delta}\right),$$

Adding this term to the righthand side in Eq. (EC.B.4), simplifying and collecting constants completes the proof. □

We can now prove the main result of the section.

Proof of Theorem 4.3. We first treat the case $\frac{16L^2}{C\gamma}\geq 1$. Then, Lemmas B.6 and B.7 bound each of the maximal deviations in Lemma 4.1. Instantiating them with $\delta\to\delta/2$, adding their righthand sides and applying the union bound thus bounds the sub-optimality. Collecting dominant terms yields

$$\mathsf{SubOpt}_K(\alpha_h^{\mathsf{S\text{-}SAA}},h) \;\leq\; \mathrm{A}\cdot\frac{L\sqrt{C}}{\sqrt{\gamma}}\cdot N^{1/4}\frac{\lambda_{\max}^{5/4}}{\lambda_{\min}}d^{7/4}\cdot\log^{7/4}\left(\frac{1}{\delta}\right)\cdot\frac{\log^3(K)}{\sqrt{K}}.$$

Now, in the case that $\frac{16L^2}{C\gamma}<1$, we can always increase L to the larger Lipschitz constant $\frac{\sqrt{C\gamma}}{4}$. We can then apply the case above yielding a similar bound but with leading term $C/4$. Taking the maximum of both bounds proves the theorem. □

B.4. Proof of Theorem 4.4: Shrunken-SAA with Fixed Anchors for Discrete Problems

We first use Corollary 4.1 proven in Section 4.3 to prove the following bounds on the maximal deviations of interest via Theorem 4.1.

LEMMA B.8 **(Bound on Maximal Deviation for True Performance)**. *Under the assumptions of Corollary 4.1, there exists a constant* A *such that with probability at least* $1-\delta$,

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

Proof. 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,

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

It follows from Theorem 4.1, that there exists a constant A_1 such that with probability at least $1-\delta$,

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

Optimizing the choice of p, substituting and collecting constants completes the proof. □

LEMMA B.9 **(Bound on Maximal Deviation for LOO Performance)**. *Under the assumptions of Corollary 4.1, there exists a constant* A *such that with probability at least* $1-\delta$,

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

Proof. The proof follows that of Lemma B.8 closely. We first bound the variable J in Eq. (4.3) corresponding to the process $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0):\alpha\geq 0\right\}$ with the envelope given by Lemma 4.2. By Corollary 4.1,

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

Hence,

$$\begin{aligned}\sqrt[p]{\mathbb{E}[J^p]} \;&\leq\; 9C\frac{1}{N_{\max}}\frac{\lambda_{\max}}{\lambda_{\min}}\sqrt{K}\sqrt{\log\left(2d\sum_{k=1}^{K}|\mathcal{X}_k|\right)}\cdot\sqrt[p]{\mathbb{E}[\hat{N}_{\max}^p]}\\ &\leq\; \mathrm{A}_0\cdot C\frac{\lambda_{\max}}{\lambda_{\min}}\cdot\sqrt{\log\left(2d\sum_{k=1}^{K}|\mathcal{X}_k|\right)}\cdot 6^{1/p}p\log K\sqrt{K}\end{aligned}$$

where A_0 is a universal constant.

By Theorem 4.1 there exists a constant A_1 such that with probability at least $1-\delta$,

$$\sup_{\alpha\geq 0}\left|\frac{1}{K}\sum_{k=1}^{K} Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)-\mathbb{E}[Z_k^{\mathsf{LOO}}(\alpha,\boldsymbol{p}_0)]\right| \leq \mathrm{A}_1\left(\frac{6\cdot 5}{\delta}\right)^{1/p} p^{3/2}\cdot C\frac{\lambda_{\max}}{\lambda_{\min}}\cdot\frac{\log K}{\sqrt{K}}\cdot\sqrt{\log\left(2d\sum_{k=1}^{K}|\mathcal{X}_k|\right)}$$

Optimizing p and collecting constants proves the lemma. □

We can now prove the main result of the sectiosn.

Proof of Theorem 4.4. Lemmas B.8 and B.9 bound the maximal deviations in Lemma 4.1. Instantiating them for $\delta\to\delta/2$, adding their righthand sides and applying the union bound bounds the sub-optimality. Collecting dominant terms proves the result. □

B.5. Proof of Theorem 4.5: Shrunken-SAA with Data-Driven Anchors for Discrete Problems

As a first step towards our proof, we bound the cardinality of $\left\{\mathbf{Z}(\alpha,\boldsymbol{q}):\alpha\geq 0,\boldsymbol{q}\in\Delta^d\right\}$ and $\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}):\alpha\geq 0,\boldsymbol{q}\in\Delta^d\right\}$. As argued in the main text, to bound $\left|\left\{\mathbf{Z}(\alpha,\boldsymbol{q}):\alpha\geq 0,\boldsymbol{q}\in\Delta^d\right\}\right|$ it suffices to count the number of j-dimensional fully-specified polyhedron in the arrangement induced by Eq. (4.8).

Counting the polyhedra induced by hyperplane arrangements is a classical problem in geometry. For example, it is well-known that the number of d-dimensional, fully-specified polyhedra in a hyperplane arrangement with m hyperplanes in $\mathbb{R}^d$ is at most $\sum_{i=0}^{d}\binom{m}{i}$ (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 $\mathbb{R}^d$.

Lemma B.10 **(Number of Fully-Specified Polyhedra)**. *In a hyperplane arrangement with m hyperplanes in $\mathbb{R}^d$, the number of fully-specified polyhedra is at most*

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

Proof of Lemma B.10 Each fully-specified polyhedron has some dimension, $0\leq j\leq d$. We will count the number of such fully-specified polyhedra by counting for each dimension j.

Fix some $0\leq j\leq d$. Notice that each j-dimensional polyhedron lives in a j-dimensional subspace defined by $d-j$ linearly independent hyperplanes from the arrangement. There are at most $\binom{m}{d-j}$ ways to choose these linearly independent $d-j$ hyperplanes. Next project the remaining hyperplanes onto this subspace which yields at most $m-d+j$ 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 $m-d+j$ hyperplanes in j-dimensional space is at most

$\sum_{i=0}^{j}\binom{m-d+j}{i}$. In summary, it follows that there are at most $\binom{m}{d-j}\sum_{i=0}^{j}\binom{m-d+j}{i}$ 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

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

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

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

where the last equality is again the binomial expansion. □

Using this lemma, we can prove the following:

LEMMA B.11 **(Size of Discrete Solutions Sets for Data-Driven Anchors)**.

$$\left|\left\{\mathbf{Z}(\alpha,\boldsymbol{q}):\alpha\ge 0,\boldsymbol{q}\in\Delta^d\right\}\right|\le\left(\sum_{k=1}^{K}\left|\mathcal{X}\right|_k^{\,2}\right)^d,\quad \left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}):\alpha\ge 0,\boldsymbol{q}\in\Delta^d\right\}\right|\le d^d\left(\sum_{k=1}^{K}\left|\mathcal{X}\right|_k^{\,2}\right)^d.$$

Proof of Lemma B.11. Recall there are $m=\sum_{k=1}^{K}\binom{|\mathcal{X}_k|}{2}$ hyperplanes in the arrangement Eq. (4.8) and the number of fully-specified polyhedra in this arrangement upper-bounds $\left|\left\{\mathbf{Z}(\alpha,\boldsymbol{q}):\alpha\ge 0,\boldsymbol{q}\in\Delta^d\right\}\right|$. Noting $1+2m=1+\sum_{k=1}^{K}|\mathcal{X}_k|\left(|\mathcal{X}_k|-1\right)\le\sum_{k=1}^{K}|\mathcal{X}_k|^2$ yields the first bound.

A similar argument can be used to bound $\left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}):\alpha\ge 0,\boldsymbol{q}\in\Delta^d\right\}\right|$. In particular,

$$\begin{aligned}\left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}):\alpha\ge 0,\boldsymbol{q}\in\Delta^d\right\}\right| &\le \left|\left\{(\boldsymbol{x}_k(\alpha,\boldsymbol{q},\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i))_{k=1,i=1}^{K,\;\;d}\;:\;\boldsymbol{q}\in\Delta_d,\alpha\ge 0\right\}\right|\\ &\le \left|\left\{(\boldsymbol{x}_k(\alpha(\boldsymbol{\theta}),\boldsymbol{q}(\boldsymbol{\theta}),\hat{\boldsymbol{m}}_k-\boldsymbol{e}_i))_{k=1,i=1}^{K,\;\;d}\;:\;\boldsymbol{\theta}\in\mathbb{R}^d_+\right\}\right|. \qquad \text{(EC.B.5)}\end{aligned}$$

We then consider the arrangement generated by

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

for all $k=1,\ldots,K$, $i,j=1,\ldots,|\mathcal{X}_k|$ with $i\ne j$, and $l=1,\ldots d$. Notice there are $d\sum_{k=1}^{k}\binom{|\mathcal{X}_k|}{2}$ such hyperplanes. Moreover, $\left|\left\{\mathbf{Z}^{\mathsf{LOO}}(\alpha,\boldsymbol{q}):\alpha\ge 0,\boldsymbol{q}\in\Delta^d\right\}\right|$ is upper-bounded by the number of fully-specified polyhedra in this arrangement. Note that $1+2d\sum_{k=1}^{k}\binom{|\mathcal{X}_k|}{2}=1+d\sum_{k=1}^{K}|\mathcal{X}_k|\left(|\mathcal{X}_k|-1\right)\le d\sum_{k=1}^{K}|\mathcal{X}_k|^2$. Plugging in this value into Lemma B.10 yields the second bound above. □

Importantly, both bounds in Lemma B.11 are polynomial in K whenever $\left|\mathcal{X}^k\right|$ are bounded over k and d is fixed.

We next use Lemma B.11 to bound the maximal deviations of interest via Theorem 4.1.

LEMMA B.12 **(Bound on Maximal Deviation for True Performance for General Anchors)**. *Under the assumptions of Theorem 4.5, there exists a constant* A *such that with probability at least* $1-\delta$*,*

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

Proof of Lemma B.12 By Lemmas B.11 and 4.2 and since $\left(\sum_{k=1}^K\left|\mathcal{X}^k\right|^2\right)^d\leq\left(\sum_{k=1}^K\left|\mathcal{X}^k\right|\right)^{2d}$,

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

The remainder of the proof follows that of Lemma B.8. □

LEMMA B.13 **(Bound on Maximal Deviation for LOO Performance for General Anchors)**. *Under the assumptions of Theorem 4.5, there exists a constant* A *such that with probability at least* $1-\delta$*,*

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

Proof of Lemma B.13. By Lemmas B.11 and 4.2 and since $d^d\left(\sum_{k=1}^K\left|\mathcal{X}^k\right|^2\right)^d\leq\left(d\sum_{k=1}^K\left|\mathcal{X}^k\right|\right)^{2d}$,

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

The rest follows as in the proof of Lemma B.9. □

We can now prove the main result of the section.

Proof of Theorem 4.5. We apply our usual strategy. Note Lemmas B.12 and B.13 bound the two maximal deviations in Lemma 4.1 respectively. Instantiating them for $\delta\to\delta/2$, adding the right hand sides and applying the union bound yields a bound on the sub-optimality. Collecting dominant terms yields the result. □

Appendix C: Deferred Proofs from Section 5

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

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

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

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

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

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 (c.f. 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 $\alpha > 0$,

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

(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*

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

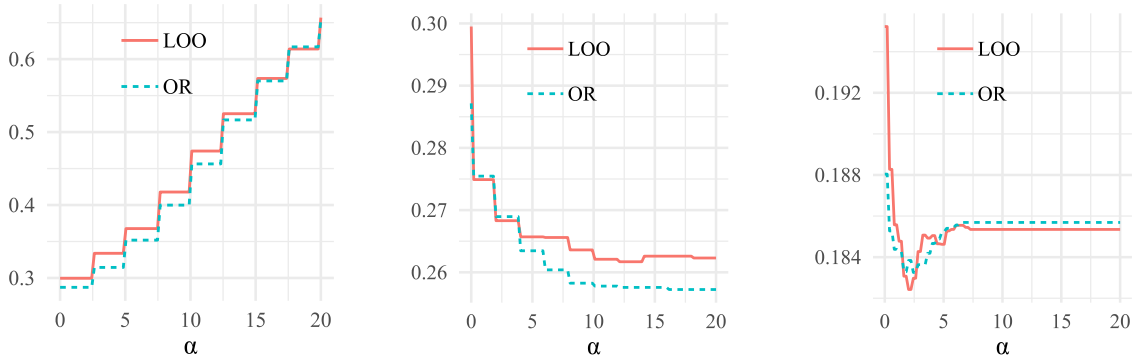
where $\hat{\mu}_k$ is the sample mean for the k^{th} subproblem.

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.

Appendix E: Additional Figures and Computational Details

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



(a) $\alpha^{\text{S-SAA}}_{\boldsymbol{p}_0} = \alpha^{\text{OR}}_{\boldsymbol{p}_0} = 0$ (b) $\alpha^{\text{S-SAA}}_{\boldsymbol{p}_0} \approx 12.12$, $\alpha^{\text{OR}}_{\boldsymbol{p}_0} \approx 16.16$ (c) $\alpha^{\text{S-SAA}}_{\boldsymbol{p}_0} \approx 2.02$, $\alpha^{\text{OR}}_{\boldsymbol{p}_0} \approx 2.42$

Figure EC.1 **LOO and Oracle Curves**. We consider $K = 1000$ 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 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.

$\boldsymbol{p}_k$ 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 Results

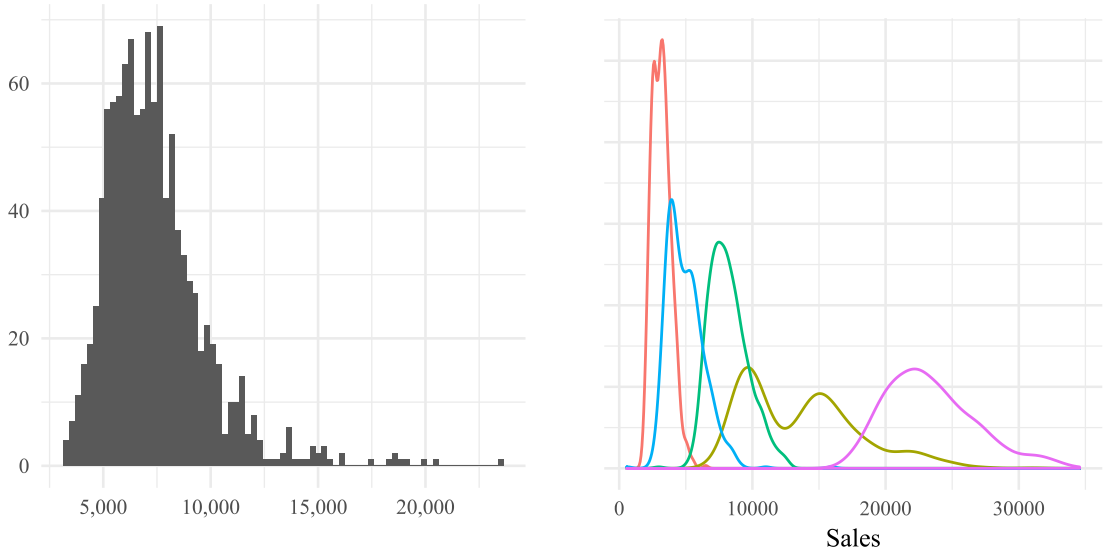
On average, less than 2.5% of stores are open on weekends, and hence we drop all weekends from our dataset. Similarly, the data exhibits a mild upward linear trend at a rate of 215 units a year (approximately 3.7% increase per year), with a p-value $< .001$. This trend is likely due to inflation and 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 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.

E.4. Additional Figures from Section 6.

The first panel of Fig. EC.2 shows the average daily demand by store for each of the $1,115$ stores in our dataset.

Figure EC.3 shows the standard deviation of each of our methods on simulated data from Section 6.2 as a function of K.



(a) **Distribution of Average Daily Demand** (b) **Demand Distributions by Store**

Figure EC.2 **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.

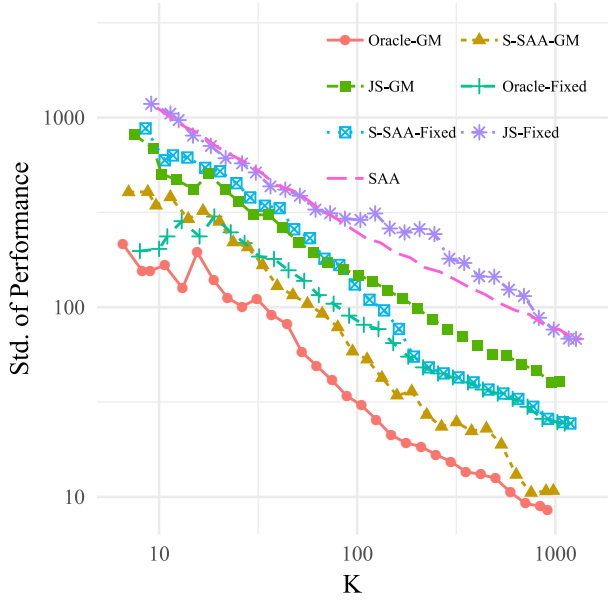


Figure EC.3 **Standard Deviation of Performance** For each method, the standard deviation of converges to zero because performance concentrates at its expectation as $K \to \infty$. Notice that our Shrunken-SAA methods are less variable than the decoupled SAA solution because pooling increases stability.

Figure EC.4 shows the average amount of pooling by method by K on our simulated data set from Section 6.2.

Figure EC.5 shows results from our historical backtest in Section 6.5 with $d = 50$ and $d = \infty$.

Figure EC.6 shows results from our historical backtest in Section 6.6 with $\hat{N}_k = 20$ or $\hat{N}_k = 40$, both fixed.

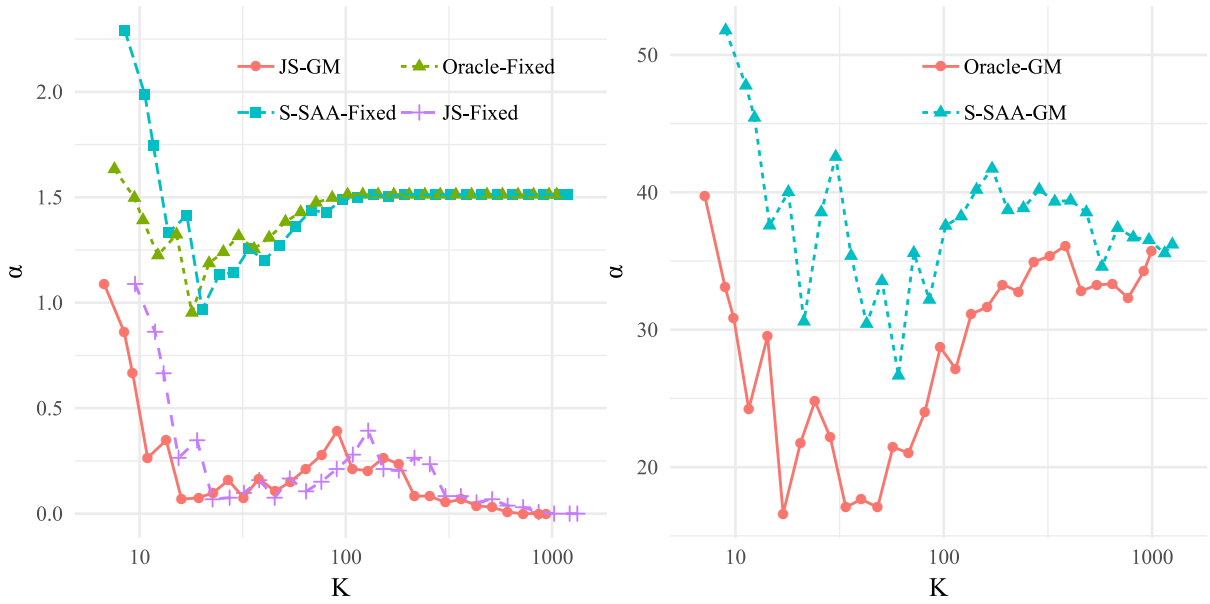
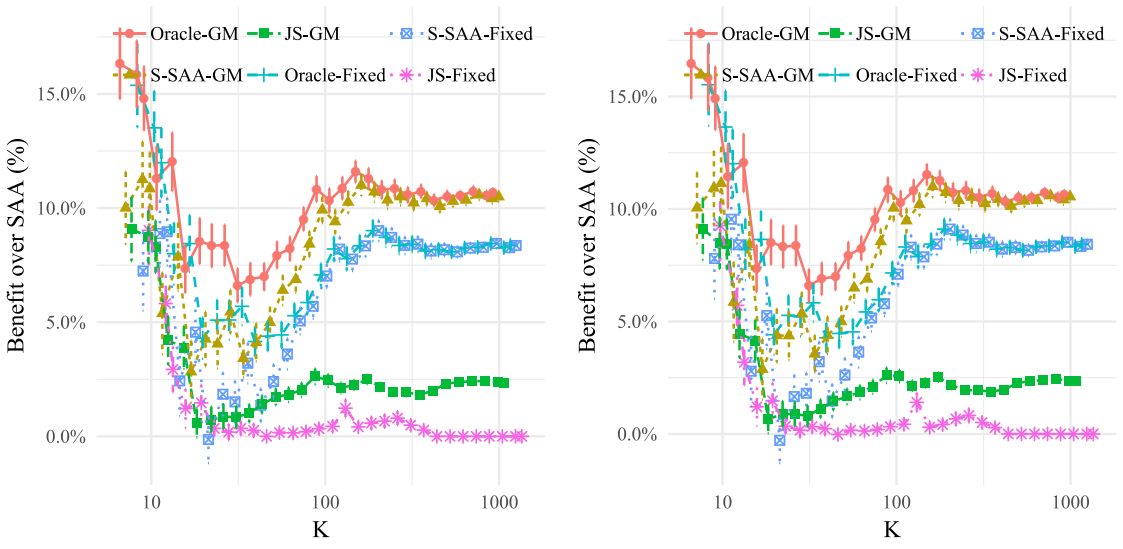
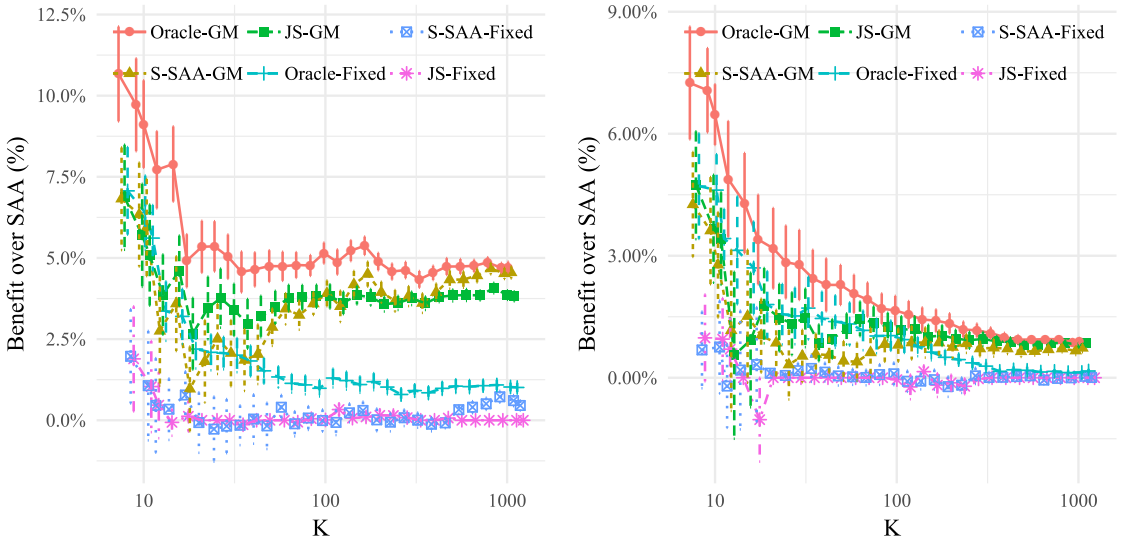


Figure EC.4 **Amount of Pooling by Method** We plot the amount of data-pooling (α) for each of the above methods. When shrinking to $\hat{\boldsymbol{p}}^{\mathrm{GM}}$, both the oracle and our Shrunken-SAA method shrinks very aggressively. Plotted separately for clarity.



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

Figure EC.5 **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 leave-one-out calculation, i.e., $d = \infty$. Error bars show ± 1 standard error. The differences between the plots are essentially indescernible.



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

Figure EC.6 **Dependence on N.** Evaluated on historical data with $d = \infty$. Error bars show ± 1 standard error.