Small-Data, Large-Scale Linear Optimization with Uncertain Objectives

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

Optimization applications often depend upon a huge number of uncertain parameters. In many contexts, however, the amount of relevant data per parameter is small, and hence, we may only have imprecise estimates. We term this setting – where the number of uncertainties is large, but all estimates have low precision – the "small-data, large-scale regime." We formalize a model for this new regime, focusing on optimization problems with uncertain linear objectives. We show that common data-driven methods, such as sample average approximation, data-driven robust optimization, and certain regularized policies, may perform poorly in this new setting. We then propose a novel framework for selecting a data-driven policy from a given policy class. Like the aforementioned data-driven methods, our new policy enjoys provably good performance in the large-sample regime. Unlike these methods, we show that in the small-data, large-scale regime, our data-driven policy performs comparably to an oracle best-in-class policy under some mild conditions. We strengthen this result for linear optimization problems and two natural policy classes: the first inspired by the empirical Bayes literature and the second by regularization techniques. For both classes, the suboptimality gap between our proposed policy and the oracle policy decays exponentially fast in the number of uncertain parameters, even for a fixed amount of data. Thus, these policies retain the strong large-sample performance of traditional methods, and additionally enjoy provably strong performance in the small-data, large-scale regime. Numerical experiments confirm the significant benefits of our methods.

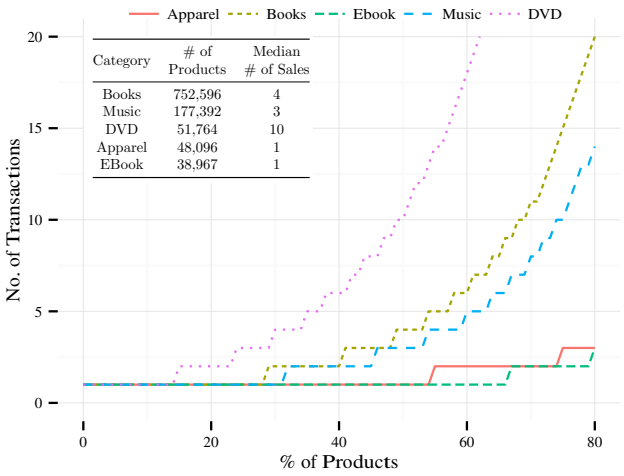
Key words: Data-driven optimization. Small-data, large-scale regime. Stein's unbiased risk estimation.
History: This paper was first submitted in Oct. 2017. A revision was submitted in Feb. 2019.

1. Introduction

We live in a small-data world. Popular press about the age of "Big Data" notwithstanding, many real-world decision-making problems exhibit a huge number of uncertain parameters with a small amount of relevant data per parameter. Consider the following two examples.

Inventory Management for Low-Demand Products: Large online retailers carry millions of products, but most products have few sales per quarter. Figure 1 summarizes the number of sales per product for the five most popular product categories based on data for a large e-retailer. The vast majority of products have fewer than 5 sales. Consequently, although there are *many* products with uncertain demands, there are *limited* demand data per product.

Vehicle Routing after an Accident: Ride-sharing platforms like Uber and logistics services like UPS frequently update routing decisions based on near real-time traffic data across millions of road segments. In the wake of a large accident or weather disruption, however, historical traffic



Category	# of Products	Median # of Sales
Books	752,596	4
Music	177,392	3
DVD	51,764	10
Apparel	48,096	1
EBook	38,967	1

Figure 1 **Number of Sales per Product.** Data taken from a leading internet retailer for all products that sold at least one unit in the top five categories between July and September 2005. Except for the DVD category, the median number of sales for each product is at most five.

patterns often shift dramatically; there may only be a few hours of relevant data on the new, post-disruption travel times. Again, although there are *many* road segments with uncertain travel times, there are *limited* traffic data per segment.

Some reflection suggests that the "small-amount-of-relevant-data" phenomenon in these examples is at least partially driven by the nature of modern decision-making under uncertainty. It is not unusual for real applications to require making thousands of decisions simultaneously, in time-changing environments with only low-precision estimates. This combination of features – highly granular decision-making, time-changing environments, and low-precision estimates – combine to drive the small-data, large-scale phenomenon. Their ubiquity suggests that a host of other applications, such as new-user product recommendations and disaster response operations, may also exhibit these features. We term this decision-making setting – many uncertain parameters, each with limited relevant data and, hence, an imprecise estimate – the *small-data, large-scale* regime.

By contrast, many traditional, data-driven optimization methods are theoretically justified by studying their performance in the large-sample regime, where the number of uncertain parameters is fixed, but we have access to increasing amounts of data, and hence increasingly precise estimates of all parameters. For example, the Sample Average Approximation (SAA) approach is well known to converge to the full-information optimal performance in the large-sample regime (Shapiro et al. 2009, Kleywegt et al. 2002). This type of large-sample performance guarantee shows that these methods will perform well when the amount of data is large relative to the number of uncertainties. However, it is unclear how these methods may perform in the small-data, large-scale regime.

At a high level, the main messages of this paper are as follows. First, the small-data, large-scale regime is structurally different from the large-sample regime, and, consequently, traditional methods may perform quite poorly in this regime. Second, it is possible to design novel methods that retain the strong large-sample performance guarantees of traditional methods but that *additionally* have provably good and empirically strong performance in the small-data, large-scale regime.

Since the class of all small-data, large-scale optimization problems is too broad to treat in a single paper, we focus on problems with known feasible regions but an uncertain, linear objective:

$$\mathrm{P}^n: \quad Z^*(\boldsymbol{\mu}) \equiv \max_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}.$$

Here $\mathcal{X} \subseteq \mathbb{R}^n$ is a known, non-empty, compact, potentially non-convex set and the objective coefficients $\boldsymbol{\mu}$ are **unknown**. Instead, we are given noisy estimates $\hat{\mu}_j$ for each j, each of which was formed using only a *small* amount of *relevant* data. (A precise model for $\hat{\boldsymbol{\mu}}$ is given in Section 2). Let $\boldsymbol{x}^*(\boldsymbol{\mu})$ denote an optimal solution to P^n.

Admittedly, Problem P^n is not general enough to cover sophisticated models for the aforementioned inventory management and vehicle routing applications. That said, we do consider P^n to be a fundamental building block for these applications. In particular, P^n does subsume transportation and shortest-path problems as special cases (Bertsimas and Tsitsiklis 1997, Chapt. 7).

Our Contributions and Main Results: Using Problem P^n, we highlight unique features and challenges of the small-data, large-scale regime. In particular, in contrast to the large-sample regime, the classical Sample Average Approximation (SAA) method can perform arbitrarily badly (Example 2.6), and no data-driven optimization procedure can guarantee more than constant fraction of the full-information optimum value $Z^*(\boldsymbol{\mu})$ (Theorem 2.7).

Since consistently attaining full-information performance is provably impossible in the small-data, large-scale regime by Theorem 2.7, we instead restrict attention to classes of data-driven policies and seek methods for identifying a member whose performance is comparable to a certain oracle policy. This oracle has access to the true value of $\boldsymbol{\mu}$ in P^n, but is restricted to use a policy from the class, thus formalizing the notion of a "best-in-class" policy.

We propose a novel framework for this task and show that if the given policy class satisfies a certain uniform convergence criteria given in Theorem 3.5, then our proposed policy performs comparably to the oracle policy. Specifically, we prove that when the estimates $\hat{\mu}_j$ are Gaussian and $n \to \infty$ in P^n, the performance of our proposed policy converges to that of the oracle policy, even if the amount of relevant data per uncertainty remains fixed. Moreover, when $\hat{\mu}_j$ have non-Gaussian distributions, the difference in performance between our proposed policy and the oracle policy converges to a constant that measures the degree of non-normality and this constant does not depend on the parameters of P^n.

We then specialize and strengthen our general result to linear optimization problems, i.e., where $\mathcal{X}$ is polyhedral (cf. Eq. (4.1)), focusing on two specific policy classes. The first, which we call "Bayes-Inspired" policies, is motivated by the empirical Bayes and compound estimation literature in statistics; see Zhang (2003), Efron (2012), and the references therein. These

methods target applications such as microarray analysis, genomics, and compressed sensing, all of which involve simultaneously solving thousands of separate inference problems with limited data. Thus, these methods are particularly suited for the small-data, large-scale regime. Indeed, simple "estimate-then-optimize" policies leveraging empirical Bayes estimators already outperform SAA in the small-data, large-scale regime (cf. Sec. 4). However, we observe empirically that these policies do not typically achieve near oracle performance, because they fail to exploit the particular optimization structure. By contrast, by specializing our general framework, we propose a new policy that achieves near oracle performance for large n, and, thus, necessarily outperforms these "estimate-then-optimize" variants. We strengthen our previous general-purpose results by providing an explicit, non-asymptotic bound on the suboptimality gap that converges to the aforementioned "non-normality" constant. This bound converges exponentially fast in the number of uncertain parameters n, even for a fixed amount of data per parameter (Theorem 4.3).

The second policy class that we consider, which we call "Regularization-Inspired" policies, is motivated by the growing literature on incorporating regularizers into the SAA problem to improve performance and computational tractability; see Nesterov (2005), Negahban et al. (2012), and the references therein. We focus on a weighted ℓ_2-regularizer. Based upon a well-known equivalence between regularization and robustness, these policies are equivalent to policies obtained by solving a robust optimization problem with an ellipsoidal uncertainty set (Lemma E.1). We prove that common cross-validation techniques – which are routinely used to specify the regularization parameter in the large-sample regime – do not achieve near oracle performance (Theorem 5.1). Similarly, we illustrate empirically that policies based on probabilistic feasibility guarantees – a standard approach to sizing uncertainty sets in robust optimization – also do not achieve near oracle performance. By contrast, by specializing our general framework, we propose a new policy that achieves near oracle performance for large n, and, thus, may outperform these approaches. We again improve upon our previous general-purpose result by proving a non-asymptotic bound on its suboptimality gap that converges exponentially fast to zero (if the $\hat{\mu}_j$ are Gaussian) and converges exponentially fast to the aforementioned "non-normality" constant (Theorem 5.2), otherwise.

In all three of our small-data, large-scale performance guarantees, the suboptimality of our policy depends on the degree of "non-normality" of the estimators $\hat{\mu}_j$. Methods that either explicitly or implicitly assume Gaussian inputs are common in the high-dimensional statistics and compound decision-making literature, at least partially because many simple estimators $\hat{\mu}_j$ are approximately normally distributed (see Examples 2.2 and 2.5) . Consequently, many authors derive estimation procedures that are provably optimal under Gaussian assumptions, and then argue such procedures have good *practical* performance, even when strict normality does not hold; see, e.g., Xie et al. (2012), Mukherjee et al. (2015), Efron and Morris (1973, 1975), Donoho and Johnstone (1995),

Morris (1983). We adopt a similar perspective. Our bounds make theoretically precise the sense in which our policy can be expected to perform well as normality is violated, and, in Section F.5, we provide a preliminary empirical assessment of our methods under non-normality.

Importantly, we stress that our small-data, large-scale performance guarantees for our proposed policies come at essentially "no statistical cost" in the large-sample regime. Specifically, we prove an explicit, non-asymptotic suboptimality bound relative to the full-information optimum $Z_n^*(\boldsymbol{\mu})$ for our policy (Theorem 3.6). In the large-sample setting with i.i.d. observations, this bound converges to zero at a rate comparable to SAA. In practice, the performance may even be much better than SAA (see, e.g., Sec. F.4). In this sense, we argue our methods retain the strong large-sample properties of SAA and also enjoy additional small-data, large-scale performance guarantees.

Finally, we present a simulation study calibrated to real data for managing a portfolio of online advertisements. Our methods outperform traditional methods and "estimate-then-optimize" methods from high-dimensional statistics so long as the number of uncertainties is sufficiently large.

Relationship to Prior Work: Problem P^n is a special case of the stochastic optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^n} \quad \mathbb{E}\left[c(\boldsymbol{x},\boldsymbol{\xi})\right], \tag{1.1}$$

in which $c(\boldsymbol{x},\boldsymbol{\xi}) = \frac{1}{n}\boldsymbol{\xi}^\top\boldsymbol{x}$ and $\mathbb{E}[\boldsymbol{\xi}] = \boldsymbol{\mu}$. There is now a rich literature in data-driven optimization on solving data-driven versions of Eq. (1.1) when the distribution of $\boldsymbol{\xi}$ is unknown, but one has access to a dataset $\hat{\boldsymbol{\xi}}^1,\ldots,\hat{\boldsymbol{\xi}}^S$ drawn i.i.d as $\boldsymbol{\xi}$.[1] Our model for $\hat{\boldsymbol{\mu}}$ approximates this data-generation mechanism (cf. Example 2.2 below), but also captures other settings (e.g., Example 2.5).

As mentioned, many methods for addressing data-driven versions of Equation (1.1) satisfy large-sample performance guarantees similar to those of SAA. Such methods include Robust SAA (Bertsimas et al. 2018b), regularized SAA variants (Negahban et al. 2012), stochastic gradient descent and its variants (Nemirovski et al. 2009, Lan 2012, Nesterov 2009), and distributional robust optimization (Delage and Ye 2010, Esfahani and Kuhn 2018, Gupta 2019). Many of these methods seek to improve upon SAA by establishing guarantees beyond large-sample performance. For example, many data-driven robust and distributionally robust optimization methods can be tuned to ensure that for finite data, solutions will satisfy certain probabilistic feasibility guarantees; see Bertsimas et al. (2018a,b) for recent references. Meanwhile, gradient-based methods with suitable step-sizes can offer probabilistic optimality guarantees for finite samples. Finally, certain regularizers are

[1] The majority of the literature treats an expected value of objective as in Eq. (1.1), often because the selected solution will be held fixed for a long-time. Alternatively, one could study the out-of-sample performance as random variable, i.e., $c(\hat{\boldsymbol{x}},\overline{\boldsymbol{\xi}})$ where $\hat{\boldsymbol{x}}$ depends on the data and $\overline{\boldsymbol{\xi}}$ is an i.i.d. copy of $\boldsymbol{\xi}$. For Problem P_n with n large, however, this out-of-sample performance is an average of n terms and hence converges to its expectation whenever these terms are sufficiently independent by strong law of large numbers, further motivating our focus on expected values.

known to yield either consistent estimates of the unknown parameter or consistent predictions if the underlying parameter satisfies certain a priori structure such as sparsity; see Bickel et al. (2009), Candès and Recht (2009) and Wainwright (2009) for representative results and Negahban et al. (2012) for additional references. These results do not directly extend to the small-data, large-scale regime.

There is also a separate literature studying Equation (1.1) in the high-dimensional statistics and machine learning communities. In this setting, $\boldsymbol{\xi}$ is often, but not always, an $n+1$ dimensional vector, consisting of an n-dimensional feature component and a 1-dimensional response component, and Equation (1.1) represents a prediction problem. Most relevantly for our work, this literature has placed special emphasis on studying the performance of methods under various asymptotic scalings of n (the number of decision variables) and S (the number of samples), especially the large-sample regime ($\mathsf{S} \to \infty$ and n fixed) and the high-dimensional regime (both $\mathsf{S} \to \infty$ and $n \to \infty$). The celebrated VC-dimension theory (Vapnik 1999) and algorithm stability theory (Bousquet and Elisseeff 2002) both provide general frameworks for analyzing approaches to Equation (1.1) in these regimes. Many works exist that strengthen these performance guarantees for special cases of Equation (1.1) (see e.g., Bühlmann and Geer (2011) and references therein for lasso-regression, and Belloni and Chernozhukov (2011) for penalized quantile regression). Ban and Rudin (2014) connects this literature with the operations literature, highlighting the benefits of leveraging feature vectors in the particular case of the newsvendor problem in both the large-sample and high-dimensional regimes. Subsequently, other authors have explored incorporating feature information in operations research problems both in terms of developing high-quality estimates to plug-in to an optimization formulation (Ferreira et al. 2015, Ban et al. 2018, Chen et al. 2015) and approaches that blend estimation and optimization (Bertsimas and Kallus 2019, Elmachtoub and Grigas 2017).

One can view our work as complementing this statistical literature in the setting where S is fixed and $n \to \infty$ (again, see, Example 2.2), although all our results are stated in a non-asymptotic framework. From a technical point of view, although we leverage some standard tools, e.g., pseudo-dimension (an extension of VC dimension), the structure of P^n requires novel and specialized analysis, especially for the polyhedral feasibility sets considered in Sections 4 and 5. One cannot simply apply the VC-theory "out of the box" because these constraints introduce a non-trivial dependence structure, and the requisite quantities are thus *not* simple averages of independent, random functions. Moreover, we focus on proving "best-in-class" optimality results, while much of the high-dimensional literature instead focuses on generalization results which bound the difference between in-sample and out-of-sample performance.

Our approach to computing asymptotically best-in-class policies uses a novel adaptation of Stein's unbiased risk estimation (SURE) (Stein 1981), a common approach to model selection in

statistics; for representative examples, see Donoho and Johnstone (1995), Tibshirani and Taylor (2012) and Candes et al. (2013). To the best of our knowledge, however, SURE has never been leveraged for more general data-driven optimization problems. The key idea of our adaptation is to replace the standard SURE estimator with an asymptotically unbiased approximation constructed so that the approximation error vanishes as the number of uncertainties grows large. In this sense, our approach is similar in spirit to that of Mukherjee et al. (2015), who also develop an asymptotic approximation to SURE. However, Mukherjee et al. (2015) heavily leverage the specific structure of the check-loss function, while our approach applies generally to optimization problems with linear objectives and any class of policies that satisfies the requisite uniform convergence criteria.

2. Formulation and Properties of the Small-Data, Large-Scale Regime

Throughout the remainder of the paper, vectors will be denoted in bold, while scalars appear in a regular font. We use $\phi(\cdot)$ and $\Phi(\cdot)$ to denote the density and cumulative distribution functions of the standard normal random variable. We write $\hat{\mu}_j \sim \mathcal{N}(m, v)$ to indicate that the random variable $\hat{\mu}_j$ is normally distributed with mean m and variance v.

Importanty, we adopt the following assumption on the estimates $\hat{\mu}_j$.

Assumption 2.1 (Model for $\hat{\boldsymbol{\mu}}$) For each $j = 1, \ldots, n$, $\hat{\mu}_j$ is unbiased, i.e., $\mathbb{E}[\hat{\mu}_j] = \mu_j$, and has *known* precision ν_j, i.e., $\mathbb{E}[(\hat{\mu}_j - \mu_j)^2] = 1/\nu_j$.

Recall that precision is the reciprocal of the variance. Similar assumptions, i.e., mean-zero noise with some measure of dispersion, are very common in the high-dimensional statistics literature (Johnstone 2015) and often used in the robust optimization literature to assess the performance of uncertainty sets; see, e.g., Bertsimas and Sim (2004), Ben-Tal and Nemirovski (2000), Chen et al. (2007). The precisions ν_j implicitly measure the amount of relevant data used in constructing $\hat{\mu}_j$. We use them to define the small-data, large-scale regime in general settings. As a first step, consider the following motivating example.

Example 2.2 (Finite Observation Model) Suppose that, for each $j = 1, \ldots, n$, we observe $\mathsf{S}_j \geq 1$ i.i.d. random variables $\xi_j^1, \ldots, \xi_j^{\mathsf{S}_j}$, each with mean μ_j and known precision ν_0. Since the μ_j are potentially unrelated and all draws are independent, the data $\xi_j^1, \ldots, \xi_j^{\mathsf{S}_j}$ provide no information for estimating μ_k whenever $k \neq j$. Thus, the sample mean $\hat{\mu}_j = \frac{1}{\mathsf{S}_j} \sum_{\ell=1}^{\mathsf{S}_j} \xi_j^\ell$ is arguably the most natural, unbiased estimate of μ_j, and has precision $\nu_j \equiv \mathsf{S}_j \nu_0$.

In this example, the small-data large-scale regime corresponds to the setting where n is very large, *but* S_j *is fixed and small for each* j. In particular, although the total amount of data $\sum_{j=1}^n \mathsf{S}_j \to \infty$ as $n \to \infty$, the precision ν_j remains bounded for each j. Thus, the additional data do not provide

additional precision in estimating μ_j, i.e., $\hat{\mu}_j \not\to_p \mu_j$ as $n \to \infty$. In this sense there is a small amount of relevant data for estimating each μ_j. By contrast, the large-sample regime in this example corresponds to the setting where n is fixed and S_j is large for each j. In particular, for each j, $\nu_j \to \infty$ as $\mathsf{S}_j \to \infty$, and $\hat{\mu}_j \to_p \mu_j$ as $\mathsf{S}_j \to \infty$ by the law of large numbers.[2]

Note the distinct behavior of the ν_j in the two regimes above. Our definition of the small-data, large-scale regime is inspired by this difference. Specifically, for any $\hat{\boldsymbol{\mu}} \in \mathbb{R}^n$ and $\boldsymbol{\nu} \in \mathbb{R}^n_+$, let the tuple $(\mathrm{P}^n, \hat{\boldsymbol{\mu}}, \boldsymbol{\nu})$ denote an instance of P^n with unbiased estimators $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^\top$ with corresponding precisions $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^\top$.

Definition 2.3 (Small-Data, Large-Scale Regime) We say a sequence of instances $\left\{(\mathrm{P}^n, \hat{\boldsymbol{\mu}}^n, \boldsymbol{\nu}^n) : \hat{\boldsymbol{\mu}}^n \in \mathbb{R}^n, \boldsymbol{\nu}^n \in \mathbb{R}^n_+, n \geq 2\right\}$ is in the *small-data, large-scale* regime if there exists a $\nu_{\mathsf{max}} < \infty$ such that $\nu^n_j \leq \nu_{\mathsf{max}}$ for all $j = 1, \ldots, n$ and $n = 2, \ldots, \infty$.[3]

In other words, in the small-data, large-scale regime, all estimates have bounded precision. For comparison, we also define the large-sample regime. In this case, we consider a sequence of *estimators* that are improving in quality as we collect more data. This sequence requires an extra indexing variable S (mnemonically, S stands for the number of samples).

Definition 2.4 (Large-Sample Regime) For a fixed n, we say a sequence of instances $\left\{(\mathrm{P}^n, \hat{\boldsymbol{\mu}}^{\mathsf{S}}, \boldsymbol{\nu}^{\mathsf{S}}) : \hat{\boldsymbol{\mu}}^{\mathsf{S}} \in \mathbb{R}^n, \boldsymbol{\nu}^{\mathsf{S}} \in \mathbb{R}^n_+, \mathsf{S} \geq 1\right\}$ is in the *large-sample* regime if $\lim_{\mathsf{S}\to\infty} \min_{j=1,\ldots,n} \nu^{\mathsf{S}}_j = \infty$.

We stress that in the large-sample regime, the problem P^n is fixed, but the precision of each estimate $\hat{\boldsymbol{\mu}}^{\mathsf{S}}$ increases with S. In contrast, in the small-data, large-scale regime, the dimension of the problem P^n increases with n while the precision remains bounded. We prefer modeling in terms of precisions as above instead of number of samples S because it allows us to extend the idea of a "small amount of relevant data" to other settings in a unified way. Consider the following:

Example 2.5 (Linear Regression) Suppose for each j, we observe $(\xi_j, \boldsymbol{f}_j) \in \mathbb{R}^{p+1}$ where $\mathbb{E}[\xi_j] = \mu_j$, and $\boldsymbol{f}_j \in \mathbb{R}^p$ is an auxiliary feature vector. For example, in our vehicle routing example, $\boldsymbol{f}_j$ may capture the speed limit, number of lanes, and distance of the road segment. In such a setting, we might posit a linear regression model, i.e.,

$$\xi_j = \boldsymbol{\beta}^T \boldsymbol{f}_j + \epsilon_j, \quad \text{with } \mathbb{E}[\epsilon_j] = 0 \text{ and } \mathbb{E}[\epsilon_j^2] = \sigma^2, \quad j = 1, \ldots, n,$$

[2] As an aside, we mention that by the central limit theorem, we expect this estimator to be approximately Gaussian if the ξ^k_j are not too heavy-tailed and S_j is not too small.

[3] We focus on $n \geq 2$ to avoid some trivial cases.

and then estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ obtained by ordinary least squares. We can map this example to our model by letting $\hat{\mu}_j = \hat{\boldsymbol{\beta}}^\top \boldsymbol{f}_j$. Standard regression results show that $\hat{\mu}_j$ is unbiased with precision $1/\nu_j = \sigma^2 \boldsymbol{f}_j{}^\top (\boldsymbol{F}^\top \boldsymbol{F})^{-1}_{jj} \boldsymbol{f}_j$ where $\boldsymbol{F} \in \mathbb{R}^{n \times p}$ is the matrix with rows given by $\boldsymbol{f}_j{}^\top$.

Intuitively, some components of μ_j may have a smaller amount of "relevant" data, because if some features are very rare, coordinates j depending on these features will have "less" relevant data. The small-data, large-scale regime intuitively corresponds to settings where n is large and most coordinates depend on a small number of rare features. Making this idea formal in terms of n and p requires additionally specifying structure on the matrix $\boldsymbol{F}$, e.g., its eigenspectrum. By contrast, the precisions ν_j, as above, provide a simple way to define the small-data, large-scale regime and large-sample regime for such examples.[4]

Finally, we define a *data-driven policy* $\boldsymbol{x}(\cdot)$ for P^n to be a function mapping $(\hat{\boldsymbol{\mu}}, \boldsymbol{\nu}, \mathcal{X})$ to a feasible solution $\boldsymbol{x}(\hat{\boldsymbol{\mu}}, \boldsymbol{\nu}, \mathcal{X}) \in \mathcal{X}$. To simplify notation, we typically only emphasize the dependence on the random variable $\hat{\boldsymbol{\mu}}$, writing $\boldsymbol{x}(\hat{\boldsymbol{\mu}})$ instead of $\boldsymbol{x}(\hat{\boldsymbol{\mu}}, \boldsymbol{\nu}, \mathcal{X})$. Implicitly, $\boldsymbol{x}(\hat{\boldsymbol{\mu}})$ depends on n.

We are interested in studying the performance of various data-driven policies under both regimes defined above. In the remainder of this section, we highlight the significant differences between these regimes by proving that certain fundamental results for data-driven optimization in the large-sample regime do *not* carry over to the small-data, large-scale regime for P^n.

2.1. The Performance of SAA in the Small-Data, Large-Scale Regime

The Sample Average Approximation (SAA) method is perhaps the most ubiquitous and well-studied approach to data-driven optimization. In our context, the SAA policy proxies the unknown $\boldsymbol{\mu}$ by its unbiased estimate $\hat{\boldsymbol{\mu}}$ and optimizes against this proxy:

$$\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}}) \;\in\; \arg\max_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{n} \hat{\boldsymbol{\mu}}^\top \boldsymbol{x} \;,$$

(compare also to Example 2.2). Under mild assumptions in the large-sample regime, $\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}}) \to \boldsymbol{x}^*(\boldsymbol{\mu})$ almost surely (Kleywegt et al. 2002). Other authors, e.g., Lim et al. (2011), have shown the fragility of the SAA method in other contexts. We demonstrate that in the small-data, large-scale regime, SAA can perform *arbitrarily* poorly.

Example 2.6 (SAA Performs Poorly in the Small-Data, Large-Scale Regime) Fix any $0 < \alpha < 0.075$, and let $\mathcal{X} = \{\boldsymbol{x} \in [0,1]^n : \frac{1}{n}\sum_{j=1}^n x_j \le \alpha\}$. Let $\hat{\mu}_j \sim \mathcal{N}(\mu_j, 1/\nu_j)$, independently, with

$$\mu_j = \begin{cases} 0 & \text{if } j \text{ is odd,} \\ 1 & \text{if } j \text{ is even,} \end{cases} \qquad \nu_j = \begin{cases} 1 & \text{if } j \text{ is odd,} \\ \nu & \text{if } j \text{ is even.} \end{cases}$$

[4] As an aside, we mention that Frees (1991) proves that under mild assumptions, the $\hat{\mu}_j$ in this model are approximately normally distributed.

In words, we would like to identify αn high-reward items using only the noisy estimates $\hat{\mu}_j$. The full-information optimum value is α; that is, $Z_n^*(\boldsymbol{\mu}) = \alpha$.

By inspection, $x_j^{SAA} = \mathbb{I}(\hat{\mu}_j > q^n)\mathbb{I}(\hat{\mu}_j \geq 0)$, where q^n is the $\lfloor n\alpha \rfloor^{\text{th}}$ largest value among $\hat{\mu}_1, \ldots, \hat{\mu}_n$, except possibly for one fractional component. Note q^n is the $(1-\alpha)^{\text{th}}$ quantile of the empirical distribution function: $q^n = \inf\{x : \frac{1}{n}\sum_{j=1}^n \mathbb{I}(\hat{\mu}_j \leq x) \geq 1-\alpha\}$. Because the empirical distribution converges uniformly to the true distribution, the sample quantile q^n also converges to the true quantile (Van der Vaart 2000, Lemma 21.2), i.e., $q^n \to q_\nu$ as $n \to \infty$, where q_ν solves

$$\begin{aligned} \frac{1}{2}\Pr\left\{1 + \frac{\zeta}{\sqrt{\nu}} \leq q_\nu\right\} + \frac{1}{2}\Pr\{\zeta \leq q_\nu\} = 1-\alpha \quad &\iff \quad \frac{1}{2}\Phi((1-q_\nu)\sqrt{\nu}\,) + \frac{1}{2}\Phi(-q_\nu) = \alpha \\ &\iff \quad \Phi((1-q_\nu)\sqrt{\nu}\,) + \Phi(-q_\nu) = 2\alpha\ , \end{aligned}$$

and ζ is a standard normal random variable. Consider the function $f_\nu(q) = \Phi((1-q)\sqrt{\nu}\,) + \Phi(-q)$. The function $f_\nu(\cdot)$ is strictly decreasing with $f_\nu(-\infty) = 2$ and $f_\nu(\infty) = 0$. Note that $f_\nu(1.01) = \Phi(-0.01\sqrt{\nu}) + \Phi(-1.01) \geq \Phi(-1.01) > 0.1562 > 2\alpha$, where the last inequality follows because $\alpha < 0.075$. Since ν is arbitrary, this means that $q_\nu > 1.01$ for all $\nu > 0$.

Thus, the SAA performance satisfies

$$\frac{1}{n}\sum_{j=1}^n \mu_j x_j^{SAA} = \frac{1}{2}\frac{2}{n}\sum_{k=1}^{n/2} \mathbb{I}(\hat{\mu}_{2k} \geq q^n)\mathbb{I}(\hat{\mu}_j \geq 0) \ \to_{a.s.}\ \frac{1}{2}\Phi((1-q_\nu)\sqrt{\nu}\,), \qquad \text{as } n \to \infty.$$

Note that $0 \leq \frac{1}{2}\Phi((1-q_\nu)\sqrt{\nu}) \leq \frac{1}{2}\Phi(-0.01\sqrt{\nu})$ because $q_\nu > 1.01$. Therefore, as $\nu \to \infty$, the above limit converges to zero. Thus, in the small-data, large-scale limit with large enough ν, the SAA solution will have performance close to 0. For comparison, randomly choosing αn of the indices has expected performance $\alpha/2 > 0$. Moreover, even for finite n and reasonably small ν, the performance can be quite bad. For example, when $\nu = 2$, $n = 100$, and $\alpha = .05$, SAA achieves about 80% of the full-information optimum, while our proposed method from Section 4 achieves about 98%.

It may not be surprising that SAA performs badly in the previous example because it does not leverage any information about the ν_j. Since the odd items have a lower precision (higher variance) than the even items, the $\hat{\mu}_j$ with odd j frequently appear better than the $\hat{\mu}_j$ with even j, despite having a lower mean. Other authors have observed similar poor performance of SAA in finite sample, inspiring techniques that utilize the ν_j, such as robust optimization and regularization, but these approaches typically do not directly leverage the scale of the optimization. In Section 5, we discuss refinements of these approaches for the small-data, large-scale regime.

2.2. Full-Information Performance is Unachievable in Small-Data, Large-Scale Regime

The main result of this section is as follows:

Theorem 2.7 (Unattainability of Full-Information Optimality) *Let* $\boldsymbol{x}(\hat{\boldsymbol{\mu}})$ *denote* any *data-driven policy. Then, for each* $n \geq 2$*, there exist instances of* P^n *with* $\mathcal{X} = [0,1]^n$*,* $\boldsymbol{\mu} \in \{-1,+1\}^n$*,* $\hat{\mu}_j \sim \mathcal{N}(\mu_j, 1)$ *for all* j*, and* $\hat{\mu}_1, \ldots, \hat{\mu}_n$ *are independent such that*

$$\frac{\mathbb{E}\left[\frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\mu}})\right]}{Z^*(\boldsymbol{\mu})} \;<\; 0.842.$$

This upper bound is not tight. Nonetheless, a consequence of the theorem is that given any data-driven policy $\boldsymbol{x}(\cdot)$, one can construct a sequence of instances $\{(\mathrm{P}^n, \hat{\boldsymbol{\mu}}^n, \boldsymbol{\nu}^n) : n \geq 2\}$, in the small-data, large-scale regime such that the expected performance of $\boldsymbol{x}(\cdot)$ is at most a constant fraction of the full-information optimal performance for each instance.

See Appendix B for a proof. Loosely, the proof proceeds by generating random instances of P^n by sampling $\boldsymbol{\mu}$ uniformly from $\{-1,+1\}^n$. The worst-case expected performance of $\boldsymbol{x}(\hat{\boldsymbol{\mu}})$ across these instances is bounded by its average expected performance across these instances. To complete the theorem, we compute an upper bound on this average expected performance. We note that the constant 0.842 arises from the way in which we generate the random instance in the proof. In particular, a different sampling procedure for $\boldsymbol{\mu} \in \{-1,+1\}^n$ might yield tighter upper bounds.

3. Selecting Policies in the Small-Data, Large-Scale Regime

In light of Theorem 2.7, constructing policies that consistently achieve full-information optimal performance in the small-data, large-scale regime is impossible. Consequently, we restrict attention to a class of policies and focus on selecting a member that performs nearly as well as an oracle policy that knows the true value of $\boldsymbol{\mu}$ in advance, but is constrained to use a policy within the class. If the class of policies is sufficiently rich, we expect this oracle policy, and, hence, our selected policy, to have good practical performance in applications. Specifically, let $\boldsymbol{x}(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}})$ be a data-driven policy indexed by a parameter $\boldsymbol{\theta} \in \Theta$, and let $\mathcal{X}^\Theta(\hat{\boldsymbol{\mu}}) = \{\boldsymbol{x}(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}}) \in \mathcal{X} : \boldsymbol{\theta} \in \Theta\}$ be a set of such policies.

Definition 3.1 (Oracle Policy) Let

$$\boldsymbol{\theta}^{\mathsf{OR}} = \boldsymbol{\theta}^{\mathsf{OR}}(\mathrm{P}^n, \hat{\boldsymbol{\mu}}, \boldsymbol{\nu}) \in \arg\max_{\boldsymbol{\theta} \in \Theta} \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}})$$

We define the policy $\hat{\boldsymbol{\mu}} \mapsto \boldsymbol{x}\left(\boldsymbol{\theta}^{\mathsf{OR}}, \hat{\boldsymbol{\mu}}\right)$ to be an *oracle policy* for Problem P^n and policy class $\mathcal{X}^\Theta(\hat{\boldsymbol{\mu}})$.

Note that $\boldsymbol{\theta}^{\mathsf{OR}}$ is random because it depends on $\hat{\boldsymbol{\mu}}$. The performance of every policy in $\mathcal{X}^\Theta(\hat{\boldsymbol{\mu}})$ is bounded above by the performance of $\boldsymbol{x}(\boldsymbol{\theta}^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ almost surely in $\hat{\boldsymbol{\mu}}$; in this sense, the oracle policy serves as a benchmark. However, $\boldsymbol{x}(\boldsymbol{\theta}^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ is not a "valid" data-driven policy because identifying $\boldsymbol{\theta}^{\mathsf{OR}}$ requires knowing $\boldsymbol{\mu}$; in this sense, it is an "oracle." Indeed, it is not obvious that there exist data-driven policies which attain this benchmark without knowing $\boldsymbol{\mu}$.

We introduce two example policy classes that we study throughout the remainder:

Example 3.2 (Bayes-Inspired Policies) In Section 4, we use a Bayesian argument to motivate the class of policies

$$\mathcal{X}^{Bayes}(\hat{\boldsymbol{\mu}}) \equiv \big\{\boldsymbol{x}(\tau,\hat{\boldsymbol{\mu}}) \in \mathcal{X} : \tau \geq 0\big\}, \quad \text{where} \quad \boldsymbol{x}(\tau,\hat{\boldsymbol{\mu}}) \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\sum_{j=1}^{n} \frac{\nu_j}{\nu_j+\tau}\hat{\mu}_j x_j. \tag{3.1}$$

To avoid ambiguity, we index these policies by τ instead of $\boldsymbol{\theta}$. The oracle policy $\boldsymbol{x}(\tau^{\mathsf{OR}},\hat{\boldsymbol{\mu}})$ is defined by $\tau^{\mathsf{OR}} \in \arg\max_{\tau\geq 0} \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{\mu}^\top \boldsymbol{x}(\tau,\hat{\boldsymbol{\mu}})$. We discuss properties of this oracle policy in Section 4.

Example 3.3 (Regularization Policies) In Section 5 we motivate and study the class of policies

$$\mathcal{X}^{Reg}(\hat{\boldsymbol{\mu}}) \equiv \big\{\boldsymbol{x}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}) : \Gamma \in [\Gamma_{\mathsf{min}},\Gamma_{\mathsf{max}}]\big\}, \quad \text{where} \quad \boldsymbol{x}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}) \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x} - \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{2n}\sum_{j=1}^{n}\frac{x_j^2}{\nu_j}. \tag{3.2}$$

Here, $\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}$ are user-specified parameters. Again, to avoid ambiguity, we index this policy class by Γ instead of $\boldsymbol{\theta}$. Its oracle policy $\boldsymbol{x}^{\mathcal{R}}(\Gamma^{\mathsf{OR}},\hat{\boldsymbol{\mu}})$ is defined by $\Gamma^{\mathsf{OR}} \in \arg\max_{\Gamma\in[\Gamma_{\mathsf{min}},\,\Gamma_{\mathsf{max}}]} \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}})$.

Viewing P^n through the lens of a policy class also clarifies why SAA sometimes performs poorly in the small-data, large-scale regime.

Example 3.4 (SAA Revisited) Suppose $\boldsymbol{x}^{SAA} \in \mathcal{X}^{\Theta}(\hat{\boldsymbol{\mu}})$ almost surely, i.e., there exists $\boldsymbol{\theta}^{SAA} \in \Theta$ such that $\boldsymbol{x}(\boldsymbol{\theta}^{SAA},\hat{\boldsymbol{\mu}}) = \boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})$. Observe that $\boldsymbol{\theta}^{SAA} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})$ because

$$\hat{\boldsymbol{\mu}}^\top \boldsymbol{x}(\boldsymbol{\theta}^{SAA},\hat{\boldsymbol{\mu}}) \;=\; \hat{\boldsymbol{\mu}}^\top \boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}}) \;=\; \max_{\boldsymbol{x}\in\mathcal{X}} \hat{\boldsymbol{\mu}}^\top \boldsymbol{x} \;\geq\; \sup_{\boldsymbol{\theta}\in\Theta} \hat{\boldsymbol{\mu}}^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}),$$

where the last inequality follows because $\boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}) \in \mathcal{X}$ for all $\boldsymbol{\theta} \in \Theta$.

Thus, SAA can be seen as finding the policy member that maximizes $\frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})$. By contrast, the oracle policy seeks to maximize the true reward $\frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})$. Because of the dependence of $\boldsymbol{x}(\theta,\hat{\boldsymbol{\mu}})$ on $\hat{\boldsymbol{\mu}}$, the SAA objective is biased. In the large-sample regime, this bias often vanishes asymptotically. However, in the small-data, large-scale regime, the bias is non-negligible, partially explaining SAA's poor performance. We emphasize that if the policy class is large enough that $\boldsymbol{x}^{SAA} \in \mathcal{X}^{\Theta}(\hat{\boldsymbol{\mu}})$ almost surely, its oracle policy necessarily performs at least as well as the SAA.

3.1. Our Bias-Corrected Policy

The main result of this section is a general framework for selecting a member of $\mathcal{X}^{\theta}(\hat{\boldsymbol{\mu}})$ whose performance is "comparable" to $\boldsymbol{x}(\boldsymbol{\theta}^{\mathsf{OR}},\hat{\boldsymbol{\mu}})$, in the sense that we expect under mild conditions the suboptimality gap to be small in the small-data, large-scale regime. The key idea is to construct an approximation to $\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})$ that does not require knowing $\boldsymbol{\mu}$, and, then, to select $\hat{\boldsymbol{\theta}}$ to optimize this approximation.

One such approximation might be $\hat{\boldsymbol{\mu}}^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})$ because $\hat{\boldsymbol{\mu}}$ is an unbiased estimator of $\boldsymbol{\mu}$. However, as noted in Example 3.4, this naive approximation is biased and should be corrected. Our bias-correction is inspired by Stein's Lemma (Ross et al. 2011), which states that for any (differentiable)

function $f(\cdot)$ and $\zeta \sim \mathcal{N}(\mu, 1/\nu)$, we have $\mathbb{E}[(\zeta-\mu)f(\zeta)] = 1/\nu \cdot \mathbb{E}\left[\frac{\partial}{\partial\zeta}f(\zeta)\right]$. Considering the j^{th} element of the bias $\mathbb{E}\left[(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})\right]$, this roughly suggests

$$\mathbb{E}[(\hat{\mu}_j-\mu_j)x_j(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})] \;\approx\; \frac{1}{\nu_j}\mathbb{E}\left[\frac{\partial}{\partial\hat{\mu}_j}x_j(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})\right] \;\approx\; \frac{1}{2h\sqrt{\nu_j}}\mathbb{E}\left[x_j(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}+h\boldsymbol{e}_j/\sqrt{\nu_j}) - x_j(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}-h\boldsymbol{e}_j/\sqrt{\nu_j})\right],$$

where $\boldsymbol{e}_j$ is the standard unit vector in the j^{th} direction, and $0<h<1$ is a user-defined bandwidth parameter. For clarity, the last line approximates the derivative by a first-order finite difference with step $h/\sqrt{\nu_j}$, and the derivative in the second line may not be well-defined since the solution $\boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})$ need not be differentiable (or even continuous). Lemma C.2 in the appendix resolves these issues by generalizing Stein's Lemma to approximately Gaussian random variables and approximate derivatives.

In any case, this heuristic argument suggests the bias-correction

$$B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}}) \;\equiv\; \frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\left[x_j\Big(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}+\frac{h}{\sqrt{\nu_j}}\boldsymbol{e}_j\Big) - x_j\Big(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}-\frac{h}{\sqrt{\nu_j}}\boldsymbol{e}_j\Big)\right] .$$

Approximating $\frac{1}{n}\boldsymbol{\mu}^\top\boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})$ by $\frac{1}{n}\hat{\boldsymbol{\mu}}^\top\boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}) - B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}})$, we let

$$\hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\hat{\boldsymbol{\mu}}^\top\boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}) - B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}}). \tag{3.3}$$

3.2. Performance in the Small-Data, Large-Scale Regime

The suboptimality of $\boldsymbol{x}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\mu}})$ with respect to $\boldsymbol{x}(\boldsymbol{\theta}^{\mathsf{OR}},\hat{\boldsymbol{\mu}})$ will depend on the quality of the above approximations, which in turn depends on the policy class $\mathcal{X}^\Theta(\hat{\boldsymbol{\mu}})$ and the estimator $\hat{\boldsymbol{\mu}}$. We next prove a general purpose bound that provides intuition into the types of policy classes and estimators for which this suboptimality gap is likely small in the small-data, large-scale regime. The key idea is that our bias-corrected criteria for $\hat{\boldsymbol{\theta}}$ is very close to the true criteria that defines $\boldsymbol{\theta}^{\mathsf{OR}}$ in this regime.

Theorem 3.5 (A General Bound on the Sub-Optimality Gap) *Suppose Assumption 2.1 holds and there exists σ^2 such that for each j, the random variable $(\hat{\mu}_j-\mu_j)\sqrt{\nu}_j$ has a density $\phi_j(\cdot)$ and is sub-Gaussian with variance proxy at most σ^2.*[5] *Then,*

$$\underbrace{\frac{1}{n}\boldsymbol{\mu}^\top\boldsymbol{x}(\boldsymbol{\theta}^{\mathsf{OR}},\hat{\boldsymbol{\mu}}) - \frac{1}{n}\boldsymbol{\mu}^\top\boldsymbol{x}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\mu}})}_{\text{Sub-Optimality Gap}}$$
$$\leq \underbrace{\frac{4(h^{-1}+24\sigma^2)}{\sqrt{\nu_{\min}}}\times\mathsf{TV}\times\log\Big(\frac{e}{\mathsf{TV}}\Big)}_{\text{Degree of Non-Normality}} + \underbrace{\frac{4h^2}{\sqrt{\nu_{\min}}}}_{\text{Derivative Approximation}}$$
$$+\underbrace{\sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}\Big|(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top\boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}) - \mathbb{E}[(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top\boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})]\Big| + \sup_{\boldsymbol{\theta}\in\Theta}\Big|B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}}) - \mathbb{E}[B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}})]\Big|}_{\text{Maximal Stochastic Deviations}},$$

[5] Recall, a mean-zero random variable ξ is sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[\exp(t\xi)] \leq \exp(t^2\sigma^2/2)$ for all $t\in\mathbb{R}$ (Wainwright 2015). Estimators are frequently modeled as sub-Gaussian.

where $\mathsf{TV} \equiv \frac{1}{2n}\sum_{j=1}^{n} \|\phi_j - \phi\|_1$ *is a number between zero and one measuring the average total variation distance between* $\phi_j(\cdot)$ *and the standard normal density* $\phi(\cdot)$.

Intuitively, Theorem 3.5 suggests $\boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})$ performs comparably to $\boldsymbol{x}(\boldsymbol{\theta}^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ in the small-data, large-scale regime whenever 1) the $\hat{\mu}_j$ are nearly gaussian, 2) the $\hat{\mu}_j$ are sufficiently independent and 3) $\mathcal{X}^{\Theta}(\hat{\boldsymbol{\mu}})$ is not "too complex" a function class, as measured, e.g., by its VC-dimension.

To see this, note first that Theorem 3.5 decomposes the suboptimality gap into three terms. The first term is a deterministic constant measuring the degree of non-normality in the $\hat{\mu}_j$ and, notably, does not depend on the instance P^n. It is small whenever TV is small. Moreover, when each $\hat{\mu}_j$ is Gaussian, then $(\hat{\mu}_j - \mu_j)\sqrt{\nu}_j$ is a standard normal random variable with $\phi_j(\cdot) = \phi(\cdot)$, so $\mathsf{TV} = 0$ and the first term on the right-hand-side of the theorem vanishes.[6] The total variation TV thus measure the average degree of non-normality of the $\hat{\mu}_j$. In particular, TV will be small when each $\hat{\mu}_j$ is approximately Gaussian even if $\hat{\boldsymbol{\mu}}$ is *not* jointly multivariate Gaussian.

The second term is also a deterministic constant stemming from our Approximate Stein Lemma (cf. Lemma C.2 in the appendix) and can be made small through a suitable choice of bandwidth h. We discuss this choice in more detail below.

The third term, consisting of the two suprema, is more subtle and is a random variable. The argument of each suprema can be written as the average of n, mean-zero random variables. Intuition suggests that for a fixed $\boldsymbol{\theta}$, these terms will concentrate at zero (their mean) *provided that* the $\hat{\mu}_j$ are sufficiently independent as $n \to \infty$.[7] If, additionally, the policy class $\mathcal{X}^{\Theta}(\hat{\boldsymbol{\mu}})$ is not too complex, i.e., it has low VC-dimension or metric entropy, this concentration occurs uniformly for all $\theta \in \Theta$ (see, e.g., Pollard (1990) or Van der Vaart (2000)). Importantly, this convergence holds as $n \to \infty$, even if the ν_j remain bounded, i.e., it holds in the small-data, large-scale regime. Commonly used policy classes in optimization, including those from Sections 4 and 5, often have low metric entropy, suggesting this convergence holds for a wide variety of policy classes.

Stepping back, if the $\hat{\mu}_j$ are approximately gaussian, and sufficiently independent, then we expect, intuitively, that all three terms in the suboptimality bound will be small in the small-data, large-scale regime for many policy classes. Making this intuition formal requires a careful analysis of the dependency in $x_j(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}})$ across j and computing the "complexity" of $\mathcal{X}^{\Theta}(\hat{\boldsymbol{\mu}})$. We carry out this analysis in Sections 4 and 5 for our Bayes-Inspired and Regularization policy classes in the special case of linear optimization problems with independent $\hat{\mu}_j$. This analysis confirms these policies have near-oracle performance in the small-data, large-scale regime in this setting.

[6] We adopt the usual convention that $0 \cdot \log 0 = 0$ by continuity.

[7] For example, standard central-limit-theorem results still hold when the summands are mildly dependent. Intuitively, similar phenomena should hold here.

Theorem 3.5 also highlights the tradeoff in specifying the bandwidth h. As $h \to 0$, $\sup_{\boldsymbol{\theta}\in\Theta} \left| B(\boldsymbol{\theta}, h, \hat{\boldsymbol{\mu}}) - \mathbb{E}[B(\boldsymbol{\theta}, h, \hat{\boldsymbol{\mu}})] \right|$ grows because $B(\boldsymbol{\theta}, h, \hat{\boldsymbol{\mu}})$ scales with $\frac{1}{h}$, and if $\mathsf{TV} \neq 0$, the "Degree of Non-Normality term" also grows. Thus, a good bandwidth must balance between these two terms and the error from derivative approximation in Theorem 3.5. One heuristic for selecting a bandwidth when $\mathsf{TV} = 0$ might be to select h large enough that $\boldsymbol{\theta} \mapsto B(\boldsymbol{\theta}, h, \hat{\boldsymbol{\mu}})$ is smooth in $\boldsymbol{\theta}$, since ideally the above suprema is small and $\boldsymbol{\theta} \mapsto \mathbb{E}[B(\boldsymbol{\theta}, h, \hat{\boldsymbol{\mu}})]$ is a smooth function of $\boldsymbol{\theta}$. Sections 4 and 5 provide more precise guidance on selecting a good bandwidth for those classes.

A potential drawback of our method (cf. Eq. (3.3)) is that computing $B(\boldsymbol{\theta}, h, \hat{\boldsymbol{\mu}})$ for each $\boldsymbol{\theta}$ seemingly involves solving $2n$ instances of P^n to determine $\boldsymbol{x}\left(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}} + \frac{h}{\sqrt{\nu_j}}\boldsymbol{e}_j\right)$ and $\boldsymbol{x}\left(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}} - \frac{h}{\sqrt{\nu_j}}\boldsymbol{e}_j\right)$ for each j. (See, e.g., our example policy classes above.) For large n, this may be prohibitive. In Sections 4 and 5, we utilize the structure of the policy class to circumvent this issue, developing a more computationally efficient bias correction.

In summary, although Theorem 3.5 provides a framework for analyzing (cf. Eq. (3.3)), at this level of generality, its usefulness is primarily foundational. It highlights the key conditions on the estimator and policy class required for good performance, and, also provides the "roadmap" for our more specialized analysis of specific policy classes Sections 4 and 5. We expect future research might also leverage Theorem 3.5 to study alternate optimization problems and policy classes by showing the relevant maximal stochastic deviations are small.

3.3. Performance in the Large-Sample Regime

Before specializing and strengthening Theorem 3.5, we prove that, despite being motivated by the small-data, large-scale regime, $\hat{\boldsymbol{\theta}}$ has good theoretical performance in the large-sample regime. The key insight is that the bias-correction is small in the large-sample regime. Thus, $\boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})$ approximately optimizes the SAA objective provided $X_n^\Theta(\hat{\boldsymbol{\mu}})$ contains $x^{SAA}(\hat{\boldsymbol{\mu}})$ as a member.

LEMMA 3.1 **(Bound to SAA Performance)**. *Suppose $\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}}) \in \mathcal{X}^\Theta(\hat{\boldsymbol{\mu}})$ almost surely. Then,*

$$0 \leq \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \left(\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}}) - \boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})\right) \leq \frac{1}{h\nu_{\mathsf{min}}}, \qquad \textit{where } \nu_{\mathsf{min}} = \min_{j=1,\ldots,n} \nu_j.$$

Under fairly general assumptions, SAA solutions perform comparably to the full-information optimum in the large-sample regime, i.e., as ν_{min} grows large. We use Lemma 3.1 to prove our policy also performs well in these settings:

Theorem 3.6 (Bound to Full-Information Optimum) *Suppose $\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}}) \in \mathcal{X}^\Theta(\hat{\boldsymbol{\mu}})$ almost surely. Then, under Assumption 2.1,*

$$\mathbb{E}\left[\left|Z^*(\boldsymbol{\mu}) - \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})\right|\right] \leq \frac{2 + h^{-1}}{\sqrt{\nu_{\mathsf{min}}}}.$$

Theorem 3.6 is a non-asymptotic result but is perhaps best understood by considering a sequence of instances of $\{(\mathrm{P}^n, \hat{\boldsymbol{\mu}}^{\mathsf{S}}, \boldsymbol{\nu}^{\mathsf{S}}) : \mathsf{S} \geq 1\}$ (indexed by S) in the large-sample regime. Since the bandwidth h scales with n, it is fixed in the large-sample regime, but $\nu^{\mathsf{S}}_{\mathsf{min}} \to \infty$ as $\mathsf{S} \to \infty$. Consequently, Theorem 3.6 proves $\boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})$ converges to full-information optimal performance in the large-sample regime, even if $\hat{\boldsymbol{\mu}}$ are potentially highly non-normal. We summarize these ideas in the following corollary, which proves that our policy also performs well in the large-sample regime.

Corollary 3.7 (Best of Both Regimes) *Consider a sequence $\{(P^n, \hat{\boldsymbol{\mu}}^{\mathsf{S}}, \boldsymbol{\nu}^{\mathsf{S}}) : \mathsf{S} \geq 1\}$ of instances in the large-sample regime, each satisfying Assumption 2.1. For each P^n, let $\hat{\boldsymbol{\theta}}^{\mathsf{S}}$ be a solution to Eq.* (3.3)*. Assume for every S that $\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}}^{\mathsf{S}}) \in \mathcal{X}^{\Theta}(\hat{\boldsymbol{\mu}}^{\mathsf{S}})$, almost surely. Then, as $\mathsf{S} \to \infty$, the performance $\frac{1}{n}\boldsymbol{\mu}^{\top}\boldsymbol{x}(\hat{\boldsymbol{\theta}}^{\mathsf{S}}, \hat{\boldsymbol{\mu}}^{\mathsf{S}})$ converges in $\mathcal{L}_1$ to the full-information optimum Z^*.*

Corollary 3.7 provides a strong argument in favor of the policy $\boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})$. Indeed, a potential criticism of our approach is that it limits attention to the class of policies $\mathcal{X}^{\Theta}(\hat{\boldsymbol{\mu}})$, which may not contain the full-information optimum, while SAA (and similar approaches) is essentially nonparametric, considering all possible policies. Theorem 3.6 and Corollary 3.7 prove that despite restricting to a policy class, our proposed policy achieves performance comparable to the full-information optimum in the large-sample regime, provided the policy class is rich enough to contain $\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})$. We consider this a fairly minor condition, and is satisfied, for example by $\mathcal{X}^{Bayes}(\hat{\boldsymbol{\mu}})$ and $\mathcal{X}^{Reg}(\hat{\boldsymbol{\mu}})$ when $\Gamma_{\mathsf{min}} = 0$. Moreover, if, as in Example 2.2, $\nu_j = O(\mathsf{S})$ for all j, then the rate of convergence to full-information optimum is is $O\big(1/\sqrt{\mathsf{S}}\big)$, which is comparable to the convergence rate of the SAA solution (Ruszczynski and Shapiro 2003, Kleywegt et al. 2002).

4. Bayes-Inspired Policies over Polyhedral Feasible Regions

In this section, we specialize and strengthen Theorem 3.5 for the policy class in Example 3.2 with a polyhedral feasible region. We first motivate this policy class.

Consider a Bayesian approach to the decision problem $(\mathrm{P}^n, \hat{\boldsymbol{\mu}}, \boldsymbol{\nu})$. In this approach, we first assume the unknown $\boldsymbol{\mu}$ is drawn as a realization from a *known* prior distribution π, e.g., we might assume $\mu_j \sim \mathcal{N}(0, 1/\tau_0)$ for some $\tau_0 \geq 0$, independently across j . We then assume the likelihood $\hat{\boldsymbol{\mu}} \,|\, \boldsymbol{\mu}$ follows a known likelihood distribution, e.g., that $\hat{\mu}_j \,|\, \mu_j \sim \mathcal{N}(\mu_j, 1/\nu_j)$, independently across j. In principle, we can then compute the Bayes-optimal policy with respect to π using the data $\hat{\boldsymbol{\mu}}$. For Gaussian priors and likelihoods, conjugacy gives the posterior mean $\mathbb{E}^{\pi}[\mu_j \,|\, \hat{\boldsymbol{\mu}}] = \frac{\nu_j}{\nu_j + \tau_0}\hat{\mu}_j$, and it follows that $x(\tau_0, \hat{\boldsymbol{\mu}})$ defined in Example 3.2 is a Bayes-optimal policy.

Bayes policies enjoy strong performance guarantees. For example, Bayes policies are always admissible, i.e., no other data-driven policy pareto-dominates a Bayes optimal policy (Berger 2013).

Under mild assumptions, Bayes optimal policies also form an essentially complete class; that is, any data-driven policy is weakly dominated by a Bayes optimal policy for some prior (Berger 2013).

Nonetheless, Bayes policies require strong assumptions, e.g., that one knows the precise form of the prior. We can partially mitigate these drawback by not fixing a prior, but instead considering a class of policies, each of which is Bayes optimal for some prior, and then seeking its oracle member. Indeed, this perspective motivates our policy class $\mathcal{X}^{Bayes}(\hat{\boldsymbol{\mu}})$, i..e, it is the set of Bayes-optimal policies when the μ_j are drawn from some mean-zero Gaussian prior, and the $\hat{\mu}_j \,|\, \mu_j$ are Gaussian, justifying our terminology "Bayes-Inspired."[8]

We argue that even if one does not adopt the Bayesian perspective, i.e., one does *not* believe $\boldsymbol{\mu}$ is drawn from some mean-zero, Gaussian prior, $\mathcal{X}^{Bayes}(\hat{\boldsymbol{\mu}})$ is still a rich and interesting class of policies for two reasons. First, $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ has an intuitive structure. Each estimate $\hat{\mu}_j$ is shrunk towards zero. Components with lower precision are shrunk more aggressively because if $\nu_j < \nu_k$, then $\frac{\nu_j}{\nu_j+\tau} \leq \frac{\nu_k}{\nu_k+\tau}$. The parameter τ controls the degree of shrinkage.

Second, a host of popular estimators for $\boldsymbol{\mu}$ have the general form $\left(\frac{\nu_j}{\nu_j+\tau}\hat{\mu}_j \,:\, j=1,\ldots,n\right)$ for some data-driven value of τ when $\hat{\boldsymbol{\mu}}$ is Gaussian. These estimators often have provably good theoretical properties, and exhibit good empirical performance in applications, even when the underlying Gaussian assumptions might be violated; see the references below. The class $\mathcal{X}^{Bayes}(\hat{\boldsymbol{\mu}})$ thus contains the corresponding "estimate-then-optimize" policy for each of these estimators.[9] Consequently, $\boldsymbol{x}(\tau^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ must perform at least as well as these estimate-then-optimize policies. We consider this to be a strong argument for good performance of the oracle policy in a broad range of applications. Examples of such estimators and their properties when $\hat{\boldsymbol{\mu}}$ is Gaussian include:

- **Maximum Likelihood Estimation:** The value $\tau = 0$ yields the estimator $\hat{\boldsymbol{\mu}}$, which is the maximum likelihood estimator for $\boldsymbol{\mu}$ and also the minimum variance unbiased estimator. Moreover, by construction, $\boldsymbol{x}(0, \hat{\boldsymbol{\mu}})$ equals the sample average approximation $\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})$ because $\boldsymbol{x}(0, \hat{\boldsymbol{\mu}}) = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x}$. Since the SAA policy is a member of $\mathcal{X}^{Bayes}(\hat{\boldsymbol{\mu}})$, it follows that its performance must be no better than that of the oracle policy. In fact, as shown in Example 4.1, its performance can be much worse.
- **James-Stein Shrinkage:** The value $\tau^{JS} = 1 - \left(1 - \frac{n-2}{\|\hat{\boldsymbol{\mu}}\|_2^2}\right)^{-1}$ yields the James-Stein estimator. When $\nu_j = 1$ for all $j = 1,\ldots,n$ and $n \geq 3$, Stein (1981) proved that this biased estimate has smaller mean-squared error than $\hat{\boldsymbol{\mu}}$.

[8] Of course, since we treat $\boldsymbol{\mu}$ as an unknown constant and not a realization from a prior in our analysis, there is technically no notion of a Bayes-optimal policy, hence necessitating the qualifier "Inspired."

[9] For clarity, "estimate-then-optimize" policies are policies computed by first estimating $\boldsymbol{\mu}$ using some statistical criteria (e.g., maximum likelihood estimation), and then plugging in that estimator for $\boldsymbol{\mu}$ in P^n and solving.

- **Empirical Bayes:** The values τ^{MM} and τ^{MLE}, defined by

$$\tau^{MM} \equiv \left(\frac{1}{n}\sum_{j=1}^{n}\left(\hat{\mu}_j^2 - 1/\nu_j\right)\right)^{-1} \quad \text{and} \quad \tau^{MLE} \text{ is a solution to } \sum_{j=1}^{n}\frac{\hat{\mu}_j^2 - 1/\tau^{MLE} - 1/\nu_j}{(1/\tau^{MLE}+1/\nu_j)^2} \;=\; 0,$$

  correspond, respectively, to the empirical Bayes moment-matching and empirical Bayes maximum likelihood estimators (Xie et al. 2012)[10]. Generally speaking, parametric empirical Bayes methods assume that the data were generated from a Bayesian hierarchical model and then use the marginal distribution of the data to fit parameters of the prior. For our specific case, this amounts to assuming that μ_j is a realization of a mean-zero Gaussian prior, and then using either the method of moments or maximum-likelihood estimation on $\hat{\boldsymbol{\mu}}$ to estimate τ; see Xie et al. (2012) for details. If the assumed hierarchical model is actually valid, the two methods both converge in performance as $n \to \infty$ to the Bayes optimal procedure.
- **SURE Estimation:** Finally, Xie et al. (2012) propose the choice τ^{SURE}, which solves[11]

$$\sum_{j=1}^{n}\left(\frac{\nu_j^{-2}\hat{\mu}_j^2}{(1/\nu_j+1/\tau^{SURE})^3} - \frac{\nu_j^{-2}}{(1/\nu_j+1/\tau^{SURE})^2}\right) = 0.$$

  Xie et al. (2012) prove that as $n \to \infty$, this estimator achieves the minimum mean-squared error among all estimators of the form $\left(\frac{\nu_j}{\nu_j+\tau}\hat{\mu}_j \;:\; j=1,\ldots,n\right)$ almost surely.

By construction, $\boldsymbol{x}(\tau^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ is no worse than the above estimate-then-optimize policies, *whether or not $\boldsymbol{\mu}$ is a realization from a Gaussian prior*.

A potential criticism of estimate-then-optimize policies is that the estimation phase involves a purely statistical criterion that is agnostic to the down-stream optimization (Elmachtoub and Grigas 2017, Liyanage and Shanthikumar 2005). By contrast, $\boldsymbol{x}(\tau^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ does leverage optimization structure. Consequently, as shown in the next example, it may perform strictly better than these estimate-then-optimize policies.

Example 4.1 (Benefits over Estimate-Then-Optimize) Consider the ranking problem from Example 2.6 with $\alpha = 0.05$, $\hat{\mu}_j \sim \mathcal{N}(\mu_j, 1/\nu_j)$ independently across j and

$$(\mu_j, \nu_j) = \begin{cases} (0.0,\ 0.1) & \text{if} \quad 1 \le j \le \lfloor n/3 \rfloor, \\ (0.3,\ 4.0) & \text{if} \quad \lfloor n/3 \rfloor < j \le \lfloor 2n/3 \rfloor, \\ (1.0,\ 1.0) & \text{if} \quad \lfloor 2n/3 \rfloor < j \le n. \end{cases}$$

We have three types of items in roughly equal proportions: "Low" items have low value ($\mu_j = 0$) and low precisions ($\nu_j = 0.1$), so their noisy estimates frequently make them appear attractive. "Medium" items have medium value and very high precision ($\mu_j = 0.3$ and $\nu_j = 4$). "High" items

[10] If the equation has no solution, take $\tau^{MLE} = 0$.

[11] If the equation has no solution, take $\tau^{SURE} = 0$.

have good value and medium precision, with $\mu_j = 1$ and $\nu_j = 1$. Note that these values of μ_j do not resemble a draw from a Gaussian prior, and that a good solution will contain many High items.

The first panel of Figure 2 shows the performance of $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ for varying τ, a single realization of $\hat{\boldsymbol{\mu}}$, and $n = 512$ and $n = 2^{17}$. The second panel shows the performance of a subset of the above estimate-then-optimize policies, $\boldsymbol{x}(\tau^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$, and our proposed policy $\boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})$ (defined in the next section). See Appendix F for a larger figure including all the above estimate-then-optimize policies.

The example highlights the tradeoffs induced by τ. Note that for any τ, $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ corresponds to ranking the rewards $\frac{\nu_j}{\nu_j+\tau}\hat{\mu}_j$ and selecting the top 5% of items. When $\tau = 0$, we have no shrinkage, the Low items appear very attractive, and $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ chooses many Low items. This is illustrated in Figure 3(a), which shows the distribution of the reward and that "Low" items (in red) comprise the tail of the distribution. If τ is too large, such as $\tau = 4.94$, there is "too much" shrinkage, and the Medium items (in green) comprise the tail of the distribution, as shown in Figure 3(c). For intermediate values of τ, say $\tau = 0.577$, High items appear the most attractive, c.f., Figure 3(b).

This tradeoff in turn drives the performance of the various methods. Indeed, in Figure 3(d), we see that for this example, τ^{OR} is fairly small but strictly positive. Hence, SAA ($\tau = 0$) shrinks too little, and the estimate-then-optimize policies generally shrink too much. More specifically, the estimate-then-optimize policies seek τ values that yield low mean-squared error, a statistical objective. Because of the form of our optimization problem, however, lower mean-squared error does not translate into better performance.

By contrast, we observe that our proposed policy $\boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})$ (defined in the next section) converges in performance to $\boldsymbol{x}(\tau^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ as $n \to \infty$. As we will see in the next section, this policy *does* explicitly leverage the optimization structure.

4.1. A Near-Oracle Policy for the Bayes-Inspired Class

Instead of applying Theorem 3.5 verbatim to bound the suboptimality, we utilize the special structure of $\mathcal{X}^{Bayes}(\hat{\boldsymbol{\mu}})$ to construct a more computationally efficient bias correction and focus on a special case of P^n to develop further insight. Specifically, we will assume that $\mathcal{X}$ is polyhedral, i.e.,

$$\mathcal{X} = \left\{ \boldsymbol{x} \in [0,1]^n : \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{A}_j x_j \leq \boldsymbol{b} \right\}, \tag{4.1}$$

where $\boldsymbol{b} \in \mathbb{R}^m$ is an arbitrary vector, so that P^n represents an arbitrary linear optimization problem over a bounded feasible region. Linear optimization is a fundamental class of problems subsuming many applications. Any linear optimization with a bounded feasible region can be written in the form of Problem P^n, after potentially scaling and shifting to place the feasible region in the unit box $[0,1]^n$. In particular, for a fixed n, scaling the constraints by $1/n$ is without loss of generality, and our results below are stated for fixed n.

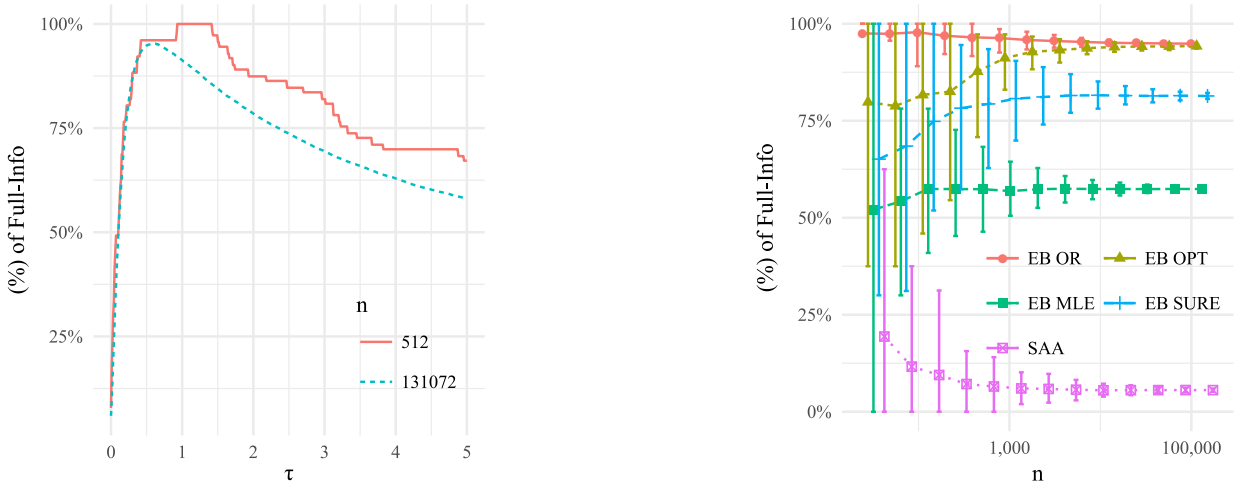


Figure 2 **Benefits over Estimate-Then-Optimize in Example 4.1.** The left panel shows the performance of $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ for varying τ, $n = 2^{17} = 131,072$ (dotted blue line) and $n = 2^9 = 512$ (solid red line). The function is discontinuous but becomes smoother as $n \to \infty$. The right panel shows the performance of various data-driven policies from $\mathcal{X}^{Bayes}(\hat{\boldsymbol{\mu}})$ for varying n: "SAA" is the sample average approximation, "EB OR" is the oracle benchmark, "EB MLE" is the empirical Bayes maximum likelihood estimate, and "EB SURE" is the Stein's Unbiased Risk Estimate. "EB OPT" (defined in Section 4) is our proposed procedure. Only EB OPT achieves best-in-class performance.

We stress that $\mathcal{X}$, and, hence, $\boldsymbol{A}$ and $\boldsymbol{b}$, are known a priori. However, with no further assumptions, $\mathcal{X}$ may be empty, rendering P^n trivial. To avoid technicalities, we will assume P^n is strictly feasible:

DEFINITION 4.1 (s_0-STRICT FEASIBILITY). We say that $\mathcal{X}$ is *s_0-strictly feasible* if there exists $s_0 > 0$ and $\boldsymbol{x}_0 \in \mathcal{X}$ such that $\frac{1}{n}\sum_{j=1}^n \boldsymbol{A}_j x_j^0 + s_0 \boldsymbol{e} \leq \boldsymbol{b}$.

Assuming strict feasibility is slightly stronger than assuming $\mathcal{X}$ is non-empty, but it can often be achieved via a small perturbation of $\boldsymbol{A}$ and $\boldsymbol{b}$.[12]

To define our specialized policy, we first reinterpret $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ as the solution to

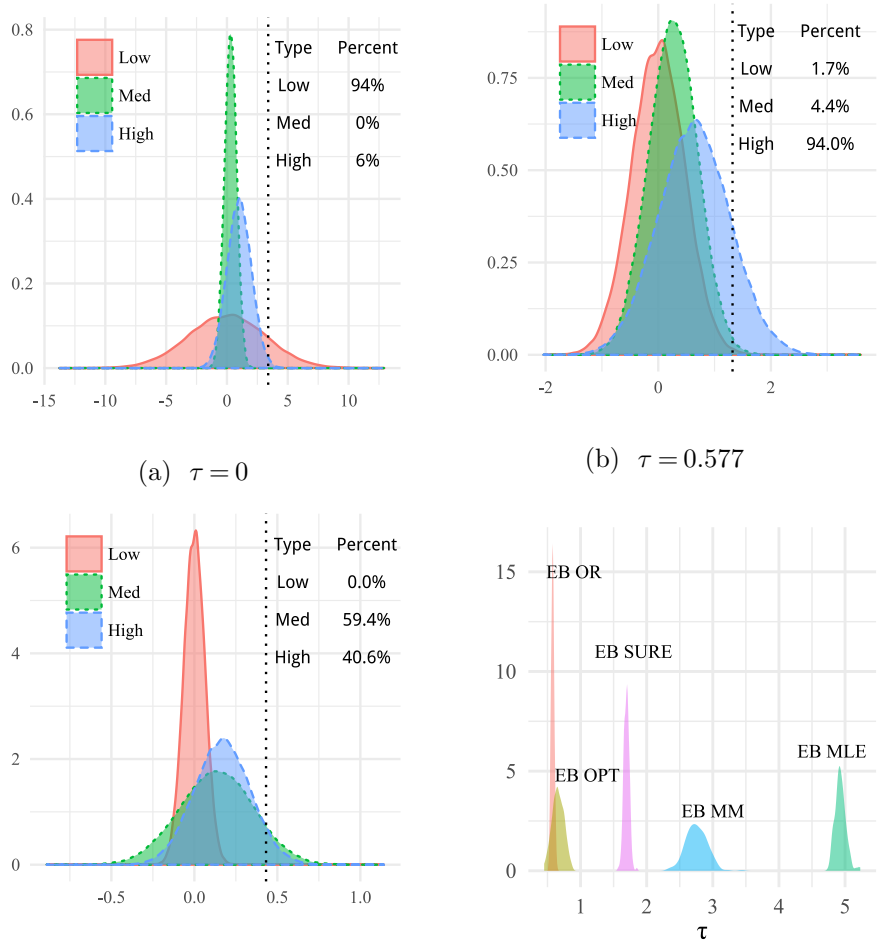
$$\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}}) \in \arg\max_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{n}\sum_{j=1}^n r_j(\tau, \hat{\mu}_j) x_j, \quad \text{where} \quad r_j(\tau, \hat{\mu}_j) = \frac{\nu_{\min} + \tau}{\nu_{\min}} \times \frac{\nu_j}{\nu_j + \tau}\hat{\mu}_j. \tag{4.2}$$

The objective coefficients in Equations (3.1) and (4.2) differ by a positive scaling, so the two definitions of $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ are equivalent. However, an advantage of Equation (4.2) is that the variance of $r_j(\tau, \hat{\mu}_j)$ is bounded below by $1/\nu_j$, which simplifies the analysis.

Under Equation (4.1), the dual to Equation (4.2) is

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^m_+} D_{\hat{\boldsymbol{\mu}}}(\boldsymbol{\lambda}, \tau), \quad \text{where} \quad D_{\hat{\boldsymbol{\mu}}}(\boldsymbol{\lambda}, \tau) \equiv \boldsymbol{b}^\top \boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^n (r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda})^+.$$

[12] For example, if $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ encoded an inequality constraint, e.g., $\boldsymbol{a}_0^\top \boldsymbol{x} \leq b_0$ and $-\boldsymbol{a}_0^\top \boldsymbol{x} \leq -b_0$, we might relax these constraints to $\boldsymbol{a}_0^\top \boldsymbol{x} \leq b_0 + \delta$ and $-\boldsymbol{a}_0^\top \boldsymbol{x} \leq -b_0 + \delta$ to achieve strict feasibility. The extent to which such a relaxation is an acceptable approximation is obviously application dependent.



(a) $\tau = 0$ (b) $\tau = 0.577$

(c) $\tau = 4.94$ (d) Density of estimator

Figure 3 **Effects of Shrinkage on Solution in Example 4.1.** Figures 3(a), 3(b), and 3(c) show smoothed histograms of $\nu_j \hat{\mu}_j/(\nu_j + \tau)$ for various τ along a single sample path, $n = 2^{17} = 131,072$. When $\alpha = .05$, all items to the right of the dotted black line are chosen. Although a small amount of shrinkage increases the number of High items chosen, excessive shrinkage causes many Medium items to be eventually chosen. Figure 3(d) shows the density of the fitted τ for various methods over 200 simulations. Estimate-then-optimize methods generally over-shrink for this example.

Let $\boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}})$ be an optimal solution. Now we can define our alternate bias correction: For any bandwidth $0 < h < 1$, define $h_j(\tau) = \frac{\nu_{\min}+\tau}{\nu_{\min}} \times \frac{h\sqrt{\nu_j}}{\nu_j+\tau} = r_j(\tau, h/\sqrt{\nu_j})$, and let

$$B^{Bayes}(\tau, h, \hat{\boldsymbol{\mu}}) \equiv \frac{1}{n} \sum_{j=1}^{n} \frac{1}{2h\sqrt{\nu_j}} \mathbb{I}(\left| r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}}) \right| \le h_j(\tau)),$$

and define our candidate policy $\boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})$ by

$$\hat{\tau} \in \arg\max_{\tau \ge 0} \ \frac{1}{n} \hat{\boldsymbol{\mu}}^\top \boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}}) - B^{Bayes}(\tau, h, \hat{\boldsymbol{\mu}}). \tag{4.3}$$

Note that computing $\hat{\tau}$ does **not** require knowledge of $\boldsymbol{\mu}$. We motivate this particular form of the bias-correction in the next section. Note that we can compute $\hat{\tau}$ above using a grid search.

We next bound the suboptimality $\boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})$ against the oracle for $\mathcal{X}^{Bayes}(\hat{\boldsymbol{\mu}})$. To develop explicit tail bounds on the stochastic terms for all finite n, we require the following assumptions.

Assumption 4.2 (Independent Near-Gaussians) *In addition to Assumption 2.1, assume that*

i) (Independence) *The random variables $\hat{\mu}_1, \ldots, \hat{\mu}_n$ are independent.*

ii) (Bounded precisions) *There exists $\nu_{\mathsf{min}}, \nu_{\mathsf{max}}$ such that $0 < \nu_{\mathsf{min}} \le \nu_j \le \nu_{\mathsf{max}} < \infty$ for all $j = 1, \ldots, n$.*

iii) (Sub-Gaussian) *There exists a positive constant σ such that for all $j = 1, \ldots, n$, $(\hat{\mu}_j - \mu_j)\sqrt{\nu_j}$ is sub-Gaussian with variance proxy at most σ^2.*

iv) (Bounded and Positive Density) *For all $j = 1, \ldots, n$, $(\hat{\mu}_j - \mu_j)\sqrt{\nu_j}$ admits a density $\phi_j(\cdot)$ that is bounded and strictly positive over any finite interval; that is, $\max_{j=1,\ldots,n} \sup_{t \in \mathbb{R}} \phi_j(t) < \infty$ and $\min_{j=1,\ldots,n} \inf_{t:|t| \le T} \phi_j(t) > 0$ for all $T > 0$.*

The above assumptions impose rather mild requirements on the distribution of $\hat{\mu}_j$. When $\hat{\mu}_j$ are Gaussian, then $\phi_j(\cdot)$ is the density of the standard Gaussian, which is bounded above by $1/\sqrt{2\pi}$ and bounded below by $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{T^2}{2}\right)$ over an interval $[-T, T]$. Intuitively, if the $\hat{\mu}_j$ are approximately normally distributed, we expect these conditions to also hold as well. We can now strengthen Theorem 3.5 for $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ by bounding the maximal stochastic deviations:

Theorem 4.3 (Finite-Dimension Bound for the Bayes-Inspired Policy Class) *Consider P^n under Assumption 4.2 with $\mathcal{X}$ given by Equation (4.1) such that $\mathcal{X}$ is s_0-strictly feasible and $m \ge 1$. Let C_μ, C_A be such that $|\mu_j| \le C_\mu$ and $\|\boldsymbol{A}_j\|_\infty \le C_A$ for all $j = 1, \ldots, n$, and $\beta > 0$ be such that the minimum eigenvalue of the matrix $\frac{1}{n}\sum_{j=1}^n \boldsymbol{A}_j \boldsymbol{A}_j^\top \in \mathbb{R}^{m \times m}$ is at least β.*

Then, for all $0 < \delta < 1$, $0 < h < 1$, there exist positive constants C_1, C_2 not depending on $\{n, m, \delta, h\}$, such that

$$0 \le \frac{1}{n}\boldsymbol{\mu}^\top \left(\boldsymbol{x}(\tau^{\mathsf{OR}}, \hat{\boldsymbol{\mu}}) - \boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})\right) \le 2C_1 \Bigg(\underbrace{\frac{\mathsf{TV}\log(e/\mathsf{TV})}{h}}_{\substack{\text{Degree of} \\ \text{Non-Normality}}} + \underbrace{\frac{\delta}{h}}_{\substack{\text{Approximating} \\ \text{Dual Solution}}} + \underbrace{h^2}_{\substack{\text{Approximating} \\ \text{Stein's Lemma}}} \Bigg) + \underbrace{2R}_{\substack{\text{Stochastic} \\ \text{Errors}}},$$

where for any $\epsilon > \frac{m^2 \log n}{n^{3/2} \log(m+1)}$,

$$\mathbb{P}\{R > 6\epsilon\} \le 130 \left[\exp\left(\frac{-C_2 \delta \sqrt{n}}{\sqrt{m}\log(m+1)}\right) \log\left(1 + \frac{\sqrt{m}\log(m+1)}{C_2 \delta \sqrt{n}}\right) \; + \; \exp\left(\frac{-C_2 \epsilon h \sqrt{n}}{\log(m+1)}\right) \right],$$

and $\mathsf{TV} = \frac{1}{2n}\sum_{j=1}^n \|\phi_j - \phi\|_1$ is the average total variation distance between $\phi_j(\cdot)$ and the standard normal density $\phi(\cdot)$.

Theorem 4.3 decomposes the suboptimality gap into a deterministic error arising from various approximation steps in our proof and a stochastic error with bounded tails. The result is not asymptotic; it holds for a fixed instance of $(\mathrm{P}^n, \hat{\boldsymbol{\mu}}, \boldsymbol{\nu})$, i.e., for finite n. Similar to Theorem 3.5, the bound suggests our method will perform well in the small-data, large-scale regime if the $\hat{\mu}_j$ are nearly Gaussian. Specifically, for a well-chosen sequence of bandwidths h_n depending on n, the performance of our policy converges to that of the oracle asymptotically, as in the following corollary whose proof is given in Appendix D.3.

Corollary 4.4 (Almost Sure Convergence to Oracle) *Consider a sequence $\{(P^n, \hat{\boldsymbol{\mu}}^n, \boldsymbol{\nu}^n) : n \geq 2\}$ of instances in the small-data, large-scale regime, each satisfying the hypothesis of Theorem 4.3 such that the parameters do not grow with n; that is, m is constant, $\|\boldsymbol{A}_j\| \leq C_A$, $\left|\mu_j^n\right| \leq C_\mu$, and $\nu_{\mathsf{min}} \leq \nu_j^n \leq \nu_{\mathsf{max}}$ for all $1 \leq j \leq n$. Suppose further that the smallest eigenvalue of $\frac{1}{n}\sum_{j=1}^n \boldsymbol{A}_j^n \boldsymbol{A}_j^{n\top} \in \mathbb{R}^{m\times m}$ is at least β for all n. Let h_n be bandwidth parameters chosen such that $h_n \to 0$ and $h_n\sqrt{n} \to \infty$. Finally, for each P^n, let $\hat{\tau}^n$ be the solution to Equation (4.3) with bandwidth h_n. If each $\hat{\mu}_j$ is Gaussian, then the policy $\boldsymbol{x}(\hat{\tau}^n, \hat{\boldsymbol{\mu}})$ performs as well as the oracle policy for $\mathcal{X}^{Bayes,n}(\hat{\boldsymbol{\mu}})$, almost surely as $n \to \infty$.*

For the Gaussian case with $\mathsf{TV} = 0$, by matching the orders of the deterministic and stochastic errors in Theorem 4.3, one can show that the rate of convergence for the bound is optimized when $h_n = O(n^{-1/6})$, in which case the suboptimality gap converges to zero as $O_p(n^{-1/3})$. Similar arguments hold when some parameters grow mildly with n, such as $m = O(\log n)$, with slightly modified bandwidth sequences. Regardless of the precise scaling, Theorem 4.3 provides good evidence that our procedure based on $\hat{\tau}$ should have a strong performance for instances with large, finite n as long as the other parameters are not too large. We confirm this claim numerically in Section 6.

In the remainder of this section, we motivate our alternate bias correction in Equation (4.3) and outline the proof of Theorem 4.3 by considering the asymptotic regime where $n \to \infty$ while other parameters stay fixed. See Appendix D.1 for a formal proof for finite n, including explicit values for the constants C_1 and C_2.

4.2. Proof Outline for Theorem 4.3 and the Alternate Bias Correction

The proof of Theorem 4.3 follows the framework of Theorem 3.5; i.e., we first seek to prove that

$$\sup_{\tau \geq 0} \frac{1}{n} \left| \sum_{j=1}^n (\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j) x_j(\tau, \hat{\boldsymbol{\mu}}) - \mathbb{E}[(\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j) x_j(\tau, \hat{\boldsymbol{\mu}})] \right| \to_p 0. \tag{4.4}$$

Establishing this convergence involves two key ideas:

1) Rounding the Primal Solution (Lemma D.6): We first use linear optimization duality to rewrite $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ more simply. By complementary slackness, for all but possibly m components where the x_j are fractional, we have that

$$x_j(\tau, \hat{\mu}_j) = \mathbb{I}(r_j(\tau, \hat{\mu}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}})). \tag{4.5}$$

These fractional terms contribute at most $\frac{m}{n} \max_{j=1,\ldots,n} |\mu_j - \hat{\mu}_j|$ to the total. Since the maximum of sub-Gaussian random variables concentrates at its mean, which grows like $O(\sqrt{\log n})$, the error from rounding the primal solution is small:

$$\frac{1}{n}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}}) = \frac{1}{n} \sum_{j=1}^{n} (\mu_j - \hat{\mu}_j) \mathbb{I}(r_j(\tau, \hat{\mu}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}})) + o_p(1). \tag{4.6}$$

2) Approximating the Dual Solution (Lemma D.7): The sum in Equation (4.6) consists of *dependent* random variables because $\boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}})$ depends on the entire vector $\hat{\boldsymbol{\mu}}$. This dependence poses a technical challenge in establishing the requisite convergence. The second key idea approximates $\boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}})$ by the solution of an "average" dual problem to break this dependence. Define

$$\min_{\boldsymbol{\lambda} \in \mathbb{R}^m_+} D(\boldsymbol{\lambda}, \tau), \quad \text{where} \quad D(\boldsymbol{\lambda}, \tau) \equiv \boldsymbol{b}^\top \boldsymbol{\lambda} + \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[(r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda})^+],$$

and let $\boldsymbol{\lambda}(\tau)$ be an optimal solution. Intuitively, $\boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}}) = \boldsymbol{\lambda}(\tau) + o_p(1)$ for large n (Lemma D.5), so that the error from approximating the dual solution is also small:

$$\frac{1}{n}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}}) = \frac{1}{n} \sum_{j=1}^{n} (\mu_j - \hat{\mu}_j) \mathbb{I}(r_j(\tau, \hat{\mu}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)) + o_p(1).$$

Importantly, replacing $\boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}})$ with $\boldsymbol{\lambda}(\tau)$ transforms the sum into a sum of *independent* random variables. We can now use standard uniform law of large number results (Lemma D.8) to establish

$$\frac{1}{n}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}}) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left[(\mu_j - \hat{\mu}_j) \mathbb{I}(r_j(\tau, \hat{\mu}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau))\right] + o_p(1). \tag{4.7}$$

Finally, by "unwinding" the above approximations, we can also show that

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left[(\mu_j - \hat{\mu}_j) \mathbb{I}(r_j(\tau, \hat{\mu}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau))\right] = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[(\hat{\mu}_j - \mu_j) x_j(\tau, \hat{\boldsymbol{\mu}})] + o_p(1),$$

which proves Equation (4.4).

If we sought to apply Theorem 3.5 directly, we would next need to prove

$$\frac{1}{n} \sum_{j=1}^{n} \frac{1}{2h\sqrt{\nu}_j} \left[x_j\left(\tau, \hat{\boldsymbol{\mu}} + \frac{h}{\sqrt{\nu}_j} \boldsymbol{e}_j\right) - x_j\left(\tau, \hat{\boldsymbol{\mu}} - \frac{h}{\sqrt{\nu}_j} \boldsymbol{e}_j\right) \right] \to_p 0.$$

This convergence can be established using similar arguments as above using the average dual optimal solution, but, it is easier to take a different approach. Specifically, instead of applying

our approximate Stein's Lemma to $x_j(\cdot)$ as in Theorem 3.5, we apply it to right-hand side of Equation (4.7). This leads to the third key idea in the proof.

3) A Custom Bias Correction: Let $\zeta_j = (\hat{\mu}_j - \mu_j)\sqrt{\nu_j}$ and let $f_j : \mathbb{R} \to \mathbb{R}$ be defined by $f_j(t) = \frac{-1}{\sqrt{\nu_j}}\mathbb{I}(r_j(\tau, \mu_j + t/\sqrt{\nu_j}) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau))$ for all $t \in \mathbb{R}$. Note that ζ_j is mean-zero, has precision 1, and is sub-Gaussian with parameter σ^2. Furthermore,

$$\mathbb{E}\left[(\mu_j - \hat{\mu}_j)\mathbb{I}(r_j(\tau, \hat{\mu}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau))\right] = \mathbb{E}\left[\zeta_j f_j(\zeta_j)\right].$$

Using Lemma C.2, we approximate $\mathbb{E}[\zeta_j f_j(\zeta_j)]$ by a first-order finite difference with step h, so

$$\begin{aligned}\mathbb{E}[\zeta_j f_j(\zeta_j)] &\approx \frac{1}{2h}\mathbb{E}[f_j(\zeta_j + h) - f_j(\zeta_j - h)] \\ &= \frac{-\left(\mathbb{I}\left\{r_j\left(\tau, \hat{\mu}_j + h/\sqrt{\nu_j}\right) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)\right\} - \mathbb{I}\left\{r_j\left(\tau, \hat{\mu}_j - h/\sqrt{\nu_j}\right) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)\right\}\right)}{2h\sqrt{\nu_j}}.\end{aligned}$$

Since $\hat{\mu}_j \mapsto r_j(\tau, \hat{\mu}_j)$ is linear, $r_j(\tau, \hat{\mu}_j + h/\sqrt{\nu}_j) = r_j(\tau, \hat{\mu}_j) + r_j(\tau, h/\sqrt{\nu}_j) = r_j(\tau, \hat{\mu}_j) + h_j(\tau)$, and similarly, $r_j(\tau, \hat{\mu}_j - h/\sqrt{\nu}_j) = r_j(\tau, \hat{\mu}_j) - h_j(\tau)$. Thus,

$$\begin{aligned}&\mathbb{I}\left\{r_j(\tau, \hat{\mu}_j + h/\sqrt{\nu}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)\right\} - \mathbb{I}\left\{r_j\left(\tau, \hat{\mu}_j - h/\sqrt{\nu_j}\right) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)\right\} \\ &= \mathbb{I}\left\{r_j(\tau, \hat{\mu}_j) + h_j(\tau, h/\sqrt{\nu}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)\right\} - \mathbb{I}\left\{r_j\left(\tau, \hat{\mu}_j\right) - h_j(\tau, h/\sqrt{\nu_j}) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)\right\} \\ &= \mathbb{I}\left\{r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau) > -h_j(\tau, h/\sqrt{\nu}_j)\right\} - \mathbb{I}\left\{r_j\left(\tau, \hat{\mu}_j\right) \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau) > +h_j(\tau, h/\sqrt{\nu_j})\right\} \\ &= \mathbb{I}\left\{\left|r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)\right| \le h_j(\tau)\right\}.\end{aligned} \tag{4.8}$$

In light of Equation (4.7), this suggests the alternate bias correction

$$\frac{1}{n}\sum_{j=1}^{n} \frac{1}{2h\sqrt{\nu_j}}\mathbb{E}\left[\mathbb{I}(\left|r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)\right| \le h_j(\tau))\right].$$

Estimating the expectation by its sample average (Lemma D.10) and replacing $\boldsymbol{\lambda}(\tau)$ with $\boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}})$ (Lemma D.11) gives the bias correction term appearing in Equation (4.3).

Note that an advantage of this custom bias correction term over the explicit finite difference procedure of Theorem 3.5 is that evaluating the bias correction for a fixed τ only involves determining $\boldsymbol{\lambda}(\tau, \hat{\boldsymbol{\mu}})$, which is typically obtained for "free" as a by-product of computing $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$. Thus, we need only solve one optimization per τ, instead of $2n$.

The detailed proof of Theorem 4.3 in Appendix D provides explicit (finite n) bounds on each of the $o_p(1)$ remainder terms to show that they are uniformly small. As an aside, we note that one can also intuitively derive $B^{Bayes}(\tau, h, \hat{\boldsymbol{\mu}})$ by simply plugging in the approximation Eq. (4.5) into our original bias-correction $B(\theta, h, \hat{\boldsymbol{\mu}})$ and simplifying along the lines of Eq. (4.8).

4.3. Performance in the Large-Sample Regime

Despite the slight differences between $B^{Bayes}(\tau, h, \hat{\boldsymbol{\mu}})$ and our original bias correction $B(\theta, h, \hat{\boldsymbol{\mu}})$, the proofs of Lemma 3.1 and Theorem 3.6 carry through with almost no adjustments, yielding:

Corollary 4.5 (Bound to Full-Information Optimum) *Let $\hat{\tau}$ be a solution to Equation* (4.3)*. Then, under Assumption 2.1,*

$$\mathbb{E}\left[\frac{1}{n}\left|\boldsymbol{\mu}^\top\left(\boldsymbol{x}^*(\boldsymbol{\mu}) - \boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})\right)\right|\right] \leq \frac{2 + h^{-1}}{\sqrt{\nu_{\mathsf{min}}}}.$$

Following an argument identical to the one for Corollary 3.7, we can also prove that $\boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})$ achieves full-information optimal performance in the large-sample regime. We omit the details for brevity.

5. Regularization Policies over Polyhedral Feasible Regions

Another approach to improving the finite-sample behavior of SAA is to use a regularizer. We next tailor our framework in Equation (3.3) to our Regularization Policies introduced in Example 3.3:

$$\mathcal{X}^{Reg}(\hat{\boldsymbol{\mu}}) \equiv \left\{\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) : \Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}]\right\}, \quad \text{where} \quad \boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x} - \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{2n}\sum_{j=1}^{n}\frac{x_j^2}{\nu_j}.$$

Here Γ controls the amount of regularization. Our choice to scale Γ by $\frac{\sqrt{\nu_{\mathsf{min}}}}{2n}$ is without loss of generality but simplifies our large-sample analysis in Section E.5. This regularizer penalizes the SAA objective, discouraging solutions that contain many low-precision components. Choosing low-precision components causes SAA to perform poorly in Example 2.6. Thus, we intuitively expect regularized solutions with well-chosen Γ to improve upon SAA.

We note that $\mathcal{X}^{Reg}(\hat{\boldsymbol{\mu}})$ can equivalently be cast as robust optimal solutions over uncertainty sets given by ellipsoids. See Ben-Tal and Nemirovski 2002 for an overview of robust optimization and Appendix E.1 for discussion and a formal statement.

5.1. Cross-Validation Approaches to Selecting Γ

The most common data-driven approach to selecting the regularization parameter in the large-sample regime is to use some form of cross-validation (Friedman et al. 2001). Unfortunately, cross-validation procedures may not be well-defined in the small-data, large-scale regime, and, even when they are well-defined, may not have perform comparably to the oracle. We prove this claim for the canonical examples of leave-one-out (LOO) and K-fold cross-validation. Informally, K-fold cross-validation divides the data into K roughly equal portions or *folds*, and then iteratively forms the policy class using all the data except the k^{th} fold, and evaluates these policies on the "left-out" fold. It then selects $\Gamma^{\mathsf{K-fold}}$ to maximize the average performance, where averaging is performed over all K possible choices of the left-out fold. In practice, K is typically taken to be 2, 5 or 10.

LOO validation is the special case when $K = \mathsf{S}$, and each component has the same amount of data, i.e., $\mathsf{S}_j = \mathsf{S}$. Hold-out-validation is the special case when $K = 2$ and one maximizes the performance when leaving out the first fold. Notice, K-fold cross-validation is not defined when $K > \mathsf{S}$, and in the small-data, large-scale regime, S may be very small, e.g., less than 10, so that not all forms of cross-validation may even be defined.

We only analyze K-fold cross-validation in the specific setting of Example 2.2 for the special case that $\mathsf{S}_j = \mathsf{S} \geq K$ for all j and S is a multiple of K for simplicity. Let

$$\overline{\mu}_j^k \equiv \frac{K}{\mathsf{S}} \sum_{l=(k-1)\mathsf{S}/K+1}^{k\mathsf{S}/K} \hat{\mu}_j^l, \qquad \overline{\mu}_j^{-k} \equiv \frac{K}{\mathsf{S}(K-1)} \sum_{\substack{1 \leq l \leq (k-1)\mathsf{S}/K, \\ k\mathsf{S}/K < l \leq \mathsf{S}}} \hat{\mu}_j^l$$

be sample averages of the j^{th} coordinate in, and excluding, the k^{th} fold, respectively. Then

$$\boldsymbol{x}^{\mathsf{K}}(\Gamma, \overline{\boldsymbol{\mu}}^{-k}) \in \arg\max_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{n} \boldsymbol{x}^\top \overline{\boldsymbol{\mu}}^{-k} - \frac{\Gamma}{2n} \frac{1}{\sqrt{\mathsf{S}\nu_0}} \sqrt{\frac{K}{K-1}} \sum_{j=1}^n x_j^2$$

is the analogue of Equation (3.2) for this setting that excludes the k^{th} fold and accounts for the adjusted precision. The K-fold cross-validation solution selects

$$\Gamma^{\mathsf{K-fold}} \in \arg\max_{\Gamma \in [\Gamma_{\min}, \Gamma_{\max}]} \frac{1}{K} \sum_{k=1}^K \frac{1}{n} \overline{\boldsymbol{\mu}}^{k\top} \boldsymbol{x}^{\mathsf{k}}(\Gamma, \hat{\boldsymbol{\mu}}^{-k}) \tag{5.1}$$

and implements $\boldsymbol{x}^{\mathcal{R}}(\Gamma^{\mathsf{K-fold}}, \hat{\boldsymbol{\mu}})$. Notice that the implemented policy uses the full data (all folds). The following theorem proves that, there exist instances where 5-fold, 10-fold and LOO validation will not achieve oracle performance in the small-data, large-scale regime.

Theorem 5.1 (Leave-One-Out and K-fold Cross-Validation Are Not Best-In-Class) *There exists a sequence of instances* $\left\{(P^n, \hat{\boldsymbol{\mu}}^n, \boldsymbol{\nu}^n) : \hat{\boldsymbol{\mu}}^n \in \mathbb{R}^n, \boldsymbol{\nu}^n \in \mathbb{R}^n_+, n \geq 2\right\}$ *in the small-data, large-scale regime with* $\hat{\boldsymbol{\mu}}^n$ *as in Example 2.2 with* $\mathsf{S}_j = \mathsf{S} \equiv 10$ *for all* j*, such that*

$$\lim_{n\to\infty} \frac{\frac{1}{n}\boldsymbol{\mu}^{n\top} \boldsymbol{x}^{\mathcal{R}}(\Gamma^{\mathsf{K-fold},n}, \hat{\boldsymbol{\mu}})}{\frac{1}{n}\boldsymbol{\mu}^{n\top} \boldsymbol{x}^{\mathcal{R}}(\Gamma^{\mathsf{OR},n}, \hat{\boldsymbol{\mu}})} < 0.03 ,$$

for any $K \in \{2, 5, 10\}$*. Here* $\Gamma^{\mathsf{K-Fold},n}$ *and* $\Gamma^{\mathsf{OR},n}$ *are the* $K-fold$ *and oracle* Γ *for the* n^{th} *instance.*

In other words, none *of* 5*-fold,* 10*-fold, hold-out, or LOO validation achieve oracle performance in the small-data, large-scale regime for this instance.*

The proof of the theorem is constructive (see Appendix). Intuitively, K-fold cross validation does not achieve oracle performance because, when S is small, $\boldsymbol{x}^{\mathsf{k}}(\Gamma, \overline{\boldsymbol{\mu}}^{-k})$ may be very different from $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$. For example, if $\mathsf{S} = 2$, leaving out one data point amounts to throwing away 50% of the data. Hence, the right hand side of Equation (5.1) is a poor approximation to $\boldsymbol{\mu}^\top \boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$. If in addition the true oracle curve $\Gamma \mapsto \boldsymbol{\mu}^\top \boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ is not very flat near its optimum, then this poor approximation will generally not yield a near minimizer. One can see this intuition directly in the proof of Theorem 5.1, in particular, in Fig. EC.1.

5.2. A Near-Oracle Policy for the Regularization-Inspired Class

We next use Theorem 3.5 to identify a policy that performs comparably to the oracle. As in Section 4.1, we exploit the structure of the policy class to develop a more computationally efficient bias correction. Lemma E.2 (see appendix) shows the dual problem to Equation (3.2) is

$$\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}) \in \arg\ \min_{\boldsymbol{\lambda}\geq 0}\ \boldsymbol{b}^\top\boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^{n} w_j(\Gamma,\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda})\ ,$$

$$\text{where}\qquad w_j(\Gamma,t) = \begin{cases} 0 & \text{if}\quad t<0, \\ \frac{\nu_j}{2\Gamma\sqrt{\nu_{\mathsf{min}}}}t^2 & \text{if}\quad 0\leq t\leq \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\nu_j}, \\ t-\frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{2\nu_j} & \text{if}\quad \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\nu_j}<t\ . \end{cases} \tag{5.2}$$

Intuitively, the function $w_j(\Gamma,t)$ is a smoothed approximation of the hinge function $\left(t-\frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\nu_j}\right)^+$. Lemma E.2 also shows that $x_j^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}})$ can be written explicitly in terms of the dual solution as

$$x_j^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}) = \frac{\nu_j}{\Gamma\sqrt{\nu_{\mathsf{min}}}}\left(\left[\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}})\right]^+ - \left[\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}) - \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\nu_j}\right]^+\right).$$

Substituting this expression into our standard bias-correction and taking the limit as $h\to 0$ suggests the alternate, more computationally efficient correction

$$B_n^{Reg}(\Gamma,\hat{\boldsymbol{\mu}}) \equiv \frac{1}{\Gamma n\sqrt{\nu_{\mathsf{min}}}}\sum_{j=1}^{n}\mathbb{I}\left(0\leq \hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}) \leq \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\nu_j}\right).$$

Finally, for any $0<\Gamma_{\mathsf{min}}<\Gamma_{\mathsf{max}}<\infty$, define

$$\hat{\Gamma}\in \arg\max_{\Gamma_{\mathsf{min}}\leq\Gamma\leq\Gamma_{\mathsf{max}}} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top\boldsymbol{x}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}) - B_n^{Reg}(\Gamma,\hat{\boldsymbol{\mu}}). \tag{5.3}$$

Under assumptions similar to Theorem 4.3, we can strengthen Theorem 3.5 for $\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma},\hat{\boldsymbol{\mu}})$ by bounding the maximal stochastic deviations.

Theorem 5.2 (Finite-Dimension Bound for the Regularization Policy Class) *Consider P^n under Assumption 4.2 with $\mathcal{X}$ as in Equation (4.1) with $m\geq 1$ and $\mathcal{X}$ is s_0-strictly feasible. Let C_μ, C_A be such that $|\mu_j|\leq C_\mu$ and $\|\boldsymbol{A}_j\|_\infty\leq C_A$ for all $j=1,\ldots,n$, and $\beta>0$ be such that the minimum eigenvalue of $\frac{1}{n}\sum_{j=1}^n \boldsymbol{A}_j\boldsymbol{A}_j^\top\in\mathbb{R}^{m\times m}$ is at least β. Assume $0<\Gamma_{\mathsf{min}}<\Gamma_{\mathsf{max}}<\infty$. Then, for each $0<\delta<1$, there exist positive constants C_1,C_2,C_3,C_4 not depending on $\{n,m,\delta\}$, such that*

$$0\leq \frac{1}{n}\boldsymbol{\mu}^\top(\boldsymbol{x}^{\mathcal{R}}(\Gamma^{\mathsf{OR}},\hat{\boldsymbol{\mu}}) - \boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma},\hat{\boldsymbol{\mu}}) \leq \underbrace{C_1\delta}_{\substack{\text{Approximating}\\ \text{Dual Solution}}} + \underbrace{C_1\mathsf{TV}\log(e/\mathsf{TV})}_{\substack{\text{Degree}\\ \text{Non-Normality}}} + \underbrace{2R_n}_{\substack{\text{Stochastic}\\ \text{Errors}}}\ ,$$

where for any $\epsilon>\delta^2/\sqrt{n}$,

$$\mathbb{P}\{R_n>4\epsilon\}\ \leq\ C_2\left(\exp\left(\frac{-C_3\delta\sqrt{n}}{\sqrt{m}\log(m+1)}\right)\log\left(1+\frac{\sqrt{m}\log(m+1)}{C_3\delta\sqrt{n}}\right)+\exp\left(\frac{-C_4\epsilon\sqrt{n}}{\log(m+1)}\right)\right), \tag{5.4}$$

and $\mathsf{TV}=\frac{1}{2n}\sum_{j=1}^n\|\phi_j-\phi\|_1$ is the average total variation distance between $\phi_j(\cdot)$ and the standard normal density $\phi(\cdot)$.

Like Theorem 4.3, Theorem 5.2 holds for a fixed instance $(\mathrm{P}^n, \hat{\boldsymbol{\mu}}, \boldsymbol{\nu})$, i.e., finite n. The structure of the bound, however, provides insight into the performance in the small-data, large scale regime, and shows that the performance of our policy converges to that of the oracle when the dimension increases and $\hat{\mu}_j$ is gaussian. This result is stated in the following corollary:

Corollary 5.3 (Almost Sure Convergence to Oracle) *Consider a sequence of instances of* $\{(P^n, \hat{\boldsymbol{\mu}}^n, \boldsymbol{\nu}^n) : n \geq 2\}$ *in the small-data, large-scale regime, with each instance satisfying Assumption 4.2 with* $\mathcal{X}_n$ *given by Equation* (4.1) *and each* $\mathcal{X}_n$ *satisfying Assumption 4.1 for a common* s_0. *Suppose further that the parameters do not grow with* n, *i.e.,* m *is constant,* $\|\boldsymbol{A}_j\| \leq C_A$, $\left|\mu_j^n\right| \leq C_\mu$, *and* $\nu_{\mathsf{min}} \leq \nu_j^n \leq \nu_{\mathsf{max}}$ *for all* $1 \leq j \leq n$ *and that the smallest eigenvalue of* $\frac{1}{n}\sum_{j=1}^n \boldsymbol{A}_j^n \boldsymbol{A}_j^{n\top} \in \mathbb{R}^{m\times m}$ *is at least* β *for all* n. *For each* P_n, *let* $\hat{\Gamma}^n$ *be given by Equation* (5.3). *If each* $\hat{\mu}_j$ *is Gaussian, then the policy* $\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma}^n, \hat{\boldsymbol{\mu}})$ *performs as well as the oracle policy for* $\mathcal{X}^{Reg,n}(\hat{\boldsymbol{\mu}})$, *almost surely as* $n \to \infty$.

We stress that it is unclear a priori whether $\boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})$ or $\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma}, \hat{\boldsymbol{\mu}})$ yields better performance in Problem P^n since it is not clear which benchmark, $\boldsymbol{x}(\tau^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ or $\boldsymbol{x}^{\mathcal{R}}(\Gamma^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$, is superior. In general, we find that the difference is application specific.

Like Theorem 4.3, the proof of Theorem 5.2 also follows Theorem 3.5 and uses an "average" dual problem to break the dependence between terms (Lemma E.6). Practically, the restriction $0 < \Gamma_{\mathsf{min}} < \Gamma_{\mathsf{max}} < \infty$ is mild; we expect in practice to optimize $\hat{\Gamma}$ by searching over a finite grid. It is an open question if Theorem 5.2 can be strengthened to allow $\Gamma_{\mathsf{min}} = 0$ or $\Gamma_{\mathsf{max}} = \infty$.

Although we focused above on the small-data, large-scale regime, our policy also achieves full-information optimality in the large-sample regime. (Note Theorem 3.6 is not applicable when $\Gamma_{\mathsf{min}} > 0$ because $\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})$ may not be in $\mathcal{X}^{Reg}(\hat{\boldsymbol{\mu}})$). See Appendix E.5 for theorem and discussion.

6. Numerical Experiments

We next study the empirical performance of our methods in the context of a specific application: the online-advertising portfolio optimization problem (OAPOP). Our goals are two-fold: 1) quantify the potential benefits of our Bayes-Inspired and Regularization-Inspired best-in-class policies over traditional variants for this application and 2) assess the robustness of our results to increasingly large departures from normality.

We focus on the OAPOP because we view it as typical of the small-data, large-scale optimization regime. Loosely speaking, the OAPOP involves an advertiser seeking to allocate a finite budget among various "targeting items" to maximize her return. Targeting items may represent keywords, impressions, cookies, and websites and may span different advertising channels and platforms. In practice, an advertiser must also estimate the expected cost and expected revenue for each targeting item before electing an allocation. Pani et al. (2017) provide a thorough overview of the OAPOP,

including its pivotal role in the online-advertising industry and the recent surge of decision-support software products for the problem, such as Adobe Marketing Cloud and Google's DoubleClick. Most importantly, the authors observe that a typical instance of the OAPOP may involve tens of thousands of targeting items and that because the underlying problem parameters shift rapidly, estimates are necessarily very noisy – the two defining features of the small-data, large-scale regime.

Pani et al. (2017) argue that despite the many complexities of the online-advertising industry, the OAPOP can be modeled effectively as an offline, fractional, multi-choice knapsack problem. The choices correspond to different bid levels for each targeting item, while the weights and rewards correspond to the expected costs and revenues. The authors propose a customized algorithm for massively large instances where the expected returns and expected cost of each item are known. We adopt a similar perspective, complementing their work by focusing instead on instances where the returns are not known, but rather imprecisely estimated. As in Rusmevichientong and Williamson (2006), we consider only a single, representative bid level. The resulting problem is an offline, fractional knapsack problem with uncertain rewards. It can be written in the form of Problem P^n with $\mathcal{X} = \left\{\boldsymbol{x} \in [0,1]^n : \frac{1}{n}\boldsymbol{c}^\top \boldsymbol{x} \leq 1\right\}$ for some fixed cost-vector $\boldsymbol{c} \in \mathbb{R}^n$.

We simulate a variety of instances of the OAPOP and apply our data-driven procedures. Our precise simulation procedure for $\boldsymbol{\mu}$, $\hat{\boldsymbol{\mu}}$, and $\boldsymbol{\nu}$ is in Appendix F.2 and closely follows the procedure of Pani et al. (2017). Those authors calibrated this procedure to match a real-world dataset drawn from Google's Keyword Planner tool for "medium-high volume keywords from a wide variety of industry categories including retail (apparel, footwear, etc.), insurance, and financial services" (Pani et al. 2017, pg. 23). Overall, although simulated, we believe our instances to be realistic in structure.

Throughout our experiments, SAA is the sample average approximation policy; EB OR is the oracle policy $\boldsymbol{x}(\tau^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$; EB OPT is our proposed policy $\boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})$; EB MLE and EB MM are the estimate-then-optimize policies based on empirical Bayes maximum likelihood and moment-matching estimates, respectively; and SURE is the estimate-then-optimize policy based on the SURE estimate (c.f. Section 4). Similarly, Reg OR is the oracle policy $\boldsymbol{x}^{\mathcal{R}}(\Gamma^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ and Reg OPT is our proposed policy $\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma}, \hat{\boldsymbol{\mu}})$. RO 1% and RO 5% are robust optimization policies using elliptical uncertainty sets for two different choices of r (c.f. Appendix E.1.). Specifically, we use the "safe-approximation" guideline for robustness (Chapter 2 in Ben-Tal and Nemirovski 2002), and for $\epsilon = 0.01, 0.05$, we let $r = \sqrt{2\log(1/\epsilon)}$. For both the Bayes-Inspired and Regularization-Inspired classes, we also compare to hold-out (denoted "HO"), 5-fold (denoted "K5") and leave-one-out (denoted "LOO") cross-validation. We note that cross-validation strategies are seemingly rare in the empirical bayes literature, but can be defined analogously as in the regularization case. We also remind the reader that that K-fold cross validation is not well defined when $S < K$.

The code for all experiments written in the Julia programming language (Bezanson et al. (2017)) is available at ***BLINDED FOR REVIEW.*** When computing $\hat{\tau}^n$, we take $h_n = n^{-1/6}$ and searched exhaustively over $\tau \in [0, 5]$ using a grid of size .01. When computing τ^{OR}, we use a parametric linear programming algorithm to find the exact optimum over $\mathbb{R}_+$. Similarly, when computing both $\hat{\Gamma}$ and Γ^{OR}, we search exhaustively over $\Gamma \in [1, 100]$ using a grid of size .5. No special effort was devoted to tuning these parameters.

Before proceeding, we summarize our main insights as follows:

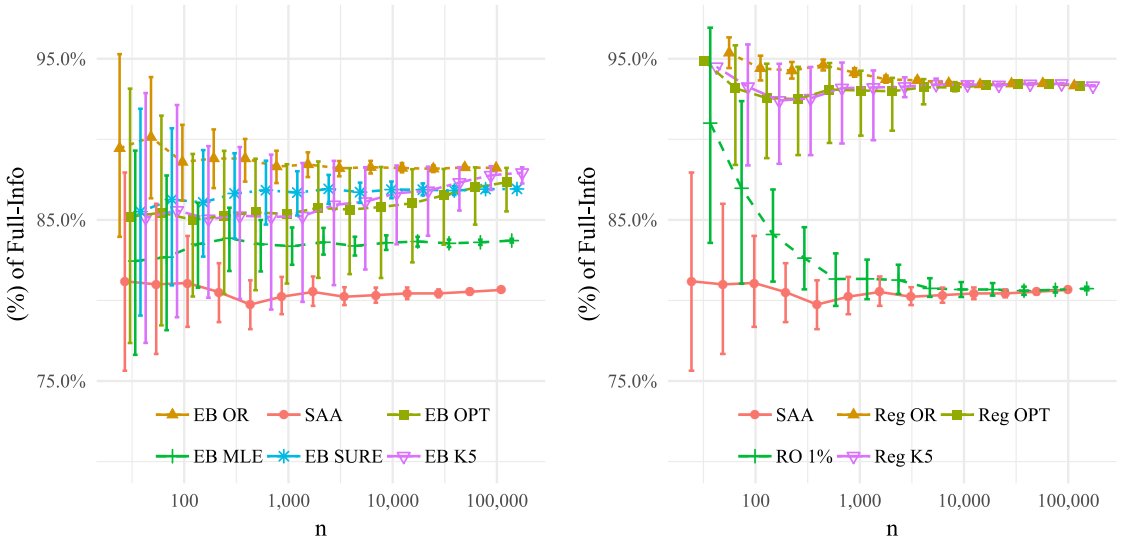
1. For large n, our small-data, large-scale methods (EB OPT, Reg OPT) offer significant benefits over SAA. They also offer a smaller benefit over estimate-then-optimize methods that do not leverage the optimization structure.
2. For this particular application, cross-validation methods perform well, often comparable to the oracle, in contrast to their theoretical analysis. This distinction arises because the OAPOP problem is very flat around its optimum.
3. For small to moderate n, our methods exhibit somewhat more variability than estimate-then-optimize methods methods and, as a consequence, can have poorer performance. Their variability, however, is comparable to cross-validation procedures.
4. Overall, the performance of our methods seems robust to mild departures from normality.

6.1. Finite-dimensional Behavior (finite n)

We first study the behavior of our approach as n grows when $\hat{\mu}_j$ is Gaussian, i.e., $\mathsf{TV} = 0$. Figure 4 shows the performance of various data-driven methods across 200 simulations for increasing n. We have split the methods into two plots – Bayes-Inspired policies and Regularization-Inspired policies – and included a subset of policies for readability. A larger plot with all policies is in Appendix F.3.

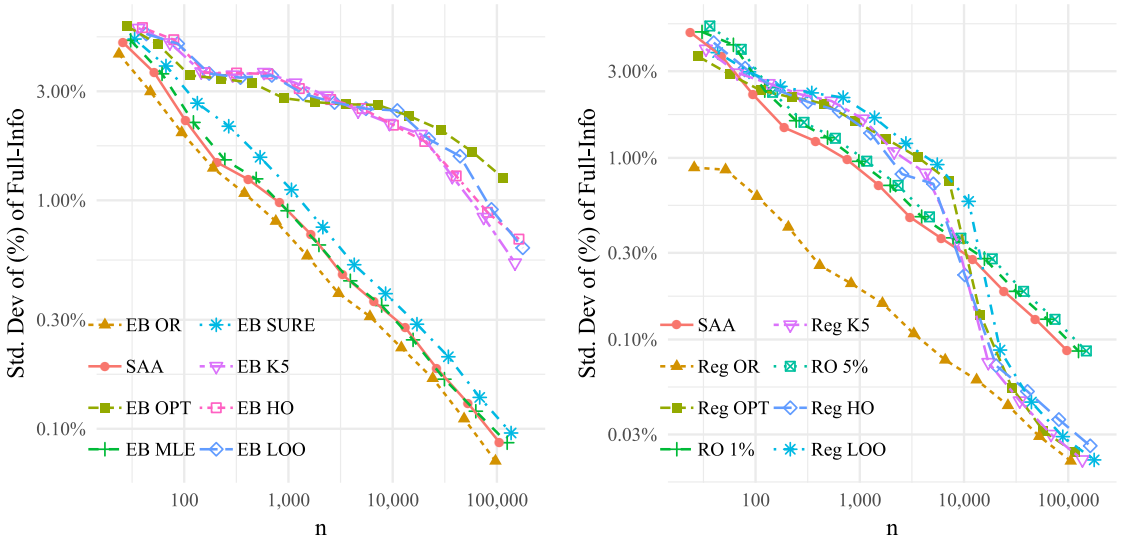
With respect to the Bayes-Inspired policies, several features are immediately clear. First, as expected, our proposed policy EB OPT offers significantly better performance than estimate-then-optimize policies and SAA, especially as n grows large. By contrast, cross-validation procedures, which do leverage the optimization structure, have very similar performance in this example. A drawback of both EB OPT and the cross-validation procedures is that they are more variable than competitors. This feature is more pronounced in Figure 5, where we plot the standard deviation of the performance (relative to the full-information optimum) of each method along the 200 simulations. Nonetheless, from Figure 4, we would argue that the benefits in average performance outweigh the extra variability for $n > 30,000$, a fairly reasonable number for this application.

Regularization-Inspired policies exhibit similar performance. Our proposed policy Reg OPT converges quickly to oracle performance. Cross-validation approaches are highly competitive. By



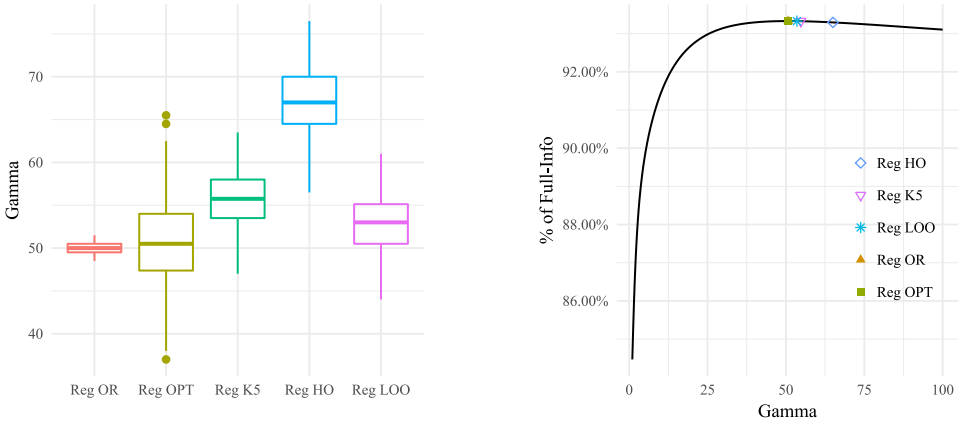
(a) Bayes-Inspired Policies (b) Regularization-Inspired Policies

Figure 4 **Relative Performance by Policy for OAPOP.** The left panel plots the performance of $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ for various n and data-driven procedures for choosing τ from Section 4. The right panel plots the performance of $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ along the same sample paths for data-driven procedures for choosing Γ from Section 5. The error bars represent 10% and 90% quantiles over 200 simulations.



(a) Bayes-Inspired Policies (b) Regularization-Inspired Policies

Figure 5 **Standard Deviation of Performance for OAPOP.** The left panel plots the standard deviation of $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ for various n and data-driven procedures for choosing τ from Section 4. The right panel plots the standard deviation of $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ along the same sample paths for data-driven procedures for choosing Γ from Section 5. Note the log-log scales.



(a) Distribution of Γ's by method

(b) $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$

Figure 6 **Explaining Performance of Cross-Validation for OAPOP.** The left panel plots the distribution of the optimizing Γ for various methods across the 200 simulations. The right panel plots the oracle curve $\Gamma \to \boldsymbol{\mu}^\top \boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ for a single realization, when $n = 2^{17}$, with optimizing Γ of other methods indicated.

contrast, the robust optimization policies with radii specified according to the safe-approximation guideline have significantly worse performance, converging to SAA as $n \to \infty$.

In the case of Regularization-Inspired policies, the strong performance of cross-validation sharply contrasts with Theorem 5.1. We believe this performance is somewhat application specific. Recall that cross-validation policies perform poorly when i) the cross-validation estimates for each fold differ greatly from original policy class (i.e. $\boldsymbol{x}^{\mathrm{k}}(\Gamma, \overline{\boldsymbol{\mu}}^{-k})$ differs from $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ in Eq. (5.1)) and ii) the oracle curve $\Gamma \mapsto \boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ is not too flat at its optimum. For our OAPOP instances, the first condition holds. This is somewhat difficult to see by examining the curves directly (left panel of Figure EC.5 in Appendix F.3), but it is more evident by noting that the optimizing Γ's for these curves converge to different values as $n \to \infty$ (right panel of Figure EC.5). Indeed, Figure 6 below shows boxplots of the distribution of the various estimated Γ's when $n = 2^{17}$, highlighting that the cross-validation procedures over-regularize in this example, while Reg Opt converges to Γ^{OR}.

The key observation, however, is that the second condition does not hold for these instances: the oracle curve is very flat; see right panel of Figure 6. Hence, even though cross-validation procedures select the "wrong" Γ, this error manifests as a neglible amount of sub-optimality. (Contrast this curve to the oracle curve in Theorem 5.1, seen in Figure EC.1.) In our opinion, this feature explains the strong performance of K-fold cross-validation in this application, and, Theorem 5.1 shows this feature does not always hold. By contrast, our Reg Opt policy is guaranteed to achieve oracle performance in the small-data, large-scale regime under the conditions in Theorem 5.2.

A very similar analysis can be performed for the cross-validation policies in the Bayes-Inspired policy class. See Figures EC.6 and EC.7 in Appendix F.3 for the details. There, too, we notice that

the oracle curve is quite flat at its optimum (in contrast to, e.g., Example 4.1) partially explaining the strong performance of cross-validation for these instances.

As an aside, we note that the optimal policies EB OPT and Reg OPT typically have very different structures. For example, on a typical path with $n = 2^{17}$, more than 50% of the positive values of $\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma}, \hat{\boldsymbol{\mu}})$ (Reg OPT) are fractional. By contrast, $\boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})$ (EB OPT) has at most one fractional value by construction, regardless of the size of n. Depending on the intended application, this difference in structure may be pertinent.

6.2. Other Experiments

Appendices F.4 and F.5 present additional experiments assessing the relative performance of our methods when S is large but finite, and when $\hat{\boldsymbol{\mu}}$ is non-gaussian, respectively. Generally, we find that EB Opt and Reg Opt are competitive with cross-validation methods, and generally outperform SAA and estimate-then-optimize methods when S is large, even when estimators are non-gaussian. In particular, our method is robust to some non-normality, when noise is still sub-Gaussian and admits a density (see Appendix for details).

7. Conclusion and Future Directions

Motivated by emerging optimization applications where the amount of relevant data per parameter is small, we proposed a novel method for dealing with linear optimization in the small-data, large-scale regime. In contrast to the large-sample regime, the full-information optimum is not achievable in this context, and, so, we must focus on finding a best-in-class policy. As a benchmark, we consider an oracle policy that knows the underlying parameters in advance but is constrained to use a policy from a specific class.

By correcting for the bias in the estimated objective value, we developed a novel framework for designing a policy whose performance converges to this oracle benchmark and applied the framework to two important classes of policies: Bayes-Inspired and Regularization-Inspired policies. Numerical results show that our policies are robust and perform well across a broad range of problem instances, surpassing traditional methods in the literature.

Our framework can be applied to an arbitrary class of policies. However, to ensure convergence, we need to establish uniform convergence properties of the bias correction terms. It would be interesting to explore other classes of policies and optimization problems for which such uniform convergence can be established. This paper focused on linear optimization, but the spirit of our approach can be extended to nonlinear and discrete optimizations as well. This is a challenging but potentially exciting research area. It is our hope that this research galvanizes the community to consider optimization problems within this important small-data, large-scale regime.

Acknowledgments

We would like to thank the Area Editor, Professor Yinyu Ye, the associate edtior, and the three referees for their detailed and thoughtful comments, which substantially improved our exposition and technical results. The authors were partially supported by the National Science Foundation under Grants CMMI-1661732 and CMMI-1824860.

References

Ban, G.-Y., J. Gallien, A. Mersereau. 2018. Dynamic procurement of new products with covariate information: The residual tree method. To appear in *Manufacturing & Service Operations Management*.

Ban, G.-Y., C. Rudin. 2014. The big data newsvendor: Practical insights from machine learning. *Operations Research* **67**(1) 90–108.

Belloni, A., V. Chernozhukov. 2011. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* **39**(1) 82–130.

Ben-Tal, A., A. Nemirovski. 2000. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming* **88**(3) 411–424.

Ben-Tal, A., A. Nemirovski. 2002. Robust optimization–methodology and applications. *Mathematical Programming* **92**(3) 453–480.

Berger, J. O. 2013. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, NY.

Bertsimas, D., V. Gupta, N. Kallus. 2018a. Data-driven robust optimization. *Mathematical Programming* **167**(2) 235–292.

Bertsimas, D., V. Gupta, N. Kallus. 2018b. Robust Sample Average Approximation. *Mathematical Programming* **171**(1-2) 217–282.

Bertsimas, D., N. Kallus. 2019. From predictive to prescriptive analytics. To appear in *Management Science*.

Bertsimas, D., M. Sim. 2004. The price of robustness. *Operations Research* **52**(1) 35–53.

Bertsimas, D., J. N. Tsitsiklis. 1997. *Introduction to Linear Optimization*. Athena Scientific, Belmont, MA.

Bezanson, J., A. Edelman, S. Karpinski, V. B. Shah. 2017. Julia: A fresh approach to numerical computing. *SIAM Review* **59**(1) 65–98.

Bickel, P. J., Y. Ritov, A. B. Tsybakov. 2009. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37**(4) 1705–1732.

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

Bühlmann, P., S. A. Geer. 2011. *Statistics for High-Dimensional Data*. Springer-Verlag, Berlin, Heidelberg.

Candès, E. J., B. Recht. 2009. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics* **9**(6) 717.

Candes, E. J., C. A. Sing-Long, J. D. Trzasko. 2013. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing* **61**(19) 4643–4657.

Chen, X., Z. Owen, C. Pixton, D. Simchi-Levi. 2015. A statistical learning approach to personalization in revenue management. Working Paper, MIT.

Chen, X., M. Sim, P. Sun. 2007. A robust optimization perspective on stochastic programming. *Operations Research* **55**(6) 1058–1071.

Delage, E., Y. Ye. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* **58**(3) 595–612.

Donoho, D. L., I. M. Johnstone. 1995. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**(432) 1200–1224.

Efron, B. 2012. *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge, UK.

Efron, B., C. Morris. 1973. Stein's estimation rule and its competitors — An empirical bayes approach. *Journal of the American Statistical Association* **68**(341) 117–130.

Efron, B., C. Morris. 1975. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association* **70**(350) 311–319.

Elmachtoub, A. N, P. Grigas. 2017. Smart "predict, then optimize". Working Paper, Columbia University. arXiv preprint arXiv:1710.08005.

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.

Ferreira, K. J., B. H. A. Lee, D. Simchi-Levi. 2015. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* **18**(1) 69–88.

Frees, E. W. 1991. Linear regression and U-statistics. *Sankhyā: The Indian Journal of Statistics, Series A* **53**(1) 84–96.

Friedman, J., T. Hastie, R. Tibshirani. 2001. *The Elements of Statistical Learning*. Springer-Verlag, New York, NY.

Gupta, V. 2019. Near-optimal Bayesian ambiguity sets for distributionally robust optimization. To appear in *Management Science*. Available at `http://www.optimization-online.org/DB_FILE/2015/07/4983.pdf`.

Johnstone, I. M. 2015. *Gaussian Estimation: Sequence and Wavelet Models*. Working Draft. URL `http://statweb.stanford.edu/~imj/GE09-08-15.pdf`.

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.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, *IJCAI'95*, vol. 2. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1137–1145.

Lan, G. 2012. An optimal method for stochastic composite optimization. *Mathematical Programming* **133**(1) 365–397.

Lim, A. E .B., J. G. Shanthikumar, G. Y. Vahn. 2011. Conditional value-at-risk in portfolio optimization: Coherent but fragile. *Operations Research Letters* **39**(3) 163–171.

Liyanage, L. H., J. G. Shanthikumar. 2005. A practical inventory control policy using operational statistics. *Operations Research Letters* **33**(4) 341–348.

Morris, C. N. 1983. Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association* **78**(381) 47–55.

Mukherjee, G., L. D. Brown, P. Rusmevichientong. 2015. Empirical Bayes prediction for the multivariate newsvendor loss function. Working Paper, USC. arXiv preprint arXiv:1511.00028.

Negahban, S., B. Yu, M. J. Wainwright, P. K. Ravikumar. 2012. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27**(4) 538–557.

Nemirovski, A., A. Juditsky, G. Lan, A. Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* **19**(4) 1574–1609.

Nesterov, Y. 2005. Smooth minimization of non-smooth functions. *Mathematical Programming* **103**(1) 127–152.

Nesterov, Y. 2009. Primal-dual subgradient methods for convex problems. *Mathematical Programming* **120**(1) 221–259.

Pani, A., S. Raghavan, M. Sahin. 2017. Large-scale advertising portfolio optimization in online marketing. *Working Paper* URL `http://terpconnect.umd.edu/~raghavan/preprints/lsoapop.pdf`.

Pollard, D. 1990. Empirical processes: Theory and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, vol. 2. Institute of Mathematical Statistics, i–86.

Ross, Nathan, et al. 2011. Fundamentals of stein's method. *Probability Surveys* **8** 210–293.

Rusmevichientong, P., D. P. Williamson. 2006. An adaptive algorithm for selecting profitable keywords for search-based advertising services. *Proceedings of the 7th ACM Conference on Electronic Commerce*. ACM, 260–269.

Ruszczynski, A.P., A. Shapiro. 2003. *Stochastic Programming*, vol. 10. Elsevier Amsterdam.

Shapiro, A., D. Dentcheva, A. Ruszczyński. 2009. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics.

Stein, C. M. 1981. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9**(6) 1135–1151.

Tibshirani, R. J., J. Taylor. 2012. Degrees of freedom in lasso problems. *The Annals of Statistics* **40**(2) 1198–1232.

Van der Vaart, A. W. 2000. *Asymptotic Statistics*. No. 3 in Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK.

Vapnik, V. N. 1999. An overview of statistical learning theory. *IEEE transactions on neural networks* **10**(5) 988–999.

Wainwright, M. J. 2009. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory* **55**(12) 5728–5741.

Wainwright, M. J. 2015. Lecture notes. `http://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf`. [Online; accessed Feb 2016].

Xie, X., S. C. Kou, L. D. Brown. 2012. Sure estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association* **107**(500) 1465–1479.

Zhang, C.-H. 2003. Compound decision theory and empirical Bayes methods. *The Annals of Statistics* **31**(2) 379–390.

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: Small-Data Linear Optimization

Appendix A: Background Results on Uniform Laws of Large Numbers (ULLN)

Consider independent random variables $Z_1, \ldots, Z_n$ taking values in some abstract space Ξ. Let $\mathcal{T}$ be an arbitrary indexing set, and let $f_j(t, Z_j)$ be a sequence of functions $f_j : \mathcal{T} \times \Xi \to \mathbb{R}$. For a fixed $t \in \mathcal{T}$, the sum $\frac{1}{n}\sum_{j=1}^n f_j(t, Z_j) - \mathbb{E}[f_j(t, Z_j)]$ is a mean-zero random variable that, under suitable regularity conditions, will converge in probability to 0 as $n \to \infty$ by the law of large numbers. Uniform laws of large numbers establish conditions under which this convergence happens *uniformly* over $\mathcal{T}$, that is, conditions under which

$$\sup_{t\in\mathcal{T}} \frac{1}{n}\left|\sum_{j=1}^n f_j(t, Z_j) - \mathbb{E}[f_j(t, Z_j)]\right| \to_p 0 \quad \text{as } n \to \infty. \tag{A.1}$$

From an optimization perspective, such convergence results imply that for large n, minimizers of the sample average are approximate minimizers of the expected average; see Lemma C.1. There exists a well-developed literature on uniform laws of large numbers; see, for example, Pollard (1990) and Van der Vaart (2000). To keep our paper self-contained, we summarize a few key results. Our exposition and notation mirrors those of Pollard (1990). These results are not the tightest possible but are sufficient for our purposes.

For any $\mathcal{F} \subseteq R^n$, define the packing number $M(\epsilon, \mathcal{F})$ to be the largest number m such that there exist m points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{F}$ with $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 > \epsilon$, $1 \le i < j \le m$. For each fixed $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, let $\mathcal{F}(\boldsymbol{Z}) = \left\{ \big(f_1(t, Z_1), f_2(t, Z_2), \ldots, f_n(t, Z_n)\big) \in \mathbb{R}^n : t \in \mathcal{T} \right\} \subseteq \mathbb{R}^n$ be the set of n-dimensional vectors as t varies over the indexing set $\mathcal{T}$. Finally, let $F_j(Z_j)$ be a corresponding set of envelope functions, that is, functions satisfying $|f_j(t, Z_j)| \le F_j(Z_j)$ for all $t \in \mathcal{T}$ and $Z_j \in \Xi$. We write $\boldsymbol{F}(\boldsymbol{Z})$ to denote the vector whose j^{th} component is $F_j(Z_j)$.

Let $\Psi(t) = \frac{1}{5}\exp(t^2)$, and, for any real-valued random variable Z, define the Orlicz norm $\|Z\|_\Psi$ as follows: $\|Z\|_\Psi \equiv \inf\{C > 0 : \mathbb{E}[\Psi(|Z|/C)] \le 1\}$. Random variables with a finite Orlicz norm are sub-Gaussian (Pollard 1990). The following lemma summarizes well-known facts about $\|\cdot\|_\Psi$:

Lemma A.1 (Properties of the Orlicz Norm)

i) *For any constant C', $\|C'\|_\Psi \le |C'|$.*

ii) *If Z is a mean-zero, sub-Gaussian random variable with variance proxy at most σ^2, then $\|Z\|_\Psi \le 2\sigma$.*

iii) *For $j = 1, \ldots, n$, let Z_j, be a mean-zero, sub-Gaussian random variable with variance proxy at most σ^2. Then, $\|\max_{j=1,\ldots,n} Z_j\|_\Psi \le 2\sigma\sqrt{2 + \log n}$.*

iv) *Let* $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ *be a vector of independent, mean-zero sub-Gaussian random variables each with variance proxy at most* σ^2*. If* $n > 1$*, then* $\| \|\boldsymbol{Z}\|_2 \|_\Psi \leq \sigma\sqrt{2n}$*.*

Proof: The first claim follows immediately from the definition.

For the second claim, using the tail expectation formula, we obtain

$$\begin{aligned}\mathbb{E}\left[\exp\left(Z^2/C^2\right)\right] &= \int_0^\infty \mathbb{P}\left(\exp\left(Z^2/C^2\right) > t\right) dt = 1 + \int_1^\infty \mathbb{P}\left(|Z| > |C|\sqrt{\log(t)}\right) dt \\ &\leq 1 + 2\int_1^\infty \exp\left(-\frac{C^2\log(t)}{2\sigma^2}\right) dt,\end{aligned}$$

where the last inequality follows because Z is sub-Gaussian. This integral converges if $C > \sqrt{2}\sigma$, yielding $\mathbb{E}\left[\exp\left(Z^2/C^2\right)\right] \leq 1 + \frac{4\sigma^2}{C^2 - 2\sigma^2}$. For $C = 2\sigma$, we get $\mathbb{E}\left[\exp\left(Z^2/C^2\right)\right] \leq 3$, so $\|Z\|_\Psi \leq 2\sigma$.

For the third claim, (Pollard 1990, Lemma 3.2, pg. 10) proves $\| \max_{j=1,\ldots,n} Z_j \|_\Psi \leq \sqrt{2 + \log n} \max_{j=1,\ldots,n} \|Z_j\|_\Psi$. Applying our previous result for sub-Gaussian random variables proves the claim.

For the final claim, using independence of $Z_1, \ldots, Z_n$,

$$\mathbb{E}\left[\exp\left(\left\|\frac{\boldsymbol{Z}}{\sigma\sqrt{2n}}\right\|_2^2\right)\right] = \mathbb{E}\left[\exp\left(\sum_{j=1}^n \frac{Z_j^2}{2\sigma^2 n}\right)\right] = \prod_{j=1}^n \mathbb{E}\left[\exp\left(\frac{Z_j^2}{2\sigma^2 n}\right)\right] \leq \left(1 - \frac{1}{n}\right)^{-n/2},$$

where the last inequality follows from (Wainwright 2015, Thm. 2.1 Part IV, pg. 16) and uses the fact that $\frac{1}{n} < 1$. This function is decreasing in n, and at $n = 2$, it equals 2, which is less than 5. This completes the proof. □

By specializing the results of Pollard (1990), we have the following:

Theorem A.2 *Suppose that there exist constants* A, W *(not depending on* ϵ*) such that for each* $\boldsymbol{Z} \in \Xi^n$*,*

$$M\left(\epsilon \|\boldsymbol{F}(\boldsymbol{Z})\|_2, \mathcal{F}(\boldsymbol{Z})\right) \leq A\epsilon^{-W}. \tag{A.2}$$

Let $V(A, W) \equiv \frac{W + 2\log A}{W + \sqrt{\log A}}$*. If* $\| \|\boldsymbol{F}(\boldsymbol{Z})\|_2 \|_\Psi \leq K$*, then*

$$\mathbb{P}\left\{\sup_{t\in\mathcal{T}} \left\| \sum_{j=1}^n f_j\left(t, Z_j\right) - \mathbb{E}\left[f_j(t, Z_j)\right] \right\| > t \right\} \leq 25\exp\left(\frac{-t}{9\,K\,V(A,W)}\right).$$

Proof: We sketch how the results are obtained as a special case of Pollard (1990). It follows from Equation (7.4) of Pollard (1990) that

$$\mathbb{P}\left\{\sup_{t\in\mathcal{T}} \left\| \sum_{j=1}^n f_j\left(t, Z_j\right) - \mathbb{E}\left[f_j(t, Z_j)\right] \right\| > t \right\} \leq 25\exp\left(\frac{-t}{\|J_n(\boldsymbol{Z})\|_\Psi}\right),$$

where $J_n(\boldsymbol{Z}) \equiv 9\int_0^{\sup\{\|\boldsymbol{f}\|\,:\,\boldsymbol{f}\in\mathcal{F}(\boldsymbol{Z})\}} \sqrt{\log M(\epsilon,\mathcal{F}(\boldsymbol{Z}))}\; d\epsilon$. Thus, it suffices to shows that the $\|J_n(\boldsymbol{Z})\|_\Psi \le 9\,K\,V(A,W)$. As discussed in Equation (7.7) in Chapter 7 of Pollard (1990), the entropy integral can be bounded in terms of the envelopes $\boldsymbol{F}(\boldsymbol{Z})$, A, and W as follows:

$$J_n(\boldsymbol{Z}) \le 9\|\boldsymbol{F}(\boldsymbol{Z})\|_2 \int_0^1 \sqrt{W\log(1/\epsilon)+\log A}\; d\epsilon.$$

Make the change of variables $\sqrt{W\log(1/\epsilon)+\log A} \mapsto u$, so $\epsilon = \exp\left(-(u^2-\log A)/W\right)$ with $d\epsilon = -\frac{2u}{W}\exp\left(-(u^2-\log A)/W\right)du$. Thus,

$$\begin{aligned}\int_0^1 \sqrt{W\log(1/\epsilon)+\log A}\; d\epsilon &= \int_{\sqrt{\log A}}^\infty \frac{2u^2}{W}\exp\left(-(u^2-\log A)/W\right)du \\ &= \sqrt{W}A^{1/W}\int_{\sqrt{\frac{\log A}{W}}}^\infty 2t^2\exp(-t^2)dt,\end{aligned}$$

where the last equality follows from the change of variable $u/\sqrt{W}\mapsto t$. Using integration by parts, the last integral is equal to $\sqrt{\log A} + \frac{\sqrt{\pi W}}{2}A^{1/W}\mathrm{Erfc}\left(\sqrt{\frac{\log A}{W}}\right)$, where $\mathrm{Erfc}(s)\equiv\frac{2}{\sqrt{\pi}}\int_s^\infty \exp(-t^2)dt$ is the complementary error function. Substitute the standard bound, $\mathrm{Erfc}(s)\le\frac{2}{\sqrt{\pi}}\exp(-s^2)/s$, and simplify to yield $\int_0^1\sqrt{W\log(1/\epsilon)+\log A}\,d\epsilon \le V(A,W)$. Thus, $J_n(\boldsymbol{Z})\le 9V(A,W)\|\boldsymbol{F}(\boldsymbol{Z})\|_2$, so $\|J_n(\boldsymbol{Z})\|_\Psi \le 9\,V(A,W)\,K$, and this completes the proof. □

As noted in the theorem below, Equation (A.2) of Theorem A.2 is satisfied by sets with bounded pseudo-dimension (Pollard 1990). Recall that the set $\mathcal{F}(\boldsymbol{Z})$ has pseudo-dimension at most V if for any $\boldsymbol{c}\in\mathbb{R}^n$ and subset of indices $\mathcal{J}\subseteq\{1,\ldots,n\}$ with $|\mathcal{J}|=V+1$,

$$\left|\left\{\left(\,\mathbb{I}(f_j(t,Z_j)>c_j)\;|\;j\in\mathcal{J}\,\right)\in\{0,1\}^{|\mathcal{J}|}\;:\;t\in\mathcal{T}\right\}\right| \;<\; 2^{V+1}.$$

Here $\left(\,\mathbb{I}(f_j(t,Z_j)>c_j)\;|\,j\in\mathcal{J}\,\right)$ denotes a binary vector of dimension $|\mathcal{J}|$. If the above inequality holds instead with equality, we say that $\boldsymbol{c}$ is a witness to the shattering of $\mathcal{J}$.

Theorem A.3 *If $\mathcal{F}(\boldsymbol{Z})$ has pseudo-dimension at most $V\ge 2$, then $M\left(\epsilon\|\boldsymbol{F}(\boldsymbol{Z})\|_2,\mathcal{F}(\boldsymbol{Z})\right)$ satisfies Equation* (A.2) *of Theorem A.2, with $W=4V$, $A=V^{6V}$, and $V(A,W)\le 1+3\log(V)$.*

Proof: Again, we specialize results from Pollard (1990). Specifically, tracking the constants from Theorems 4.7, 4.8, and 4.10 of Pollard (1990), we find that the theorem holds for $A\ge(1+V)^2/C^2$ and $W=4V$, where $C=\min_{t\ge1}\sqrt{t}/(1+2\log t)^V$. By differentiating and substituting, we find $C=(4V)^{-V}\exp(V-\frac{1}{4})$, so that

$$\log C = -V(\log(4)-1)-V\log V-\frac{1}{4}\ge -2V\log V,$$

where the inequality follows by comparing the derivatives of both sides for $V\ge2$. Then,

$$\log\left(\frac{(1+V)^2}{C^2}\right) = 2\log(V+1)-2\log C \le 2\log(V+1)+4V\log V\le 6V\log(V),$$

where the last inequality again follows by comparing derivatives. This proves the claim for A. To bound $V(A,W)$, note $\sqrt{\log A}>0$, so that $V(A,W)=\frac{W+2\log A}{W+\sqrt{\log A}} \le 1+\frac{2\log A}{W} \le 1+3\log(V)$. □

Appendix B: Proofs of the Results in Section 2

We now present proofs of the results in Section 2.

B.1. Proof of Theorem 2.7:

As mentioned, the proof of Theorem 2.7 involves generating a random instance of P^n. To that end, let π be a probability distribution on $\mathcal{M} \subseteq \mathbb{R}^n$. Consider the hierarchical Bayes model where the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ follows a prior distribution π and

$$W_j | \boldsymbol{Y} \sim \mathcal{N}(Y_j, 1/\nu_j), \text{ independently, } j = 1, \ldots, n. \tag{B.1}$$

Consider then the Bayesian decision problem $\max_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{n} \boldsymbol{Y}^\top \boldsymbol{x}$ where $\boldsymbol{Y}$ is unobserved and $\boldsymbol{W}$ is data. This problem is a random instance of P^n, where the true values of the unknown parameter $\boldsymbol{\mu}$ are considered random. The expected performance of $\boldsymbol{x}(\cdot)$ in P^n for a fixed realization of $\boldsymbol{\mu}$ equals $\mathbb{E}\left[\frac{1}{n} \boldsymbol{Y}^\top \boldsymbol{x}(\boldsymbol{W}) \ \middle| \ \boldsymbol{Y} = \boldsymbol{\mu}\right]$.

A straightforward computation shows that the Bayes-optimal solution with respect to π is

$$\boldsymbol{x}^{Bayes}(\pi, \boldsymbol{W}) \ \in \ \arg\max_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}[Y_j | \boldsymbol{W}] x_j.$$

Before we prove Theorem 2.7, we use this Bayes optimal solution to compute an upper bound on the expected revenue of any data-driven policy.

Lemma B.1 (Bayes Policies Bound Worst-Case Performance) *Let $\boldsymbol{x}(\cdot)$ denote any data-driven policy. For any compact set $\mathcal{M} \subset \mathbb{R}^n$ and prior distribution $\pi : \mathcal{M} \to \mathbb{R}_+$,*

$$\inf_{\boldsymbol{\mu} \in \mathcal{M}} \mathbb{E}\left[\frac{1}{n} \boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\mu}})\right] \ \leq \ \mathbb{E}\left[\frac{1}{n} \boldsymbol{Y}^\top \boldsymbol{x}(\boldsymbol{W})\right] \ \leq \ \mathbb{E}\left[\frac{1}{n} \boldsymbol{Y}^\top \boldsymbol{x}^{Bayes}(\pi, \boldsymbol{W})\right] \ .$$

Proof: Note $\inf_{\boldsymbol{\mu} \in \mathcal{M}} \mathbb{E}\left[\frac{1}{n} \boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\mu}})\right] = \inf_{\boldsymbol{\mu} \in \mathcal{M}} \mathbb{E}\left[\frac{1}{n} \boldsymbol{Y}^\top \boldsymbol{x}(\boldsymbol{W}) \,|\, \boldsymbol{Y} = \boldsymbol{\mu}\right]$. Thus, the first inequality follows because the worst-case reward is less than or equal to the average reward over $\mathcal{M}$. For the second,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n} \boldsymbol{Y}^\top \boldsymbol{x}(\boldsymbol{W})\right] \ &= \ \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} Y_j x_j(\boldsymbol{W})\right] \ = \ \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[Y_j | \boldsymbol{W}\right] x_j(\boldsymbol{W})\right] \\
&\leq \ \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[Y_j | \boldsymbol{W}\right] x_j^{Bayes}(\pi, \boldsymbol{W})\right] \ = \ \mathbb{E}\left[\frac{1}{n} \boldsymbol{Y}^\top \boldsymbol{x}^{Bayes}(\pi, \boldsymbol{W})\right] \ ,
\end{aligned}$$

where the inequality follows from the definition of a Bayes optimal policy. □

Lemma B.1 asserts that for any $\boldsymbol{\mu}$, the Bayes-optimal performance in the system Eq. (B.1) upper bounds the worst-case expected performance of *all other* data-driven policies in our original system. We note that this result is actually a special case of a considerably more general, classical result in statistical minimax decision theory (Wald 1947, 1949).

We can now prove the theorem.

Proof of Theorem 2.7: Write

$$\inf_{\boldsymbol{\mu}\in\{-1,+1\}^n} \frac{\frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\mu}})}{Z^*(\boldsymbol{\mu})} \;=\; 1 - \sup_{\boldsymbol{\mu}\in\{-1,+1\}^n} \frac{Z^*(\boldsymbol{\mu}) - \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\mu}})}{Z^*(\boldsymbol{\mu})} \;\le\; 1 - \sup_{\boldsymbol{\mu}\in\{-1,+1\}^n} Z^*(\boldsymbol{\mu}) - \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\mu}}),$$

where the inequality follows because $\mathcal{X}_n = [0,1]^n$, so $Z^*(\boldsymbol{\mu}) \le 1$ for all $\boldsymbol{\mu} \in \{-1,+1\}^n$. We will lower bound this supremum.

Take a Rademacher prior π for $\boldsymbol{Y}$; that is, for all j, $\mathbb{P}\{Y_j = 1\} = \mathbb{P}\{Y_j = -1\} = \frac{1}{2}$, and $Y_1, \ldots, Y_n$ are independent. Then,

$$\begin{aligned}
\sup_{\boldsymbol{\mu}\in\{-1,+1\}^n} Z^*(\boldsymbol{\mu}) - \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\mu}}) &= \sup_{\boldsymbol{\mu}\in\{-1,+1\}^n} \left\{ \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}^*(\boldsymbol{\mu}) - \mathbb{E}\left[\frac{1}{n}\boldsymbol{Y}^\top \boldsymbol{x}(\boldsymbol{W}) \,|\, \boldsymbol{Y} = \boldsymbol{\mu} \right] \right\} \\
&\ge \mathbb{E}\left[\frac{1}{n}\boldsymbol{Y}^\top \boldsymbol{x}^*(\boldsymbol{Y}) - \frac{1}{n}\boldsymbol{Y}^\top \boldsymbol{x}(\boldsymbol{W}) \right] \qquad \text{(B.2)} \\
&= \mathbb{E}\left[\frac{1}{n}\boldsymbol{Y}^\top \boldsymbol{x}^*(\boldsymbol{Y}) \right] - \mathbb{E}\left[\frac{1}{n}\boldsymbol{Y}^\top \boldsymbol{x}(\boldsymbol{W}) \right] \\
&\ge \mathbb{E}\left[\frac{1}{n}\boldsymbol{Y}^\top \boldsymbol{x}^*(\boldsymbol{Y}) \right] - \mathbb{E}\left[\frac{1}{n}\boldsymbol{Y}^\top \boldsymbol{x}^{Bayes}(\pi, \boldsymbol{W}) \right], \qquad \text{(B.3)}
\end{aligned}$$

where inequality (B.2) follows because the supremum exceeds the average, and inequality (B.3) follows from Lemma B.1.

By inspection $x_j^*(\boldsymbol{Y}) = \mathbb{I}(Y_j \ge 0)$, so $\mathbb{E}\left[\frac{1}{n}\boldsymbol{Y}^\top \boldsymbol{x}^*(\boldsymbol{Y})\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Y_j^+\right] = \frac{1}{2}$.

It only remains to upper bound the Bayesian policy. Again, by inspection, $x_j^{Bayes}(\pi, \boldsymbol{W}) = \mathbb{I}(\mathbb{E}[Y_j \,|\, \boldsymbol{W}] \ge 0)$. Note that

$$\mathbb{P}\{Y_j = +1 \,|\, W_j\} = \frac{\phi(W_j - 1)}{\phi(W_j - 1) + \phi(W_j + 1)} \quad \text{and} \quad \mathbb{P}\{Y_j = -1 \,|\, W_j\} = \frac{\phi(W_j + 1)}{\phi(W_j - 1) + \phi(W_j + 1)} \;,$$

so

$$\mathbb{E}\left[Y_j | \boldsymbol{W}\right] = \mathbb{E}\left[Y_j | W_j\right] = \frac{\phi(W_j - 1) - \phi(W_j + 1)}{\phi(W_j - 1) + \phi(W_j + 1)} = \tanh(W_j) \;,$$

and, therefore $\mathbb{E}\left[Y_j | \boldsymbol{W}\right] \ge 0$ if and only $W_j \ge 0$. Thus, $x_j^{Bayes}(\pi, \boldsymbol{W}) = \mathbb{I}(W_j \ge 0)$, and

$$\frac{1}{n}\sum_{j=1}^n \mathbb{E}\left[Y_j | \boldsymbol{W}\right] x_j^{Bayes}(\pi, \boldsymbol{W}) = \frac{1}{n}\sum_{j=1}^n \frac{\phi(W_j - 1) - \phi(W_j + 1)}{\phi(W_j - 1) + \phi(W_j + 1)} \mathbb{I}(W_j \ge 0).$$

Finally, note that the density function of each W_j is $t \mapsto \frac{1}{2}\left(\phi(t-1) + \phi(t+1)\right)$, so by symmetry,

$$\begin{aligned}
\frac{1}{n}\mathbb{E}\left[\sum_{j=1}^n \mathbb{E}\left[Y_j | \boldsymbol{W}\right] x_j^{Bayes}(\pi, \boldsymbol{W}) \right] \;&=\; \mathbb{E}\left[\frac{\phi(W_1 - 1) - \phi(W_1 + 1)}{\phi(W_1 - 1) + \phi(W_1 + 1)} \mathbb{I}(W_1 \ge 0) \right] \\
&=\; \frac{1}{2}\int_0^\infty (\phi(w-1) - \phi(w+1)) dw \\
&=\; \frac{\Phi(1) - \Phi(-1)}{2} \;,
\end{aligned}$$

which implies that $\sup_{\boldsymbol{\mu}\in\{-1,+1\}^n} \left\{ Z^*(\boldsymbol{\mu}) - \frac{1}{n}\mathbb{E}\left[\boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\mu}})\right] \right\} \ge \frac{1}{2} - \frac{\Phi(1) - \Phi(-1)}{2} = \Phi(-1)$. Thus,

$$\inf_{\boldsymbol{\mu}\in\{-1,+1\}^n} \frac{\frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\mu}})}{Z^*(\boldsymbol{\mu})} \le 1 - \Phi(-1) < 0.842 \;.$$

□

Appendix C: Proof of the Results in Section 3

We now present proofs of the results in Section 3. Often we will optimize an approximation to a target function instead of optimizing the target function directly. The following lemma quantifies the sub-optimality induced in the solution by the approximation. We will use this lemma repeatedly throughout.

Lemma C.1 (Uniform Approximation) *Let $f_1 : T \mapsto \mathbb{R}$ and $f_2 : T \mapsto \mathbb{R}$ be two functions, and let $t_1 \in \arg\max_{t\in T} f_1(t)$ and $t_2 \in \arg\max_{t\in T} f_2(t)$ be their respective maximizers. Then,*

$$0 \leq f_1(t_1) - f_1(t_2) \leq 2 \sup_{t\in T} |f_1(t) - f_2(t)| \ .$$

Proof: The first inequality follows from optimality of t_1. For the second, note that

$$f_1(t_1) - f_1(t_2) = f_1(t_1) - f_2(t_1) + f_2(t_1) - f_2(t_2) + f_2(t_2) - f_1(t_2) \leq 2 \sup_{t\in T} |f_1(t) - f_2(t)| \ ,$$

where we use the optimality of t_2 to drop the second term. □

Proof of Lemma 3.1: As in Example 3.4, we have that there exists $\boldsymbol{\theta}^{SAA} \in \Theta$ such that $\boldsymbol{x}(\boldsymbol{\theta}^{SAA}, \hat{\boldsymbol{\mu}}) = \boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})$ and $\boldsymbol{\theta}^{SAA} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x}(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}})$. Furthermore, the bias correction $B(\boldsymbol{\theta}, h, \hat{\boldsymbol{\mu}})$ is uniformly bounded because $x_j(\cdot) \in [0, 1]$ and

$$\sup_{\theta\geq\Theta} \left| \frac{1}{n} \sum_{j=1}^{n} \frac{1}{2h\sqrt{\nu_j}} \left[x_j\Big(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}} + \frac{h}{\sqrt{\nu_j}}\boldsymbol{e}_j\Big) - x_j\Big(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}} - \frac{h}{\sqrt{\nu_j}}\boldsymbol{e}_j\Big) \right] \right| \ \leq \ \frac{1}{2h\sqrt{\nu_{\min}}}.$$

Thus, the objective of Eq. (3.3) is a uniform approximation to the objective of $\sup_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x}(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}})$. Applying Lemma C.1 completes the proof. □

Proof of Lemma 3.6 By definition of $\boldsymbol{x}^*(\boldsymbol{\mu})$, $Z^*(\boldsymbol{\mu}) - \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}}) \ = \ \frac{1}{n}\boldsymbol{\mu}^\top(\boldsymbol{x}^*(\boldsymbol{\mu}) - \boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})) \geq 0$. Moreover,

$$\begin{aligned}
\boldsymbol{\mu}^\top(\boldsymbol{x}^*(\boldsymbol{\mu}) - \boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})) &= (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{x}^*(\boldsymbol{\mu}) + \hat{\boldsymbol{\mu}}^\top \left(\boldsymbol{x}^*(\boldsymbol{\mu}) - \boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})\right) \\
&\quad + \hat{\boldsymbol{\mu}}^\top \left(\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}}) - \boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}})\right) + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \boldsymbol{x}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\mu}}) \\
&\leq 2 \sup_{\boldsymbol{x}\in\mathcal{X}} \left|(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{x}\right| + \frac{n}{h\sqrt{\nu_{\min}}} \ \leq \ 2\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_1 + \frac{n}{h\sqrt{\nu_{\min}}},
\end{aligned}$$

where the first inequality uses the optimality of $\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})$ and Lemma 3.1, and the second inequality follows from Cauchy-Schwarz inequality and $\mathcal{X} \subseteq [0, 1]^n$. By Jensen's Inequality,

$$\mathbb{E}[\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_1 \ = \sum_{j=1}^{n} \mathbb{E}[|\hat{\mu}_j - \mu_j|] \ = \ \sum_{j=1}^{n} \mathbb{E}\left[\sqrt{(\hat{\mu}_j - \mu_j)^2}\right] \ \leq \ \frac{n}{\sqrt{\nu_{\min}}} \ ,$$

and putting everything together proves the lemma. □

C.1. Proof of Theorem 3.5

The proof of Theorem 3.5 makes use of Lemma C.1 (above). Indeed, this lemma suggests we bound $2\sup_{\theta\in\Theta}\left|\frac{1}{n}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top \boldsymbol{x}(\theta,\hat{\boldsymbol{\mu}}) - B(\theta,h,\hat{\boldsymbol{\mu}})\right|$ to prove Theorem 3.5.

The key idea in bounding this suprema is an approximate version of Stein's Lemma. For any differentiable function f, Stein's Lemma asserts $\mathbb{E}[\zeta f(\zeta)] = \mathbb{E}[f'(\zeta)]$ whenever ζ is a standard normal random variable. In Appendix B, we prove an extension for random variables that are approximately Gaussian, and replace the derivative by a first order finite difference. We consider almost everywhere (rather than everywhere) differentiable functions in order to handle discontinuous functions such as indicators, such as in the proof of Theorem 4.3 in Section 4.

Lemma C.2 (Approximate Stein's Lemma) *Let $0 < h < 1$ and ξ be a mean-zero, sub-Gaussian random variable with variance proxy at most σ^2. Suppose that $\mathbb{E}[\xi^2]=1$ and that ξ admits a density, denoted $\overline{\phi}(\cdot)$. Suppose further that f is almost everywhere differentiable. Then,*

$$\left|\mathbb{E}[\xi f(\xi)] - \mathbb{E}\left[\frac{1}{2h}(f(\xi+h)-f(\xi-h))\right]\right| \le 4\|f\|_\infty h^2 + \|\overline{\phi}-\phi\|_1\|f\|_\infty\left(h^{-1}+24\sigma^2-\log\left(\|\overline{\phi}-\phi\|_1\right)\right).$$

When $\|\overline{\phi}-\phi\|_1 = 0$, i.e., when ξ is Gaussian, the difference in expectations is bounded by $4h^2\|f\|_\infty$. If, in addition, $h\to 0$, the difference converges to 0, recovering the original Stein Lemma. The proof of this lemma is given in Section C.2.

Equipped with these two lemmas, we can prove Theorem 3.5.

Proof of Theorem 3.5: By definition,

$$\boldsymbol{\theta}_n^{\mathsf{OR}} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}) \qquad \text{and} \qquad \hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}) - B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}}) \ .$$

By Lemma C.1, it suffices to bound $\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})) - B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}})\right|$. By triangle inequality,

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}) - B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}})\right| &\le \sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}) - \frac{1}{n}\mathbb{E}\Big[(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top \boldsymbol{x}(\theta,\hat{\boldsymbol{\mu}})\Big]\right| \\ &\quad + \sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\mathbb{E}\Big[(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})\Big] - \mathbb{E}\Big[B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}})\Big]\right| \\ &\quad + \sup_{\boldsymbol{\theta}\in\Theta}\left|\mathbb{E}\Big[B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}})\Big] - B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}})\right|.\end{aligned}$$

We focus on the second term. For each $\boldsymbol{\theta}\in\Theta$,

$$\begin{aligned}&\left|\frac{1}{n}\mathbb{E}\Big[(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top \boldsymbol{x}(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})\Big] - \mathbb{E}\Big[B(\boldsymbol{\theta},h,\hat{\boldsymbol{\mu}})\Big]\right| \\ &\quad\le \frac{1}{n}\sum_{j=1}^n\left|\mathbb{E}\Big[(\hat{\mu}_j-\mu_j)x_j(\boldsymbol{\theta},\hat{\boldsymbol{\mu}})\Big] - \mathbb{E}\left[\frac{1}{2h\sqrt{\nu_j}}\left(x_j\Big(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}+\frac{h}{\sqrt{\nu_j}}\boldsymbol{e}_j\Big) - x_j\Big(\boldsymbol{\theta},\hat{\boldsymbol{\mu}}-\frac{h}{\sqrt{\nu_j}}\boldsymbol{e}_j\Big)\right)\right]\right| \\ &\quad= \frac{1}{n}\sum_{j=1}^n\left|\mathbb{E}\left[\zeta_j f_j(\zeta_j)\right] - \mathbb{E}\left[\frac{1}{2h}\left(f_j(\zeta_j+h)-f_j(\zeta_j-h)\right)\right]\right| \ ,\end{aligned}$$

where for all j, $\zeta_j = (\hat{\mu}_j - \mu_j)\sqrt{\nu_j}$ and $f_j : \mathbb{R} \to \mathbb{R}$ is defined by

$$f_j(t) = \frac{1}{\sqrt{\nu_j}}\mathbb{E}\left[x_j\Big(\boldsymbol{\theta}, \Big(\frac{t}{\sqrt{\nu_j}} + \mu_j\Big)\boldsymbol{e}_j + \sum_{\ell \neq j} \hat{\mu}_\ell \boldsymbol{e}_\ell\Big) \;\Big|\; \hat{\mu}_j = \frac{t}{\sqrt{\nu_j}} + \mu_j \right].$$

Note that ζ_j has mean-zero, variance 1, and is sub-Gaussian with variance proxy at most σ^2. Moreover, $\|f\|_\infty \leq \frac{1}{\sqrt{\nu_j}}$ since $x_j(\cdot) \in [0,1]$. By Lemma C.2, we have that

$$\begin{aligned}
&\left| \frac{1}{n}\mathbb{E}[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \boldsymbol{x}(\boldsymbol{\theta}, \hat{\boldsymbol{\mu}}))] - \mathbb{E}[B(\boldsymbol{\theta}, h, \hat{\boldsymbol{\mu}})] \right| \\
&\quad \leq \frac{1}{n}\sum_{j=1}^n \frac{4h^2}{\sqrt{\nu_j}} + \frac{1}{\sqrt{\nu_j}}\|\phi_j - \phi\|_1 \left(\frac{1}{h} + 24\sigma^2 - \log(\|\phi_j - \phi\|_1) \right) \\
&\quad \leq \frac{4h^2}{\sqrt{\nu_{\mathsf{min}}}} + \frac{(h^{-1} + 24\sigma^2)}{\sqrt{\nu_{\mathsf{min}}}} \cdot \frac{1}{n}\sum_{j=1}^n \|\phi_j - \phi\|_1 - \frac{1}{n\sqrt{\nu_{\mathsf{min}}}} \sum_{j=1}^n \|\phi_j - \phi\|_1 \log(\|\phi_j - \phi\|_1)
\end{aligned}$$

We can "clean up" the bound slightly. Let $c = h^{-1} + 24\sigma^2$, so that $h < 1$ implies $c > 1$. Note $t \mapsto -t\log(t)$ is concave, so by Jensen's inequality $-\frac{1}{n}\sum_{j=1}^n \|\phi_j - \phi\|_1 \log(\|\phi_j - \phi\|_1) \leq -2\mathsf{TV}\log(2\mathsf{TV}) \leq 2\mathsf{TV}\log\left(\frac{1}{2\mathsf{TV}}\right) \leq 2c\mathsf{TV}\log\left(\frac{e}{\mathsf{TV}}\right)$. Substituting above and simplifying yields

$$\frac{4h^2}{\sqrt{\nu_{\mathsf{min}}}} + \frac{2c\mathsf{TV}}{\sqrt{\nu_{\mathsf{min}}}} + \frac{2c\mathsf{TV}\log(e/\mathsf{TV})}{\sqrt{\nu_{\mathsf{min}}}}.$$

Finally note that $0 \leq \mathsf{TV} \leq 1$ implies that $\mathsf{TV} \leq \mathsf{TV}\log(e/\mathsf{TV})$, so this quantity is at most $\frac{4h^2}{\sqrt{\nu_{\mathsf{min}}}} + \frac{4c\mathsf{TV}\log(e/\mathsf{TV})}{\sqrt{\nu_{\mathsf{min}}}}$. Replacing the value of c yields the result. □

C.2. Proof of Lemma C.2

Proof: The lemma is trivial if $\|f\|_\infty = \infty$, so assume that $\|f\|_\infty < \infty$. We first prove the lemma in the special case that ξ has a standard normal distribution. In this case, $\overline{\phi} = \phi$, so $\|\phi - \overline{\phi}\|_1 = 0$, and the last two terms of the bound are zero.

Let $K(t) = \frac{1}{2}\mathbb{I}(|t| \leq 1)$ denote the box kernel. Write

$$\begin{aligned}
\int_{z\in\mathbb{R}} \phi(z)\frac{1}{2h}(f(z+h) - f(z-h))dz &= \int_{z\in\mathbb{R}} \phi(z) \int_{t\in\mathbb{R}} f'(t) K\left(\frac{t-z}{h}\right) h^{-1} dt dz \\
&= \int_{t\in\mathbb{R}} f'(t) \int_{z\in\mathbb{R}} \phi(z) K\left(\frac{z-t}{h}\right) h^{-1} dz dt \\
&= \int_{t\in\mathbb{R}} f'(t)\frac{1}{2h}(\Phi(t+h) - \Phi(t-h))dt \\
&= -\int_{t\in\mathbb{R}} f(t)\frac{1}{2h}(\phi(t+h) - \phi(t-h))dt.
\end{aligned}$$

The second equality follows from Fubini's theorem, since $\left|\phi(z)f'(t)K\left(\frac{t-z}{h}\right)h^{-1}\right| \leq \phi(z)h^{-1}f'(t)$, which is integrable. The last equality is integration by parts.

Now bound the difference in expectations from the lemma as

$$\int_{t\in\mathbb{R}} f(t)\left(t\phi(t)+\frac{1}{2h}(\phi(t+h)-\phi(t-h))\right)dt \le \|f\|_\infty \int_{t\in\mathbb{R}}\left|t\phi(t)+\frac{1}{2h}(\phi(t+h)-\phi(t-h))\right|dt.$$
$$= \|f\|_\infty \int_{t\in\mathbb{R}}\left|\phi'(t)-\frac{1}{2h}(\phi(t+h)-\phi(t-h))\right|dt,$$

where the last equality follows from $-\phi'(t)=t\phi(t)$.

From a Taylor expansion and mean-value theorem, we have for some $t_1\in[t,t+h]$, $t_2\in[t-h,t]$,

$$\left|\phi'(t)-\frac{1}{2h}(\phi(t+h)-\phi(t-h))\right| \le \frac{h^2}{2\cdot 3!}\left|\phi^{(3)}(t_1)-\phi^3(t_2)\right| \le \frac{h^2}{3!}\sup_{s\in[t-h,t+h]}\left|\phi^{(3)}(s)\right|,$$

where $\phi^{(3)}(s)$ is the third derivative of the normal density. A direct computation shows that $\left|\phi^{(3)}(s)\right| = \phi(s)\,|s|\,|3-s^2|$. Thus, using $0<h<1$ and that $\phi(s)$ is decreasing in $|s|$, we bound

$$\left|\phi'(t)-\frac{1}{2h}(\phi(t+h)-\phi(t-h))\right| \le \frac{h^2}{2\cdot 3!}\phi(|t|-1)(|t|+1)(3+(|t|+1)^2).$$

Numerically integrating the right-hand side over t shows that it is at most $4h^2$, proving the lemma when ξ is a standard normal.

We now consider the case that ξ is not normal. Let $\zeta\sim\mathcal{N}(0,1)$. We will bound the error incurred by replacing ξ by ζ in the expectations of the lemma. Specifically, for any $T>0$, write

$$\begin{aligned}
\left|\mathbb{E}\big[\xi f(\xi)\big]-\mathbb{E}\big[\zeta f(\zeta)\big]\right| &\le \left|\mathbb{E}\big[\xi f(\xi)\cdot\mathbb{I}\{|\xi|\le T\}\big]-\mathbb{E}\big[\zeta f(\zeta)\cdot\mathbb{I}\{|\zeta|\le T\}\big]\right| \\
&\quad + \left|\mathbb{E}\big[\xi f(\xi)\cdot\mathbb{I}\{|\xi|> T\}\big]\right| + \left|\mathbb{E}\big[\zeta f(\zeta)\cdot\mathbb{I}\{|\zeta|> T\}\big]\right| \\
&= \left|\int_{t:|t|\le T} t f(t)(\overline{\phi}(t)-\phi(t))dt\right| \\
&\quad + \left|\mathbb{E}\big[\xi f(\xi)\cdot\mathbb{I}\{|\xi|> T\}\big]\right| + \left|\mathbb{E}\big[\zeta f(\zeta)\cdot\mathbb{I}\{|\zeta|> T\}\big]\right| \\
&\le \int_{t:|t|\le T} |t f(t)|\left|\overline{\phi}(t)-\phi(t)\right|dt \\
&\quad + \mathbb{E}\big[\,|\xi f(\xi)|\cdot\mathbb{I}\{|\xi|> T\}\big] + \mathbb{E}\big[\,|\zeta f(\zeta)|\cdot\mathbb{I}\{|\zeta|> T\}\big] \\
&\le T\|f\|_\infty\|\phi-\overline{\phi}\|_1 + \|f\|_\infty\mathbb{E}\big[\,|\xi|\,\mathbb{I}\{|\xi|> T\}\big] + \|f\|_\infty\mathbb{E}\big[\,|\zeta|\cdot\mathbb{I}\{|\zeta|> T\}\big]
\end{aligned}$$

We bound $\mathbb{E}\big[\,|\xi|\,\mathbb{I}\{|\xi|>T\}\big]$ using the tail integral for expectation:

$$\begin{aligned}
E\big[\,|\xi|\,\mathbb{I}\{|\xi|>T\}\big] &= \int_0^\infty \mathbb{P}\left(|\xi|\,\mathbb{I}\{|\xi|>T\}>t\right)dt \\
&= \int_0^\infty \mathbb{P}\left(|\xi|>t \text{ and } |\xi|>T\right)dt \\
&= T\mathbb{P}\left(|\xi|>T\right) + \int_T^\infty \mathbb{P}\left(|\xi|>t\right)dt,
\end{aligned}$$

where the first equality follows because t is nonnegative, and the second equality follows from splitting the integral at T. Since ξ is sub-Gaussian, we can bound both probabilities,

$$\begin{aligned} E\big[\,|\xi|\,\mathbb{I}\{|\xi|>T\}\,\big] &\le 2Te^{-\frac{T^2}{2\sigma^2}} + 2\int_T^\infty e^{-\frac{t^2}{2\sigma^2}}dt \\ &= 2Te^{-\frac{T^2}{2\sigma^2}} + 2\sigma\sqrt{2\pi}(1-\Phi(T/\sigma)) \\ &\le 2Te^{-\frac{T^2}{2\sigma^2}} + 2\sigma\sqrt{2\pi}e^{-\frac{T^2}{2\sigma^2}}, \end{aligned}$$

where second equality follows by evaluating the Gaussian integral and the last from a standard tail bound for the normal cdf $\Phi(\cdot)$.

Next, we claim that ζ is sub-Gaussian with variance proxy σ^2. Indeed, ζ is sub-Gaussian with variance proxy 1. However, since the variance of ξ is 1, it must be that $\sigma^2>1$ (see Wainwright (2015)), proving the claim. It follows that $E\big[\,|\zeta|\,\mathbb{I}\{|\zeta|>T\}\,\big]$ is also bounded by $2Te^{-\frac{T^2}{2\sigma^2}} + 2\sigma\sqrt{2\pi}e^{-\frac{T^2}{2\sigma^2}}$.

In summary, we have shown that for any $T>0$,

$$\big|\mathbb{E}\big[\xi f(\xi)\big] - \mathbb{E}\big[\zeta f(\zeta)\big]\big| \le T\|f\|_\infty\|\phi-\overline{\phi}\|_1 + 4\|f\|_\infty e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi}).$$

On the other hand, $\big|\frac{1}{2h}\left(f(t+h)-f(t-h)\right)\big| \le \frac{\|f\|_\infty}{h}$, so that

$$\left|\mathbb{E}\left[\frac{1}{2h}\left(f(\xi+h)-f(\xi-h)\right)\right] - \mathbb{E}\left[\frac{1}{2h}\left(f(\zeta+h)-f(\zeta-h)\right)\right]\right| \le \frac{\|f\|_\infty}{h}\|\phi-\overline{\phi}\|.$$

Combining, we conclude that for any $T\ge 0$,

$$\begin{aligned} &\mathbb{E}\big[\xi f(\xi)\big] - \mathbb{E}\left[\frac{1}{2h}\left(f(\xi+h)-f(\xi-h)\right)\right] \\ &\quad\le \mathbb{E}\big[\zeta f(\zeta)\big] - \mathbb{E}\left[\frac{1}{2h}\left(f(\zeta+h)-f(\zeta-h)\right)\right] + \|f\|_\infty\left((h^{-1}+T)\|\phi-\overline{\phi}\|_1 + 4e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi})\right) \\ &\quad\le 4h^2\|f\|_\infty + \|f\|_\infty\left((h^{-1}+T)\|\phi-\overline{\phi}\|_1 + 4e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi})\right), \end{aligned}$$

where the last line follows from the special case of normally distributed random variables.

Thus, to complete the lemma, it suffices to show that

$$\min_{T>0}\left\{(h^{-1}+T)\|\phi-\overline{\phi}\|_1 + 4e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi})\right\} \le \|\overline{\phi}-\phi\|_1\left(h^{-1}+24\sigma^2-\log\left(\|\overline{\phi}-\phi\|_1\right)\right).$$

This optimization does not admit a simple closed-form solution, so we instead first upper bound the objective before optimizing. To this end, write

$$\begin{aligned} 4e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi}) &= \exp\left(-\frac{T^2}{2\sigma^2}+\log(4T+4\sigma\sqrt{2\pi})\right) \\ &\le \exp\left(-\frac{T^2}{2\sigma^2}+4T+4\sigma\sqrt{2\pi}-1\right) \qquad\qquad (\text{since } \log(t)<t-1) \\ &= \exp\left(-\frac{T^2}{2\sigma^2}+5T+4\sigma\sqrt{2\pi}-1-T\right) \\ &\le \exp\left(\frac{25}{2}\sigma^2+4\sigma\sqrt{2\pi}-1-T\right), \end{aligned}$$

where we've used the fact that the quadratic $-\frac{T^2}{2\sigma^2}+5T$ is maximized at $T^*=5\sigma^2$. We further bound this quantity noting that since $\sigma^2>1$, $\frac{25}{2}\sigma^2+4\sigma\sqrt{2\pi}-1\leq(\frac{25}{2}+4\sqrt{2\pi})\sigma^2\leq 23\sigma^2$. Substituting above we have that

$$\min_{T>0}\left\{(h^{-1}+T)\|\phi-\overline{\phi}\|_1+4e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi})\right\} \;\leq\; \min_{T>0}\left\{(h^{-1}+T)\|\phi-\overline{\phi}\|_1+e^{-T+23\sigma^2}\right\}.$$

The solution to this second optimization can be found by differentiation, and is $T^*=23\sigma^2-\log\left(\|\phi-\overline{\phi}\|_1\right)$. Substituting in shows,

$$\min_{T>0}\left\{(h^{-1}+T)\|\phi-\overline{\phi}\|_1+4e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi})\right\}\leq\|\overline{\phi}-\phi\|_1\left(1+h^{-1}+23\sigma^2-\log\left(\|\overline{\phi}-\phi\|_1\right)\right).$$

Upperbounding 1 by σ^2 proves the theorem. □

Appendix D: Proofs of the Results in Section 4

In Section D.1 below, we prove Theorem 4.3. The proof requires some auxiliary results, which are proven in Section D.2.

D.1. Proof of Theorem 4.3.

In this section, we say a constant C is *dimension-independent* if C does *not* depend on $\{n,m,\delta,h\}$ but may depend on $\{\nu_{\min},\nu_{\max},C_\mu,C_A,\beta,s_0,\phi_{\min},\phi_{\max}\}$. The constants C_1,C_2 in Theorem 4.3 are dimension-independent. By Lemma C.1, it suffices to bound $\sup_{\tau\geq 0}\left|\hat{\boldsymbol{\mu}}^\top\boldsymbol{x}(\tau,\hat{\boldsymbol{\mu}})-B_n^{Bayes}(\tau,h,\hat{\boldsymbol{\mu}})-\boldsymbol{\mu}^\top\boldsymbol{x}(\tau,\hat{\boldsymbol{\mu}})\right|$. By the triangle inequality,

$$\begin{aligned}\sup_{\tau\geq 0}\left|\frac{1}{n}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top\boldsymbol{x}(\tau,\hat{\boldsymbol{\mu}})-B_n^{Bayes}(\tau,h,\hat{\boldsymbol{\mu}})\right|\leq\;&\text{Error from Rounding Primal Solution}\\&+\text{Error from Approximating Dual Solution}\\&+\text{Error from ULLN for Dual Approximation}\\&+\text{Error from Approximating Stein's Lemma}\\&+\text{Error from ULLN for Bias Approximation}\\&+\text{Error from Approximating Dual Solution in Bias}\end{aligned}$$

where

Error from Rounding Primal Solution:

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^n(\hat{\mu}_j-\mu_j)\left(x_j(\tau,\hat{\boldsymbol{\mu}})-\;I(r_j(\tau,\hat{\mu}_j)>\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}}))\right)\right|$$

Error from Approximating Dual Solution:

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^n(\hat{\mu}_j-\mu_j)\left(\mathbb{I}(r_j(\tau,\hat{\mu}_j)>\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}}))-\mathbb{I}(r_j(\tau,\hat{\mu}_j)>\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))\right)\right|$$

Error from ULLN for Dual Approximation:

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}_j-\mu_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j)>\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))-\mathbb{E}[(\hat{\mu}_j-\mu_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j)>\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))]\right|$$

Error from Approximating Stein's Lemma:

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[(\hat{\mu}_j-\mu_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j)>\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))]-\frac{1}{2h\sqrt{\nu_j}}\mathbb{P}\{\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\}\right|$$

Error from ULLN for Bias Approximation:

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu}_j}\left(\mathbb{P}\left\{\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\right\}-\mathbb{I}(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau))\right)\right|$$

Error from Approximating Dual Solution in Bias:

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu}_j}\left(\mathbb{I}(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau))-\mathbb{I}(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\right|\leq h_j(\tau))\right)\right|$$

Lemmas D.6, D.7, D.8, D.9, D.10, and D.11 below bound each of these sources of error. In particular, there exists a positive, dimension-independent constant C_1 such that

$$\sup_{\tau\geq 0}\left|\frac{1}{n}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top\boldsymbol{x}(\tau,\hat{\boldsymbol{\mu}})-B_n^{Bayes}(\tau,h,\hat{\boldsymbol{\mu}})\right| \quad \leq C_1\left(h^2+\frac{\delta}{h}+\frac{\mathsf{TV}}{h}-\mathsf{TV}\log(\mathsf{TV})\right)+R_0+R_1+R_2+R_3+R_4$$

where $R_0,\ldots,R_4$ are the stochastic remainders from these lemmas, and C_1 is the maximum of the relevant constants from these lemmas. In this bound, we have also used the fact that $h<1$ to bound $\delta<\delta/h$ and $\mathsf{TV}<\mathsf{TV}/h$.

Next, define the dimension-independent constant

$$\lambda_{\mathsf{max}}\equiv\frac{2}{s_0}\left(\frac{\nu_{\mathsf{max}}}{\nu_{\mathsf{min}}}\left(C_\mu+\frac{1}{\sqrt{\nu_{\mathsf{max}}}}\right)+1\right),$$

where s_0 is the slack parameter from Assumption 4.1, and the event

$$\mathcal{E}=\left\{\sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\|_1\leq\lambda_{\mathsf{max}},\ \sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})-\boldsymbol{\lambda}(\tau)\|\leq\delta,\ \text{ and }\ \sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau)\|_1\leq\lambda_{\mathsf{max}}\right\},\tag{D.1}$$

Then Lemmas D.6, D.7, D.8, D.9, D.10, and D.11 also provide explicit, positive, dimension-independent constants $C_3,\ldots,C_7$ such that

$$\begin{aligned}
&\mathbb{P}\{R_0+R_1+R_2+R_3+R_4>6\epsilon\}\\
&\quad\leq\mathbb{P}\{\mathcal{E}^c\}+\mathbb{P}\{R_0+R_1+R_2+R_3+R_4>6\epsilon\text{ and }\mathcal{E}\}\\
&\quad\leq 85\exp\left(-\frac{C_3\delta\sqrt{n}}{\sqrt{m}\log(m+1)}\right)\log\left(1+\frac{\sqrt{m}\log(m+1)}{C_3\delta\sqrt{n}}\right)+5\exp\left(-\frac{n^2\epsilon^2\nu_{\mathsf{min}}}{32\sigma^2m^2\log n}\right)\\
&\qquad+25\exp\left(-\frac{C_4\epsilon\sqrt{n}}{\log(m+1)}\right)+25\exp\left(-\frac{C_5\epsilon\sqrt{n}}{\log(m+1)}\right)+25\exp\left(-\frac{C_6\epsilon h\sqrt{n}}{\log(m+1)}\right)\\
&\qquad+50\exp\left(-\frac{C_7\epsilon h\sqrt{n}}{\log(m+1)}\right).
\end{aligned}$$

We can simplify this bound by letting $C_2 = \min(C_3, \ldots, C_7, \frac{\nu_{\min}}{32\sigma^2})$ and using the fact that $h < 1$ to combine the last 4 terms. Moreover,

$$\epsilon \geq \frac{m^2 \log n}{n^{3/2} \log(m+1)} \quad \text{if and only if} \quad \frac{\epsilon^2 n^2}{m^2 \log n} \geq \frac{\epsilon \sqrt{n}}{\log(m+1)}.$$

Consequently, we can upperbound the second term by $5 \exp\left(-\frac{C_2 \epsilon h \sqrt{n}}{\log(m+1)}\right)$. Combining yields,

$$\sup_{\tau \geq 0} \left| \frac{1}{n} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}}) - B_n^{Bayes}(\tau, h, \hat{\boldsymbol{\mu}}) \right| \leq C_1 \left(h^2 + \frac{\delta}{h} + \frac{\mathsf{TV}}{h} - \mathsf{TV} \log(\mathsf{TV}) \right) + R,$$

where

$$\mathbb{P}\{R > 6\epsilon\} \leq 130 \left(\exp\left(\frac{-C_2 \delta \sqrt{n}}{\sqrt{m} \log(m+1)} \right) \log\left(1 + \frac{\sqrt{m} \log(m+1)}{C_2 \delta \sqrt{n}} \right) + \exp\left(-\frac{C_2 \epsilon \sqrt{n}}{\log(m+1)} \right) \right).$$

□

D.2. Auxiliary Lemmas for Theorem 4.3.

We first establish the pseudo-dimension of several different sets that arise in the proof. This lemma will be used to prove the uniform convergence of various quantities.

Lemma D.1 (Pseudo-dimensions of Key Sets) *For each $\hat{\boldsymbol{\mu}} \in \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{c} \in \mathbb{R}^n$, $h \in \mathbb{R}$, and $K \in \mathbb{R}$, the pseudo-dimension of*

i) $\left\{ \left(\mathbb{I}(r_j(\tau, \hat{\mu}_1) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > c_j) \mid j = 1, \ldots, n \right) \in \{0,1\}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\}$ *is at most* $2m+2$,

ii) $\left\{ \left(r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \mid j = 1, \ldots, n \right) \in \mathbb{R}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\}$ *is at most* $2m+2$,

iii) $\left\{ \left(\mathbb{I}\left(\left| r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq c_j \right) \in \{0,1\}^n \mid j = 1, \ldots, n \right) \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\}$ *is at most* $10(2m+2)$,

iv) $\left\{ \left((\hat{\mu}_j - \mu_j) \mathbb{I}(r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > c_j) \mid j = 1, \ldots, n \right) \in \mathbb{R}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\}$ *is at most* $2m+2$,

v) $\left\{ \left(\mathbb{I}\left(\left| r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq r_j(\tau, h) + K \right) \mid j = 1, \ldots, n \right) \in \{0,1\}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\}$ *is at most* $10(2m+2)$.

vi) $\left\{ \left(\mathbb{I}\left(r_j(\tau, h) \leq \left| r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq r_j(\tau, h) + K \right) \mid j = 1, \ldots, n \right) \in \{0,1\}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\}$ *is at most* $100(2m+2)$.

Proof: i) Note that $\mathbb{I}\left(r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > c_j \right) = \mathbb{I}(\nu_j(\hat{\mu}_j - c_j) - (\frac{\nu_j \hat{\mu}_j}{\nu_{\min}} - c_j)\tau - \nu_j \boldsymbol{A}_j^\top \boldsymbol{\lambda} - \tau \boldsymbol{A}_j^\top \boldsymbol{\lambda} > 0)$, and that

$$\left\{ \left(\nu_j(\hat{\mu}_j - c_j) - \left(\frac{\nu_j \hat{\mu}_j}{\nu_{\min}} - c_j \right) \tau - \nu_j \boldsymbol{A}_j^\top \boldsymbol{\lambda} - \tau \boldsymbol{A}_j^\top \boldsymbol{\lambda} \mid j = 1, \ldots, n \right) \in \mathbb{R}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\}$$

lives in an affine subspace of dimension at most $2m+2$. By Lemma 4.4 in Pollard (1990), the pseudo-dimension of the set $\left\{ \left(\mathbb{I}(r_j(\tau, \hat{\mu}_1) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > c_j) \mid j = 1, \ldots, n \right) \in \{0,1\}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\}$ is therefore at most $2m+2$.

ii) Suppose by contradiction the statement were false. Then, there exists $\mathcal{J} \subseteq \{1,\ldots,n\}$ with $|\mathcal{J}| = 2m+3$ and $\boldsymbol{c} \in \mathbb{R}^n$ such that

$$\left| \left\{ \left(\mathbb{I}(r_j(\tau,\hat{\mu}_j) - A_j^\top \boldsymbol{\lambda} > c_j) \mid j \in \mathcal{J}\right) \in \{0,1\}^{|\mathcal{J}|} \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\} \right| = 2^{2m+3}$$

We then claim that for the set defined in part i) with this $\boldsymbol{c}$, $\mathbf{0}$ is a witness to the shattering of $\mathcal{J}$. In particular, observe, that

$$\begin{aligned}
&\left|\left\{ \left(\mathbb{I}\left(\mathbb{I}(r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > c_j) > 0\right) \mid j \in \mathcal{J}\right) \in \{0,1\}^{|\mathcal{J}|} \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \right\}\right| \\
&\quad = \left|\left\{ \ \left(\mathbb{I}(r_j(\tau,\hat{\mu}_j) - A_j^\top \boldsymbol{\lambda} > c_j) \mid j \in \mathcal{J}\right) \in \{0,1\}^{|\mathcal{J}|} \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \ \right\}\right| \ = \ 2^{2m+3},
\end{aligned}$$

which contradicts part i).

iii) For any $(\tau, \boldsymbol{\lambda})$, the binary vector $\left(\mathbb{I}(\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}\right| \le c_j) \mid j = 1,\ldots,n\right)$ is the pointwise minimum of $\left(\mathbb{I}(r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \le c_j) \mid j = 1,\ldots,n\right)$ and $\left(\mathbb{I}(-r_j(\tau,\hat{\mu}_j) + \boldsymbol{A}_j^\top \boldsymbol{\lambda} \le c_j) \mid j = 1,\ldots,n\right)$. By part i), the pseudo-dimension of $\left\{\left(\mathbb{I}(r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > c_j) \mid j = 1,\ldots,n\right) \in \{0,1\}^n : \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R}\right\}$ is at most $2m+2$. Analogous arguments show the pseudo-dimesion of these two sets are at most $2m+2$. By Lemma 5.1 in Pollard (1990), the pseudo-dimension of the pointwise minimum is at most $10(2m+2)$.

iv) Suppose by contradiction the statement were false. Then, there exists $\mathcal{J} \subseteq \{1,\ldots,n\}$ with $|\mathcal{J}| = 2m+3$ and $\overline{\boldsymbol{c}} \in \mathbb{R}^n$ such that

$$\left|\left\{ \ \left(\ \mathbb{I}\Big((\hat{\mu}_j - \mu_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j) - A_j^\top \boldsymbol{\lambda} > c_j) \ > \ \overline{c}_j\Big) \ \Big| \ j \in \mathcal{J} \ \right) \in \{0,1\}^{|\mathcal{J}|} : \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R} \ \right\}\right| = 2^{2m+3}$$

This would imply that the vector $(\overline{c}_j/(\hat{\mu}_j - \mu_j) \mid j \in \mathcal{J})$ witnesses the shattering of $\mathcal{J}$ for the set defined in part i), a contradiction.

v) This set is the pointwise minimum of the following two subsets:

$$\begin{aligned}
&\left\{\left(\mathbb{I}(r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \ge -r_j(\tau,h) - K) \mid j = 1,\ldots,n\right) \in \{0,1\}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R}\right\} \\
&\left\{\left(\mathbb{I}(r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \le r_j(\tau,h) + K) \mid j = 1,\ldots,n\right) \in \{0,1\}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R}\right\}
\end{aligned}$$

Consider the first set. Since $r_j(\tau,\hat{\mu}_j) + r_j(\tau,h) = r_j(\tau,\hat{\mu}_j + h)$, this set is of the form considered in part i), so its pseudo-dimension is at most $2m+2$. An analogous argument holds for the second set. By Lemma 5.1 in Pollard (1990), the pointwise minimum has pseudo-dimension at most $10(2m+2)$.

vi) This set is the pointwise minimum of the following two subsets:

$$\begin{aligned}
&\left\{\left(\mathbb{I}\left(\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}\right| \le r_j(\tau,h) + K\right) \mid j = 1,\ldots,n\right) \in \{0,1\}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R}\right\} \\
&\left\{\left(\mathbb{I}\left(r_j(\tau,h) \le \left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}\right|\right) \mid j = 1,\ldots,n\right) \in \{0,1\}^n \ : \ \boldsymbol{\lambda} \in \mathbb{R}^m, \tau \in \mathbb{R}\right\}.
\end{aligned}$$

Part v) proves the first set has pseudo-dimension at most $10(2m+2)$. An analogous argument applies to the second. By Lemma 5.1 in Pollard (1990), the pseudo-dimension of the pointwise minimum is at most $100(2m+2)$. □

We also require the following concentration result. Note that $\left|\frac{1}{n}\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})^\top \boldsymbol{x}(\tau,\hat{\boldsymbol{\mu}})\right| \le \frac{1}{n}\sum_{j=1}^n |r_j(\tau,\hat{\boldsymbol{\mu}})|$ (since $x_j(\tau,\hat{\boldsymbol{\mu}}) \in [0,1]$), so this result effectively bounds the objective value of Eq. (3.1).

Lemma D.2 (Concentration of $\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1$) *For all $\tau \ge 0$,*

i) $\frac{1}{n}\sum_{j=1}^n \mathbb{E}[|r_j(\tau,\hat{\boldsymbol{\mu}})|] \le \frac{\nu_{\mathsf{max}}}{\nu_{\mathsf{min}}}\left(C_\mu + 1/\sqrt{\nu_{\mathsf{max}}}\right)$.

ii) $\mathbb{P}\left(\sup_{\tau\ge 0}\left|\frac{1}{n}\sum_{j=1}^n \left(|r_j(\tau,\hat{\boldsymbol{\mu}})| - \mathbb{E}[|r_j(\tau,\hat{\boldsymbol{\mu}})|]\right)\right| > t\right) \le 2\exp\left(-\frac{n\,\nu_{\mathsf{min}}^3\, t^2}{72\,\nu_{\mathsf{max}}^2\,\sigma^2}\right)$.

Proof: Notice $0 < \frac{\nu_j}{\nu_j+\tau}\cdot\frac{\nu_{\mathsf{min}}+\tau}{\nu_{\mathsf{min}}} \le \frac{\nu_j}{\nu_{\mathsf{min}}}$. Hence,

$$|r_j(\tau,\hat{\boldsymbol{\mu}})| \le \frac{\nu_j}{\nu_{\mathsf{min}}}|\hat{\mu}_j| \le \frac{\nu_j}{\nu_{\mathsf{min}}}\left(|\hat{\mu}_j-\mu_j| + |\mu_j|\right) \le \frac{\nu_j}{\nu_{\mathsf{min}}}\left(|\hat{\mu}_j-\mu_j| + C_\mu\right),$$

where second inequality is the triangle inequality. Then, by Jensen's inequality,

$$\mathbb{E}[|\hat{\mu}_j-\mu_j|] = \mathbb{E}\left[\sqrt{(\hat{\mu}_j-\mu_j)^2}\right] \le \sqrt{\mathbb{E}\left[(\hat{\mu}_j-\mu_j)^2\right]} = 1/\sqrt{\nu_j}.$$

Combining shows $\mathbb{E}[|r_j(\tau,\hat{\boldsymbol{\mu}})|] \le \frac{\nu_j}{\nu_{\mathsf{min}}}\left(C_\mu + 1/\sqrt{\nu_j}\right) \le \frac{\nu_{\mathsf{max}}}{\nu_{\mathsf{min}}}\left(C_\mu + 1/\sqrt{\nu_{\mathsf{max}}}\right)$. Averaging over j proves the first claim.

For the second claim, first consider the special case $\tau = 0$, i.e., $\mathbb{P}\left(\frac{1}{n}\left|\sum_{j=1}^n \left(|\hat{\mu}_j| - \mathbb{E}[|\hat{\mu}_j|]\right)\right| > t\right)$. We will first prove that $|\hat{\mu}_j| - \mathbb{E}[|\hat{\mu}_j|]$ is sub-Gaussian with variance proxy at most $36\sigma^2/\nu_j$. Let $\tilde{\mu}_j$ be an i.i.d. copy of $\hat{\mu}_j$. Then, for any $a > 0$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(a\left(|\hat{\mu}_j| - \mathbb{E}[|\hat{\mu}_j|]\right)^2\right)\right] &\le \mathbb{E}\left[\exp\left(a\left(|\hat{\mu}_j| - |\tilde{\mu}_j|\right)^2\right)\right] && \text{(Jensen's Inequality)}\\
&\le \mathbb{E}\left[\exp\left(a\left(|\hat{\mu}_j - \tilde{\mu}_j|\right)^2\right)\right] && \text{(Triangle-Inequality)}\\
&= 1 + \int_1^\infty \mathbb{P}\left(e^{a(\hat{\mu}_j-\tilde{\mu}_j)^2} > t\right) dt && \text{(Tail Formula for Expectation)}\\
&= 1 + \int_1^\infty \mathbb{P}\left(|\hat{\mu}_j - \tilde{\mu}_j| > \sqrt{\frac{\log t}{a}}\right) dt
\end{aligned}$$

Note that because $\hat{\mu}_j$ and $\tilde{\mu}_j$ are i.i.d., $\hat{\mu}_j - \tilde{\mu}_j$ is mean-zero, and sub-Gaussian with variance proxy at most $2\sigma^2/\nu_j$. From the usual sub-Gaussian tail-bound then,

$$\mathbb{E}\left[\exp\left(a\left(|\hat{\mu}_j| - \mathbb{E}[|\hat{\mu}_j|]\right)^2\right)\right] \le 1 + 2\int_1^\infty e^{-\nu_j\frac{\log t}{4a\sigma^2}} dt = 1 + 2\int_1^\infty t^{\frac{-\nu_j}{4a\sigma^2}} dt$$

For $a = \nu_j/12\sigma^2$, this integral converges and can be evaluated explicitly, yielding $\mathbb{E}\left[\exp\left(a\left(|\hat{\mu}_j| - \mathbb{E}[|\hat{\mu}_j|]\right)^2\right)\right] \le 2$. By (Rivasplata 2012, Theorem 3.1, Part 3)), it follows that $|\hat{\mu}_j| - \mathbb{E}[|\hat{\mu}_j|]$ is sub-Gaussian with variance proxy at most $3/a = 36\sigma^2/\nu_j$.

The standard sub-Gaussian concentration shows

$$P\left(\left|\frac{1}{n}\sum_{j=1}^n \left(|\hat{\mu}_j| - \mathbb{E}\left[|\hat{\mu}_j|\right]\right)\right| > t\right) \le 2\exp\left(-\frac{n\,\nu_{\mathsf{min}}\,t^2}{72\,\sigma^2}\right) \tag{D.2}$$

Finally, since

$$||r_j(\tau,\hat{\mu}_j)|-\mathbb{E}[|r_j(\tau,\hat{\mu}_j)|]| \;\leq\; \frac{\nu_j}{\nu_{\min}}\cdot\Big||\hat{\mu}_j|-\mathbb{E}[|\hat{\mu}_j|]\Big| \;\leq\; \frac{\nu_{\max}}{\nu_{\min}}\cdot\Big||\hat{\mu}_j|-\mathbb{E}[|\hat{\mu}_j|]\Big|,$$

we have

$$\begin{aligned}\mathbb{P}\left(\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\big(|r_j(\tau,\hat{\boldsymbol{\mu}})|-\mathbb{E}\left[|r_j(\tau,\hat{\boldsymbol{\mu}})|\right]\big)\right|>t\right) &\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}\big(|\hat{\mu}_j|-\mathbb{E}\left[|\hat{\mu}_j|\right]\big)\right|>\frac{\nu_{\min}t}{\nu_{\max}}\right)\\ &\leq 2\exp\left(-\frac{n\,\nu_{\min}^3\,t^2}{72\,\nu_{\max}^2\,\sigma^2}\right).\end{aligned}$$

This completes the proof.

□

As the first step in the proof of Theorem 4.3, we observe that the solutions to the original dual problem $\min_{\boldsymbol{\lambda}\geq 0} D_{\hat{\boldsymbol{\mu}}}(\boldsymbol{\lambda},\tau)$ and the "average" dual problem $\min_{\boldsymbol{\lambda}\geq 0} D(\boldsymbol{\lambda},\tau)$ are uniformly bounded, and are uniformly close to one another with high probability. Recall from the proof of Theorem 4.3 the dimension-independent constant

$$\lambda_{\max}\equiv\frac{2}{s_0}\left(\frac{\nu_{\max}}{\nu_{\min}}\left(C_\mu+\frac{1}{\sqrt{\nu_{\max}}}\right)+1\right),$$

and event

$$\mathcal{E}=\left\{\sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\|_1\leq\lambda_{\max},\ \sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})-\boldsymbol{\lambda}(\tau)\|\leq\delta,\ \text{ and }\ \sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau)\|_1\leq\lambda_{\max}\right\}.$$

We will show that $\mathcal{E}$ occurs with high probability in three steps. In Lemma D.3, we show that the optimal solutions to both dual problems are uniformly bounded by $\lambda_{\max}$ with high probability. We then prove the strong convexity of $\boldsymbol{\lambda}\mapsto D(\boldsymbol{\lambda},\tau)$ (Lemma D.4), and we will use that result to show that $\mathcal{E}$ occurs with high probability (Lemma D.5).

Lemma D.3 (Optimal Dual Variables Bounded)

i) $\sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau)\|_1 \;<\; \lambda_{\max}$.

ii) $\mathbb{P}\left\{\sup_{\tau\geq 0}\|\boldsymbol{\lambda}_n(\tau,\hat{\boldsymbol{\mu}})\|_1\geq\lambda_{\max}\right\} \;\leq\; 2\exp\left(-\frac{n\,\nu_{\min}^3}{72\,\nu_{\max}^2\,\sigma^2}\right)$.

Proof: By optimality, $D(\tau,\boldsymbol{\lambda}(\tau))\leq D(\tau,\mathbf{0})\leq\frac{1}{n}\mathbb{E}\left[\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1\right]$. Since $\boldsymbol{\lambda}(\tau)\geq\mathbf{0}$, we have $\|\boldsymbol{\lambda}(\tau)\|_1=\boldsymbol{e}^\top\boldsymbol{\lambda}(\tau)$. Thus,

$$\begin{aligned}\|\boldsymbol{\lambda}(\tau)\|_1\leq\max_{\boldsymbol{\lambda}\geq 0}\quad & \boldsymbol{e}^\top\boldsymbol{\lambda}\\ \text{s.t.}\quad & \boldsymbol{b}^\top\boldsymbol{\lambda}+\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[(r_j(\tau,\hat{\boldsymbol{\mu}})-\boldsymbol{A}_j^\top\boldsymbol{\lambda})^+\right]\leq\frac{1}{n}\mathbb{E}[\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1].\end{aligned}$$

Since $\mathcal{X}$ is s_0-strictly feasible, by Lagrangian duality

$$\begin{aligned}
\|\boldsymbol{\lambda}(\tau)\|_1 &\leq \max_{\boldsymbol{\lambda}\geq 0} \quad \boldsymbol{e}^\top\boldsymbol{\lambda} + \frac{1}{s_0}\left(\frac{1}{n}\mathbb{E}[\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1] - \boldsymbol{b}^\top\boldsymbol{\lambda} - \frac{1}{n}\sum_{j=1}^n \mathbb{E}[(r_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda})^+]\right)\\
&= \max_{\boldsymbol{\lambda}\geq 0} \quad \boldsymbol{e}^\top\boldsymbol{\lambda} + \frac{1}{s_0}\left(\frac{1}{n}\mathbb{E}[\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1] - \boldsymbol{b}^\top\boldsymbol{\lambda} - \frac{1}{n}\sum_{j=1}^n \mathbb{E}[\max_{x_j\in[0,1]} x_j(r_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda})]\right)\\
&\leq \max_{\boldsymbol{\lambda}\geq 0} \quad \boldsymbol{e}^\top\boldsymbol{\lambda} + \frac{1}{s_0}\left(\frac{1}{n}\mathbb{E}[\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1] - \boldsymbol{b}^\top\boldsymbol{\lambda} - \frac{1}{n}\sum_{j=1}^n \mathbb{E}[x_j^0(r_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda})]\right)\\
&= \max_{\boldsymbol{\lambda}\geq 0} \quad \left(\boldsymbol{e} - \frac{1}{s_0}\boldsymbol{b} + \frac{1}{ns_0}\boldsymbol{A}\boldsymbol{x}^0\right)^\top\boldsymbol{\lambda} + \frac{1}{ns_0}\left(\mathbb{E}[\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1] - \mathbb{E}\left[\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})^\top\boldsymbol{x}^0\right]\right)
\end{aligned}$$

By Assumption 4.1, $\frac{1}{n}\boldsymbol{A}\boldsymbol{x}^0 + s_0\boldsymbol{e} \leq \boldsymbol{b} \iff \boldsymbol{e} - \frac{1}{s_0}\boldsymbol{b} + \frac{1}{ns_0}\boldsymbol{A}\boldsymbol{x}^0 \leq 0$, which implies that $\boldsymbol{\lambda} = \boldsymbol{0}$ is optimal for this last optimization problem. Thus, for all $\tau \geq 0$,

$$\begin{aligned}
\|\boldsymbol{\lambda}(\tau)\|_1 &= \frac{1}{ns_0}\left(\mathbb{E}[\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1] - \mathbb{E}\left[\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})^\top\boldsymbol{x}^0\right]\right),\\
&\leq \frac{2}{ns_0}\mathbb{E}[\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1] && \text{(since } \boldsymbol{x}^0 \in [0,1]^n\text{)}\\
&\leq \lambda_{\mathsf{max}} && \text{(by Lemma D.2).}
\end{aligned}$$

This proves the first statement.

For the second statement, we follow an identical sequence of steps using $D_{\hat{\boldsymbol{\mu}}}(\tau, \boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}}))$ to conclude that $\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\|_1 \leq \frac{2}{ns_0}\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1$. Then, by definition of λ_{max} and Lemma D.2,

$$\begin{aligned}
\mathbb{P}\left(\sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\|_1 > \lambda_{\mathsf{max}}\right) &\leq \mathbb{P}\left(\sup_{\tau\geq 0}\frac{1}{n}\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1 > \frac{\nu_{\mathsf{max}}}{\nu_{\mathsf{min}}}\left(C_\mu + 1/\sqrt{\nu_{\mathsf{max}}}\right) + 1\right)\\
&\leq \mathbb{P}\left(\sup_{\tau\geq 0}\frac{1}{n}\left(\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1 - \mathbb{E}\left[\|\boldsymbol{r}(\tau,\hat{\boldsymbol{\mu}})\|_1\right]\right) > 1\right)\\
&\leq 2\exp\left(-\frac{n\,\nu_{\mathsf{min}}^3}{72\,\nu_{\mathsf{max}}^2\,\sigma^2}\right)
\end{aligned}$$

This completes the proof. □

Lemma D.4 (Strong Convexity of Average Dual) *There exists a dimension-independent constant $\kappa > 0$ such that the function $\boldsymbol{\lambda} \mapsto D(\tau,\boldsymbol{\lambda})$ is κ-strongly convex for all $\boldsymbol{\lambda} \in \mathbb{R}_+^m$ such that $\|\boldsymbol{\lambda}\|_1 \leq \lambda_{\mathsf{max}}$.*

Proof: The Hessian (in terms of $\boldsymbol{\lambda}$) of $D(\tau,\boldsymbol{\lambda})$ is

$$H_{\boldsymbol{\lambda}} \circ D(\tau,\boldsymbol{\lambda}) = \frac{1}{n}\sum_{j=1}^n \frac{\partial^2}{\partial s^2}\mathbb{E}[(r_j(\tau,\hat{\mu}_j) - s)^+]\Big|_{s=\boldsymbol{A}_j^\top\boldsymbol{\lambda}} \cdot \boldsymbol{A}_j\boldsymbol{A}_j^\top.$$

To show this matrix is strictly positive definite, we first study the function $s \mapsto \mathbb{E}[(r_j(\tau,\hat{\mu}_j) - s)^+]$. Intuitively, this function is not strongly convex over the whole real line (at the extremes it

approaches linear), but it is strongly convex on any bounded interval. By differentiating under the integral sign

$$\begin{aligned}\frac{\partial^2}{\partial s^2}\mathbb{E}[(r_j(\tau,\hat{\mu}_j)-s)^+] &= \mathbb{P}\left(r_j(\tau,\hat{\mu}_j)=s\right)\\ &= \mathbb{P}\left(\hat{\mu}_j=\frac{\nu_{\mathsf{min}}}{\nu_{\mathsf{min}}+\tau}\frac{\nu_j+\tau}{\nu_j}s\right)\\ &= \mathbb{P}\left(\sqrt{\nu_j}(\hat{\mu}_j-\mu_j)=\left(\frac{\nu_{\mathsf{min}}}{\nu_{\mathsf{min}}+\tau}\frac{\nu_j+\tau}{\nu_j}s-\mu_j\right)\sqrt{\nu_j}\right).\end{aligned}$$

Furthermore, for $\|\lambda\|_1 \leq \lambda_{\mathsf{max}}$,

$$\left|\sqrt{\nu_j}\left(\frac{\nu_{\mathsf{min}}}{\nu_{\mathsf{min}}+\tau}\frac{\nu_j+\tau}{\nu_j}s-\mu_j\right)\right|\Bigg|_{s=\boldsymbol{A}_j^\top\boldsymbol{\lambda}} \leq \sqrt{\nu_j}\left(\left|\boldsymbol{A}_j^\top\boldsymbol{\lambda}\right|+|\mu_j|\right) \leq \sqrt{\nu_{\mathsf{max}}}\left(C_A\lambda_{\mathsf{max}}+C_\mu\right),$$

by the Cauchy-Schwarz inequality and assumptions on parameters. By Assumption 4.2, then,

$$\frac{\partial^2}{\partial s^2}\mathbb{E}[(r_j(\tau,\hat{\mu}_j)-s)^+]\Bigg|_{s=\boldsymbol{A}_j^\top\boldsymbol{\lambda}} = \mathbb{P}\left(\hat{\mu}_j=\frac{\nu_{\mathsf{min}}}{\nu_{\mathsf{min}}+\tau}\frac{\nu_j+\tau}{\nu_j}s\right)\Bigg|_{s=\boldsymbol{A}_j^\top\boldsymbol{\lambda}} \geq \phi_{\mathsf{min}}\left(\sqrt{\nu_{\mathsf{max}}}\left(C_A\lambda_{\mathsf{max}}+C_\mu\right)\right) > 0.$$

Now let $\sigma_{min}(\cdot)$ denote the minimal eigenvalue of a symmetric, positive definite matrix. Then, in light of the above,

$$\begin{aligned}\sigma_{min}\left(H_{\boldsymbol{\lambda}}\circ D(\tau,\boldsymbol{\lambda})\right) &\geq \phi_{\mathsf{min}}\left(\sqrt{\nu_{\mathsf{max}}}\left(C_A\lambda_{\mathsf{max}}+C_\mu\right)\right)\sigma_{min}\left(\frac{1}{n}\sum_{j=1}^n \boldsymbol{A}_j\boldsymbol{A}_j^\top\right)\\ &\geq \phi_{\mathsf{min}}\left(\sqrt{\nu_{\mathsf{max}}}\left(C_A\lambda_{\mathsf{max}}+C_\mu\right)\right)\beta.\end{aligned}$$

Take κ to be $\phi_{\mathsf{min}}\left(\sqrt{\nu_{\mathsf{max}}}\left(C_A\lambda_{\mathsf{max}}+C_\mu\right)\right)\beta$ to complete the proof.

□

We can now prove that $\mathcal{E}$ defined in Equation (D.1) occurs with high probability.

Lemma D.5 (Uniform Convergence of Dual Solutions) *There exists a positive, dimension-independent constant C such that*

$$\mathbb{P}\{\mathcal{E}^c\} \leq 86\exp\left(\frac{-C\delta\sqrt{n}}{\sqrt{m}\log(m+1)}\right)\log\left(1+\frac{\sqrt{m}\log(m+1)}{C\delta\sqrt{n}}\right).$$

Proof: If $\sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\|_1 \leq \lambda_{\mathsf{max}}$, then, combining strong convexity (Lemma D.4) with $D_{\hat{\boldsymbol{\mu}}}(\tau,\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})) \leq D_{\hat{\boldsymbol{\mu}}}(\tau,\boldsymbol{\lambda}(\tau))$, write

$$D(\tau,\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})) - D_{\hat{\boldsymbol{\mu}}}(\tau,\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})) - \left(D(\tau,\boldsymbol{\lambda}(\tau)) - D_{\hat{\boldsymbol{\mu}}}(\tau,\boldsymbol{\lambda}(\tau))\right) \geq \frac{\kappa}{2}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})-\boldsymbol{\lambda}(\tau)\|_2^2. \tag{D.3}$$

For k in the integers, define the set of functions

$$S_k = \left\{\boldsymbol{h}:\mathbb{R}_+\mapsto\mathbb{R}_+^m \mid 2^{k-1} \leq \sqrt{n}\sup_{\tau\geq 0}\|\boldsymbol{h}(\tau)-\boldsymbol{\lambda}(\tau)\|_2 < 2^k\right\}.$$

If $\boldsymbol{\lambda}(\cdot, \hat{\boldsymbol{\mu}}) \in S_k$, then Equation (D.3) implies

$$\sup_{\tau, \boldsymbol{h}(\cdot) \in S_k} |D(\tau, \boldsymbol{h}(\tau)) - D_{\hat{\boldsymbol{\mu}}}(\tau, \boldsymbol{h}(\tau)) - (D(\tau, \boldsymbol{\lambda}(\tau)) - D_{\hat{\boldsymbol{\mu}}}(\tau, \boldsymbol{\lambda}(\tau)))| \;\geq\; \frac{\kappa 2^{2k-2}}{2n}. \tag{D.4}$$

We first bound the probability of this event by bounding the increments of the stochastic process

$$(\tau, \boldsymbol{h}(\cdot)) \mapsto D_{\hat{\boldsymbol{\mu}}}(\tau, \boldsymbol{\lambda}(\tau)) - D_{\hat{\boldsymbol{\mu}}}(\tau, \boldsymbol{h}(\tau)) - \big(D(\tau, \boldsymbol{\lambda}(\tau)) - D(\tau, \boldsymbol{h}(\tau))\big).$$

By writing out the definitions of $D_{\hat{\boldsymbol{\mu}}}$ and D, we recognize this stochastic process as the difference between an empirical average and its expectation. We will apply the first part of Theorem A.2.

Let $F_j(\hat{\boldsymbol{\mu}}_j) = C_A 2^k \sqrt{m/n}$ for all j. Note that $F_j(\cdot)$ is a constant function. We will show that it is an envelope. Note that

$$\begin{aligned}
&\big|[r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)]^+ - [r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{h}(\tau)]^+\big| \\
&\qquad \leq \big|\boldsymbol{A}_j^\top (\boldsymbol{\lambda}(\tau) - \boldsymbol{h}(\tau))\big| \;\leq\; \|\boldsymbol{A}_j\|_2 2^k/\sqrt{n} \;\leq\; C_A 2^k \sqrt{m/n} \;=\; F_j(\hat{\boldsymbol{\mu}}_j),
\end{aligned}$$

so that $\|\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2\|_\Psi \leq C_A 2^k \sqrt{m}$ by Lemma A.1. Let

$$\begin{aligned}
\mathcal{F}_1 &\equiv \left\{ \big([r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)]^+ - [r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{h}(\tau)]^+ \mid j = 1, \ldots, n\big) \in \mathbb{R}^n \;:\; \tau \in \mathbb{R}, \boldsymbol{h}(\cdot) \in S_k \right\} \\
\mathcal{F}_2 &\equiv \left\{ \big([r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}]^+ - [r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{h}]^+ \mid j = 1, \ldots, n\big) \in \mathbb{R}^n \;:\; \tau \in \mathbb{R}, \boldsymbol{\lambda} \in \mathbb{R}^m, \boldsymbol{h} \in \mathbb{R}^m \right\}.
\end{aligned}$$

We next show that Equation (A.2) holds for $\mathcal{F}_1$ and determine an upper bound for $V(A, W)$. Observe that since $\mathcal{F}_1 \subseteq \mathcal{F}_2$, it follows that $M(\epsilon \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}_1) \leq M(\epsilon \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}_2)$. To bound the packing number for $\mathcal{F}_2$, note that $\mathcal{F}_2$ is a pointwise difference of sets of the form,

$$\mathcal{F}_3 \equiv \left\{ \big([r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}]^+ \mid j = 1, \ldots, n\big) \in \mathbb{R}^n \;:\; \tau \in \mathbb{R}, \boldsymbol{\lambda} \in \mathbb{R}^m \right\}.$$

so by Page 22 in Pollard (1990), $M(\epsilon \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}_2) \leq M(\epsilon \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2/4, \mathcal{F}_3)^2$.

The function $(f_1, \ldots, f_n) \mapsto (f_1^+, \ldots, f_n^+)$ is a contraction mapping from $\mathbb{R}^n \mapsto \mathbb{R}^n$. Thus, by (Pollard 1990, pg. 23-24), the packing numbers of $\mathcal{F}_3$ are upperbounded by the packing numbers of

$$\left\{ \big([r_j(\tau, \hat{\mu}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}] \mid j = 1, \ldots, n\big) \in \mathbb{R}^n \;:\; \tau \in \mathbb{R}, \boldsymbol{\lambda} \in \mathbb{R}^m \right\}.$$

Part ii) of Lemma D.1 shows the pseudo-dimension of this last set is at most $2m + 2$. Theorem A.3 shows that for $V = 2m + 2$, $W_m = 4V$, $A_m = V^{6V}$, we have $M(\epsilon \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2/4, \mathcal{F}_3) \leq A_m 2^{2W_m} \epsilon^{-W_m}$. Substituting back yields $M(\epsilon \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}_1) \leq A_m^2 2^{4W_m} \epsilon^{-2W_m}$, which proves that $\mathcal{F}_1$ satisfies Equation (A.2) with $A = A_m^2 2^{4W_m}$ and $W = 2W_m$.

To bound $V(A, W)$, note $\sqrt{\log A} > 0$ so that

$$V(A, W) \leq 1 + \frac{2 \log A}{W} \leq 1 + \frac{4 \log A_m + 8 W_m \log 2}{2 W_m} \leq 1 + 4 \log 2 + 3 \log(V) \leq 1 + 7 \log 2 + 3 \log(m + 1).$$

The last expression is at most $12\log(m+1)$ for $m \geq 1$. Finally, applying Theorem A.2 to Equation (D.4) proves

$$\mathbb{P}\left\{\sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\|_1 \leq \boldsymbol{\lambda}_{max} \text{ and } \lambda(\cdot,\hat{\boldsymbol{\mu}}) \in S_{k,n}\right\} \leq 25\exp\left(\frac{-C_1 2^k}{\sqrt{m}\log(m+1)}\right).$$

where $C_1 = \frac{\kappa}{2^3\cdot 9\cdot 12 C_A}$.

We use the above bound to decompose the $\mathbb{P}\{\mathcal{E}^c\}$ into "peels" indexed by k:

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\| \leq \lambda_{\max} \text{ and } \sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau)-\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\|_2 \geq \delta\right\}\\
&\leq \sum_{k=\lceil\log_2(\delta\sqrt{n})\rceil}^{\infty}\mathbb{P}\left\{\sup_{\tau\geq 0}\|\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\| \leq \lambda_{\max} \text{ and } \boldsymbol{\lambda}(\cdot,\hat{\boldsymbol{\mu}}) \in S_{k,n}\right\}\\
&\leq 25\sum_{k=\lceil\log_2(\delta\sqrt{n})\rceil}^{\infty}\exp\left(\frac{-C_1 2^k}{\sqrt{m}\log(m+1)}\right)\\
&\leq 25\int_{\log_2(\delta\sqrt{n})}^{\infty}\exp\left(\frac{-C_1 2^x}{\sqrt{m}\log(m+1)}\right)dx \;=\; \frac{25}{\log 2}\int_{\frac{C_1\delta\sqrt{n}}{\sqrt{m}\log(m+1)}}^{\infty}\exp(-u)\frac{du}{u},
\end{aligned}$$

where the last inequality follows by making the change of variables $u = \frac{C_1}{\sqrt{m}\log(m+1)}2^x$. We recognize the last integral as the exponential integral which admits the bound

$$\int_x^{\infty}\exp(-t)\frac{dt}{t} \leq \exp(-x)\log(1+1/x) \quad \text{for } x>0.$$

Applying this bound and combining with Lemma D.3 yields:

$$\mathbb{P}\{\mathcal{E}^c\} \leq \frac{25}{\log 2}\exp\left(\frac{-C_1\delta\sqrt{n}}{\sqrt{m}\log(m+1)}\right)\log\left(1+\frac{\sqrt{m}\log(m+1)}{C_1\delta\sqrt{n}}\right)+2\exp\left(-\frac{n\,\nu_{\min}^3}{72\,\nu_{\max}^2\sigma^2}\right).$$

To "clean up" the right-hand side, observe that for $0\leq\delta\leq 1$, $\frac{\delta}{\sqrt{m}\log(m+1)} \leq 1$ and, for $n\geq 2$, $\sqrt{n}\leq n$, so we can bound the second term by $2\exp\left(-\frac{\nu_{\min}^3\delta\sqrt{n}}{72\nu_{\max}^2\sigma^2\sqrt{m}\log(m+1)}\right)$. Then, let $C = \min\left(C_1, \frac{\nu_{\min}^3}{72\nu_{\max}^2\sigma^2}\right)$ and note that $25/\log(2)+2<86$ to prove the lemma. □

We can now proceed to bound the various terms in the proof of Theorem 4.3.

Lemma D.6 (Rounding the Primal Solution)

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}_j-\mu_j)\left(x_j(\tau,\hat{\boldsymbol{\mu}}) - \; I(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^{\top}\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}}))\right)\right| \leq R_0,$$

where $\mathbb{P}\{R_0\geq\epsilon\} \leq 5\exp\left(\frac{-n^2\epsilon^2\nu_{\min}}{32\sigma^2 m^2\log n}\right)$.

Proof: By complementary slackness, $x_j(\tau,\hat{\boldsymbol{\mu}}) = \mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^{\top}\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}}))$ except possibly for m fractional terms. These fractional terms contribute at most $\frac{m}{n}\max_j|\hat{\mu}_j-\mu_j| \leq \frac{m}{n\sqrt{\nu_{\min}}}\max_j|\zeta_j|$ where ζ_j

are each standardized random variables with sub-Gaussian parameter σ^2. Let $R_0 = \frac{m}{n\sqrt{\nu_{\min}}} \max_j |\zeta_j|$. Then, by Markov's inequality, for any $C > 0$,

$$\begin{aligned}\mathbb{P}(R_0 > t) = \mathbb{P}\left(\max_j |\zeta_j| > \frac{tn\sqrt{\nu_{\min}}}{m}\right) &= \mathbb{P}\left(\Psi\left(\max_j |\zeta_j|/C\right) > \Psi\left(\frac{tn\sqrt{\nu_{\min}}}{Cm}\right)\right) \\ &\leq 5\exp\left(\frac{-n^2t^2\nu_{\min}}{m^2C^2}\right) \cdot \mathbb{E}\left[\Psi\left(\frac{\max_j |\zeta_j|}{C}\right)\right],\end{aligned}$$

where again $\Psi(t) = \frac{1}{5}\exp(t^2)$ defines an Orlicz-norm. Let $C = 2\sigma\sqrt{2+\log n}$. Then, by part iii) of Lemma A.1 it follows that

$$\mathbb{P}(R_0 > \epsilon) \leq 5\exp\left(\frac{-n^2\epsilon^2\nu_{\min}}{4\sigma^2m^2(2+\log n)}\right).$$

We can simplify this result slightly since $n \geq 2$ implies $(2+\log n) \leq 8\log n$. Substituting this upper bound and simplifying completes the lemma. □

Lemma D.7 (Approximating the Dual Solution) *Recall $\mathcal{E}$ defined in Equation* (D.1)*. There exist positive dimension-independent constants C_1, C_2 such that,*

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}(\mu_j - \hat{\mu}_j)\Big(\mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})) \;-\; \mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))\Big)\right| \;\leq\; C_1\delta + R_1 \qquad \text{(D.5)}$$

where $\mathbb{P}\{\mathcal{E} \text{ and } R_1 > \epsilon\} \leq 25\exp\left(\frac{-C_2\epsilon\sqrt{n}}{\log(m+1)}\right)$.

Proof: Restrict attention to paths where $\mathcal{E}$ occurs. Only terms where the two indicator functions differ in Equation (D.5) contribute to the sum. If the first indicator is 1 while the second is zero, then

$$\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}}) < r_j(\tau,\hat{\mu}_j) \leq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau) \implies \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau) - C_A\delta \leq r_j(\tau,\hat{\mu}_j) \leq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau) + C_A\delta.$$

If the first indicator is zero, while the second is one, a similar implication holds. Consequently,

$$\left|(\mu_j - \hat{\mu}_j)\Big(\mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})) \;-\; \mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))\Big)\right| \leq |\mu_j - \hat{\mu}_j|\,\mathbb{I}\left(\left|r_j(\tau,\hat{\boldsymbol{\mu}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta\right).$$

Next, note $\frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\min}}{\nu_{\min}+\tau} \leq 1$ for all $\tau \geq 0$. Hence,

$$\begin{aligned}\left|r_j(\tau,\hat{\boldsymbol{\mu}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta &\iff \left|\hat{\mu}_j - \frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\min}}{\nu_{\min}+\tau}\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta\frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\min}}{\nu_{\min}+\tau} \\ &\implies |\hat{\mu}_j| \leq \frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\min}}{\nu_{\min}+\tau}\left(C_A\delta + \left|\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\right) \\ &\implies |\hat{\mu}_j| \leq C_A\delta + \left|\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \\ &\implies |\hat{\mu}_j| \leq C_A(\delta + \lambda_{\max}),\end{aligned}$$

Let $K_0 \equiv C_\mu + C_A(\lambda_{\max} + \delta)$. Then, these implications prove that if $\left|r_j(\tau, \hat{\boldsymbol{\mu}}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)\right| \leq C_A\delta$, then

$$|\mu_j - \hat{\mu}_j| \;\leq\; C_\mu + |\hat{\mu}_j| \;\leq\; C_\mu + C_A(\lambda_{\max} + \delta) \;=\; K_0.$$

We combine these observations to simplify our original supremum:

$$\begin{aligned}
&\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^n(\mu_j - \hat{\mu}_j)\Big(\mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})) \;-\; \mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))\Big)\right| \\
&\qquad\leq\; K_0\sup_{\tau\geq 0}\frac{1}{n}\sum_{j=1}^n \mathbb{I}(\left|r_j(\tau,\hat{\boldsymbol{\mu}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta) \\
&\qquad\leq\; K_0\sup_{\tau\geq 0}\frac{1}{n}\sum_{j=1}^n \mathbb{P}\{\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta\} \\
&\qquad\qquad + K_0\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^n \mathbb{I}(\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta) - \mathbb{P}\left\{\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta\right\}\right|.
\end{aligned}$$

We will now bound each of these two supremums. Let $\zeta_j = \sqrt{\nu_j}(\hat{\mu}_j - \mu_j)$ be a standardized increment. Then, $\mathbb{P}\{\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta\} = \mathbb{P}\{|\zeta_j - s| \leq C_A\delta\frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\min}+\tau}{\nu_{\min}}\sqrt{\nu_j}\}$, where $s = \sqrt{\nu_j}\left(\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\min}+\tau}{\nu_{\min}} - \mu_j\right)$. For any $\tau \geq 0$, $\frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\min}+\tau}{\nu_{\min}}\sqrt{\nu_j} \leq \sqrt{\nu_{\max}}$. By Assumption 4.2, this probability is thus at most $2C_A\phi_{max}\sqrt{\nu_{\max}}$, and thus the first supremum is bounded by $C_1\delta$ where $C_1 = 2K_0C_A\phi_{max}\sqrt{\nu_{\max}}$.

We use Theorem A.2 to bound the probability that the second supremum exceeds ϵ. Take the envelopes $F_j(\hat{\mu}_j)$ to be K_0 so that $\|\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2\|_\Psi = K_0\sqrt{n}$ by Lemma A.1. Part iii) of Lemma D.1 shows that the pseudo-dimension of $\left\{\left(\mathbb{I}(\left|r_j(\tau,\hat{\boldsymbol{\mu}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta) \mid j = 1,\ldots,n\right) \in \mathbb{R}^n : \tau \geq 0\right\}$ is at most $10(2m+2)$, so that, for $m \geq 1$,

$$V(A,W) \leq 1 + 3\log(20) + 3\log(m+1) \leq 18\log(m+1),$$

where the last inequality follows because $m \geq 1$. Applying Theorem A.2 shows that

$$\begin{aligned}
&\mathbb{P}\left\{K_0\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^n \mathbb{I}(\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta) - \mathbb{P}\left\{\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \leq C_A\delta\right\}\right| > \epsilon\right\} \\
&\qquad\leq 25\exp\left(\frac{-C_2\epsilon\sqrt{n}}{\log(m+1)}\right),
\end{aligned}$$

where $C_2 = (9K_0\cdot 18)^{-1}$. This completes the proof. □

Lemma D.8 (ULLN for Dual Approximation) *There exists a positive, dimension-independent constant C such that*

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^n(\mu_j - \hat{\mu}_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)) - \mathbb{E}\left[(\mu_j - \hat{\mu}_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))\right]\right| \leq R_2$$

where $\mathbb{P}\{R_2 > \epsilon\} \leq 25\exp\left(-\frac{C\epsilon\sqrt{n}}{\log(m+1)}\right)$.

Proof: We apply Theorem A.2 with envelopes $F_j(\hat{\mu}_j) \equiv |\mu_j - \hat{\mu}_j|$. From Lemma A.1, $\|\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2\|_\Psi \le \sigma\sqrt{\frac{2}{\nu_{\min}}}\sqrt{n}$. Part iv) of Lemma D.1 shows the the pseudo-dimension of the set $\left\{\Big((\mu_j - \hat{\mu}_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\tau)) \mid j = 1,\ldots,n\Big) \in \mathbb{R}^n \ : \ \tau \ge 0\right\}$ is at most $2m+2$. By Theorem A.3, its packing numbers satisfy Equation (A.2), and, since $m \ge 1$,

$$V(A,W) \le 1 + 3\log(2) + 3\log(m+1) \le 8\log(m+1).$$

Substituting in these numbers proves the lemma for $C = \frac{\sqrt{\nu_{\min}}}{9\cdot 8\sigma\sqrt{2}}$. □

Lemma D.9 (Approximating Stein's Lemma) *For any $0 < h < 1$, let $h_j(\tau) = \frac{(\nu_{\min}+\tau)\sqrt{\nu_j}}{\nu_{\min}(\nu_j+\tau)}h = r_j(\tau, h/\sqrt{\nu_j})$. Then, for each $j = 1,\ldots,n$,*

$$\left|\mathbb{E}\left[(\mu_j - \hat{\mu}_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))\right] + \frac{1}{2h\sqrt{\nu_j}}\mathbb{P}\left\{\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \le h_j(\tau)\right\}\right| \le \frac{4h^2}{\sqrt{\nu_{\min}}} + \frac{\|\phi_j - \phi\|_1}{\sqrt{\nu_{\min}}}\left(h^{-1} + 24\sigma^2 - \log\left(\|\phi_j - \phi\|_1\right)\right).$$

Moreover,

$$\frac{1}{n}\sum_{j=1}^n \left|\mathbb{E}\left[(\mu_j - \hat{\mu}_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))\right] + \frac{1}{2h\sqrt{\nu_j}}\mathbb{P}\left\{\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \le h_j(\tau)\right\}\right| \le \frac{4h^2}{\sqrt{\nu_{\min}}} + \frac{2\mathsf{TV}}{\sqrt{\nu_{\min}}}\left(h^{-1} + 24\sigma^2 - \log\left(2\mathsf{TV}\right)\right).$$

Proof: Let $\zeta_j = \sqrt{\nu_j}(\hat{\mu}_j - \mu_j)$, and define the function

$$f(t) = \frac{-1}{\sqrt{\nu_j}}\mathbb{I}(r_j(\tau, t/\sqrt{\nu_j} + \mu_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)).$$

Then, $\mathbb{E}\left[(\mu_j - \hat{\mu}_j)\mathbb{I}(r_j(\tau,\hat{\mu}_j) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau))\right] = \mathbb{E}[\zeta_j f(\zeta_j)]$. We apply Lemma C.2 to this function. It is bounded by $1/\sqrt{\nu_j}$. Moreover

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{2h}\big(f(\zeta_j + h) - f(\zeta_j - h)\big)\right] &= \frac{1}{2h\sqrt{\nu_j}}\mathbb{E}\Big[\mathbb{I}\left(r_j(\tau,\zeta_j/\sqrt{\nu_j} + \mu_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau) > -r_j(\tau,h/\sqrt{\nu_j})\right) \\
&\qquad - \mathbb{I}\left(r_j(\tau,\zeta_j/\sqrt{\nu_j} + \mu_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau) > r_j(\tau,h/\sqrt{\nu_j})\right)\Big] \\
&= \frac{1}{2h\sqrt{\nu_j}}\mathbb{P}\Big\{\left|r_j(\tau,\zeta_j/\sqrt{\nu_j} + \mu_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \le r_j(\tau,h/\sqrt{\nu_j})\Big\} \\
&= \frac{1}{2h\sqrt{\nu_j}}\mathbb{P}\Big\{\left|r_j(\tau,\hat{\mu}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right| \le h_j(\tau)\Big\}.
\end{aligned}$$

Applying Lemma C.2 yields the first result. The second follows by summing and applying Jensen's inequality. □

Lemma D.10 (ULLN for Bias Approximation) *There exists a positive, dimension-independent constant C such that*

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\left(\mathbb{I}(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau))-\mathbb{P}\left\{\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\right\}\right)\right|<R_3$$

where $\mathbb{P}\{R_3>\epsilon\}\leq 25\exp\left(\frac{-C\epsilon h\sqrt{n}}{\log(m+1)}\right)$.

Proof: We again apply Theorem A.2. Take the envelope to be $\frac{1}{2h\sqrt{\nu_{\min}}}$, so that $\|\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2\|_\Psi \leq \frac{\sqrt{n}}{2h\sqrt{\nu_{\min}}}$. By Lemma D.1 part v), the pseudo-dimension of $\left\{\left(\ \mathbb{I}(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau))\ \mid j=1,\ldots,n\ \right)\in\mathbb{R}^n\ :\ \boldsymbol{\lambda}\in\mathbb{R}^m,\tau\in\mathbb{R}\right\}$ is at most $10(2m+2)$. By Theorem A.3, Equation (A.2) is satisfied and, for $m\geq 1$,

$$V(A,W)\leq 1+3\log(20)+3\log(m+1)\leq 18\log(m+1).$$

Applying the theorem yields the result for $C=\frac{2\nu_{\min}}{9\cdot 18}$. □

Lemma D.11 (Approximating Dual Solution in Bias) *There exist positive, dimension-independent constants C_1, C_2 such that*

$$\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\left(\mathbb{I}(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau))-\mathbb{I}(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\right|\leq h_j(\tau))\right)\right|$$

is at most $C_1\frac{\delta}{h}+R_4$, where

$$\mathbb{P}\{R_4>2\epsilon\ \textit{and}\ \mathcal{E}\}\leq 50\exp\left(-\frac{C_2h\epsilon\sqrt{n}}{\log(m+1)}\right).$$

Proof: Restrict attention to paths where $\mathcal{E}$ occurs. Only terms where the indicators differ contribute to the sum. We have two cases: If the first indicator is 1 and the second is 0, then

$$h_j(\tau)\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\right|\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|+C_A\delta,$$

so that $h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)$. Similarly, if the second indicator is 1 and the first is 0, then,

$$h_j(\tau)\geq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\right|\geq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|-C_A\delta,$$

so that $h_j(\tau)\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq C_A\delta+h_j(\tau)$. Combining the above inequalities, we obtain

$$\begin{aligned}\mathbb{I}\Big(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\Big)-\mathbb{I}\Big(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\right|\leq h_j(\tau)\Big)\\ \leq\ \mathbb{I}\Big(h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\Big)\\ +\ \mathbb{I}\Big(h_j(\tau)\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq C_A\delta+h_j(\tau)\Big).\end{aligned}$$

Thus, by the triangle inequality,

$$\begin{aligned}
&\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\left(\mathbb{I}(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau))-\mathbb{I}(\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau,\hat{\boldsymbol{\mu}})\right|\leq h_j(\tau))\right)\right| \\
&\quad\leq \sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\mathbb{I}\left(h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\right)\right| \\
&\qquad+\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\mathbb{I}\left(h_j(\tau)\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq C_A\delta+h_j(\tau)\right)\right|.
\end{aligned}$$

We will focus on bounding $\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\mathbb{I}\left(h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\right)\right|$. The same argument applies to the other term. Note that

$$\begin{aligned}
&\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\mathbb{I}\Big(h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\Big)\right| \\
&\quad\leq \sup_{\tau\geq 0}\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\mathbb{P}\Big\{h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\Big\} \\
&\quad+\sup_{\tau\geq 0}\left\{\frac{1}{2h\sqrt{\nu_{\mathsf{min}}}}\left|\sum_{j=1}^{n}\mathbb{I}(h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau))\right.\right. \\
&\qquad\left.\left.-\mathbb{P}\Big\{h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\Big\}\right|\right\}
\end{aligned}$$

Now consider $\mathbb{P}\Big\{h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\Big\}$. Rewrite the probability in terms of the standardized increment $\zeta_j=\sqrt{\nu_j}(\hat{\mu}_j-\mu_j)$, yielding $\mathbb{P}\left\{h-C_A\delta\sqrt{\nu_j}\frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\mathsf{min}}}{\nu_{\mathsf{min}}+\tau}\leq|\zeta_j-s|\leq h\right\}$ where $s=\sqrt{\nu_j}\left(\frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\mathsf{min}}}{\nu_{\mathsf{min}}+\tau}\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right)$. This probability is bounded by the probability that ζ_j belongs to an interval of length at most $2C_A\delta\sqrt{\nu_j}\frac{\nu_j+\tau}{\nu_j}\frac{\nu_{\mathsf{min}}}{\nu_{\mathsf{min}}+\tau}\leq 2C_A\sqrt{\nu_{\mathsf{max}}}$. Thus, the first supremum is bounded by $\frac{C_1\delta}{h}$ where $C_1=2C_A\phi_{max}\sqrt{\nu_{\mathsf{max}}}$..

We bound the second supremum using Theorem A.2. Take the envelopes to be $\frac{1}{2h\sqrt{\nu_{\mathsf{min}}}}$. Lemma D.1 bounds the pseudo-dimension of the relevant set to be at most $100(2m+2)$. Thus, for $m\geq 1$

$$V(A,W)\leq 1+3\log(200)+3\log(m+1)\leq 28\log(m+1)$$

and by Theorem A.2,

$$\mathbb{P}\left\{\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\mathbb{I}\left(h_j(\tau)-C_A\delta\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq h_j(\tau)\right)\right|\;>\;\frac{C_1\delta}{h}+\epsilon\right\}$$

is at most $25\exp\left(-\frac{C_2\epsilon h\sqrt{n}}{\log(m+1)}\right)$, for $C_2=\frac{2\sqrt{\nu_{\mathsf{min}}}}{9\cdot 28}$. An analogous argument shows that

$$\mathbb{P}\left\{\sup_{\tau\geq 0}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{1}{2h\sqrt{\nu_j}}\mathbb{I}\left(h_j(\tau)\leq\left|r_j(\tau,\hat{\mu}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\tau)\right|\leq C_A\delta+h_j(\tau)\right)\right|\;>\;\frac{C_1\delta}{h}+\epsilon\right\}$$

is at most $25\exp\left(-\frac{C_2\epsilon h\sqrt{n}}{\log(m+1)}\right)$, and combining the two bounds completes the proof. □

D.3. Proof of Corollary 4.4

Proof: If we take $\delta_n = \sqrt{h_n}/n^{1/4}$. Then, $\delta_n\sqrt{n} = \sqrt{h_n\sqrt{n}} \to \infty$ and $\delta_n/h_n = 1/\sqrt{h_n\sqrt{n}} \to 0$. With these choices and the normality of $\hat{\mu}_j$, the deterministic errors in Theorem 4.3 tends to zero, and the stochastic error R tends to 0 in probability, so the suboptimality gap between our procedure and the oracle procedure converges to zero. Moreover, for this scaling, these probabilities are summable, and by Borel-Cantelli's Lemma, the suboptimality gap also converges to zero almost surely, proving the desired result. □

Appendix E: Proof of the Results in Section 5 for the Regularization Policies

In this section, we first show our regularization policies can be reinterpreted via the lens of robust optimization. We then provide the proof of Theorem 5.2, which is given in Section E.2. The proof depends on Lemma E.2 below and the auxiliary results given in Section E.3. We also discuss performance of our regularized policy in the large-sample regime in Section E.5.

E.1. Relationship to Robust Optimization

The class $\mathcal{X}^{Reg}(\hat{\boldsymbol{\mu}})$ can alternatively be motivated through the lens of robust optimization. The robust optimization approach to P^n creates an uncertainty set $\mathcal{U}(\hat{\boldsymbol{\mu}})$ and then solves

$$\boldsymbol{x}\left(\mathcal{U}(\hat{\boldsymbol{\mu}})\right) \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} \min_{\boldsymbol{u}\in\mathcal{U}(\hat{\boldsymbol{\mu}})} \frac{1}{n}\boldsymbol{u}^\top\boldsymbol{x}. \tag{E.1}$$

There are a variety of proposals in the literature for constructing $\mathcal{U}(\hat{\boldsymbol{\mu}})$ that leverage a priori knowledge on the distribution of $\hat{\boldsymbol{\mu}}$ to ensure that the resulting solution $\boldsymbol{x}\left(\mathcal{U}(\hat{\boldsymbol{\mu}})\right)$ enjoys desirable statistical properties; see Bertsimas et al. (2018) for a recent treatment.

When $\hat{\boldsymbol{\mu}}$ is independent and multivariate gaussian, a natural choice for $\mathcal{U}(\hat{\boldsymbol{\mu}})$ might be the ellipse $\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r) = \{\boldsymbol{\mu} : \sum_{j=1}^n \nu_j(\mu_j - \hat{\mu}_j)^2 \le r^2\}$ because it corresponds to the level set of the relevant normal distribution. Setting r to be the $1-\epsilon$ quantile of a χ^2 random variable with n degrees of freedom guarantees that $\boldsymbol{\mu} \in \mathcal{U}(\hat{\boldsymbol{\mu}})$ with probability at least $1-\epsilon$. Many authors advocate for elliptical uncertainty sets (with different values of r) more generally, even when $\hat{\boldsymbol{\mu}}$ is non-gaussian. Ben-Tal and Nemirovski (2000) and Gupta (2019) provide some probabilistic justifications in frequentist and Bayesian settings, respectively. The wide success of elliptical uncertainty sets with various radii within the robust optimization community suggests that the class of policies $\{\boldsymbol{x}\left(\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r)\right) : r \ge 0\}$ is an interesting class for P^n and that its oracle policy should perform well in practice. Interestingly, this policy class essentially coincides with our proposed regularization class.

Lemma E.1 (Correspondence between Regularization and Uncertainty Sets) *For each* $\hat{\boldsymbol{\mu}} \in \mathbb{R}^n$, $\{\boldsymbol{x}\left(\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r)\right) : r \ge 0\} = \{\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) : \Gamma \ge 0\}$.

The above correspondence has two important implications. First, it gives an alternative intuition for the policy class $\mathcal{X}_n^{Reg}(\hat{\boldsymbol{\mu}})$, supporting the idea that $\boldsymbol{x}^{\mathcal{R}}(\Gamma^{\mathsf{OR}}, \hat{\boldsymbol{\mu}})$ should have good performance in practice. Second, our search for a best-in-class policy for $\mathcal{X}_n^{Reg}(\hat{\boldsymbol{\mu}})$ can equivalently be interpreted as searching for the "best-in-class radius" for an elliptical uncertainty set. As seen in Sec. 6, the resulting radius is often quite different from those suggested by traditional robust optimization guidelines and offers significant benefits in the small-data, large-scale regime.[13]

E.2. Proof of Theorem 5.2

Recall from Section 5 that

$$\mathcal{X}^{Reg}(\hat{\boldsymbol{\mu}}) = \left\{\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) : \Gamma \geq 0\right\} \quad \text{where} \quad \boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x} - \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{2n}\sum_{j=1}^n \frac{x_j^2}{2\nu_j},$$

and for each j, define

$$w_j(\Gamma, t) = \begin{cases} 0 & \text{if} \quad t < 0, \\ \frac{\nu_j}{2\Gamma\sqrt{\nu_{\mathsf{min}}}} t^2 & \text{if} \quad 0 \leq t \leq \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\nu_j} \\ t - \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{2\nu_j} & \text{if} \quad \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\nu_j} < t \end{cases}$$

The following lemma gives and explicit formula for $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ and characterizes the associated optimal dual variables $\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$.

Lemma E.2 (Dual to Regularized Problem) *For $j = 1, \ldots, n$,*

$$x_j^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) = \frac{\nu_j}{\Gamma\sqrt{\nu_{\mathsf{min}}}} \left(\left[\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})\right]^+ - \left[\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) - \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\nu_j}\right]^+ \right).$$

Moreover, the corresponding optimal dual variables $\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ is given by

$$\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) \in \arg\min_{\boldsymbol{\lambda}\geq 0} D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \lambda), \qquad \textit{where} \qquad D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{\lambda}) \equiv \boldsymbol{b}^\top \boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^n w_j(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}),$$

Proof: The proof uses standard results in convex optimization. Dualizing the constraints in the optimization defining $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ yields:

$$\min_{\boldsymbol{\lambda}\geq 0, \boldsymbol{\theta}\geq 0} \boldsymbol{b}^\top \boldsymbol{\lambda} + \frac{1}{n}\boldsymbol{e}^\top \boldsymbol{\theta} + \frac{1}{n}\sum_{j=1}^n \max_{x_j \geq 0} \quad (\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} - \theta_j) x_j - \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{2\nu_j} x_j^2.$$

[13] Furthermore, we remark that the above correspondence is not specific to our choice of regularizer; many regularization problems admit interpretations as robust optimization problems under a well-chosen uncertainty set that depends on the specific regularizer; see Xu et al. (2009), Ben-Tal et al. (2015), Lam (2016), Fertis (2009), Bertsimas and Copenhaver (2018), and the references therein.

The optimization in x_j can be solved explicitly, yielding $x_j^* = \frac{\nu_j}{\Gamma\sqrt{\nu_{\min}}}(\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} - \theta_j)^+$. Substitute in this value. The resulting optimization becomes

$$\begin{aligned}
&\min_{\boldsymbol{\lambda}\geq 0, \boldsymbol{\theta}\geq 0} \quad \boldsymbol{b}^\top \boldsymbol{\lambda} + \frac{1}{n}\boldsymbol{e}^\top \boldsymbol{\theta} + \frac{1}{n}\sum_{j=1}^n \frac{\nu_j}{2\Gamma\sqrt{\nu_{\min}}}\left([\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} - \theta_j]^+\right)^2 \\
&= \min_{\boldsymbol{\lambda}\geq 0} \quad \boldsymbol{b}^\top \boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^n \min_{\theta_j \geq 0} \quad \theta_j + \frac{\nu_j}{2\Gamma\sqrt{\nu_{\min}}}\left([\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} - \theta_j]^+\right)^2 \\
&= \min_{\boldsymbol{\lambda}\geq 0} \quad \boldsymbol{b}^\top \boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^n w_j(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}) \ ,
\end{aligned}$$

where the last equality follows from the fact $w_j(\Gamma, t) = \min_{z\geq 0} z + \frac{\nu_j}{2\Gamma\sqrt{\nu_{\min}}}([t-z]^+)^2$ and the corresponding optimizer is given by $z^*(t) = \left[t - \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right]^+$. Finally, substituting in the optimal value of θ_j into x_j^* yields the form given in the theorem. □

We next define an "average" dual:

$$\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \in \arg\min_{\boldsymbol{\lambda}\geq 0} D^{\mathcal{R}}(\Gamma, \lambda), \text{ where } D^{\mathcal{R}}(\Gamma, \boldsymbol{\lambda}) \equiv \boldsymbol{b}^\top \boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^n \mathbb{E}[w_j(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda})].$$

The remainder of the proof follows the proof of Theorem 4.3 very closely in structure. In this section, we will say a constant C is *dimension-independent* if C does *not* depend on $\{n, m, \delta\}$ but may depend on any other problem parameters. In light of Lemma E.2, we define the function

$$g_j(\Gamma, \boldsymbol{\lambda}) = \frac{\nu_j}{\Gamma\sqrt{\nu_{\min}}}\left([\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}]^+ - \left[\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} - \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right]^+\right), \tag{E.2}$$

so that $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\mu}) = g_j(\Gamma, \lambda^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}))$.

Here is the proof of Theorem 5.2.

Proof of Theorem 5.2: It suffices to bound $\sup_{\Gamma\in[\Gamma_{\min}, \Gamma_{\max}]} \left|\frac{1}{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) - B_n^{Reg}(\Gamma, \hat{\boldsymbol{\mu}})\right|$ (see Lemma C.1). By triangle inequality,

$$\begin{aligned}
\sup_{\Gamma\in[\Gamma_{\min}, \Gamma_{\max}]} \left|\frac{1}{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \boldsymbol{x}(\Gamma, \hat{\boldsymbol{\mu}}) - B_n^{Reg}(\Gamma, \hat{\boldsymbol{\mu}})\right| \leq\ & \text{Error from Approximating Dual Solution} \\
&+ \text{Error from ULLN for Dual Approximation} \\
&+ \text{Error from Approximating Stein's Lemma} \\
&+ \text{Error from ULLN for Bias Approximation} \\
&+ \text{Error from Approximating Dual in Bias},
\end{aligned}$$

where

Error from Approximating Dual Solution:

$$\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]}\left|\frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}_j-\mu_j)\left(g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}))-g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))\right)\right|$$

Error from ULLN for Dual Approximation:

$$\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]}\left|\frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}_j-\mu_j)g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))-\mathbb{E}[(\hat{\mu}_j-\mu_j)g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))]\right|$$

Error from Stein's Lemma:

$$\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[(\hat{\mu}_j-\mu_j)g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))]-\frac{1}{\Gamma\sqrt{\nu_{\min}}n}\sum_{j=1}^{n}\mathbb{P}\left\{0\le\hat{\mu}_j-\boldsymbol{A}_j^{\top}\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right\}\right|$$

Error from ULLN for Bias Approximation:

$$\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]}\left|\frac{1}{\Gamma\sqrt{\nu_{\min}}n}\sum_{j=1}^{n}\mathbb{P}\left\{0\le\hat{\mu}_j-\boldsymbol{A}_j^{\top}\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right\}-\mathbb{I}\left(0\le\hat{\mu}_j-\boldsymbol{A}_j^{\top}\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)\right|$$

Error from Approximating Dual in Bias:

$$\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]}\left|\frac{1}{\Gamma\sqrt{\nu_{\min}}n}\sum_{j=1}^{n}\mathbb{I}\left(0\le\hat{\mu}_j-\boldsymbol{A}_j^{\top}\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)-\mathbb{I}\left(0\le\hat{\mu}_j-\boldsymbol{A}_j^{\top}\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}})\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)\right|$$

Lemmas E.6, E.7, E.8, E.9, and E.10 below bound each of these sources of error. In particular, there exists a positive, dimension-independent constant C_1 such that

$$\sup_{\Gamma\ge0}\left|\frac{1}{n}(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^{\top}\boldsymbol{x}(\Gamma,\hat{\boldsymbol{\mu}})-B(\Gamma)\right|\le C_1\left(\mathsf{TV}-\mathsf{TV}\log(\mathsf{TV})+\delta\right)+R_1+R_2+R_3+R_4,$$

where $R_1,\ldots,R_4$ are the stochastic remainders from Lemmas E.6, E.7, E.9, and E.10 and C_1 is the maximum of the relevant constants from these lemmas. Moreover, these lemmas prove that there exist positive, dimension-independent constants $C_5,\ldots,C_{13}$ such that

$$\begin{aligned}
&\mathbb{P}\{R_1+R_2+R_3+R_4>4\epsilon\}\\
&\le\ \mathbb{P}\{\mathcal{E}^c\}+\mathbb{P}\{R_1+R_2+R_3+R_4>4\epsilon\text{ and }\mathcal{E}\}\\
&\le\ C_5\exp\left(\frac{-C_6\delta\sqrt{n}}{\sqrt{m}\log(m+1)}\right)\log\left(1+\frac{\sqrt{m}\log(m+1)}{C_6\delta\sqrt{n}}\right)\\
&\quad+2\exp\left(\frac{-C_7\epsilon^2n}{\delta^2}\right)+C_8\exp\left(-C_9\epsilon\sqrt{n}\right)+C_{10}\exp\left(\frac{-C_{11}\epsilon\sqrt{n}}{\log(m+1)}\right)+C_{12}\exp\left(-\frac{C_{13}\epsilon\sqrt{n}}{\log(m+1)}\right),
\end{aligned}$$

where the event $\mathcal{E}$ is defined in the next section in Equation (E.3). We simplify this bound by first noting that for $\epsilon>\delta^2/\sqrt{n}$, we have $\frac{\epsilon^2n}{\delta^2}>\epsilon\sqrt{n}$. This simplification allows us to combine the last 4 exponential terms (replacing dimension-independent constants as appropriate). Simplifying then yields the result. □

E.3. Auxiliary Proofs for Theorem 5.2.

Just as in the proof of Theorem 4.3, we first argue that $\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}})$ and $\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)$ are uniformly close with high probability. Define the dimension-independent constant

$$\lambda^{\mathcal{R}}_{\max}\equiv\frac{2}{s_0}\left(C_\mu+\frac{1}{\sqrt{\nu_{\min}}}\right)+\frac{\Gamma_{\max}}{2\sqrt{\nu_{\min}}}+1$$

and the event

$$\mathcal{E} = \left\{ \sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}})\|_1 \le \lambda^{\mathcal{R}}_{\max}, \sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}) - \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_1 \le \delta, \right.$$
$$\left. \sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_1 \le \lambda^{\mathcal{R}}_{\max} \right\}. \quad \text{(E.3)}$$

Lemma E.3 (Dual Variables Are Bounded)

i) $\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_1 \le \lambda^{\mathcal{R}}_{\max}$

ii) $\mathbb{P}\left\{\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}})\|_1 > \lambda^{\mathcal{R}}_{\max}\right\} \le 2\exp\left(-\frac{n\nu_{\min}}{72\sigma^2}\right)$

Proof: The proof is very similar to Lemma D.3. First

$$D^{\mathcal{R}}(\Gamma,\boldsymbol{\lambda}(\Gamma)) \le D^{\mathcal{R}}(\Gamma,\mathbf{0}) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[w_j\left(\Gamma,\hat{\mu}_j\right)\right] \le \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\hat{\mu}_j^+\right] \le \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left|\hat{\mu}_j\right| \ ,$$

where the last two inequalities follow because $w_j(\Gamma,\hat{\mu}_j) \le [\hat{\mu}_j]^+ \le |\hat{\mu}_j|$ for all Γ. Hence,

$$\begin{aligned}\|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_1 \le \max_{\boldsymbol{\lambda}\ge 0} \quad & \boldsymbol{e}^\top\boldsymbol{\lambda} \\ \text{s.t.} \quad & \boldsymbol{b}^\top\boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[w_j\left(\Gamma,\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right)\right] \le \frac{1}{n}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}\|_1\right].\end{aligned}$$

Using $(\boldsymbol{x}^0, s_0)$ from Assumption 4.1 and Lagrangian duality,

$$\begin{aligned}\|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_1 &\le \max_{\boldsymbol{\lambda}\ge 0}\left\{\boldsymbol{e}^\top\boldsymbol{\lambda} + \frac{1}{s_0}\left(\frac{1}{n}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}\|_1\right] - \boldsymbol{b}^\top\boldsymbol{\lambda} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[w_j\left(\Gamma,\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right)\right]\right)\right\} \\ &= \frac{1}{ns_0}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}\|_1\right] + \max_{\boldsymbol{\lambda}\ge 0}\left\{(\boldsymbol{e} - \frac{1}{s_0}\boldsymbol{b})^\top\boldsymbol{\lambda} - \frac{1}{ns_0}\sum_{j=1}^{n}\mathbb{E}\left[w_j\left(\Gamma,\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right)\right]\right\}\end{aligned}$$

Rewrite $w_j(\cdot)$ as

$$\begin{aligned}w_j(\Gamma,\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}) &= \max_{x_j\in[0,1]}(\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda})x_j - \frac{\Gamma\sqrt{\nu_{\min}}}{2}\frac{x_j^2}{\nu_j} \\ &\ge (\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda})x_j^0 - \frac{\Gamma\sqrt{\nu_{\min}}}{2}\frac{x_j^{0\,2}}{\nu_j},\end{aligned}$$

since $x_0^j \in [0,1]$. Take expectations of both sides and substitute above:

$$\begin{aligned}\|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_1 &\le \frac{1}{ns_0}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}\|_1\right] + \max_{\boldsymbol{\lambda}\ge 0}\ (\boldsymbol{e} - \frac{1}{s_0}\boldsymbol{b})^\top\boldsymbol{\lambda} - \frac{1}{ns_0}\sum_{j=1}^{n}(\hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda})x_j^0 - \frac{\Gamma\sqrt{\nu_{\min}}}{2}\frac{x_j^{0\,2}}{\nu_j} \\ &= \frac{1}{ns_0}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}\|_1\right] + \max_{\boldsymbol{\lambda}\ge 0}\ (\boldsymbol{e} - \frac{1}{s_0}\boldsymbol{b} + \frac{1}{ns_0}\boldsymbol{A}\boldsymbol{x}^0)^\top\boldsymbol{\lambda} - \frac{1}{ns_0}\sum_{j=1}^{n}\hat{\mu}_j x_j^0 - \frac{\Gamma\sqrt{\nu_{\min}}}{2}\frac{x_j^{0\,2}}{\nu_j}\end{aligned}$$

By Assumption 4.1, $\frac{1}{n}\boldsymbol{A}\boldsymbol{x}^0 + s_0\boldsymbol{e} \leq \boldsymbol{b} \iff \boldsymbol{e} - \frac{1}{s_0}\boldsymbol{b} + \frac{1}{ns_0}\boldsymbol{A}\boldsymbol{x}^0 \leq 0$, which implies that $\boldsymbol{\lambda} = \boldsymbol{0}$ is optimal for this last optimization problem. Thus,

$$\begin{aligned}\|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_1 &\leq \frac{1}{ns_0}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}\|_1\right] - \frac{1}{ns_0}\sum_{j=1}^{n}\hat{\mu}_j x_j^0 - \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{2}\frac{{x_j^0}^2}{\nu_j} \\ &\leq \frac{2}{ns_0}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}\|_1\right] + \frac{\Gamma_{\mathsf{max}}}{2\sqrt{\nu_{\mathsf{min}}}},\end{aligned}$$

since, again, $\boldsymbol{x}^0 \in [0,1]^n$. Finally, use the fact that $\mathbb{E}[|\hat{\mu}_j|] \leq C_\mu + \frac{1}{\sqrt{\nu_j}} \leq C_\mu + \frac{1}{\sqrt{\nu_{\mathsf{min}}}}$ to simplify, yielding

$$\|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_1 \;\leq\; \frac{2}{s_0}\left(C_\mu + \frac{1}{\sqrt{\nu_{\mathsf{min}}}}\right) + \frac{\Gamma_{\mathsf{max}}}{2\sqrt{\nu_{\mathsf{min}}}} \;\leq\; \lambda^{\mathcal{R}}_{\mathsf{max}}\ .$$

This proves the first claim.

For the second, we can follow an essentially identical series of steps using $D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}))$ to prove that

$$\|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})\|_1 \leq \frac{2}{ns_0}\|\hat{\boldsymbol{\mu}}\|_1 + \frac{\Gamma_{\mathsf{max}}}{2\sqrt{\nu_{\mathsf{min}}}}.$$

From Equation (D.2), $\frac{1}{n}\|\hat{\boldsymbol{\mu}}\|_1 \leq \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[|\hat{\mu}_j|] + R$ with $\mathbb{P}(R > t) \leq 2\exp\left(-\frac{nt^2\nu_{\mathsf{min}}}{72\sigma^2}\right)$. Moreover, as in the previous part $\mathbb{E}[|\hat{\mu}_j|] \leq C_\mu + \frac{1}{\sqrt{\nu_{\mathsf{min}}}}$. Substituting and simplifying yields,

$$\|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})\|_1 \leq \frac{2}{s_0}\left(C_\mu + \frac{1}{\sqrt{\nu_{\mathsf{min}}}}\right) + \frac{\Gamma_{\mathsf{max}}}{2\sqrt{\nu_{\mathsf{min}}}} + R.$$

Using the definition of $\lambda^{\mathcal{R}}_{\mathsf{max}}$ and the tail inequality on R proves the second statement. □

Lemma E.4 (Strong Convexity of the Average Dual) *There exists a dimension-independent constant $\kappa > 0$ such that for any $\Gamma \geq \Gamma_{\mathsf{min}}$, the function $\boldsymbol{\lambda} \mapsto \Gamma D^{\mathcal{R}}(\Gamma, \boldsymbol{\lambda})$ is κ-strongly convex in $\boldsymbol{\lambda}$ for all $\boldsymbol{\lambda} \in \mathbb{R}^m_+$ such that $\|\boldsymbol{\lambda}\|_1 \leq \lambda^{\mathcal{R}}_{\mathsf{max}}$.*

Proof: Let $\zeta_j \equiv \sqrt{\nu_j}(\hat{\mu}_j - \mu_j)$. Using the definition of $w_j(\Gamma, \hat{\mu}_j - s)$ and by differentiating under the integral sign,

$$\begin{aligned}\frac{\partial^2}{\partial s^2}\mathbb{E}[\Gamma w_j(\Gamma, \hat{\mu}_j - s)] &= \frac{\nu_j}{\sqrt{\nu_{\mathsf{min}}}}\mathbb{P}\left\{0 < \hat{\mu}_j - s < \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\nu_j}\right\} \\ &= \frac{\nu_j}{\sqrt{\nu_{\mathsf{min}}}}\mathbb{P}\left\{(s-\mu_j)\sqrt{\nu_j} < \zeta_j < (s-\mu_j)\sqrt{\nu_j} + \frac{\Gamma\sqrt{\nu_{\mathsf{min}}}}{\sqrt{\nu_j}}\right\} \\ &\geq \frac{\nu_j}{\sqrt{\nu_{\mathsf{min}}}}\mathbb{P}\left\{(s-\mu_j)\sqrt{\nu_j} < \zeta_j < (s-\mu_j)\sqrt{\nu_j} + \frac{\Gamma_{\mathsf{min}}\sqrt{\nu_{\mathsf{min}}}}{\sqrt{\nu_j}})\right\}\end{aligned}$$

Furthermore,

$$\max\left\{\left|(s-\mu_j)\sqrt{\nu_j}\right|, \left|(s-\mu_j)\sqrt{\nu_j} + \frac{\Gamma_{\mathsf{min}}\sqrt{\nu_{\mathsf{min}}}}{\sqrt{\nu_j}})\right|\right\} \leq (|s| + C_\mu)\sqrt{\nu_{\mathsf{max}}} + \Gamma_{\mathsf{min}}.$$

Consequently, by Assumption 4.2, for $|s| \le C_A \lambda^{\mathcal{R}}_{\mathsf{max}}$,

$$\frac{\partial^2}{\partial s^2}\mathbb{E}[\Gamma w_j(\Gamma, \hat{\mu}_j - s)] \ge \phi_{\mathsf{min}}\left((C_A\lambda^{\mathcal{R}}_{\mathsf{max}} + C_\mu)\sqrt{\nu_{\mathsf{max}}} + \Gamma_{\mathsf{min}}\right).$$

Let $\kappa' \equiv \phi_{\mathsf{min}}\left((C_A\lambda^{\mathcal{R}}_{\mathsf{max}} + C_\mu)\sqrt{\nu_{\mathsf{max}}} + \Gamma_{\mathsf{min}}\right)$. The remainder of the proof now parallels the proof of Lemma D.4. □

Lemma E.5 (Uniform Convergence of Dual Solutions) *There exists a positive, dimension-independent constant C such that*

$$\mathbb{P}\{\mathcal{E}^c\} \le 86\exp\left(\frac{-C\delta\sqrt{n}}{\sqrt{m}\log(m+1)}\right)\log\left(1 + \frac{\sqrt{m}\log(m+1)}{C\delta\sqrt{n}}\right).$$

Proof: We will follow the approach in the proof of Lemma D.5. If $\sup_{\Gamma\in[\Gamma_{\mathsf{min}},\Gamma_{\mathsf{max}}]} \|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})\|_1 \le \lambda^{\mathcal{R}}_{max}$, then, using the fact that $D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})) \le D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))$, we have

$$\begin{aligned}&\left[D^{\mathcal{R}}(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}((\Gamma, \hat{\boldsymbol{\mu}})) - D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}))\right] - \left[D^{\mathcal{R}}(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)) - D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))\right]\\ &\quad\ge\ D^{\mathcal{R}}(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}((\Gamma, \hat{\boldsymbol{\mu}})) - D^{\mathcal{R}}(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)) \quad\ge\quad \frac{\kappa}{2\Gamma_{\mathsf{max}}}\|\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) - \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_2^2,\end{aligned}$$

where the last inequality follows from the strong convexity of $\boldsymbol{\lambda} \mapsto \Gamma D^{\mathcal{R}}(\Gamma, \boldsymbol{\lambda})$ from Lemma D.4.

For k in the integers, define the set of functions

$$S_k = \left\{\boldsymbol{h} : \mathbb{R}_+ \mapsto \mathbb{R}^m_+ \mid 2^{k-1} \le \sqrt{n}\sup_{\Gamma\in[\Gamma_{\mathsf{min}},\ \Gamma_{\mathsf{max}}]}\|\boldsymbol{h}(\Gamma) - \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\|_2 < 2^k\right\}.$$

If $\boldsymbol{\lambda}^{\mathcal{R}}(\cdot, \hat{\boldsymbol{\mu}}) \in S_k$, then it follows that

$$\sup_{\Gamma, \boldsymbol{h}(\cdot)\in S_k}\left|\left[D^{\mathcal{R}}(\Gamma, \boldsymbol{h}(\Gamma)) - D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{h}(\Gamma))\right] - \left[D^{\mathcal{R}}(\Gamma, \boldsymbol{\lambda}(\Gamma)) - D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{\lambda}(\Gamma))\right]\right| \ \ge\ \frac{\kappa 2^{2k-2}}{2n\Gamma_{\mathsf{max}}}. \tag{E.4}$$

Consider the mapping

$$(\Gamma, \boldsymbol{h}(\cdot)) \mapsto \left[D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{\lambda}(\Gamma)) - D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}(\Gamma, \boldsymbol{h}(\Gamma))\right] - \left[D^{\mathcal{R}}(\Gamma, \boldsymbol{\lambda}(\Gamma)) - D^{\mathcal{R}}(\Gamma, \boldsymbol{h}(\Gamma))\right].$$

Using the definitions of $D^{\mathcal{R}}_{\hat{\boldsymbol{\mu}}}$ and $D^{\mathcal{R}}$, this mapping is the difference between an empirical average and its expectation. We will apply the first part of Theorem A.2. Let $F_j(\hat{\boldsymbol{\mu}}_j) = C_A 2^k\sqrt{m/n}$ for all j. Note that $F_j(\cdot)$ is a constant function, and thus, $\|\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2\|_\Psi \le C_A 2^k\sqrt{m}$ by Lemma A.1. We will show that it is an envelope. By definition, $0 \le \frac{d}{dt}w_j(\Gamma, t) \le 1$ for all t, so that function $w_j(\Gamma, \cdot)$ is non-expansive, and thus,

$$\left|\left[w_j\left(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\right) - w_j\left(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{h}(\Gamma)\right)\right| \le \left|\boldsymbol{A}_j^\top\left(\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) - \boldsymbol{h}(\Gamma)\right)\right| \ \le\ \frac{\|\boldsymbol{A}_j\|_2 2^k}{\sqrt{n}} \ \le\ F_j(\hat{\boldsymbol{\mu}}_j).$$

Let

$$\begin{aligned}\mathcal{F}_1 &\equiv \left\{\left(w_j\left(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\right) - w_j\left(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{h}(\Gamma)\right) \mid j = 1, \ldots, n\right) \in \mathbb{R}^n \ :\ \Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}], \boldsymbol{h}(\cdot) \in S_k\right\}\\ \mathcal{F}_2 &\equiv \left\{\left(w_j\left(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right) - w_j\left(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top\boldsymbol{h}\right) \mid j = 1, \ldots, n\right) \in \mathbb{R}^n \ :\ \Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}],\ \ \boldsymbol{\lambda}, \boldsymbol{h} \in \mathbb{R}^m\right\}.\end{aligned}$$

Our goals are to show that Equation (A.2) holds for $\mathcal{F}_1$ and to determine an upper bound for $V(A,W)$. Observe that since $\mathcal{F}_1 \subseteq \mathcal{F}_2$, it follows that $M(\epsilon\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2,\mathcal{F}_1) \leq M(\epsilon\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2,\mathcal{F}_2)$. To bound the packing number for $\mathcal{F}_2$, note that $\mathcal{F}_2$ is a pointwise difference of sets of the form

$$\mathcal{F}_3 \equiv \left\{ \left(w_j\left(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}\right) \mid j=1,\ldots,n\right) \in \mathbb{R}^n \ : \ \Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}], \boldsymbol{\lambda} \in \mathbb{R}^m \right\}.$$

Thus, by Page 22 in Pollard (1990), $M(\epsilon\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2,\mathcal{F}_2) \leq M(\epsilon\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2/4,\mathcal{F}_3)^2$. We will now bound the pseudo-dimension of $\mathcal{F}_3$.

For $j=1,\ldots,n$, let $a_j = \sqrt{\nu_{\mathsf{min}}}/\nu_j$. Considering an arbitrary $\boldsymbol{c} \in \mathbb{R}^n$, let

$$\mathcal{F}_4 \equiv \left\{ \left(\mathbb{I}\left(w_j\left(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}\right) > c_j\right) \mid j=1,\ldots,n\right) \in \{0,1\}^n \ : \ \Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}], \boldsymbol{\lambda} \in \mathbb{R}^m \right\}.$$

It is easy to verify that the psuedo-dimensions of $\mathcal{F}_3$ and $\mathcal{F}_4$ are the same. Recall that

$$w_j(\Gamma,t) = \begin{cases} 0 & \text{if} \quad t<0, \\ \frac{t^2}{2\Gamma a_j} & \text{if} \quad 0 \leq t \leq \Gamma a_j \\ t - \frac{\Gamma a_j}{2} & \text{if} \quad \Gamma a_j < t. \end{cases},$$

For any j, consider splitting on the events $\{c_j<0\}$, $\left\{0 \leq c_j \leq \frac{\Gamma a_j}{2}\right\}$ and $\left\{c_j > \frac{\Gamma a_j}{2}\right\}$. Then,

$$\begin{aligned}
&\mathbb{I}\Big(w_j\left(\Gamma, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}\right) > c_j\Big) \\
&\quad = \mathbb{I}\left(c_j<0\right) + \mathbb{I}\left(0 \leq c_j \leq \frac{\Gamma a_j}{2}, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > \sqrt{2\Gamma a_j c_j}\right) + \mathbb{I}\left(c_j > \frac{\Gamma a_j}{2}, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > \frac{\Gamma a_j}{2} + c_j\right) \\
&\quad = \mathbb{I}\left(c_j<0\right) + \max\left\{ \mathbb{I}\left(0 \leq c_j \leq \frac{\Gamma a_j}{2}, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > \sqrt{2\Gamma a_j c_j}\right) \ , \ \mathbb{I}\left(c_j > \frac{\Gamma a_j}{2}, \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > \frac{\Gamma a_j}{2} + c_j\right) \right\} \\
&\quad = \mathbb{I}\left(c_j<0\right) + \max\Bigg\{ \quad \min\left\{ \mathbb{I}\left(0 \leq c_j \leq \frac{\Gamma a_j}{2}\right) \ , \ \mathbb{I}\left(\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > \sqrt{2\Gamma a_j c_j}\right) \right\} \ , \\
&\qquad\qquad\qquad \min\left\{ \mathbb{I}\left(c_j > \frac{\Gamma a_j}{2}\right) \qquad , \ \mathbb{I}\left(\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > \frac{\Gamma a_j}{2} + c_j\right) \right\} \quad \Bigg\}.
\end{aligned}$$

Now, let

$$\begin{aligned}
\mathcal{G}_1 &\equiv \left\{ \left(\mathbb{I}\left(0 \leq c_j \leq \frac{\Gamma a_j}{2}\right) \mid j=1,\ldots,n \right) \in \{0,1\}^n \ : \ \Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}] \right\}, \\
\mathcal{G}_2 &\equiv \left\{ \left(\mathbb{I}\left(c_j > \frac{\Gamma a_j}{2}\right) \mid j=1,\ldots,n \right) \in \{0,1\}^n \ : \ \Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}] \right\}, \\
\mathcal{G}_3 &\equiv \left\{ \left(\mathbb{I}\left(\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > \sqrt{2\Gamma a_j c_j}\right) \mid j=1,\ldots,n \right) \in \{0,1\}^n \ : \ \Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}], \boldsymbol{\lambda} \in \mathbb{R}^m \right\}, \\
\mathcal{G}_4 &\equiv \left\{ \left(\mathbb{I}\left(\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > \frac{\Gamma a_j}{2} + c_j\right) \mid j=1,\ldots,n \right) \in \{0,1\}^n \ : \ \Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}], \boldsymbol{\lambda} \in \mathbb{R}^m \right\}.
\end{aligned}$$

Note that the pseudo-dimensions of $\mathcal{G}_1$ and $\mathcal{G}_2$ are at most one, while the pseudo-dimensions of $\mathcal{G}_3$ and $\mathcal{G}_4$ are at most $m+2$. Let $\wedge$ and $\vee$ denote the pointwise minimum and maximum, respectively. Therefore, by Lemma 5.1 in Pollard (1990), the pseudo-dimension of $\mathcal{G}_1 \wedge \mathcal{G}_3$ is at most $10(m+2)$.

Similarly, the pseudo-dimension of $\mathcal{G}_2 \wedge \mathcal{G}_4$ is at most $10(m+2)$. Therefore, pseudo-dimension of $(\mathcal{G}_1 \wedge \mathcal{G}_3) \vee (\mathcal{G}_2 \wedge \mathcal{G}_4)$ is at most $100(m+2)$. Moreover, the above argument shows that a translation of $\mathcal{F}_4$ by a constant vector is a subset of $(\mathcal{G}_1 \wedge \mathcal{G}_3) \vee (\mathcal{G}_2 \wedge \mathcal{G}_4)$. Therefore, the pseudo-dimension of $\mathcal{F}_4$, and thus $\mathcal{F}_3$, is at most $100(m+2)$.

Let $V = 100(m+2)$. Therefore, it follows from Theorem A.3 that

$$M(\epsilon\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}_1) \le M(\epsilon\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}_2) \;\le\; M\left(\frac{\epsilon}{4}\|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}_3\right)^2 \le \left[V^{6V} 4^{4V} \left(\frac{1}{\epsilon}\right)^{4V}\right]^2 = V^{12V} 4^{8V} \left(\frac{1}{\epsilon}\right)^{8V},$$

and thus, $\mathcal{F}_1$ satisfies Equation (A.2) with $A = V^{12V} 4^{8V}$ and $W = 8V$. Thus,

$$V(A,W) \le 1 + \frac{2\log A}{W} \le 1 + \frac{12V\log(V) + 8V\log(4)}{4V} = 1 + 2\log(4) + 3\log(100(m+2)),$$

and the last expression is at most $29\log(m+1)$ for $m \ge 1$. Finally, applying Theorem A.2 to Equation (E.4) proves that

$$\mathbb{P}\left\{\sup_{\Gamma\ge 0}\|\boldsymbol{\lambda}(\Gamma,\hat{\boldsymbol{\mu}})\|_1 \le \boldsymbol{\lambda}_{max} \text{ and } \lambda(\cdot,\hat{\boldsymbol{\mu}}) \in S_{k,n}\right\} \le 25\exp\left(\frac{-C_1 2^k}{\Gamma_{\mathsf{max}}\sqrt{m}\log(m+1)}\right),$$

where $C_1 = \frac{\kappa}{2^3\cdot 9\cdot 29 C_A}$.

As before, we use the above bound to decompose the $\mathbb{P}\{\mathcal{E}^c\}$ into "peels:"

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{\Gamma\ge 0}\|\boldsymbol{\lambda}(\Gamma,\hat{\boldsymbol{\mu}})\| \le \lambda_{\mathsf{max}} \text{ and } \sup_{\Gamma\ge 0}\|\boldsymbol{\lambda}(\Gamma) - \boldsymbol{\lambda}(\Gamma,\hat{\boldsymbol{\mu}})\|_2 \ge \delta\right\} \\
&\le \sum_{k=\lceil\log_2(\delta\sqrt{n})\rceil}^{\infty} \mathbb{P}\left\{\sup_{\Gamma\ge 0}\|\boldsymbol{\lambda}(\Gamma,\hat{\boldsymbol{\mu}})\| \le \lambda_{\mathsf{max}} \text{ and } \boldsymbol{\lambda}(\cdot,\hat{\boldsymbol{\mu}}) \in S_{k,n}\right\} \\
&\le 25 \sum_{k=\lceil\log_2(\delta\sqrt{n})\rceil}^{\infty} \exp\left(\frac{-C_1 2^k}{\Gamma_{\mathsf{max}}\sqrt{m}\log(m+1)}\right) \\
&\le 25 \int_{\log_2(\delta\sqrt{n})}^{\infty} \exp\left(\frac{-C_1 2^x}{\Gamma_{\mathsf{max}}\sqrt{m}\log(m+1)}\right) dx \;=\; \frac{25}{\log 2}\int_{\frac{C_1\delta\sqrt{n}}{\Gamma_{\mathsf{max}}\sqrt{m}\log(m+1)}}^{\infty} \exp(-u)\frac{du}{u},
\end{aligned}$$

where the last inequality follows by making the change of variables $u = \frac{C_1}{\Gamma_{\mathsf{max}}\sqrt{m}\log(m+1)}2^x$. We recognize the last integral as the exponential integral, which admits the bound

$$\int_x^{\infty} \exp(-t)\frac{dt}{t} \le \exp(-x)\log(1+1/x) \;\text{ for } x > 0.$$

Applying this bound and combining with Lemma E.3 yields:

$$\mathbb{P}\{\mathcal{E}^c\} \le \frac{25}{\log 2}\exp\left(\frac{-C_1\delta\sqrt{n}}{\Gamma_{\mathsf{max}}\sqrt{m}\log(m+1)}\right)\log\left(1 + \frac{\Gamma_{\mathsf{max}}\sqrt{m}\log(m+1)}{C_1\delta\sqrt{n}}\right) + 2\exp\left(-\frac{n\nu_{\mathsf{min}}}{72\sigma^2}\right).$$

To "clean up" the right-hand side, observe that for $0 \le \delta \le 1$, $\frac{\delta}{\sqrt{m}\log(m+1)} \le 1$ and, for $n \ge 2$, $\sqrt{n} \le n$, so we can bound the second term by $\exp\left(-\frac{\delta\nu_{\mathsf{min}}\sqrt{n}}{72\sigma^2\sqrt{m}\log(m+1)}\right)$. Then, let $C = \min\left(\frac{C_1}{\Gamma_{\mathsf{max}}}, \frac{\nu_{\mathsf{min}}}{72\sigma^2}\right)$ and note that $25/\log(2) + 2 < 86$ to prove the lemma. □

Lemma E.6 (Approximating the Dual Solution) *There exist dimension-independent constants C_1, C_2 such that*

$$\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]}\left|\frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}_j-\mu_j)\left(g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}))-g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))\right)\right|\leq C_1\delta+R_1,$$

where $\mathbb{P}\left\{\mathcal{E}\ \text{and}\ R_1>\epsilon\right\}\leq 2\exp\left(-\frac{C_2\epsilon^2 n}{\delta^2}\right)$.

Proof: Recall from Equation (E.2) that

$$g_j(\Gamma,\boldsymbol{\lambda})=\frac{\nu_j}{\Gamma\sqrt{\nu_{\min}}}\left([\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}]^+-\left[\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}-\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right]^+\right).$$

We claim that $g_j(\Gamma,\boldsymbol{\lambda})$ is $\left(\frac{2\nu_j}{\Gamma_{\min}\sqrt{\nu_{\min}}}C_A\right)$-Lipschitz in $\boldsymbol{\lambda}$ with respect to the ℓ_1-norm. Indeed,

$$\begin{aligned}|g_j(\Gamma,\boldsymbol{\lambda}_1)-g_j(\Gamma,\boldsymbol{\lambda}_2)|&\leq\frac{\nu_j}{\Gamma\sqrt{\nu_{\min}}}\left|(\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}_1)^+-(\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}_2)^+\right|\\&\quad+\frac{\nu_j}{\Gamma\sqrt{\nu_{\min}}}\left|\left(\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}_1-\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)^+-\left(\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}_2-\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)^+\right|\\&\leq\frac{2\nu_j}{\Gamma_{\min}\sqrt{\nu_{\min}}}\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2)\right|\ \leq\ \frac{2\nu_j}{\Gamma_{\min}\sqrt{\nu_{\min}}}C_A\|\boldsymbol{\lambda}_1-\boldsymbol{\lambda}_2\|_1\ .\end{aligned}$$

Now restrict attention to paths where $\mathcal{E}$ occurs. It follows that

$$\left|\frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}_j-\mu_j)\left(g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}}))-g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))\right)\right|\ \leq\ \frac{2\delta\nu_{\max}}{\Gamma_{\min}\sqrt{\nu_{\min}}}C_A\frac{1}{n}\sum_{j=1}^{n}|\hat{\mu}_j-\mu_j|\,.$$

Write

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}|\hat{\mu}_j-\mu_j|&\leq C_\mu+\frac{1}{n}\sum_{j=1}^{n}|\hat{\mu}_j|\\&\leq C_\mu+\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[|\hat{\mu}_j|]+\frac{1}{n}\sum_{j=1}^{n}|\hat{\mu}_j|-\mathbb{E}[|\hat{\mu}_j|]\\&\leq 2C_\mu+\frac{1}{\sqrt{\nu_{\min}}}+R,\end{aligned}$$

where $\mathbb{P}(R>t)\leq 2\exp\left(-\frac{nt^2\nu_{\min}}{72\sigma^2}\right)$ by Equation (D.2). Substituting above and letting $R_1=\frac{2\delta\nu_{\max}}{\Gamma_{\min}\sqrt{\nu_{\min}}}C_AR$ proves the theorem for $C_1=\frac{2\nu_{\max}}{\Gamma_{\min}\sqrt{\nu_{\min}}}C_A\left(2C_\mu+\frac{1}{\sqrt{\nu_{\min}}}\right)$ and $C_2=\frac{\Gamma_{\min}^2\nu_{\min}^2}{4\cdot72C_A^2\sigma^2\nu_{\max}^2}$. □

Lemma E.7 (ULLN for Dual Approximation) *There exist positive, dimension-independent constants C_1, C_2 such that*

$$\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]}\left|\frac{1}{n}\sum_{j=1}^{n}(\hat{\mu}_j-\mu_j)g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))-\mathbb{E}[(\hat{\mu}_j-\mu_j)g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))]\right|\leq R,$$

where $\mathbb{P}\{R>\epsilon\}\leq C_1\exp\left(-C_2\epsilon\sqrt{n}\right)$.

Proof: We apply Theorem A.2. For an envelope, it follows from the definition of $g_j(\Gamma, \boldsymbol{\lambda})$ in Equation (E.2) that $|(\hat{\mu}_j - \mu_j) g_j(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))| \le |\hat{\mu}_j - \mu_j| \equiv F_j(\hat{\boldsymbol{\mu}})$. Then, $\| \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2 \|_\Psi \le \sigma \sqrt{\frac{2n}{\nu_{\min}}}$ by Lemma A.1.

Next, we show that the packing numbers satisfy Equation (A.2). Let $\mathcal{F}(\hat{\boldsymbol{\mu}}) = \{(\hat{\mu}_j - \mu_j) g_j(\Gamma, \boldsymbol{\lambda}) : \Gamma \in [\Gamma_{\min}, \Gamma_{\max}],\ \|\boldsymbol{\lambda}\|_1 \le \lambda^{\mathcal{R}}_{\max}\}$.

Observe that, almost everywhere,

$$\left| \frac{\partial}{\partial \Gamma} g_j(\Gamma, \hat{\mu}_j, \boldsymbol{\lambda}) \right| = \left| \frac{1}{\Gamma \sqrt{\nu_{\min}}} (\mathbb{I}(\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} > \frac{\Gamma \sqrt{\nu_{\min}}}{\nu_j}) - g_j(\Gamma, \boldsymbol{\lambda})) \right| \le \frac{1}{\Gamma_{\min} \sqrt{\nu_{\min}}}$$

$$\|\nabla_{\boldsymbol{\lambda}} g_j(\Gamma, \hat{\mu}_j, \boldsymbol{\lambda})\|_\infty = \left\| \frac{\nu_j}{\Gamma \sqrt{\nu_{\min}}} \boldsymbol{A}_j \mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda} < \frac{\Gamma \sqrt{\nu_{\min}}}{\nu_j} \right) \right\|_\infty \le \frac{\nu_{\max}}{\Gamma_{\min} \sqrt{\nu_{\min}}} C_A.$$

Consequently,

$$\begin{aligned} \left| g_j(\Gamma_1, \boldsymbol{\lambda}^1) - g_j(\Gamma_2, \boldsymbol{\lambda}^2) \right| &\le \frac{1}{\Gamma_{\min} \sqrt{\nu_{\min}}} |\Gamma_1 - \Gamma_2| + \frac{\nu_{\max}}{\Gamma_{\min} \sqrt{\nu_{\min}}} C_A \|\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2\|_1 \\ &\le C_3 (|\Gamma_1 - \Gamma_2| + \|\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2\|_1), \end{aligned}$$

where $C_3 = \frac{1}{\Gamma_{\min} \sqrt{\nu_{\min}}} + \frac{\nu_{\max}}{\Gamma_{\min} \sqrt{\nu_{\min}}} C_A$. This further implies that

$$\begin{aligned} \left\| \left((\hat{\mu}_j - \mu_j)(g_j(\Gamma_1, \boldsymbol{\lambda}^1) - g_j(\Gamma_2, \boldsymbol{\lambda}^2)) \mid j = 1, \ldots, n \right) \right\|_2 &\le C_3 (|\Gamma_1 - \Gamma_2| + \|\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2\|_1) \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \\ &= C_3 (|\Gamma_1 - \Gamma_2| + \|\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2\|_1) \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2. \end{aligned}$$

Now, returning to the packing numbers, $M(\epsilon \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}(\mu)) \le N(\frac{\epsilon}{2} \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}(\mu))$. Consider an $\frac{\epsilon}{2C_3}$ covering with respect to the ℓ_1-norm of $[\Gamma_{\min}, \Gamma_{\max}] \times \{\|\boldsymbol{\lambda}\|_1 \le \lambda^{\mathcal{R}}_{\max}\}$. Then, from above, this yields a $\frac{\epsilon}{2} \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2$ covering of $\mathcal{F}$ as was desired. Since $[\Gamma_{\min}, \Gamma_{\max}] \times \{\|\boldsymbol{\lambda}\|_1 \le \lambda^{\mathcal{R}}_{\max}\}$ is contained within an ℓ_1 ball of radius $\Gamma_{\max} + \lambda^{\mathcal{R}}_{\max} \le 2\Gamma_{\max}$, we have from a standard result that the ℓ_1-norm covering of $[\Gamma_{\min}, \Gamma_{\max}] \times \{\|\boldsymbol{\lambda}\|_1 \le \lambda^{\mathcal{R}}_{\max}\}$ is at most $\left(\frac{3 \cdot 2 \cdot 2\Gamma_{\max} C_3}{\epsilon}\right)^{m+1}$, whereby

$$M(\epsilon \|\boldsymbol{F}(\hat{\boldsymbol{\mu}})\|_2, \mathcal{F}(\mu)) \le \left(\frac{C_4 \Gamma_{\max}}{\epsilon} \right)^{m+1},$$

where $C_4 = 12 C_3$. This proves that the packing numbers satisfy Equation (A.2) with $W = m + 1$ and $\log A = (m+1) \log(C_4 \Gamma_{\max})$, so that $V(A, W) \le 1 + 2\log(\Gamma_{\max}) + 2\log(C_4)$.

Applying Theorem A.2 and collecting dimension-independent constants yields the result. □

Instead of applying Lemma C.2 directly, we will use the explicit form of $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ from Lemma E.2 to avoid the finite differencing and develop a more computationally efficient bias correction.

Lemma E.8 (Approximating Stein's Lemma) *For each $j = 1, \ldots, n$,*

$$\begin{aligned} \mathbb{E}\left[(\hat{\mu}_j - \mu_j) g_j(\Gamma, \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)) \right] = \frac{1}{\Gamma \sqrt{\nu_{\min}}} \mathbb{P}\left\{ 0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma \sqrt{\nu_{\min}}}{\nu_j} \right\} \\ + \frac{1}{\sqrt{\nu_{\min}}} \|\phi - \phi_j\|_1 \left(\frac{1}{2\Gamma} + 24\sigma^2 - \log \|\phi - \phi_j\|_1 \right), \end{aligned}$$

so that the Error from Approximating Stein's Lemma in Theorem 5.2 is at most

$$\frac{2\mathsf{TV}}{\sqrt{\nu_{\min}}}\left(\frac{1}{2\Gamma_{\min}}+24\sigma^2-\log(2\mathsf{TV})\right).$$

Proof: First consider the special case where $\hat{\mu}_j$ is gaussian so that $\|\phi_j-\phi\|_1=0$. Then, it follows from Stein's Lemma that

$$\begin{aligned}\mathbb{E}[(\hat{\mu}_j-\mu_j)g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))] &= \frac{1}{\nu_j}\mathbb{E}\left[\frac{\partial}{\partial\mu_j}g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))\right]\\ &= \frac{1}{\Gamma\sqrt{\nu_{\min}}}\mathbb{P}\left(0\le\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right),\end{aligned}$$

where the last equality follows because $g_j(\Gamma,\boldsymbol{\lambda})=\frac{\nu_j}{\Gamma\sqrt{\nu_{\min}}}\left([\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}]^+-\left[\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}-\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right]^+\right)$, so

$$\frac{\partial}{\partial\mu_j}g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))=\frac{\nu_j}{\Gamma\sqrt{\nu_{\min}}}\mathbb{I}\left(0\le\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)\quad\text{a.e.}$$

Next consider the case that $\hat{\mu}_j$ is non-gaussian. Let $\overline{\mu}_j\sim\mathcal{N}(\mu_j,1/\nu_j)$. Similar to the proof of Lemma C.2, we use the sub-Gaussian tails of $\hat{\mu}_j$ to bound the difference in expectations when replacing $\hat{\mu}_j$ by $\overline{\mu}_j$. Specifically, by following the same steps in Lemma C.2, we have for any $T>0$,

$$\begin{aligned}\left|\mathbb{E}[(\hat{\mu}_j-\mu_j)g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))]-\mathbb{E}[(\overline{\mu}_j-\mu_j)g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))]\right| &\le \frac{1}{\sqrt{\nu_j}}\left[T\|g_j\|_\infty\|\phi_j-\phi\|_1+4\|g_j\|_\infty e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi})\right]\\ &\le \frac{1}{\sqrt{\nu_{\min}}}\left[T\|\phi_j-\phi\|_1+4e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi})\right].\end{aligned}$$

On the other hand,

$$\begin{aligned}&\left|\frac{1}{\Gamma\sqrt{\nu_{\min}}}\mathbb{P}\left(0\le\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)-\frac{1}{\Gamma\sqrt{\nu_{\min}}}\mathbb{P}\left(0\le\overline{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)\right|\\ &\quad\le\frac{1}{2\Gamma\sqrt{\nu_{\min}}}\|\phi_j-\phi\|_1.\end{aligned}$$

Adding these two inequalities and applying the triangle inequality shows

$$\begin{aligned}&\left|\mathbb{E}[(\hat{\mu}_j-\mu_j)g_j(\Gamma,\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma))]-\frac{1}{\Gamma\sqrt{\nu_{\min}}}\mathbb{P}\left(0\le\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)\right|\\ &\quad\le\frac{1}{\sqrt{\nu_{\min}}}\left[\left(T+\frac{1}{2\Gamma}\right)\|\phi_j-\phi\|_1+4e^{-\frac{T^2}{2\sigma^2}}(T+\sigma\sqrt{2\pi})\right].\end{aligned}$$

We can now follow an identical argument to Lemma C.2 to upperbound this function and then optimize T by letting $\frac{1}{2\Gamma}$ play the role of h^{-1}. This argument yields the first expression in the theorem. The second follows from Jensen's inequality and the definition of TV. □

Lemma E.9 (ULLN for Bias Approximation) *There exist positive, dimension-independent constants C_1,C_2, such that*

$$\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]}\left|\frac{1}{\Gamma\sqrt{\nu_{\min}}n}\sum_{j=1}^n\mathbb{P}\left\{0\le\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right\}-\mathbb{I}\left(0\le\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)\right|,$$

where $\mathbb{P}\{R>\epsilon\}\le C_1\exp\left(-\frac{C_2\epsilon\sqrt{n}}{\log(m+1)}\right)$.

Proof: We apply Theorem A.2. Take 1 as an envelope. To bound the packing numbers, note that $\mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right) = \min\left(\mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\right), \mathbb{I}\left(\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)\right)$, so it suffices to compute the pseudo-dimension of the two families:

$$\left\{\left(\mathbb{I}\left(\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right) \mid j = 1, \ldots, n\right) \ : \ \Gamma \in [\Gamma_{\min}, \Gamma_{\max}], \|\boldsymbol{\lambda}\|_1 \le \lambda^{\mathcal{R}}_{\max}\right\},$$
$$\left\{\left(\mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\right) \mid j = 1, \ldots, n\right) \ : \ \Gamma \in [\Gamma_{\min}, \Gamma_{\max}], \|\boldsymbol{\lambda}\|_1 \le \lambda^{\mathcal{R}}_{\max}\right\}.$$

To bound the first, note that $\left\{\left(\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j} \mid j = 1, \ldots, n\right) : \Gamma \in \mathbb{R}, \boldsymbol{\lambda} \in \mathbb{R}^m\right\}$ is contained within an $m + 2$ dimensional vector subspace, and, thus, by Lemma 4.4 in Pollard (1990), the pseudo-dimension of the first set is at most $m + 2$. A similar computation holds for the second set. It follows that the pseudo-dimension of $\left\{\left(\mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right) \mid j = 1, \ldots, n\right) : \Gamma \in [\Gamma_{\min}, \Gamma_{\max}], \|\boldsymbol{\lambda}\|_1 \le \lambda^{\mathcal{R}}_{\max}\right\}$ is at most $10(m + 2)$, so that Equation (A.2) is satisfied with $V(A, W) \le 1 + 3\log(10(m + 2)) \le C_3 \log(m + 1)$. Collecting dimension-independent constants yields the result. □

Lemma E.10 (Approximating Dual in Bias) *There exist positive, dimension-independent constants C_1, C_2, C_3 such that*

$$\sup_{\Gamma \in [\Gamma_{\min}, \Gamma_{\max}]} \left| \frac{1}{\Gamma\sqrt{\nu_{\min}} n} \sum_{j=1}^{n} \mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right) - \mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right) \right|$$

is at most $C_1\delta + R$, where $\mathbb{P}\{R > \epsilon\} \le C_2 \exp\left(\frac{-C_3\epsilon\sqrt{n}}{\log(m+1)}\right)$.

Proof: Restrict attention to paths where $\mathcal{E}$ holds. The only non-zero terms are where the indicators differ. We have 4 cases corresponding to which indicator is positive and which inequality is violated in the other indicator:

- **Case 1:** $0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}$ and $\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) < 0$.
- **Case 2:** $0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}$ and $\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) > \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}$.
- **Case 3:** $0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}$ and $\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) < 0$.
- **Case 4:** $0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}$ and $\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) > \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}$.

Consider Case 1. Note $\hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) < 0 \implies \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le C_A\delta$. Thus, $\hat{\mu}_j$ belongs to an interval of length $C_A\delta$, namely, $0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le C_A\delta$. In each of the remaining cases we can also argue that in any non-zero term, $\hat{\mu}_j$ belongs to an interval of length $C_A\delta$. Combining the four cases and "distributing" the supremum, we bound the supremum in the theorem by

$$\sup_{\Gamma \in [\Gamma_{\min}, \Gamma_{\max}]} \frac{1}{\Gamma_{\min}\sqrt{\nu_{\min}} n} \sum_{j=1}^{n} \mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le C_A\delta\right) \qquad \text{(Case 1)}$$

$$+ \sup_{\Gamma \in [\Gamma_{\min}, \Gamma_{\max}]} \frac{1}{\Gamma_{\min}\sqrt{\nu_{\min}} n} \sum_{j=1}^{n} \mathbb{I}\left(\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j} - C_A\delta \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right) \qquad \text{(Case 2)}$$

$$+ \sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \frac{1}{\Gamma_{\min}\sqrt{\nu_{\min}}n} \sum_{j=1}^{n} \mathbb{I}\left(-C_A\delta \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le 0\right) \qquad \text{(Case 3)}$$

$$+ \sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \frac{1}{\Gamma_{\min}\sqrt{\nu_{\min}}n} \sum_{j=1}^{n} \mathbb{I}\left(\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j} \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le \frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j} + \delta\right). \qquad \text{(Case 4)}$$

We will bound the contributions from each case separately. For the first case, break the contribution into two supremums

$$\begin{aligned}
&\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \frac{1}{\Gamma_{\min}\sqrt{\nu_{\min}}n} \sum_{j=1}^{n} \mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le C_A\delta\right) \\
&\qquad \le \sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \frac{1}{\Gamma_{\min}\sqrt{\nu_{\min}}n} \sum_{j=1}^{n} \mathbb{P}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le C_A\delta\right) \\
&\qquad + \sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \left\{ \left| \frac{1}{\Gamma_{\min}\sqrt{\nu_{\min}}n} \sum_{j=1}^{n} \mathbb{I}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le C_A\delta\right) \right.\right. \\
&\qquad\qquad \left.\left. - \mathbb{P}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le C_A\delta\right) \right| \right\}.
\end{aligned}$$

Define the standardized increment $\zeta_j \equiv \sqrt{\nu_j}(\hat{\mu}_j - \mu_j)$. Then

$$\mathbb{P}\left(0 \le \hat{\mu}_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma) \le C_A\delta\right) = \mathbb{P}\left(0 \le \zeta_j + \sqrt{\nu_j}\left(\mu_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^{\mathcal{R}}(\Gamma)\right) \le C_A\sqrt{\nu_j}\delta\right).$$

By Assumption 4.2, this last probability is at most $\phi_{max}C_A\sqrt{\nu_{\max}}\delta$. Thus, the first supremum in the contribution from Case 1 is at most $\frac{\phi_{max}C_A\sqrt{\nu_{\max}}}{\Gamma_{\min}\sqrt{\nu_{\min}}}\delta$.

We bound the second supremum in the contribution from Case 1 by applying Theorem A.2. The envelopes are $\frac{1}{\Gamma_{\min}\sqrt{\nu_{\min}}}$. The packing numbers can be computed entirely analogously to Lemma E.9, which shows that $V(A,W) \le C_4\log(m+1)$ for some dimension-independent C_4.

Combining both supremums proves that the contributions from Case 1 are bounded by $\frac{\phi_{max}C_A\sqrt{\nu_{\max}}}{\Gamma_{\min}\sqrt{\nu_{\min}}}\delta + R_1$, where $\mathbb{P}\{R_1 > \epsilon\} \le C_5\exp\left(-\frac{C_6\epsilon\sqrt{n}}{\log(m+1)}\right)$ for dimension independent constants C_5, C_6.

The contributions from the three remaining cases are similar. Combining all of them and collecting dimension-independent constants proves the theorem. □

E.4. Proofs of Lemma E.1, Theorem 5.1, and Corollary 5.3

Proof of Lemma E.1: When $\Gamma = r = 0$, the problems defining $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ and $\boldsymbol{x}\left(\mathcal{U}(\hat{\boldsymbol{\mu}}, r)\right)$ are identical, and hence their solutions must coincide. Thus, assume $\Gamma > 0$ and $r > 0$. Then, since objective functions of the optimization problems defining $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ and $\boldsymbol{x}\left(\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r)\right)$ are both strictly convex, both optimizers are unique.

First consider the optimization defining $\boldsymbol{x}\left(\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r)\right)$. For any fixed $\boldsymbol{x}$, the inner optimization can be solved in closed-form, yielding $\min_{\boldsymbol{\mu}\in\mathcal{U}_E(\hat{\boldsymbol{\mu}},r)} \frac{1}{n}\boldsymbol{\mu}^T\boldsymbol{x} = \hat{\boldsymbol{\mu}}^T\boldsymbol{x} - \frac{r}{n}\sqrt{\sum_{j=1}^n x_j^2/\nu_j}$. Thus,

$$\boldsymbol{x}\left(\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r)\right) = \arg\max_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{n}\hat{\boldsymbol{\mu}}^T\boldsymbol{x} - \frac{r}{n}\sqrt{\sum_{j=1}^n x_j^2/\nu_j}.$$

Now, the first-order optimality conditions for this optimization problem and the problem defining $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ are, respectively,

$$\left(\hat{\boldsymbol{\mu}} - \frac{r}{\sqrt{\sum_{j=1}^n x_j \left(\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r)\right)^2/\nu_j}}\boldsymbol{V}^{-1}\boldsymbol{x}\left(\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r)\right)\right)^\top \left(\boldsymbol{x}\left(\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r)\right) - \boldsymbol{x}\right) \geq 0, \quad \forall \boldsymbol{x}\in\mathcal{X},$$

$$\left(\hat{\boldsymbol{\mu}} - \Gamma\sqrt{\nu_{\min}}\ \boldsymbol{V}^{-1}\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})\right)^\top \left(\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}}) - \boldsymbol{x}\right) \geq 0, \quad \forall \boldsymbol{x}\in\mathcal{X},$$

where $\boldsymbol{V} = \operatorname{diag}(\nu_1, \ldots, \nu_n)$. One can now verify directly that given any Γ, $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ satisfies the optimality conditions for the robust problem for the parameter $r = \Gamma\sqrt{\nu_{\min}}\sqrt{\sum_{j=1}^n x_j^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})^2/\nu_j}$. Similarly, given any r, $\boldsymbol{x}\left(\mathcal{U}_E(r, \hat{\boldsymbol{\mu}})\right)$ satisfies the optimality conditions for the regularized problem with $\Gamma = \frac{r}{\sqrt{\nu_{\min}\sum_{j=1}^n x_j \left(\mathcal{U}_E(\hat{\boldsymbol{\mu}}, r)\right)^2/\nu_j}}$. □

Proof of Theorem 5.1. We construct $\hat{\boldsymbol{\mu}}^n$ as in Example 2.2. Specifically, suppose suppose $\xi_j^l \sim \mathcal{N}(\mu_j, 1/\nu_0)$ for all $j = 1, \ldots, n$, $l = 1, \ldots, \mathsf{S}$. Then, by construction

$$\hat{\mu}_j \sim \mathcal{N}\left(\mu_j, \frac{1}{\mathsf{S}\nu_0}\right), \quad \overline{\mu}_j^{-k} \sim \mathcal{N}\left(\mu_j, \frac{1}{\mathsf{S}\nu_0}\frac{K}{K-1}\right), \quad \overline{\mu}_j^{k} \sim \mathcal{N}\left(\mu_j, \frac{1}{\mathsf{S}\nu_0}K\right). \tag{E.5}$$

To complete the instances, take $\mathcal{X}^n = [0,1]^n$, $\nu_0 = .114$, $\Gamma_{\mathsf{min}} = 10^{-6}$, $\Gamma_{\mathsf{max}} = 100$ and

$$\mu_j = \begin{cases} 0.0408 & \text{if } j \text{ is odd} \\ -1.96 & \text{if } j \text{ is even.} \end{cases}$$

The optimization defining $\boldsymbol{x}^{\mathsf{K},n}(\Gamma, \overline{\mu}^{-k})$ decouples across j yielding,

$$x_j^{\mathsf{K},n}(\Gamma, \overline{\mu}^{-k}) = \min\left(1, \max\left(0, \frac{\sqrt{\mathsf{S}\nu_0}}{\Gamma}\sqrt{\frac{K-1}{K}}\overline{\mu}_j^{-k}\right)\right),$$

so that $\Gamma^{\mathsf{K-fold},n}$ solves

$$\Gamma^{\mathsf{K-fold},n} \in \arg\max_{\Gamma\in[\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}]} \frac{1}{K}\sum_{k=1}^K \frac{1}{n}\sum_{j=1}^n \overline{\mu}_j^k \min\left(1, \max\left(0, \frac{\sqrt{\mathsf{S}\nu_0}}{\Gamma}\sqrt{\frac{K-1}{K}}\overline{\mu}_j^{-k}\right)\right). \tag{E.6}$$

Consider the k^{th} element of the outer-summand. This is an average of n, independent terms (indexed by j). Each of these terms is continuous in Γ, and the j^{th} term upperbounded by $\left|\overline{\mu}_j^k\right|$, which is integrable. Hence, by (Van der Vaart 2000, Ex. 19.8), the k^{th} element of the outer-summand converges almost surely to its expectation, uniformly over $\Gamma \in [\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}]$ as $n \to \infty$. Since K is

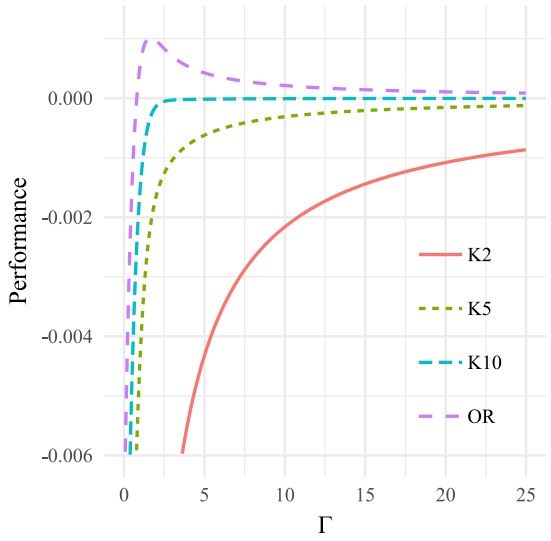


Figure EC.1 **Limiting Functions in Theorem 5.1.** "K2", "K5", "K10", "OR" refer the limit of the hold-out, 5-fold, LOO and Oracle curves as $n \to \infty$, respectively. Each curve has a unique minimizer on the region. Notice the cross-validation curves are poor approximations to the oracle curve, and the oracle curve is not very flat at its optimum, which causes the performance of $\Gamma^{\mathsf{K-fold},n}$ to differ from $\Gamma^{\mathsf{OR},n}$ as $n \to \infty$.

fixed, it follows that the overall objective also converges almost surely to its expectation. This motivates defining a limiting optimization and optimizer. For each j, let $Z_j \sim \mathcal{N}(\sqrt{\mathsf{S}\nu_0}\sqrt{\frac{K-1}{K}}\mu_j, 1)$. Then, define

$$\begin{aligned}
\Gamma^{\mathsf{K-fold},\infty} &\in \arg\max_{\Gamma\in[\Gamma_{\mathsf{min}},\Gamma_{\mathsf{max}}]} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\overline{\mu}_j^k \min\left(1, \max\left(0, \frac{\sqrt{\mathsf{S}\nu_0}}{\Gamma}\sqrt{\frac{K-1}{K}}\overline{\mu}_j^{-k}\right)\right)\right] \\
&\in \arg\max_{\Gamma\in[\Gamma_{\mathsf{min}},\Gamma_{\mathsf{max}}]} \frac{1}{n}\sum_{j=1}^{n}\mu_j\mathbb{E}\left[\min\left(1, \frac{1}{\Gamma}Z_j^+\right)\right] \\
&\in \arg\max_{\Gamma\in[\Gamma_{\mathsf{min}},\Gamma_{\mathsf{max}}]} \frac{1}{2}\mu_1\mathbb{E}\left[\min\left(1, \frac{1}{\Gamma}Z_1^+\right)\right] + \frac{1}{2}\mu_2\mathbb{E}\left[\min\left(1, \frac{1}{\Gamma}Z_2^+\right)\right],
\end{aligned}$$

where the first equality follows by the definition of Z_j and Eq. (E.5), and the second equality uses the odd-even structure of the μ_j to simplify.

One can confirm numerically that for $K \in \{2,5,10\}$ and the parameters given earlier, the above optimization has a unique minimizer $\Gamma^{\mathsf{K-fold},\infty} = \Gamma_{\mathsf{max}}$ for $K \in \{2,5,10\}$. See also Fig. EC.1. Since the optimization objective defining $\Gamma^{\mathsf{K-fold},n}$ converges uniformly to the optimization objective defining $\Gamma^{\mathsf{K-fold},\infty}$ and the latter has a unique minimizer, it follows that $\Gamma^{\mathsf{K-fold},n} \to \Gamma^{\mathsf{K-fold},\infty}$ as $n \to \infty$.

An entirely similar argument shows that

$$\Gamma^{\mathsf{OR},n} \in \arg\max_{\Gamma\in[\Gamma_{\mathsf{min}},\Gamma_{\mathsf{max}}]} \frac{1}{n}\sum_{j=1}^{n}\mu_j \min\left(1, \max\left(0, \frac{\sqrt{\mathsf{S}\nu_0}}{\Gamma}\hat{\mu}_j\right)\right).$$

We define

$$\Gamma^{\mathsf{OR},\infty} \in \arg\max_{\Gamma\in[\Gamma_{\mathsf{min}},\Gamma_{\mathsf{max}}]} \frac{1}{2}\mu_1\mathbb{E}\left[\min\left(1, \frac{1}{\Gamma}W_1^+\right)\right] + \frac{1}{2}\mu_2\mathbb{E}\left[\min\left(1, \frac{1}{\Gamma}W_2^+\right)\right],$$

where $W_j \sim \mathcal{N}(\sqrt{\mathsf{S}\nu_0}\mu_j, 1)$. Arguing identically to the above, $\Gamma^{\mathsf{OR},\infty}$ is a unique minimizer, the objective of the first optimization converges uniformly to the second, and hence, $\Gamma^{\mathsf{OR},n} \to \Gamma^{\mathsf{OR},\infty}$. Numerically, $\Gamma^{\mathsf{OR},\infty} \approx 1.64$.

To complete the proof, we note that by the above uniform convergence

$$\begin{aligned}\lim_{n\to\infty} \frac{n^{-1}\boldsymbol{\mu}^\top \boldsymbol{x}^{\mathcal{R}}(\hat{\boldsymbol{\mu}}, \Gamma^{\mathsf{OR},n})}{n^{-1}\boldsymbol{\mu}^\top \boldsymbol{x}^{\mathcal{R}}(\hat{\boldsymbol{\mu}}, \Gamma^{\mathsf{K-fold},n})} &= \frac{\mu_1 \mathbb{E}\left[\min\left(1, \frac{1}{\Gamma^{\mathsf{OR},\infty}} W_1^+\right)\right] + \mu_2 \mathbb{E}\left[\min\left(1, \frac{1}{\Gamma^{\mathsf{OR},\infty}} W_2^+\right)\right]}{\mu_1 \mathbb{E}\left[\min\left(1, \frac{1}{\Gamma_{\mathsf{max}}} W_1^+\right)\right] + \mu_2 \mathbb{E}\left[\min\left(1, \frac{1}{\Gamma_{\mathsf{max}}} W_2^+\right)\right]}, \\ &< .03,\end{aligned}$$

where the last line follows from numerical integration. For this example, observe that LOO validation correspond to $K = 10$.

Finally, Hold-out validation can be analyzed similarly to the case $K = 2$. The key observation is that each of the summands in the outer summation (over k) in Eq. (E.6) converges uniformly to its expectation. Hence, in particular, when $K = 2$, the summand corresponding to $k = 1$ converges to its expectation, which is again the objective of the optimization defining $\Gamma^{\mathsf{K-fold},\infty}$. The remainder of the proof follows the case $K = 2$.

This completes the proof. □

Proof of Corollary 5.3: Take $\delta_n \to 0$ such that $\delta_n\sqrt{n} \to \infty$. Then, both the deterministic error and the tail probability in the theorem tend to zero. In other words, the suboptimality of our policy tends to zero in the small-data, large-scale regime if the $\hat{\mu}_j$ are gaussian. Indeed, for this choice of δ_n, the tail probability is summable, and hence, by the Borel-Cantelli lemma, the convergence occurs almost surely as $n \to \infty$. □

E.5. Performance in the Large-Sample Regime

Regularization in the large-sample regime has been well-studied for a variety of statistical problems, and, in particular, it is well known that if $\Gamma^{\mathsf{S}} \to 0$ at an appropriate rate as $\mathsf{S} \to \infty$, regularized methods converge in performance to the full-information optmimum (Negahban et al. 2012). We next argue that despite the restriction of $\hat{\Gamma}$ to the interval $[\Gamma_{\mathsf{min}}, \Gamma_{\mathsf{max}}]$, $\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma}, \hat{\boldsymbol{\mu}})$ still converges in performance to the full-information optimum in the large-sample limit.

Theorem E.11 (Bound to Full-Information Optimum) *Let* $\hat{\Gamma}$ *be a solution to Equation* (5.3)*. Then, under Assumption 2.1,*

$$\mathbb{E}\left[\frac{1}{n}\left|\boldsymbol{\mu}^T(\boldsymbol{x}_n^*(\boldsymbol{\mu}) - \boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma}, \hat{\boldsymbol{\mu}}))\right|\right] \leq \frac{2}{\sqrt{\nu}_{\mathsf{min}}}\left(\Gamma_{\mathsf{max}} + \frac{1}{\Gamma_{\mathsf{min}}} + \frac{1}{2}\right).$$

Proof of Theorem E.11: Define $\Gamma^{SAA} \in \arg\max_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]} \frac{1}{n}\hat{\boldsymbol{\mu}}^\top \boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma},\hat{\boldsymbol{\mu}})$. By definition, $B_n^{Reg}(\Gamma,\hat{\mu})$ is uniformly bounded over Γ:

$$\sup_{\Gamma\in[\Gamma_{\min},\Gamma_{\max}]}\left|\frac{1}{\Gamma n\sqrt{\nu_{\min}}}\sum_{j=1}^{n}\mathbb{I}\left(0\le\hat{\mu}_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}^{\mathcal{R}}(\Gamma,\hat{\boldsymbol{\mu}})\le\frac{\Gamma\sqrt{\nu_{\min}}}{\nu_j}\right)\right|\le\frac{1}{\Gamma_{\min}\sqrt{\nu_{\min}}}\ .$$

Therefore, the objective of Equation (5.3) is a uniform approximation to the objective defining Γ^{SAA}, and by Lemma C.1, we have

$$0\le\frac{1}{n}\hat{\boldsymbol{\mu}}^\top\left(\boldsymbol{x}^{\mathcal{R}}(\Gamma^{SAA},\hat{\boldsymbol{\mu}})-\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma},\hat{\boldsymbol{\mu}})\right)\le\frac{2}{\Gamma_{\min}\sqrt{\nu_{\min}}}.$$

Note that we can equivalently write

$$\boldsymbol{x}^{\mathcal{R}}(\Gamma^{SAA},\hat{\boldsymbol{\mu}})\ \in\ \arg\max_{\boldsymbol{x}\in\mathcal{X}}\frac{1}{n}\hat{\boldsymbol{\mu}}^\top\boldsymbol{x}-\frac{\Gamma^{SAA}\sqrt{\nu_{\min}}}{n}\sum_{j=1}^{n}x_j^2\nu_j. \tag{E.7}$$

Moreover, for any $\boldsymbol{x}\in\mathcal{X}$,

$$\frac{\Gamma^{SAA}\sqrt{\nu_{\min}}}{n}\sum_{j=1}^{n}\frac{x_j^2}{\nu_j}\ \le\ \frac{\Gamma_{\max}}{n\sqrt{\nu_{\min}}}\|\boldsymbol{x}\|_2^2\ \le\ \frac{\Gamma_{\max}}{\sqrt{\nu_{\min}}},$$

where the last inequality follows because $\mathcal{X}\subseteq[0,1]^n$. Thus, the objective in (E.7) is a uniform approximation to the objective of the SAA problem, i.e., $\frac{1}{n}\hat{\boldsymbol{\mu}}^T\boldsymbol{x}$, whereby Lemma C.1 implies

$$0\le\frac{1}{n}\hat{\boldsymbol{\mu}}^\top\left(\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})-\boldsymbol{x}^{\mathcal{R}}(\Gamma^{SAA},\hat{\boldsymbol{\mu}})\right)\le\frac{2\Gamma_{\max}}{\sqrt{\nu}_{\min}}.$$

Combining, we get

$$\begin{aligned}
0\ &\le\ \boldsymbol{\mu}^\top\left(\boldsymbol{x}^*(\boldsymbol{\mu})-\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma},\hat{\boldsymbol{\mu}})\right)\\
&=\ (\boldsymbol{\mu}-\hat{\boldsymbol{\mu}})^\top\boldsymbol{x}^*(\boldsymbol{\mu})+\hat{\boldsymbol{\mu}}^\top\left(\boldsymbol{x}^*(\boldsymbol{\mu})-\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})\right)+\hat{\boldsymbol{\mu}}^\top\left(\boldsymbol{x}^{SAA}(\hat{\boldsymbol{\mu}})-\boldsymbol{x}^{\mathcal{R}}(\Gamma^{SAA},\hat{\boldsymbol{\mu}})\right)\\
&\quad+\hat{\boldsymbol{\mu}}^\top\left(\boldsymbol{x}^{\mathcal{R}}(\Gamma^{SAA},\hat{\boldsymbol{\mu}})-\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma},\hat{\boldsymbol{\mu}})\right)+(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma},\hat{\boldsymbol{\mu}})\\
&\le\ 2\sup_{\boldsymbol{x}\in\mathcal{X}}\left|(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})^\top\boldsymbol{x}\right|+\frac{2n\Gamma_{\max}}{\sqrt{\nu}_{\min}}+\frac{2n}{\Gamma_{\min}\sqrt{\nu_{\min}}}\ \le\ 2\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_1+\frac{2n}{\sqrt{\nu}_{\min}}\left(\Gamma_{\max}+\frac{1}{\Gamma_{\min}}\right),
\end{aligned}$$

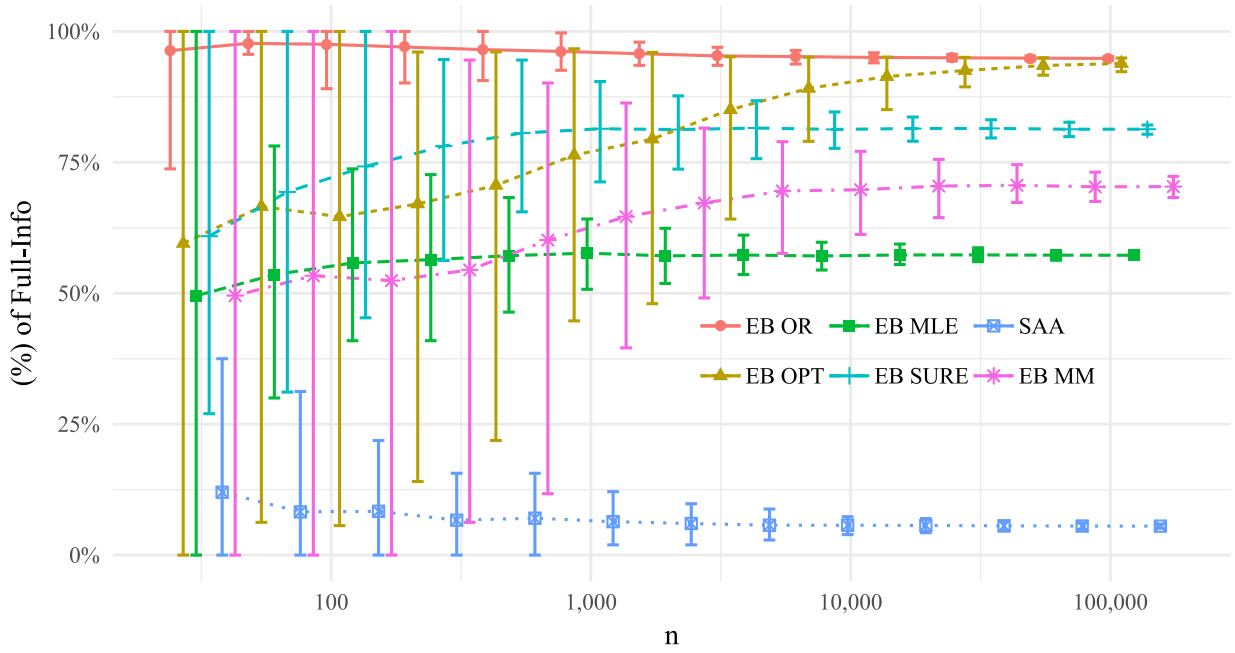
where the last inequality follows because $\mathcal{X}\in[0,1]^n$. Dividing by n and taking expectations yields

$$\mathbb{E}\left[\left|\frac{1}{n}\boldsymbol{\mu}^\top\left(\boldsymbol{x}^*(\boldsymbol{\mu})-\boldsymbol{x}(\hat{\Gamma},\hat{\boldsymbol{\mu}})\right)\right|\right]\le\frac{2}{n}\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_1\right]+\frac{2}{\sqrt{\nu}_{\min}}\left(\Gamma_{\max}+\frac{1}{\Gamma_{\min}}\right).$$

Finally,

$$\mathbb{E}\left[\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|_1\right]=\sum_{j=1}^{n}\mathbb{E}[|\hat{\mu}_j-\mu_j|]=\sum_{j=1}^{n}\mathbb{E}\left[\sqrt{(\hat{\mu}_j-\mu_j)^2}\right]\le\sum_{j=1}^{n}\sqrt{\mathbb{E}\left[(\hat{\mu}_j-\mu_j)^2\right]}\le\sum_{j=1}^{n}\frac{1}{\sqrt{\nu_j}}\le\frac{n}{\sqrt{\nu}_{\min}},$$

where the inequality follows from Jensen's inequality. Substituting above proves the theorem. □



(a) Empirical Bayes-Inspired Policies

Figure EC.2 **Performance of various data-driven procedures for Example 4.1, varying n.** The error bars represent 10% and 90% quantiles over 200 simulations. See Section 6 for a description of the methods.

Appendix F: Additional Figures and Computational Details

F.1. Additional Figures for Example 4.1.

Figure EC.2a shows the performance of all data-driven methods from the Bayes-Inspired policy for Example 4.1.

F.2. Simulating Advertising Portfolio Optimization Instances from Section 6 and Computational Details.

We interpret $\boldsymbol{\mu}$ as the expected number of clicks per targeting item, i.e., we assume that the revenue-per-click is constant. The simulation procedure from Pani et al. (2017) can then be summarized as follows: The cost and expected number of clicks for item j are, respectively,

$$c_j = 20\beta_1^j, \quad \mu_j = \beta_0^j + \beta_1^j \log\left(c_j + \exp(-\beta_0^j/\beta_1^j)\right),$$

where β_0^j, β_1^j are independent across j and their marginal distributions are $\beta_0 \sim \text{Cauchy}(7.96, 2.21)$ and $\beta_1 \sim \text{Log}(2.21, 1.43)$. The values β_0, β_1 are *dependent*, with a Gumbel copula with parameter 2.[14] Finally, any product for which $\beta_0 \notin [-700, 100]$, $\beta_1 \notin [.5, 800]$, or $\exp(\beta_0/\beta 1) > 20$ are discarded

[14] Precise parameter values are not available in the published manuscript of Pani et al. (2017) but were obtained via personal communication (email) with the authors on 8 Sept. 2017.

in the simulation, as they do not correspond well to real targeting items. Intuitively, the logarithmic dependence between μ_j and c_j reflects the typical dependence observed in real-world targeting items, while the value of c_j represents a typical, viable bid level for the item. Finally, since scaling $\boldsymbol{\mu}$ and $\boldsymbol{c}$ by a constant does not affect the relative performance of the methods, we scale both by $1/200$ for simplicity.

Pani et al. (2017) only consider the case of known rewards. Thus, we supplement the above with a procedure for generating ν_j, $\hat{\boldsymbol{\mu}}$. Intuitively, we would like to capture the phenomenon if items with a very large or very small expected-click to cost ratio probably being more rare, and, hence, more likely to have less data associated with them. They would then have lower precision estimates. To this end, let $F^{-1}_{\mu/c}(\mu_j/c_j)$ be the empirical quantile of the ratio μ_j/c_j among the n items. We take

$$\nu_j = \begin{cases} .1 & \text{if } F^{-1}_{\mu/c}(\mu_j/c_j) \leq .33 \\ 10 & \text{if } .33 < F^{-1}_{\mu/c}(\mu_j/c_j) < .66 \\ 8 & \text{o.w.} \end{cases}$$

Thus, the worst third of items have the least precision, the best third of items have medium precision, and the middle third of items have high precision. The values .1 and 10 were chosen to ensure that items with a low ratio had a reasonable probability of being mistaken for a high-ratio item. Figure EC.3 provides some graphs and summary statistics for the simulated instances.

Notice this procedure only specifies the estimates $\hat{\mu}_j$, it does not specify the "raw data" that generated these estimates. In many ways we feel this set-up well-mirrors this application; often the estimates are the outputs of sophisticated machine learning algorithms performed by a third-party (e.g., Adobe, or Google), and the raw data is not available to the decision-maker (the advertiser) at the time of targeting. Nonetheless, in order to compare our methods to K-fold cross-validation schemes, we need to also simulate "raw" data. We adopt the perspective in Example 2.2, and assume $\hat{\mu}_j = \frac{1}{\mathsf{S}} \sum_{k=1}^{\mathsf{S}} \xi_j^k$ for all j. Unless otherwise specified, we take $\mathsf{S} = 10$ and $\xi_j^k \sim \mathcal{N}(\mu_j, S/\nu_j)$. Specific experiments relax/alter these data generation assumptions (cf. Section F.4 and F.5).

F.3. Additional Computational Results from Section 6.1.

The two panels of Figure EC.4 show the performance of all data-driven methods from the Bayes-Inspired policy class and Regularization-Inspired policy class, respectively, for our online-advertising portfolio case study in Section 6.1.

Figure EC.5 examines the convergence of the optimizing Γ for various cross-validation procedures from Section 6.1.

Like our Regularization-Inspired class, cross-validation procedures for the Bayes-Inspired class perform quite well for our OAPOP instances because the oracle curve is quite flat at its optimum. Figure EC.6 below shows the convergence of the optimizing τ's for various cross-validation procedures for the instances in Section 6.1. Notice that the estimated cross validation and bias-corrected

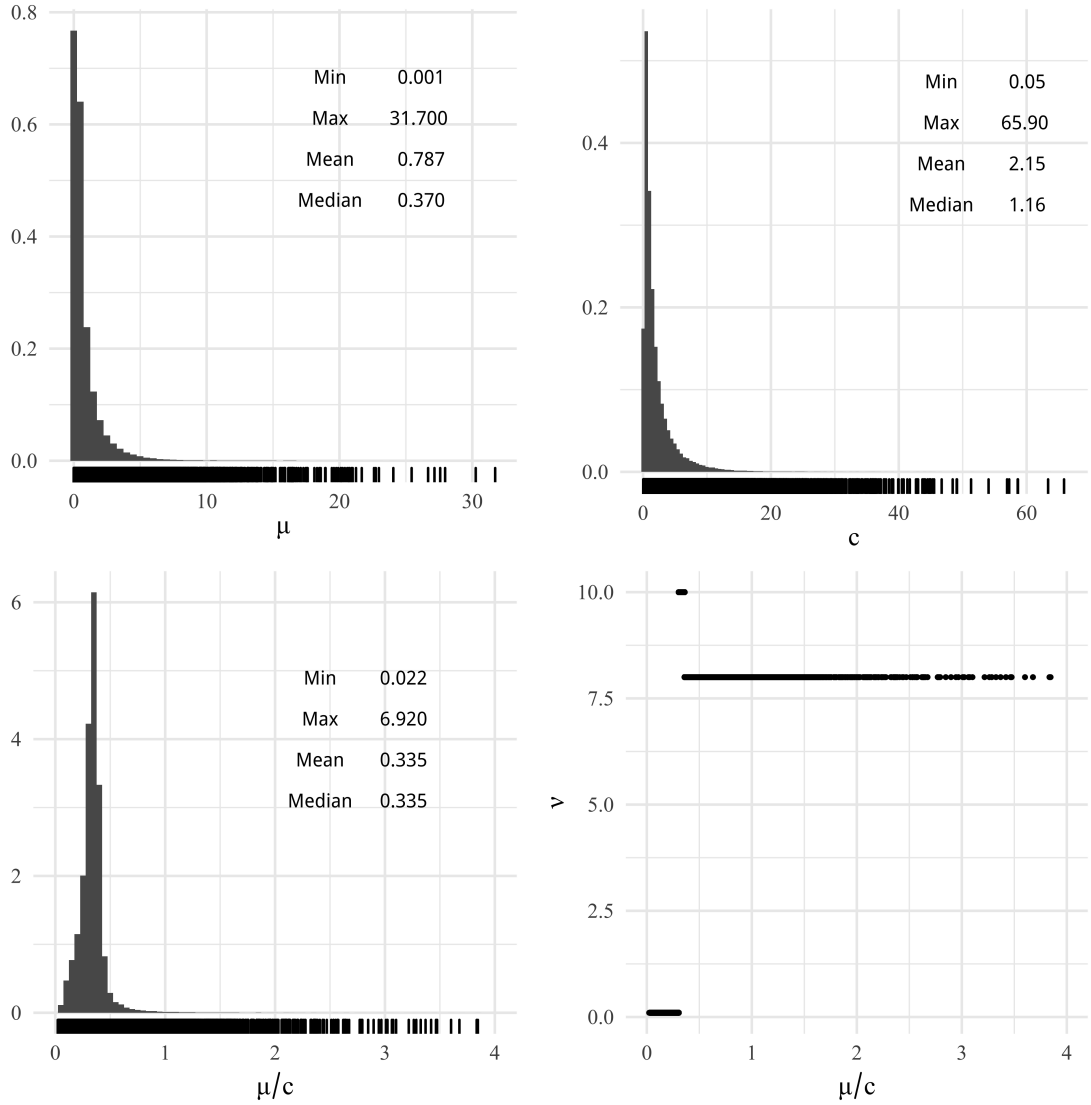
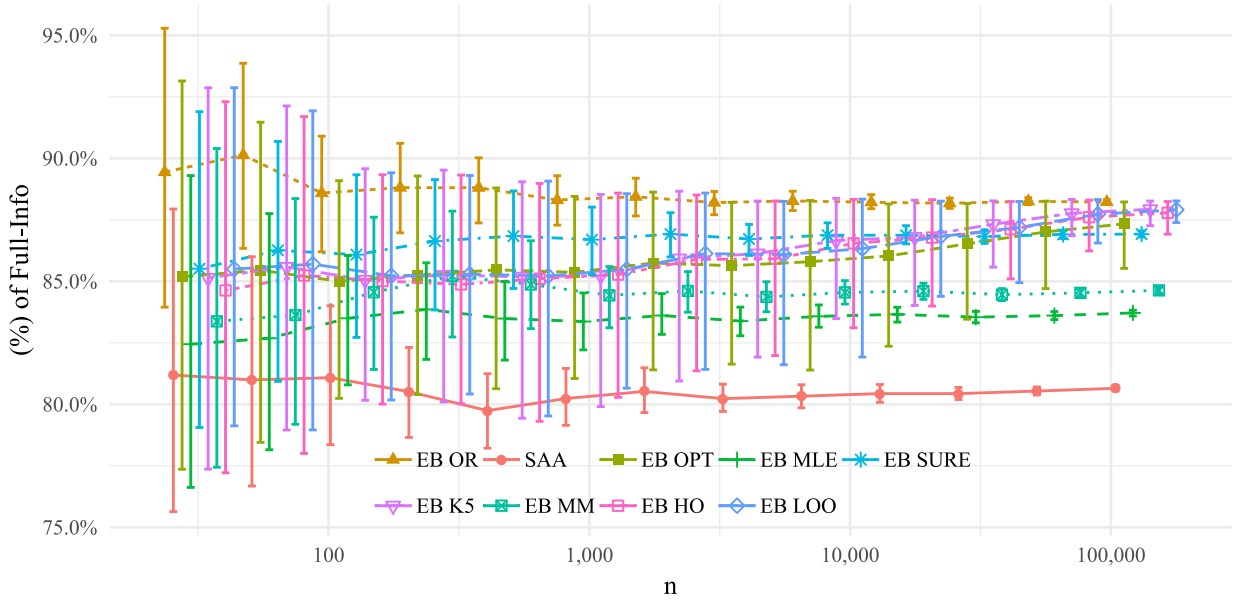


Figure EC.3 Summary Statistics of Simulated OPAOP Instances from Section 6. The histograms for $\boldsymbol{\mu}$, $\boldsymbol{c}$ and $\boldsymbol{\mu}/\boldsymbol{c}$ include rug plots on the x-axis to highlight the very long tails.

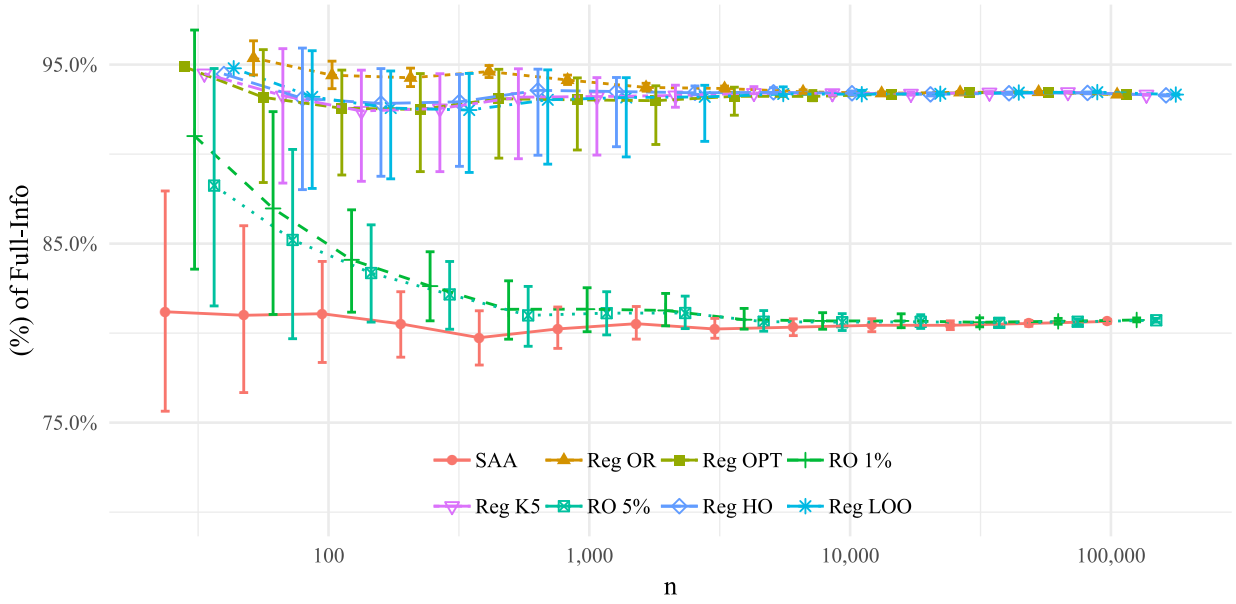
curves, themselves, are already quite variable (left panel of Figure EC.6). The result is that even for large n, e.g., $n = 2^{17}$, the distribution of the optimizing τ's is quite disperse (see left panel of Figure EC.7). Nonetheless, because the oracle curve is quite flat at its optimum, these varying τ's all yield strong performance (see right panel of Figure EC.7).

F.4. Performance in Large-Sample Regime (finite S)

In Section 6.1, we considered the case $\mathsf{S} = 10$ and $n \to \infty$. In this section we fix $n = 2^{17}$, and consider the performance of our methods as S increases. We consider a set-up similar to, but distinct from,

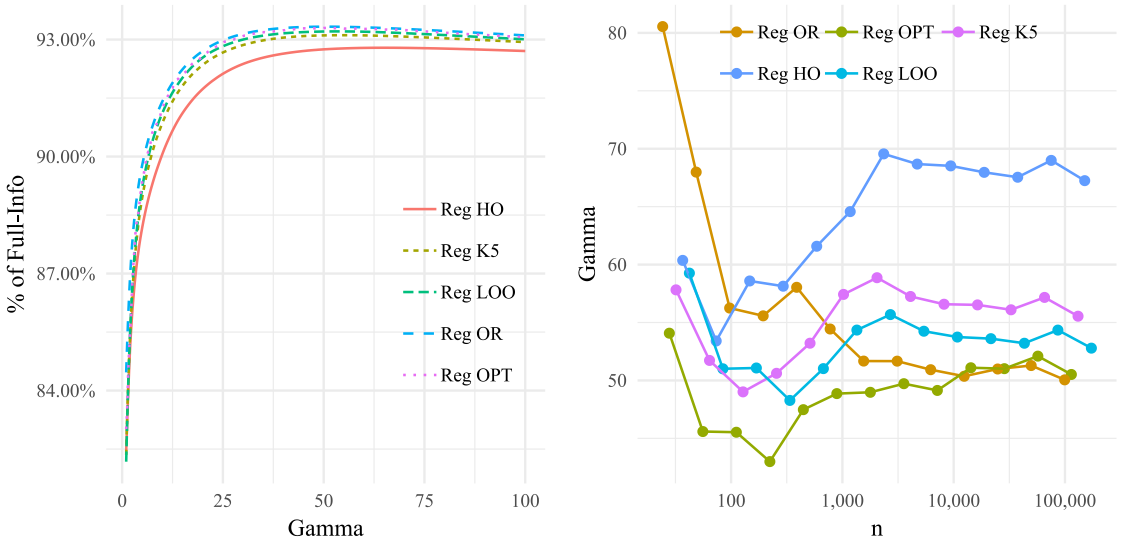


(a) Bayes-Inspired Policies



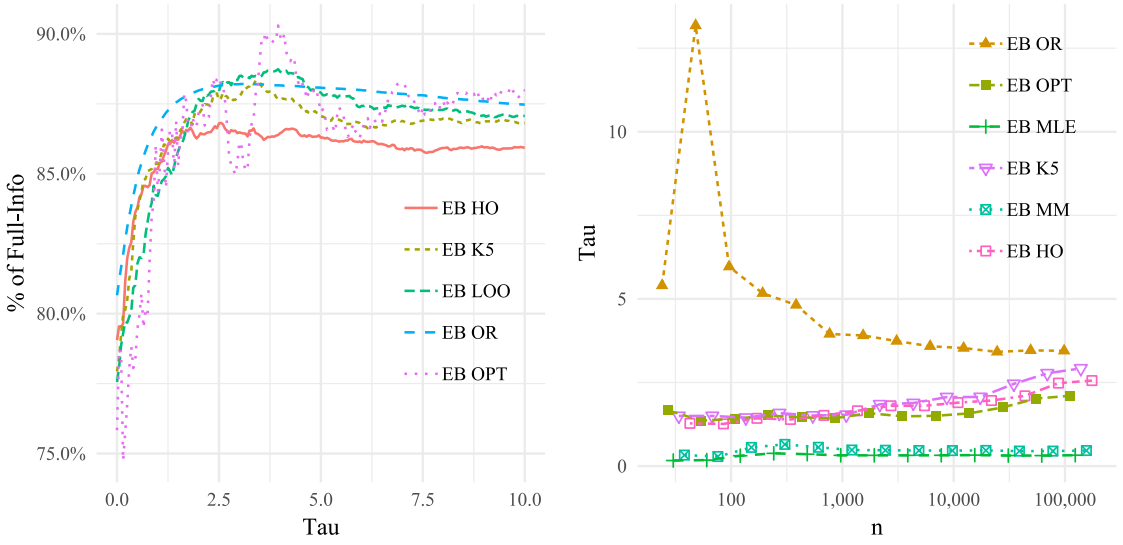
(b) Regularization-Inspired Policies

Figure EC.4 **Performance of all data-driven methods in OAPOP from Section 6.** The top panel plots the performance of $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ for various n and data-driven procedures for choosing τ from Section 4. The bottom panel plots the performance of $\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ along the same sample paths for data-driven procedures for choosing Γ from Section 5. The error bars represent 10% and 90% quantiles over 200 simulations.



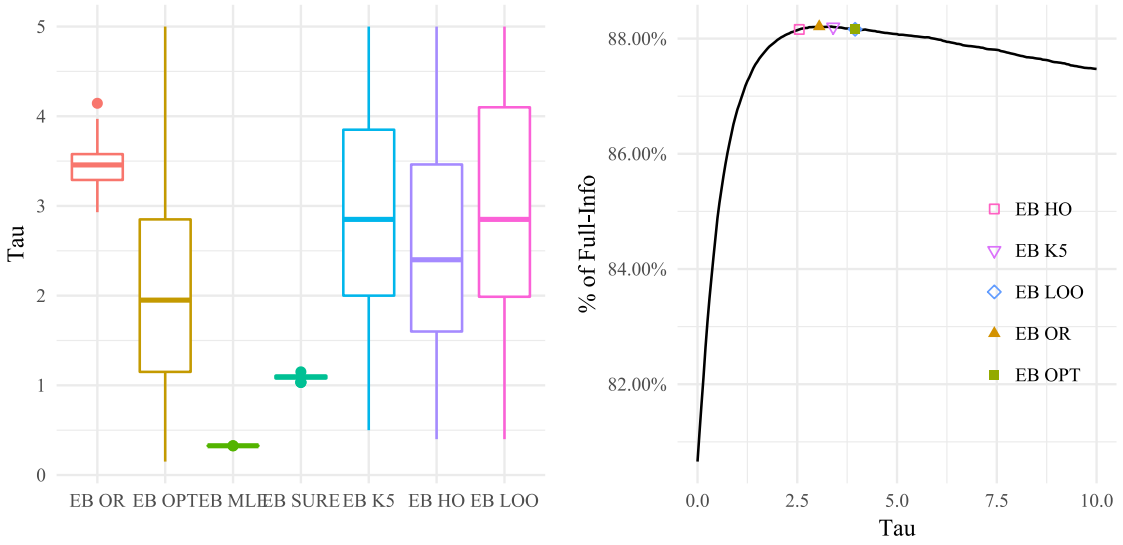
(a) Cross-Validation curves (b) Convergence of Estimates of $\Gamma^{\mathsf{OR},n}$

Figure EC.5 **Comparing Cross-Validation Policies for OAPOP.** The left panel plots the curves $\Gamma \to \frac{1}{K}\sum_{k=1}^{K} \overline{\boldsymbol{\mu}}^{k\top} \boldsymbol{x}^{k}(\Gamma, \overline{\boldsymbol{\mu}}^{-k})$ for various forms of cross-validation as well as $\Gamma \to \boldsymbol{\mu}^{\top}\boldsymbol{x}^{\mathcal{R}}(\Gamma, \hat{\boldsymbol{\mu}})$ for a single realization, when $n = 2^{17}$. The right panel plots the expected value of $\Gamma^{K-fold,n}, \hat{\Gamma}^{n}$, and $\Gamma^{\mathsf{OR},n}$ across the 200 simulations as $n \to \infty$.



(a) Cross-Validation curves (b) Convergence of Optimizing τ's

Figure EC.6 **Comparing Cross-Validation Policies for OAPOP.** The left panel plots the cross-validation curves (in τ), the target oracle curve and our bias-corrected approximation for $n = 2^{17}$. The right panel plots the average value of the optimizing τ's across the 200 simulations as $n \to \infty$.



(a) Distribution of Optimizing τ's (b) $\boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$

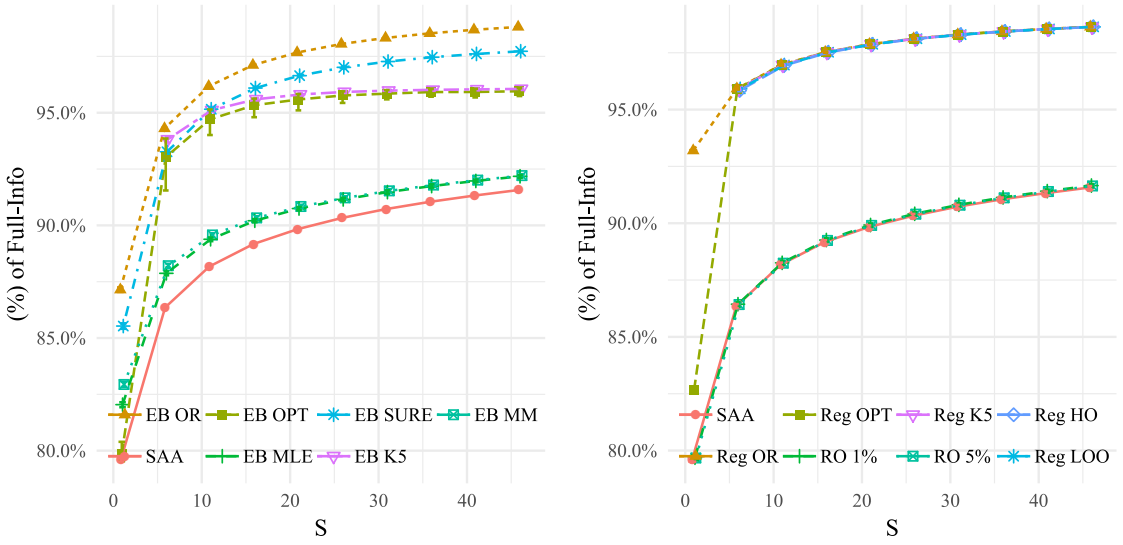
Figure EC.7 Explaining Performance of Cross-Validation for OAPOP. The left panel plots the distribution of the optimizing τ for various methods across the 200 simulations. The right panel plots the oracle curve $\tau \to \boldsymbol{\mu}^\top \boldsymbol{x}(\tau, \hat{\boldsymbol{\mu}})$ for a single realization, when $n = 2^{17}$, with optimizing τ of other methods indicated.

Example 2.2. Specifically, for $\mathsf{S} = 1, 2, \ldots,$, we take

$$\hat{\mu}_j = \frac{1}{\mathsf{S}} \sum_{k=1}^{\mathsf{S}} \left(\sqrt{\frac{1}{\nu_j}} \xi_k^j + \mu_j \right) \quad j = 1, \ldots, n, \tag{F.1}$$

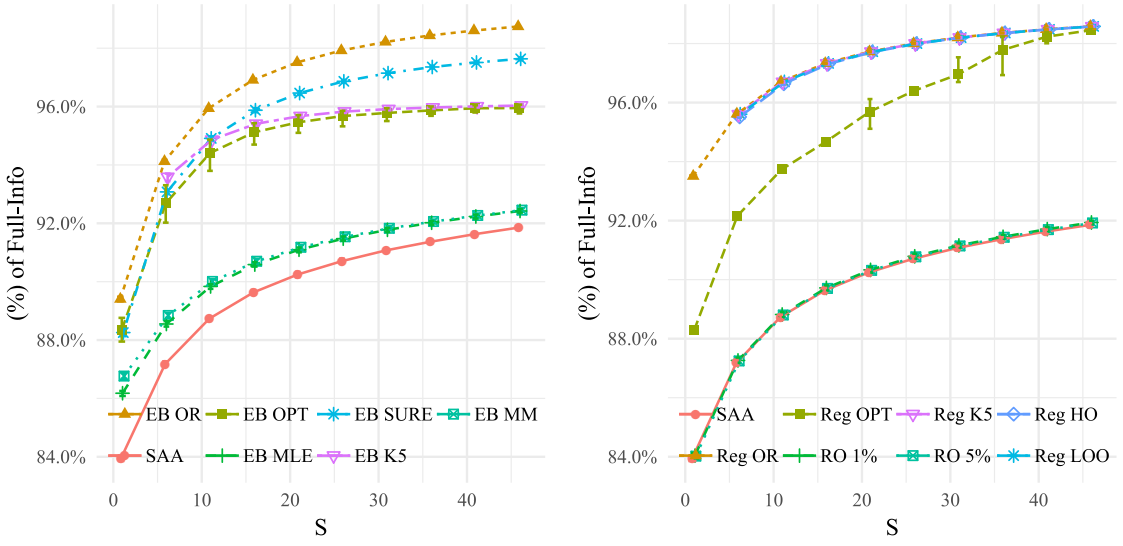
where ξ_k^j are i.i.d., random variables with mean 0 and precision 1. The effective precision of $\hat{\mu}_j$ is thus $\mathsf{S}\nu_j$, i.e., it grows as we acquire more data, and $\hat{\mu}_j$ eventually converges to a point-mass at μ_j. We consider several distributions for ξ_k^j, namely, uniform, exponential with rate 1, Student-t (with 3 degrees of freedom), and Pareto with shape 3 and unit scale. (In each case ξ_k^j is centered and scaled to have mean 0 and precision 1.) Notice these distributions include skewed, non-sub-Gaussian and heavy-tailed (having fewer than 4 moments) instances.

Figures EC.8-EC.11 below show the performance of each of our methods for these various distributions. We find several qualitative features that are consistent across these experiments. In each case, we can see that EB Opt and Reg Opt methods converge to the full-information and generally outperform SAA and most estimate-then-optimize procedures for both small and large S. A possible exception is the SURE procedure, which has excellent performance. Across the various experiments we also see that our optimization procedures have performance comparable to cross-validation procedures (even when S is small), despite the non-normality of the estimators. We believe this to be fairly strong evidence that these procedures retain the good large-sample properties of other methods.



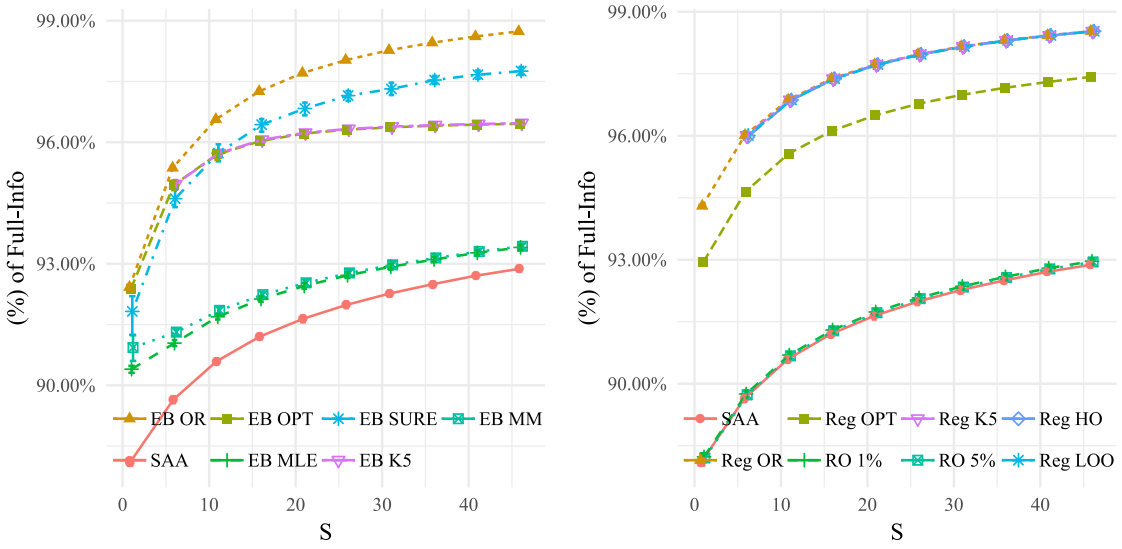
(a) Bayes-Inspired Policies (b) Regularization-Inspired Policies

Figure EC.8 **Large-Sample Performance for OAPOP (Uniform).** Each panel plots the performance of various data-driven procedures as S increases and data is drawn according to (F.1), when ξ_k^j is a centered uniform random variable. The error bars represent 10% and 90% quantiles over 200 simulations.



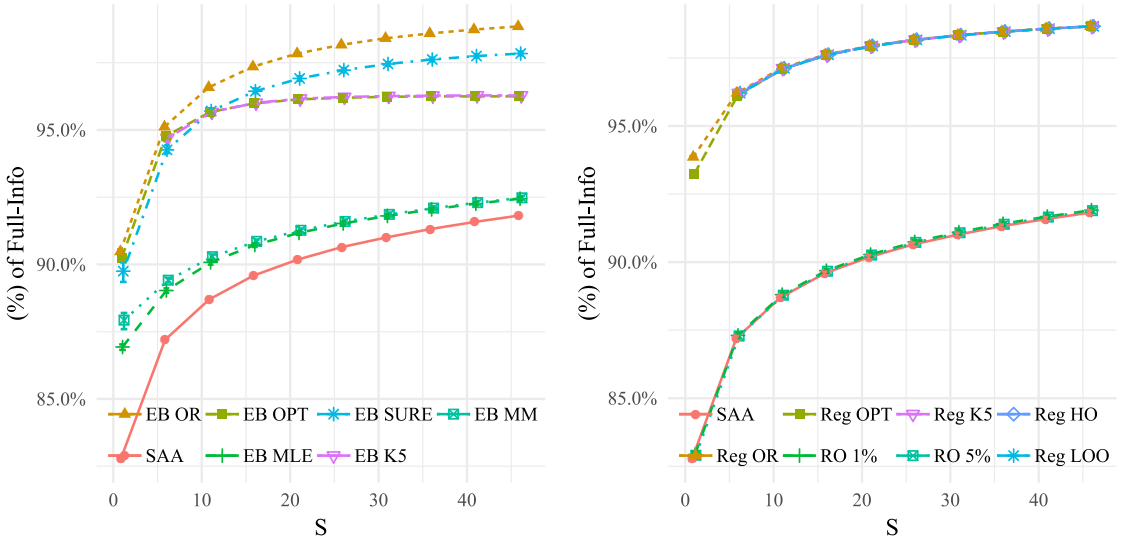
(a) Bayes-Inspired Policies (b) Regularization-Inspired Policies

Figure EC.9 **Large-Sample Performance for OAPOP (Exponential).** Each panel plots the performance of various data-driven procedures as S increases and data is drawn according to (F.1), when ξ_k^j is a centered Exponential random variable (rate = 1). The error bars represent 10% and 90% quantiles over 200 simulations.



(a) Bayes-Inspired Policies (b) Regularization-Inspired Policies

Figure EC.10 **Large-Sample Performance for OAPOP (Pareto).** Each panel plots the performance of various data-driven procedures as S increases and data is drawn according to (F.1), when ξ_k^j is a centered Pareto random variable (shape = 3, scale = 1). The error bars represent 10% and 90% quantiles over 200 simulations.



(a) Bayes-Inspired Policies (b) Regularization-Inspired Policies

Figure EC.11 **Large-Sample Performance for OAPOP (Student-t).** Each panel plots the performance of various data-driven procedures as S increases and data is drawn according to (F.1), when ξ_k^j is a centered Student-t random variable with three degrees of freedom. The error bars represent 10% and 90% quantiles over 200 simulations.

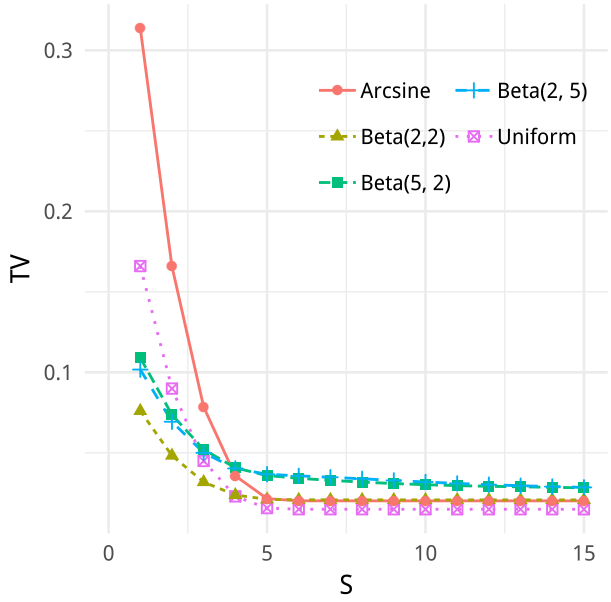


Figure EC.12 **Data Generation Procedure for Section F.5.** As S increases, the mean and standard deviation of $\hat{\mu}_j$ remain fixed for each j, but the density becomes more normal by the central limit theorem.

F.5. Robustness to Non-Normality

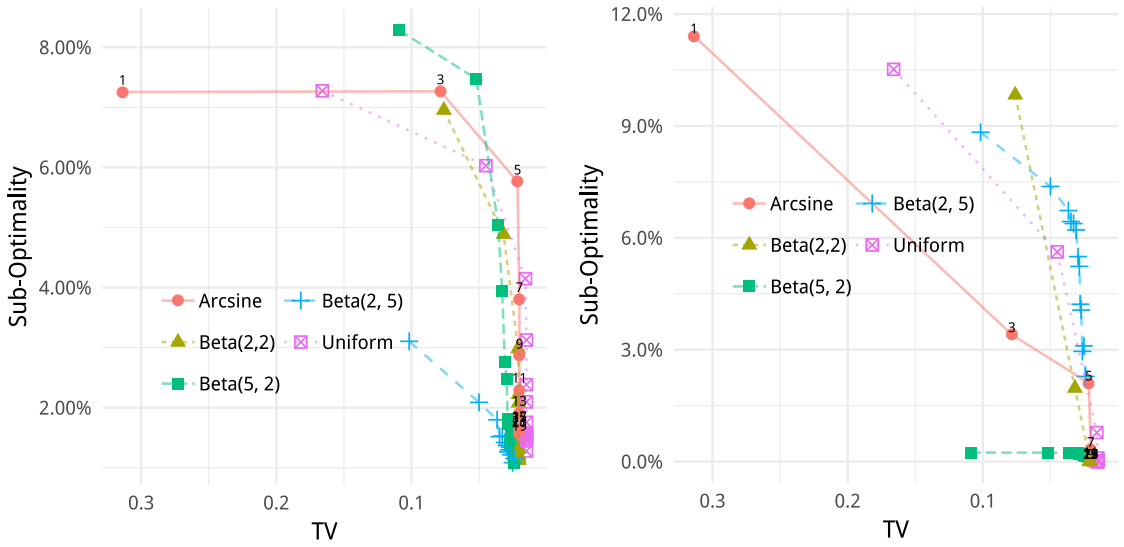
We next assess the robustness of our methods to increasing departures of normality. We consider a set-up similar to, but distinct from, Example 2.2. Specifically, for $\mathsf{S} = 1, 3, \ldots, 25$, we take

$$\hat{\mu}_j = \mu_j + \sqrt{\frac{\nu_0}{\mathsf{S}\nu_j}} \sum_{k=1}^{\mathsf{S}} \xi_k^j \quad j = 1, \ldots, n, \tag{F.2}$$

where ξ_k^j are i.i.d., mean-zero random variables with precision ν_0. A straightforward computation confirms that for any S, $\mathbb{E}[\hat{\mu}_j] = \mu_j$ and $\mathbb{E}[(\hat{\mu}_j - \mu_j)^2] = 1/\nu_j$.

We consider several different distributional choices for ξ^j, namely uniform, Beta distributions with parameters $(2,2)$ (symmetric), $(5,2)$ (negative skew) and $(2,5)$ (positive skew) and an arcsine distribution (U-shaped), each of which is first centered and normalized. Note these distributions are sub-Gaussian and admit a density. As $\mathsf{S} \to \infty$, $\hat{\mu}_j$ converges in distribution to a Gaussian, but for small S, it may be far from normal, depending on the distribution of ξ_k^j. Figure EC.12 shows the average total-variation distance, i.e., TV for each of these choices of distributions as S increases. Notice they converge quite quickly to zero, but do so at different speeds.

We re-run the experiment of Section 6.1, assuming $\hat{\mu}_j$ is drawn from the above process for various choices of ξ_j^k, $n = 2^{17}$ and increasing S. For each S and choice of distribution, we plot the average sub-optimality of our method versus the oracle performance over 200 sample paths. To make the results comparable across distributions, we plot the results against TV of $\hat{\boldsymbol{\mu}}$ instead of S. Figure EC.13 summarizes the results. Recall our theoretical results suggest each of these sub-optimality gaps in Figure EC.13 should tend to zero as $\mathsf{TV} \to 0$, and indeed, we do see such convergence across all distributions. These results suggest our method is robust to some non-normality, when noise is still sub-Gaussian and admits a density.



(a) $\boldsymbol{x}(\hat{\tau}, \hat{\boldsymbol{\mu}})$ (b) $\boldsymbol{x}^{\mathcal{R}}(\hat{\Gamma}, \hat{\boldsymbol{\mu}})$

Figure EC.13 **Robustness to Non-Normality** Average sub-optimality to the oracle performance vs. average total variation distance from a Gaussian. Performance scaled by the full-information optimum, computed over 200 sample paths, when $\hat{\boldsymbol{\mu}}$ is drawn as in Equation (F.2), $n = 2^{17}$, and S increasing.

Online Appendix References

Ben-Tal, A., D. Den Hertog, J.-P. Vial. 2015. Deriving robust counterparts of nonlinear uncertain inequalities. *Mathematical Programming* **149**(1-2) 265–299.

Ben-Tal, A., A. Nemirovski. 2000. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming* **88**(3) 411–424.

Bertsimas, D., M. S. Copenhaver. 2018. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research* **270**(3) 931–942.

Bertsimas, D., V. Gupta, N. Kallus. 2018. Data-driven robust optimization. *Mathematical Programming* **167**(2) 235–292.

Fertis, A. G. 2009. A robust optimization approach to statistical estimation problems. Ph.D. thesis, Massachusetts Institute of Technology.

Gupta, V. 2019. Near-optimal Bayesian ambiguity sets for distributionally robust optimization. To appear in *Management Science*. Available at `http://www.optimization-online.org/DB_FILE/2015/07/4983.pdf`.

Lam, H. 2016. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research* **41**(4) 1248–1275.

Negahban, S., B. Yu, M. J. Wainwright, P. K. Ravikumar. 2012. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* **27**(4) 538–557.

Pani, A., S. Raghavan, M. Sahin. 2017. Large-scale advertising portfolio optimization in online marketing. *Working Paper* URL `http://terpconnect.umd.edu/~raghavan/preprints/lsoapop.pdf`.

Pollard, D. 1990. Empirical processes: Theory and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, vol. 2. Institute of Mathematical Statistics, i–86.

Rivasplata, O. 2012. Subgaussian random variables: An expository note. *Internet publication, PDF* URL `http://www.stat.cmu.edu/~arinaldo/36788/subgaussians.pdf`.

Van der Vaart, A. W. 2000. *Asymptotic Statistics*. No. 3 in Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, UK.

Wainwright, M. J. 2015. Lecture notes. `http://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf`. [Online; accessed Feb 2016].

Wald, A. 1947. An essentially complete class of admissible decision functions. *The Annals of Mathematical Statistics* **18**(4) 549–555.

Wald, A. 1949. Statistical decision functions. *The Annals of Mathematical Statistics* **20**(2) 165–205.

Xu, H., C. Caramanis, S. Mannor. 2009. Robustness and regularization of support vector machines. *Journal of Machine Learning Research* **10**(Dec) 1485–1510.

Response to the Area Editor, AE, and Referees

Small-Data, Large-Scale Linear Optimization with Uncertain Objectives (MS-17-02522.R2)

October 23, 2019

Dear Professor Ye and the Editorial Team:

We would like to thank you, the AE, and the three referees for your most recent review of our manuscript and are happy to know we were able to successfully incorporate all of the constructive suggestions of the review team.

Following the AE's recommendation, we have addressed all of the comments by the review team. (Note, the vast majority of suggestions in this revision were editorial/typographical.) Any edits to the main document are highlighted in blue.

Thank you very much again for all of your time and consideration with our submission.

Sincerely,
The Authors

1. Response to Referee 1

We want to thank the referee for the detailed comments. As discussed below, we have addressed all the concerns in your report.

1. The abstract is too long and should be shortened to a single paragraph.

⇒ As you suggested, we have shortened the abstract to a single paragraph.

2. The paper itself is too long and does not adhere to Management Science guidelines (32 pages, 33 lines of text/1.5 spacing). My suggestions for cutting the length are:
   - The motivating examples are overly long. Given that these examples do not actually adhere to the actual problem in this paper, they can be described much more succinctly. The point you want to get across is that there are examples where you may have very little for a particular event/product/phenomenon of interest; this doesn't need so much convincing.
   - Sections 4.2 and 5.2 can be moved to an online supplement. Perhaps also Sections 5.4, 6.2 and 6.3

⇒ We have shortened the document to address the above two bullet points as follows: Following your suggestion, we have moved what was Sec. 5.1 (Robust Optimization Interpretation of Regularization Class) to Appendix E.1, moved Sec. 5.4 (Large-Sample Guarantees for Regularization Class) to Appendix E.5, moved Sec. 6.2 (Experiments in Large-Sample Regime), and 6.3 (Experiments for Robustness to Non-Normality) to Appendices F.3 and F.5, respectively. In each case we have left a sentence or two in the main text indicating the key message of that section and pointing the interested reader to the relevant appendices.

In addition, we have engaged in general tightening of the writing throughout the document. Respectfully, we have left Sec. 4.2 as is, since it is the key section that establishes intuition for the proof of the main theorem. (In previous rounds, a referee had complained that there was insufficient intuition for the proofs and that proofs were "impenetrable.") Similarly, we have left Sec. 5.2 as is since many of the editorial team had asked about the use of cross-validation, and this is the most-common way of selecting regularization parameters in practice.

In total, the main body of the paper without references has shrunken from 38.5 pages to 34 pages (excluding references). We have sincerely done our best in this regard, while retaining the crux of the excellent suggestions from the review team from previous rounds.

3. Italics are used gratuitously throughout the paper. Please reserve their usage for the most important words that need highlighting.

⇒ We have removed many instances of italics. In the revised manuscript, we highlighted the relevant (now) non-italicized words in blue.

4. Throughout the paper, "such as" should be used when giving examples, not "like," which means similar to (e.g. pp. 2 line 20 "Ride-sharing platforms like Uber").

⇒ Respectfully, we consider this an issue of style, since the meaning is entirely clear as written. Most style guides accept using "like" in this context, especially when used towards the beginning of a sentence or when the example is *not* offset by commas as a parenthetical phrase. Both mentioned instances have this feature and so we have left them as is. In other instances, we have adopted the change. See, for example, the top of page 5.

5. pp. 2, Figure 1 caption – "Apart from" not "Except [for]"

⇒ Respectfully, we disagree with this editorial suggestion. Both multiword prepositions "Apart from" and "Except for" correctly indicate exclusion in this context, however, "Apart from" can (in other contexts) also indicate a potential inclusion, e.g., "Apart from data-driven optimization, Jane also studies grammar." Because of this ambiguity, we prefer "Except for" in this caption.

6. pp. 2, lines 41 and 57: "may" not "might"

⇒ Changed.

7. pp. 3, lines 29 - 30: do you show this?

⇒ This comment refers to the fact that shortest path problem and the transportation problem are both representable as linear programs. This is a standard fact, and in the revised manuscript, we have added a citation on page 3 to (Bertsimas and Tsitsiklis 1997, Chapt. 7)

8. pp. 3, line 41, "provably impossible" – perhaps change to "with probability zero"? It is conceivable that one somehow computes the oracle solution, although this may happen with probability zero (with continuous underlying stochasticity).

⇒ We apologize for the confusion. This statement is actually not a probabilistic one; as described in the referenced Theorem 2.7, there will always exist a different problem instance (non-random) for which the policy does not achieve full-information performance. We reworded the statement to make this clearer, and added a reference to Theorem 2.7 for clarification.

9. pp. 4, line 35: please add [2] to the citation on regularization of SAA. Also pp. 5, line 52.

⇒ Respectfully, the discussion on pp. 4 line 35 and pp. 5 line 52 are about *general-purpose* regularization approaches. There of course exist many specialized regularizers approaches tailored to specific problem formulations. "Machine Learning and Portfolio Optimization" by Ban, El Karoui

and Lim is one such work, specifically aimed at portfolio optimization. If we were to include this citation, it feels appropriate then to add a host of other citations for other customized regularizers for matrix completion problems, textual analysis, causal inference problems, structured group lasso, etc. Instead, we have respectfully chosen to simply cite some seminal works on general-purpose regularization and refer the reader to references therein.

10. pp. 5, line 55: please add [1] to the citation on distributionally robust optimization.

⇒ Added. Thank you for the reminder.

11. pp. 7, lines 23-25: uniform convergence isn't a weak assumption, so I suggest re-writing the sentence so as to not suggest it is.

⇒ We were unsure how to address this comment since the requisite sentence doesn't actually use the word "weak" or suggest that uniform convergence is a weak criteria.

12. pp. 7, line 32: remove "will" from "we will".
13. pp. 9, line 51: [3] is a better reference here than Ban and Rudin (2014).
14. pp. 9, line 52: "we demonstrate" not "we prove".

⇒ Changed all three.

15. pp. 11, line 17: I'm confused by "Difference choices of $\boldsymbol{\mu}$." Isn't the theorem for all $\boldsymbol{\mu} \in \{-1, 1\}^n$?

⇒ We apologize for the confusion. We have reworded this sentence and added a sentence in next paragraph to clarify. The theorem states that *there exists* a $\boldsymbol{\mu} \in \{-1, 1\}^n$ such that the given policy $\boldsymbol{x}(\cdot)$ performs poorly. What we intended to convey is that the upper bound on the performance is not tight. Indeed, if, as part of the proof technique, one were to generate random instances in a different way one might be able to prove a stronger upper bound. (Stressing the possibility for tighter bounds in this way was suggested by one of the reviewers in the previous round.) To help clarify, we now discuss this possibility for a tighter bound *after* describing the proof technique and rephrased the sentence to make it clearer we are referring to the sampling procedure. The theorem statement previously, and currently, still stress that the result is an *existence* one, i.e., "there exists instances of P_n ...."

16. pp. 11: please don't use imprecise/emotional/unscientific words as "hopeless" (line 36). Similar for "Unfortunately" (line 57).

⇒ On pg. 11, we changed "hopeless" to "impossible", and changed "Unfortunately" to "However".

17. pp. 11, Definition 3.1.: please define what $\boldsymbol{x}(\cdot, \cdot)$ is.

⇒ This quantity is now defined in the statement just preceding Def. 3.1 on pg. 11.

18. pp. 13, lines 20-23: some of the mathematical statements would be better written on a separate line.
19. pp. 13, line 33: remove the word "loose"

⇒ Both fixed.

2. Response to Referee 2

1. P.5, line 19: The terminology of "no cost" (here and perhaps elsewhere too) may have to be clarified more precisely. I think the authors are thinking of statistical cost, but the approach does pay extra cost in computation.

⇒ Thank you for the suggestion. We added the word "statistical" per your suggestion on pg. 5. Moreover, we've placed "no statistical cost" in quotation marks in the passage to alert the reader that we do not mean it as a strictly precise term, but will clarify its precise meaning in the subsequent sentences.

2. P.6 line 57 - P.7 line 6: As the comparison with the notion of generalization errors is brought up, it may be good to clarify, at some point in the paper, the motivation of looking at the performance in the "mean" optimization rather than generalization performance. It seems the argument for the former is that the adopted policy is going to be used for a long time, so that the variability is less of a concern than the mean. (To be fair to the authors though, it seems this comparison is also present in the conventional SAA studies).

⇒ Thank you for the suggestion. We added a small discussion on pg. 5 that i) emphasizes (as you pointed out) that the majority of the literature looks at expected values and ii) that in our small-data large-scale regime for Problem P_n, the out-of-sample performance (as a random variable) converges to the expected performance (conditional on the data) under fairly mild assumptions by the strong law of large numbers. Thus, optimizing expected performance (conditional on the data) is almost the same as optimizing the out-of-sample performance.

To be more concrete, with respect to Eq. (1.1), the out-of-sample (or generalization) performance of a policy $\boldsymbol{x}(\cdot)$ is $c(\boldsymbol{x}(\hat{\boldsymbol{\xi}}^1,\ldots,\hat{\boldsymbol{\xi}}^{\mathsf{S}}),\overline{\boldsymbol{\xi}})$ where $\overline{\boldsymbol{\xi}}$ is an i.i.d. copy of $\boldsymbol{\xi}$. In the special case of Problem P_n, this is $\frac{1}{n}\sum_{j=1}^{n}\overline{\xi}_j x_j(\hat{\boldsymbol{\xi}}^1,\ldots,\hat{\boldsymbol{\xi}}^{\mathsf{S}})$. If the terms $\overline{\xi}_j$ are sufficiently independent across j and n is large (as in our small-data, large-scale regime), we would expect that, conditional on the data, this sum should converge to its expectation by the strong law of large numbers. This conditional expectation is precisely $\frac{1}{n}\boldsymbol{\mu}^\top\boldsymbol{x}(\hat{\boldsymbol{\xi}}^1,\ldots,\hat{\boldsymbol{\xi}}^{\mathsf{S}})$ which is our targeted performance throughout the paper. Notice that we condition on the data in this argument because the data, itself, causes the terms in the sum to be highly and non-trivially dependent, which is one of the technical challenges we address in the analysis of the paper when developing our high-probability bounds.

3. It seems that the notion of s_0-strict feasibility would not be able to handle equality constraint (in contrary to the claim of the authors in the response letter). As the authors rightly suggest, one can always expand the right hand side slightly to make things work, but depending on the contexts where equality constraints arise, such an adjustment may lead to insensible outcomes..

⇒ Thank you for the comment. In our response to the AE in the last round we clarify that Theorems 4.3 and 5.3 require s_0-strict-feasibility, and that polyhedra with equality constraints *do not* satisfy this condition. Please see pg. 90 Line 39 of our previous response letter where we wrote,

> "Polyhedra with equality constraints do not directly satisfy this condition.... This is a limitation of our analysis and we did our best to be as upfront about it as possible, stressing that 1) our algorithm readily applies to equality constraints (the requirement only affects the analysis) and 2) equality constraints can be modified as you suggest by a small perturbation in some cases."

We apologize if this wasn't clear in our last response letter. It *is* our intention to be as upfront about this requirement as possible in the two relevant theorems. To further clarify, we have added a footnote in this revision on pg. 20 explaining, as you point out, that in certain applications relaxing an equality constraint might not be acceptable.

Other Comments:

1. P.1, line 32: I think "regularization" is also a popular tool in statistics (I'm raising this because the authors writes that empirical Bayes comes from statistics right above..)

⇒ We've reworded this on pg. 1 so that it is clear we intend regularization in a general sense without attributing it to a particular field.

2. P.6, line 56-57: "non-trivial dependence structure": perhaps even clearer to say this dependence means it's more than just an average of independently sampled functions as in the related VC theory.

⇒ Thank you for the suggestion. We have added to the relevant passage to stress this on pg. 6

3. P.9, line 27: the notation "$(\hat{\boldsymbol{\mu}}, \boldsymbol{\nu}, \mathcal{X}) \mapsto \mathcal{X}$" is a bit odd. The left hand side of the arrow refers to the parameter while the right hand side refers to the space..

⇒ Thank you. We've adjusted the notation at the top of pg. 9.

4. P.12, line 23: "Regularization"

⇒ Fixed on pg. 12.

5. P.14, last paragraph: the requirement of "sufficiently independent" (in point 2 earlier on the same page) can be highlighted further as needed for a small stochastic deviation. Currently, it may be unclear where the sufficient independence is needed.

⇒ Thank you for the comment. We added a little to this discussion (and a footnote) on pg. 14.

6. P.17, line 17: should be "τ_0"?

⇒ We fixed this; please see the bottom of pg. 16.

7. P.18, line 25: "satisfies"

⇒ On the top of pg. 18, we reworded the awkward phrasing and instead wrote that "τ^{MLE} solves".

8. P.22, line 9: remove "." after "Equation"

⇒ We fixed this; see the bottom of pg. 20.

9. P.23, Theorem 4.3: it seems that "m" is used before it is introduced (later in the same theorem).

⇒ We now introduce m as the dimension of $\boldsymbol{b}$ on pg. 19 after Eq. (4.1), which is referenced in the theorem.

10. P.25, line 14: "$\hat{\boldsymbol{\mu}}_j$" and "$\boldsymbol{\mu}_j$" shouldn't be bolded

⇒ This is fixed; please see the bottom of pg. 24.

11. P.28, Theorem 5.2: Not an important point, but it may be better to have a class of examples that target at K-fold for a given K. It appears a little strange to focus only on the specific numbers of 2, 5 and 10.

⇒ Thank you for the comment. We indeed did attempt to construct a family of examples (indexed by K) as the reviewer suggests, but were ultimately unable to do so *cleanly*. The choices of 2, 5, and 10 refer to the overwhelmingly most popular choices of K-fold cross-validation procedures (Kohavi (1995), Friedman et al. (2001)), i.e., hold-out validation, 5-fold validation, and 10-fold validation. Thus, although a bit inelegant, our result does cover the cases of practical interest. In the end, we agree with you that the generalization to arbitrary K is perhaps not the most important point of the paper, and, hence, left the result as is. As a side note, because we have shortened our paper, Theorem 5.2 in the previous manuscript is now labeled as Theorem 5.1 on pg. 27.

12. P.30, line 41: "performance"

⇒ Fixed; see pg. 29.

3. Response to Referee 3

Section 2

Error on line 50 : shown / demonstrated?

⇒ The reviewer comment seems to refer to the sentence on pg. 9 (current pagination), which we have reworded to "demonstrate."

Section 3

I still have my doubts on the usefulness of the discussion in Section 3.1, though, as the general bias-correction term is not used in this form in either the Bayes/Robust policies discussed in Sections 4&5. However, I agree that it provides a general idea of the methods used in these sections. Hence, I do not have any further objection.

⇒ Thank you very much for your support.

When the parameters are nongaussian distributions, the difference between the oracle performance the proposed policy converges to a constant independent of the problem parameters of P^n. What is so special about the gaussian distribution that this constant is zero? Is there an intuitive way to understand this result? I suppose this is a consequence of Stein's Lemma?

⇒ Thank you for the question, and your observation is indeed correct; this is a consequence of Stein's Lemma. Stein actually proves i) that the identity $\mathbb{E}[\zeta f(\zeta)] = \mathbb{E}[f'(\zeta)]$ holds for all (weakly) differentiable functions f whenever $\zeta \sim \mathcal{N}(0,1)$, and ii) the standard normal distribution is the *only* distribution which satisfies this identity for all (weakly) differentiable f. Thus, this identity can be seen as characterizing the normal distribution, and indeed, there is an entire literature within statistics focused on proving that random variables are approximately (or asymptotically) normally distributed by showing that they approximately (or asymptotically) satisfy Stein's identity. (This technique is frequently called "Stein's Method." See Ross et al. (2011) for details.) This is in some sense what is "special" about the gaussian distribution.

That said, providing a clean intuition behind such an elegant and fundamental probability result is somewhat difficult, and, somewhat outside the scope of our paper. We've instead settled for pointing readers to an appropriate authoritative reference (See the bottom of pg. 12).

Error on line 35 : Point instead of comma after equation?

⇒ Fixed; please see the equation just above (3.3) on pg. 13.