Technical Note: Simplifying the Analysis of the Stein-Correction in the Small-Data, Large-Scale Optimization Regime

Vishal Gupta, Michael Huang, and Paat Rusmevichientong
Data Science and Operations, USC Marshall School of Business, Los Angles, CA 90089,
guptavis@usc.edu, huan076@usc.edu, rusmevic@marshall.usc.edu

This technical note provides an alternate proof of Gupta and Rusmevichientong (2021, Theorem 4.3). The proof presented here is more general, and substantively simpler. It hinges on a new approach to bounding the dependence between different components of the solution of a random linear optimization problem by considering worst-case dual realizations.

1. Introduction

In Gupta and Rusmevichientong (2021), the authors present an analysis of their Stein-Correction for debiasing in-sample performance in the small-data, large-scale optimization regime. Specifically, Theorem 4.3 of that work establishes a high-probability bound on the error of the estimator in a certain linear optimization setting. The proof focuses on the dual optimization problem and establishes that (under suitable conditions) the dual optimal solution converges to a deterministic function, uniformly over a specific policy class, at a sufficiently fast rate in the small-data, large-scale limit. As the authors explain, this argument is necessary to establish that certain sums of dependent random variables are not “too dependent,“ and hence still concentrate uniformly.

Similar in spirit, Gupta, Huang, and Rusmevichientong (2022a) present a different approach to debiasing linear optimization problems in this regime. They also study the dual optimization problem and show it satisfies an “approximate strong-convexity” property. They then leverage this property in order to establish a similar convergence of the dual optimal solutions to a deterministic function, at a sufficiently fast rate.

Both arguments are mathematically onerous and require a number of regularity assumptions. Indeed, while it is somewhat straightforward to show that the dual optimal solutions converge to deterministic value pointwise for each member of a policy class, it is much more challenging to show this convergence holds *uniformly* and occurs sufficiently fast for the remainder of the argument to go through.

The key idea of this technical note is to simplify this dual analysis by instead considering worst-case performance over all possible values for the dual optimal solution. This worst-case perspective substantially simplifies the proof without affecting the overall bound on estimation error. In our opinion, it also makes the dependence between the optimization structure and the convergence rate of the estimator more transparent.

For simplicity of exposition, we present this new proof idea in context of a generalization of Theorem 4.3 of Gupta and Rusmevichientong (2021), but the same idea can be applied to other data-driven optimization algorithms, including the so-called Variance Gradient Correction (Gupta, Huang, and Rusmevichientong, 2022a, Theorem 4.7). We leave a full discussion of these generalizations for our current working paper (Gupta, Huang, and Rusmevichientong, 2022b).

2. Model Setup and Main Result

Our problem of interest is

$$\begin{aligned} \max_{\boldsymbol{x}\in[0,1]^n} \quad & \boldsymbol{\mu}^\top \boldsymbol{x} \\ \text{s.t.} \quad & \sum_{j=1}^n \boldsymbol{A}_j x_j \le \boldsymbol{b}, \end{aligned} \tag{1}$$

where $\boldsymbol{b}\in\mathbb{R}^m$ and $\boldsymbol{A}_j\in\mathbb{R}^m$ for $j=1,\ldots,n$. We assume throughout that $n\ge 2$, Problem (1) is feasible, and that the columns $\boldsymbol{A}_j$ are in general position. Recall, a set of points in $\mathbb{R}^d$ is in general position if no k of them lie in a $(k-2)$-dimensional flat for $k=2,\ldots,d+1$. Observe that any set of columns $\boldsymbol{A}_j$ can be placed in general position by perturbing them by an arbitrarily small amount, so that this last assumption is *almost* without loss of generality.

Our data are Gaussian corruptions of the true $\boldsymbol{\mu}$, i.e.,

$$Z_j \sim \mathcal{N}(\mu_j, 1/\nu_j) \quad j=1,\ldots,n,$$

drawn independently across j, where ν_j is known for each j. (Extending the results below to the “near-Gaussian” setting of Gupta and Rusmevichientong (2021) is straightforward but tedious.) We also observe a fixed (non-random) covariate $\boldsymbol{W}_j\in\mathcal{W}$ for each $j=1,\ldots,n$, and, without loss of generality, the first component of $\boldsymbol{W}_j$ is ν_j. Finally, let $\nu_{\min}\equiv\min_{j=1,\ldots n,}\nu_j$.

We consider a class of policies indexed by functions $f \in \mathcal{F} \subseteq \mathbb{R}^{\mathbb{R}\times\mathcal{W}}$. Namely, given $f \in \mathcal{F}$, define

$$\begin{aligned} \boldsymbol{x}(f,\boldsymbol{Z}) \in \operatorname*{arg\,max}_{\boldsymbol{x}\in[0,1]^n} \quad & \sum_{j=1}^{n} f(Z_j,\boldsymbol{W}_j)x_j \\ \text{s.t.} \quad & \sum_{j=1}^{n} \boldsymbol{A}_j x_j \le \boldsymbol{b}, \end{aligned} \tag{2}$$

where ties are broken arbitrarily.

Problem (2) admits the dual linear optimization problem

$$\begin{aligned} (\boldsymbol{\lambda}(f,\boldsymbol{Z}),\boldsymbol{\theta}(f,\boldsymbol{Z})) \in \operatorname*{arg\,min}_{\boldsymbol{\lambda}\ge 0,\boldsymbol{\theta}\ge 0} \quad & \boldsymbol{b}^\top\boldsymbol{\lambda} + \boldsymbol{e}^\top\boldsymbol{\theta} \\ \text{s.t.} \quad & \theta_j \ge f(Z_j,\boldsymbol{W}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda} \quad j = 1,\ldots,n. \end{aligned} \tag{3}$$

Since the primal is bounded and feasible, a dual optimal solution exists, and we require $(\boldsymbol{\lambda}(f,\boldsymbol{Z}),\boldsymbol{\theta}(,\mathbf{Z}))$ be chosen to be a basic feasible optimal solution. Observe that in any optimal solution, $\theta_j(f,\boldsymbol{Z}) = [f(Z_j,\boldsymbol{W}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(f,\boldsymbol{Z})]^+$.

Our goal is to provide an estimator for $\boldsymbol{\mu}^\top\boldsymbol{x}(f,\boldsymbol{Z})$ and bound the error of this estimator *uniformly* over all $f \in \mathcal{F}$. To that end, we generalize the Stein Correction from Gupta and Rusmevichientong (2021) to our setting. Namely, inspired by complementary slackness, let $x_j(f,z,\boldsymbol{\lambda}) = \mathbb{I}\left\{f(z,\boldsymbol{W}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda} \ge 0\right\}$. With some overloading of notation, we let $\boldsymbol{x}(f,\boldsymbol{Z},\boldsymbol{\lambda})$ be the vector-valued function whose j^{th} component is $x_j(f,Z_j,\boldsymbol{\lambda})$.

While the vanilla Stein correction of Gupta and Rusmevichientong (2021) is formed using a central finite diference of the component functions $x_j(f,\boldsymbol{Z})$, we instead use a central finite difference of the component functions $x_j(f,Z_j,\boldsymbol{\lambda})$. Specifically, for any $h > 0$, we define

$$B(f,\boldsymbol{Z},\boldsymbol{\lambda}) \equiv \frac{1}{2h}\sum_{j=1}^{n}\frac{1}{\sqrt{\nu_j}}\left(x_j(f,Z_j+\frac{h}{\sqrt{\nu_j}},\boldsymbol{\lambda}) - x_j(f,Z_j-\frac{h}{\sqrt{\nu_j}},\boldsymbol{\lambda})\right). \tag{4}$$

We then propose estimating the unknown out of sample performance $\boldsymbol{\mu}^\top\boldsymbol{x}(f,\boldsymbol{Z})$ by the data-driven quantity $\boldsymbol{Z}^\top\boldsymbol{x}(f,\boldsymbol{Z}) - B(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z}))$.

2.1. Main Results

To rigorously state our main result, we must introduce a few more assumptions on $\mathcal{F}$. First, given $\mathcal{F}$, we define a class of binary-valued functions with domain $\mathbb{R}\times\mathcal{W}\times\mathbb{R}^m$

$$\mathcal{R}(\mathcal{F}) = \{(z,\boldsymbol{W},\boldsymbol{A}) \mapsto \mathbb{I}\left\{f(z,\boldsymbol{W}) - \boldsymbol{A}^\top\boldsymbol{\lambda} \ge 0\right\} \ : \ f \in \mathcal{F}, \boldsymbol{\lambda} \in \mathbb{R}_+^m\}. \tag{5}$$

Assumption 1 **(Bounded Complexity)**. *The class $\mathcal{F}$ is such that $\mathcal{R}(\mathcal{F})$ in Eq.* (5) *has VC dimension at most V.*

Moreover, to avoid some technical issues around non-uniqueness of the solution to Problem (1) we will make the following assumption:

Assumption 2 **(Almost Sure Dual Non-Degeneracy)**. *We have*

$$\mathbb{P}\left((\boldsymbol{\lambda}(f,\boldsymbol{Z}),\boldsymbol{\theta}(f,\boldsymbol{Z})) \text{ is a non-degenerate basic feasible solution for all } f\in\mathcal{F}\right)=1.$$

Notice that dual non-degeneracy implies primal uniqueness, so that Assumption 2 ensures $\boldsymbol{x}(f,\boldsymbol{Z})$ is uniquely defined for all $f\in\mathcal{F}$ almost surely (Bertsimas and Tsitsiklis, 1997).

Our main result is

Theorem 1 **(Uniform Bound on Estimation Error)**. *Under Assumptions 1 and 2, there exists a universal constant C such that for any $0<h<1$ and any $0<\epsilon<1$,*

$$\sup_{f\in\mathcal{F}}\underbrace{\left|\boldsymbol{Z}^\top\boldsymbol{x}(f,\boldsymbol{Z})-B(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z}))-\boldsymbol{\mu}^\top\boldsymbol{x}(f,\boldsymbol{Z})\right|}_{\textit{Estimation Error}}\le C\frac{h^2n}{\sqrt{\nu_{\min}}}+C\left(m+\frac{\sqrt{Vn}}{h}\right)\sqrt{\frac{\log n}{\nu_{\min}}}\log\left(\frac{2}{\epsilon}\right).$$

We compare Theorem 1 to Gupta and Rusmevichientong (2021, Theorem 4.3) in Section 4. For now, observe that we can optimize the rate of convergence in the theorem by choosing $h=(V/n)^{1/6}$, yielding a bound of the form

$$C\left(m+V^{1/3}n^{2/3}\right)\sqrt{\frac{\log n}{\nu_{\min}}}\log\left(\frac{2}{\epsilon}\right).$$

In practice, verifying Assumption 2 may be difficult. We can replace Assumption 2 with a simpler condition on $\mathcal{F}$. Define

$$S(f,\boldsymbol{Z})\equiv\left\{\begin{pmatrix}f(Z_j,\boldsymbol{W}_j)\\ \boldsymbol{A}_j\end{pmatrix}\ :\ j=1,\dots n\right\}\cup\{\boldsymbol{0}\}\subseteq\mathbb{R}^{m+1}.$$

Assumption 3 **(Induced Cost Vectors in General Position)**. *We have*

$$\mathbb{P}\left(S(f,\boldsymbol{Z}) \text{ are in general position for all } f\in\mathcal{F}\right)=1.$$

Theorem 2 **(Uniform Bound on Estimation Error (II))**. *Under Assumptions 1 and 3, the conclusion of Theorem 1 holds.*

Because the bounds in Theorems 1 and 2 hold uniformly, optimizing our estimator yields a nearly best-in-class policy.

Corollary 1 **(Near Best-in-Class Performance)**. *Let*

$$f^{OR} \in \arg\max_{f\in\mathcal{F}} \boldsymbol{\mu}^\top \boldsymbol{x}(f,\boldsymbol{Z}) \qquad \textit{and} \qquad \hat{f} \in \arg\max_{f\in\mathcal{F}} \boldsymbol{Z}^\top \boldsymbol{x}(f,\boldsymbol{Z}) - B(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z})),$$

with $h = (V/n)^{1/6}$. Then, under the assumptions of either Theorem 1 or Theorem 2, there exists a universal constant C such that for any $0<\epsilon<1$

$$\boldsymbol{\mu}^\top (\boldsymbol{x}(f^{OR},\boldsymbol{Z}) - \boldsymbol{x}(\hat{f},\boldsymbol{Z})) \le C\left(m + V^{1/3}n^{2/3}\right)\sqrt{\frac{\log n}{\nu_{\min}}}\log\left(\frac{2}{\epsilon}\right).$$

Proof. The proof follows directly from Theorem 1 or Theorem 2 after specifying h and invoking Gupta and Rusmevichientong (2021, Lemma C.1). □

3. Proof of Theorem 1.

Before presenting the proof, we recall the definition of the Ψ-Orlicz norm. Let $\Psi(t) = \frac{1}{5}\exp(t^2)$. Then, for any random variable Y, we define

$$\|Y\|_\Psi \equiv \inf\{C>0 : \Psi(Y/C) \le 1\}.$$

Mean-zero random variables with finite Ψ-Orlicz norm are sub-Gaussian.

We use the notation $a \lesssim b$ to mean there exists a universal constant C (not depending on problem parameters) such that $a \le Cb$.

Proof of Theorem 1. Fix some $f\in\mathcal{F}$. By triangle inequality, we have the upper bound

$$\begin{aligned}&\left|\boldsymbol{Z}^\top \boldsymbol{x}(f,\boldsymbol{Z}) - B(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z})) - \boldsymbol{\mu}^\top \boldsymbol{x}(f,\boldsymbol{Z})\right| \\ &\qquad \le \left|(\boldsymbol{Z}-\boldsymbol{\mu})^\top(\boldsymbol{x}(f,\boldsymbol{Z}) - \boldsymbol{x}(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z})))\right| + \left|(\boldsymbol{Z}-\boldsymbol{\mu})^\top \boldsymbol{x}(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z})) - B(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z}))\right|.\end{aligned}$$

The first term on the right can be bounded exactly as in (Gupta and Rusmevichientong, 2021). Specifically, by Lemma 1 below, $\boldsymbol{x}(f,\boldsymbol{Z})$ equals $\boldsymbol{x}(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z}))$ except possibly at m components, and since both $\boldsymbol{x}(f,\boldsymbol{Z})$ and $\boldsymbol{x}(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z}))$ are in $[0,1]^n$,

$$\left|(\boldsymbol{Z}-\boldsymbol{\mu})^\top(\boldsymbol{x}(f,\boldsymbol{Z}) - \boldsymbol{x}(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z})))\right| \le m\|\boldsymbol{Z}-\boldsymbol{\mu}\|_\infty.$$

Replacing this upper bound and taking the supremum over $f\in\mathcal{F}$ of both sides yields

$$\begin{aligned}&\sup_{f\in\mathcal{F}}\left|(\boldsymbol{Z}-\boldsymbol{\mu})^\top \boldsymbol{x}(f,\boldsymbol{Z}) - B(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z}))\right| \qquad (6)\\ &\qquad \le m\|\boldsymbol{Z}-\boldsymbol{\mu}\|_\infty + \sup_{f\in\mathcal{F}}\left|(\boldsymbol{Z}-\boldsymbol{\mu})^\top \boldsymbol{x}(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z})) - B(f,\boldsymbol{Z},\boldsymbol{\lambda}(f,\boldsymbol{Z}))\right|.\end{aligned}$$

Notice that the argument of the supremum on the right can be seen as a sum of n random variables, however these random variables are *dependent* because $\boldsymbol{\lambda}(f, \boldsymbol{Z})$ depends on the entire vector $\boldsymbol{Z}$. The proof of (Gupta and Rusmevichientong, 2021) thus proceeds by studying $\boldsymbol{\lambda}(f, \boldsymbol{Z})$ to bound this dependence.

We take a different approach. Instead of analyzing the function $\boldsymbol{\lambda}(f, \boldsymbol{Z})$ directly, we further upperbound the right hand side by considering the worst-case realization of $\boldsymbol{\lambda}(f, \boldsymbol{Z})$, namely we upper bound the second term on right side of Eq. (6) by

$$\sup_{f\in\mathcal{F}, \boldsymbol{\lambda}\in\mathbb{R}^m_+} \left|(\boldsymbol{Z}-\boldsymbol{\mu})^\top \boldsymbol{x}(f, \boldsymbol{Z}, \boldsymbol{\lambda}) - B(f, \boldsymbol{Z}, \boldsymbol{\lambda})\right|.$$

Applying the triangle inequality and substituting above yields the bound

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \left|(\boldsymbol{Z}-\boldsymbol{\mu})^\top \boldsymbol{x}(f, \boldsymbol{Z}) - B(f, \boldsymbol{Z}, \boldsymbol{\lambda}(f, \boldsymbol{Z}))\right| \\
&\quad\le\; m\|\boldsymbol{Z}-\boldsymbol{\mu}\|_\infty && \text{(7a)}\\
&\qquad + \sup_{f\in\mathcal{F}, \boldsymbol{\lambda}\in\mathbb{R}^m_+} \left|(\boldsymbol{Z}-\boldsymbol{\mu})^\top \boldsymbol{x}(f, \boldsymbol{Z}, \boldsymbol{\lambda}) - \mathbb{E}\left[(\boldsymbol{Z}-\boldsymbol{\mu})^\top \boldsymbol{x}(f, \boldsymbol{Z}, \boldsymbol{\lambda})\right]\right| && \text{(7b)}\\
&\qquad + \sup_{f\in\mathcal{F}, \boldsymbol{\lambda}\in\mathbb{R}^m_+} \left|B(f, \boldsymbol{Z}, \boldsymbol{\lambda}) - \mathbb{E}\left[B(f, \boldsymbol{Z}, \boldsymbol{\lambda})\right]\right| && \text{(7c)}\\
&\qquad + \sup_{f\in\mathcal{F}, \boldsymbol{\lambda}\in\mathbb{R}^m_+} \left|\mathbb{E}\left[(\boldsymbol{Z}-\boldsymbol{\mu})^\top \boldsymbol{x}(f, \boldsymbol{Z}, \boldsymbol{\lambda})\right] - \mathbb{E}\left[B(f, \boldsymbol{Z}, \boldsymbol{\lambda})\right]\right|. && \text{(7d)}
\end{aligned}$$

Writing the inner product in each of Eqs. (7b) and (7c) as a sum shows that both terms are now the maximum of a sum of *independent*, mean-zero random variables. Such sums can be analyzed using standard techniques from empirical processes. See Lemmas 2 and 3 below. This is the key idea behind our simplified proof.

The analysis of Eq. (7a) and Eq. (7d) follows (Gupta and Rusmevichientong, 2021) closely. Specifically, consider Eq. (7a). For each j, $\|Z_j - \mu_j\|_\Psi \lesssim \sqrt{1/\nu_j} \lesssim \sqrt{1/\nu_{\min}}$. Hence, by (Pollard, 1990, Lemma 3.2), $\|\|\boldsymbol{Z}-\boldsymbol{\mu}\|_\infty\|_\Psi \lesssim \sqrt{\frac{\log n}{\nu_{\min}}}$. Thus, by Markov's inequality, with probability at least $1-\epsilon$, $\|\boldsymbol{Z}-\boldsymbol{\mu}\|_\infty \lesssim \sqrt{\frac{\log n \log(2/\epsilon)}{\nu_{\min}}}$.

Similarly, expanding Eq. (7d) yields

$$\sup_{f\in\mathcal{F}, \boldsymbol{\lambda}\in\mathbb{R}^m_+} \left|\sum_{j=1}^n \mathbb{E}\left[(Z_j-\mu_j)x_j(f, Z_j, \boldsymbol{\lambda}) - \frac{1}{2h\sqrt{\nu_j}}\left(x_j(f, Z_j + \frac{h}{\sqrt{\nu_j}}, \boldsymbol{\lambda}) - x_j(f, Z_j - \frac{h}{\sqrt{\nu_j}}, \boldsymbol{\lambda})\right)\right]\right|.$$

We apply an Approximate Stein's Lemma (c.f. (Gupta and Rusmevichientong, 2021, Lemma C.2)). Specifically, let $\xi_j = \nu_j(Z_j - \mu_j)$ be a standard normal increment. Expressing the inner expectation in terms of ξ_j and applying (Gupta and Rusmevichientong, 2021, Lemma C.2) shows Eq. (7d) is at most

$$4h^2 \sum_{j=1}^{n} \sqrt{\frac{1}{\nu_j}} \le \frac{4h^2 n}{\sqrt{\nu_{\min}}},$$

by Jensen's Inequality.

Combining the above bounds with Lemmas 2 and 3 below shows that for any $0 < \epsilon < 1$, with probability at least $1 - \epsilon$,

$$\begin{aligned} &\sup_{f \in \mathcal{F}} \left| \boldsymbol{Z}^\top \boldsymbol{x}(f, \boldsymbol{Z}) - B(f, \boldsymbol{Z}, \boldsymbol{\lambda}(f, \boldsymbol{Z})) - \boldsymbol{\mu}^\top \boldsymbol{x}(f, \boldsymbol{Z}) \right| \\ &\qquad \lesssim m \sqrt{\frac{\log n \log(2/\epsilon)}{\nu_{\min}}} + \frac{h^2 n}{\sqrt{\nu_{\min}}} + \frac{\sqrt{V n \log n \log(2/\epsilon)}}{h \sqrt{\nu_{\min}}} + \sqrt{\frac{V n \log n}{\nu_{\min}}} \log\left(\frac{2}{\epsilon}\right) \\ &\qquad \lesssim \frac{h^2 n}{\sqrt{\nu_{\min}}} + \left(m + \frac{\sqrt{V n}}{h} \right) \sqrt{\frac{\log n}{\nu_{\min}}} \log\left(\frac{2}{\epsilon}\right). \end{aligned}$$

This proves the theorem. □

The proof of Theorem 2 is identical to the proof of Theorem 1 since we can use Assumption 3 instead of Assumption 2 to invoke Lemma 1. The details are omitted.

We now prove the missing lemmas from the above proof. The first lemma is a direct consequence of linear optimization duality.

LEMMA 1 **(Relating $\boldsymbol{x}(f, \boldsymbol{Z})$ and $\boldsymbol{x}(f, \boldsymbol{Z}, \boldsymbol{\lambda}(f, \boldsymbol{Z}))$).** *Under either Assumption 2 or Assumption 3,*

$$|\{j : x_j(f, \boldsymbol{Z}) \neq x_j(f, \boldsymbol{Z}, \boldsymbol{\lambda}(f, \boldsymbol{Z}))\}| \le m \qquad a.s.$$

In other words, $\boldsymbol{x}(f, \boldsymbol{Z})$ and $\boldsymbol{x}(f, \boldsymbol{Z}, \boldsymbol{\lambda}(f, \boldsymbol{Z}))$ agree in all except at most m components almost surely.

Proof. To streamline notation, fix some $f \in \mathcal{F}$ and let $\boldsymbol{f} = (f(Z_1, \boldsymbol{W}_1), \ldots, f(Z_n, \boldsymbol{W}_n))^\top$, $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}(f, \boldsymbol{Z})$ and $\boldsymbol{\theta}^* = \boldsymbol{\theta}(f, \boldsymbol{Z})$.

We first claim that $\boldsymbol{x}(f, \boldsymbol{Z})$ and $\boldsymbol{x}(f, \boldsymbol{Z}, \boldsymbol{\lambda}^*)$ can only differ at components j such that $\boldsymbol{A}_j^\top \boldsymbol{\lambda}^* = f_j$. To prove this claim, recall that at optimality, we must have $\theta_j^* = (f_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^*)^+$. Hence,

$$f_j > \boldsymbol{A}_j^\top \boldsymbol{\lambda}^* \implies \boldsymbol{\theta}^* > 0 \implies x_j(f, \boldsymbol{Z}) = 1 = x_j(f, \boldsymbol{Z}, \boldsymbol{\lambda}^*),$$

where the last implication follows by complementary slackness. Similarly,

$$f_j < \boldsymbol{A}_j^\top \boldsymbol{\lambda}^* \implies \boldsymbol{\theta}^* > f_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^* \implies x_j(f, \boldsymbol{Z}) = 0 = x_j(f, \boldsymbol{Z}, \boldsymbol{\lambda}^*),$$

where again the last implication follows by complementary slackness. Thus, to prove the lemma, it suffices to bound $\left|\{j : \boldsymbol{A}_j^\top \boldsymbol{\lambda}^* = f_j\}\right|$.

Now assume that Assumption 2 holds, i.e., that $\boldsymbol{\lambda}^*, \boldsymbol{\theta}^*$ are non-degenerate almost surely. Define the following sets

$$\begin{aligned}
\mathcal{J}_> &= \{j : f_j > \boldsymbol{A}_j^\top \boldsymbol{\lambda}^*, \theta_j^* = f_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^*\} \\
\mathcal{J}_< &= \{j : f_j < \boldsymbol{A}_j^\top \boldsymbol{\lambda}^*, \theta_j^* = 0\} \\
\mathcal{J}_= &= \{j : f_j = \boldsymbol{A}_j^\top \boldsymbol{\lambda}^*, \theta_j^* = f_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}^*, \theta_j = 0\} \\
\mathcal{I}_0 &= \{i : \lambda_i^* = 0\}.
\end{aligned}$$

Since $(\boldsymbol{\lambda}^*, \boldsymbol{\theta}^*)$ is a non-dengenerate, basic feasible solution, we must have that

$$|\mathcal{J}_>| + |\mathcal{J}_<| + 2\,|\mathcal{J}_=| + |\mathcal{I}_0| = n + m \implies |\mathcal{J}_=| + |\mathcal{I}_0| = m.$$

Hence, $|\mathcal{J}_=| \leq m$, which proves the lemma when Assumption 2 holds since $f \in \mathcal{F}$ was arbitrary.

Now assume Assumption 3 holds instead. By the same argument above, it suffices to bound $\{j : \boldsymbol{A}_j^\top \boldsymbol{\lambda}^* = f_j\}$. Suppose by contradiction this set is of size at least $m+1$. Then, after permuting the indices, we have that

$$\begin{pmatrix} f_j \\ \boldsymbol{A}_j \end{pmatrix}^\top \begin{pmatrix} -1 \\ \boldsymbol{\lambda}^* \end{pmatrix} = 0 \quad j = 1, \ldots, m+1.$$

These equalities show that the $m+2$ points

$$\left\{ \begin{pmatrix} f_j \\ \boldsymbol{A}_j \end{pmatrix} : j = 1, \ldots, m+1 \right\} \cup \{\boldsymbol{0}\}$$

lie in an m-dimensional flat and thus are not in general position, a contradiction. This concludes the lemma. □

The remaining two lemmas both follow directly from standard tools in empirical processs.

Lemma 2 **(Uniform Convergence of the In-Sample Bias)**. *Under the assumptions of Theorem 1, there exists a universal constant C such that for any $0 < \epsilon < 1$, with probability at least $1 - \epsilon$,*

$$\sup_{f \in \mathcal{F}, \boldsymbol{\lambda} \in \mathbb{R}_+^m} \left| (\boldsymbol{Z} - \boldsymbol{\mu})^\top \boldsymbol{x}(f, \boldsymbol{Z}, \boldsymbol{\lambda}) - \mathbb{E}\left[(\boldsymbol{Z} - \boldsymbol{\mu})^\top \boldsymbol{x}(f, \boldsymbol{Z}, \boldsymbol{\lambda}) \right] \right| \;\leq\; C \sqrt{\frac{V n \log n}{\nu_{\min}}} \log\left(\frac{2}{\epsilon}\right).$$

Proof. Using the definition of $x_j(f, Z_j, \boldsymbol{\lambda})$, the above supremum is

$$\sup_{f\in\mathcal{F},\boldsymbol{\lambda}\in\mathbb{R}_+^m} \left| \sum_{j=1}^n (Z_j-\mu_j)\mathbb{I}\left\{f(Z_j,\boldsymbol{W}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}\geq 0\right\} - \mathbb{E}\left[(Z_j-\mu_j)\mathbb{I}\left\{f(Z_j,\boldsymbol{W}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}\geq 0\right\}\right]\right|.$$

This is the maximal deviation of an empirical process. Notice that the vector $(|Z_1-\mu_1|,\ldots,|Z_n-\mu_n|)$ is an envelope for the process, and by (Gupta and Rusmevichientong, 2021, Lemma A.1), $\|\|\boldsymbol{Z}-\boldsymbol{\mu}\|_2\|_\Psi \lesssim \sqrt{\frac{n}{\nu_{\min}}}$.

Furthermore,

$$\begin{aligned}&\left|\left\{\left((Z_j-\mu_j)\mathbb{I}\left\{f(Z_j,\boldsymbol{W}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}\geq 0\right\}\right)_{j=1}^n \;:\; f\in\mathcal{F},\boldsymbol{\lambda}\in\mathbb{R}_+^m\right\}\right|\\ &\quad\leq\; \left|\left\{\left(\mathbb{I}\left\{f(Z_j,\boldsymbol{W}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}\geq 0\right\}\right)_{j=1}^n \;:\; f\in\mathcal{F},\boldsymbol{\lambda}\in\mathbb{R}_+^m\right\}\right|,\end{aligned}$$

and by assumption on the VC dimension of the class of functions $\mathcal{R}$, we have that the cardinality of this last set is at most $(n+1)^V$ (see (Wainwright, 2019, Prop. 4.18)). Applying the argument leading up to (Pollard, 1990, Eq. 7.4) thus shows that with probability at least $1-\epsilon$,

$$\sup_{f\in\mathcal{F},\boldsymbol{\lambda}\in\mathbb{R}_+^m} \left|(\boldsymbol{Z}-\boldsymbol{\mu})^\top\boldsymbol{x}(f,\boldsymbol{Z},\boldsymbol{\lambda}) - \mathbb{E}\left[(\boldsymbol{Z}-\boldsymbol{\mu})^\top\boldsymbol{x}(f,\boldsymbol{Z},\boldsymbol{\lambda})\right]\right| \;\lesssim\; \sqrt{\frac{Vn\log n}{\nu_{\min}}}\log\left(\frac{2}{\epsilon}\right).$$

This completes the proof. □

Lemma 3 **(Uniform Convergence of the Stein Correction)**. *Under the assumptions of Theorem 1, there exists a universal constant C such that for any $0<\epsilon<1$, with probability at least $1-2\epsilon$,*

$$\sup_{f\in\mathcal{F},\boldsymbol{\lambda}\in\mathbb{R}_+^m} |B(f,\boldsymbol{Z},\boldsymbol{\lambda}) - \mathbb{E}[B(f,\boldsymbol{Z},\boldsymbol{\lambda})]| \;\leq\; C\frac{\sqrt{Vn\log n\log(2/\epsilon)}}{h\sqrt{\nu_{\min}}}.$$

Proof. Write out the definition of $B(f,\boldsymbol{Z},\boldsymbol{\lambda})$ and apply the triangle inequality to upper bound

$$\begin{aligned}&\sup_{f\in\mathcal{F},\boldsymbol{\lambda}\in\mathbb{R}_+^m} |B(f,\boldsymbol{Z},\boldsymbol{\lambda}) - \mathbb{E}[B(f,\boldsymbol{Z},\boldsymbol{\lambda})]|\\ &\quad\leq \sup_{f\in\mathcal{F},\boldsymbol{\lambda}\in\mathbb{R}_+^m}\left|\sum_{j=1}^n \frac{1}{2h\sqrt{\nu_j}}\left(x_j(f,Z_j+\frac{h}{\sqrt{\nu_j}},\boldsymbol{\lambda}) - \mathbb{E}\left[x_j(f,Z_j+\frac{h}{\sqrt{\nu_j}},\boldsymbol{\lambda})\right]\right)\right| &\text{(8a)}\\ &\quad\quad+ \sup_{f\in\mathcal{F},\boldsymbol{\lambda}\in\mathbb{R}_+^m}\left|\sum_{j=1}^n \frac{1}{2h\sqrt{\nu_j}}\left(x_j(f,Z_j-\frac{h}{\sqrt{\nu_j}},\boldsymbol{\lambda}) - \mathbb{E}\left[x_j(f,Z_j-\frac{h}{\sqrt{\nu_j}},\boldsymbol{\lambda})\right]\right)\right| &\text{(8b)}\end{aligned}$$

We bound each of these suprema separately.

Consider Eq. (8a). This is maximum deviation of an empirical process. Notice the constant vector $(\frac{1}{2h\sqrt{\nu_1}}, \ldots, \frac{1}{2h\sqrt{\nu_1}})$ is an envelope with size at most $\frac{\sqrt{n}}{2h\sqrt{\nu_{\min}}}$. Furthermore, by the assumption on the class $\mathcal{R}$,

$$\begin{aligned}
&\left|\left\{ \left(x_j(f, Z_j + h/\sqrt{\nu_j}, \boldsymbol{\lambda})\right)_{j=1}^n \;:\; f \in \mathcal{F}, \boldsymbol{\lambda} \in \mathbb{R}_+^m \right\}\right| \\
&\quad = \left|\left\{ \left(\mathbb{I}\left\{f(Z_j + h/\sqrt{\nu_j}, \boldsymbol{W}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \ge 0\right\}\right)_{j=1}^n \;:\; f \in \mathcal{F}, \boldsymbol{\lambda} \in \mathbb{R}_+^m \right\}\right| \\
&\quad \le (n+1)^V.
\end{aligned}$$

Hence, by the argument leading up to (Pollard, 1990, Eq. 7.2), we have that for any $0 < \epsilon < 1$, with probability at least $1 - \epsilon$,

$$Eq.\ (8\text{a}) \lesssim \frac{\sqrt{Vn \log n \log(2/\epsilon)}}{h\sqrt{\nu_{\min}}}.$$

A nearly identical argument shows that with probability at least $1 - \epsilon$,

$$Eq.\ (8\text{b}) \lesssim \frac{\sqrt{Vn \log n \log(2/\epsilon)}}{h\sqrt{\nu_{\min}}}.$$

Combining proves the lemma. □

4. Comparison to Gupta and Rusmevichientong (2021)

We next compare and contrast Theorems 1 and 2 to Gupta and Rusmevichientong (2021, Theorem 4.3).

4.1. Specializing to the setting of Gupta and Rusmevichientong (2021)

Gupta and Rusmevichientong (2021, Theorem 4.3) treats a specific "Bayes-Inspired Policy Class" of the form

$$\boldsymbol{x}(\tau, \boldsymbol{Z}) \in \operatorname*{arg\,max}_{\boldsymbol{x} \in [0,1]^n} \quad \frac{\nu_{\min} + \tau}{\nu_{\min}} \sum_{j=1}^n \frac{\nu_j}{\nu_j + \tau} Z_j x_j \tag{9}$$

$$\text{s.t} \quad \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}. \tag{10}$$

We assume here that ties are broken such that $\boldsymbol{x}(\tau, \boldsymbol{Z})$ is right continuous.

These policies are subsumed within our setting by letting $W_j = 1/\nu_j$ and $\mathcal{F} = \{(z, W) \mapsto \frac{(\nu_{\min}+\tau)W^{-1}}{\nu_{\min}(W^{-1}+\tau)} z \;:\; \tau \ge 0\}$. The corresponding class $\mathcal{R}(\mathcal{F})$ can be written as

$$\left\{(z, \nu, \boldsymbol{A}) \mapsto \mathbb{I}\left\{\nu z - \nu \boldsymbol{A}^\top \boldsymbol{\lambda} - \tau \boldsymbol{A}^\top \boldsymbol{\lambda} \ge 0\right\} \;:\; \tau \ge 0, \boldsymbol{\lambda} \in \mathbb{R}_+^m.\right\}$$

Since the argument of the indicator belongs to a $2m+1$ dimensional vector space of functions, the VC dimension of $\mathcal{R}(\mathcal{F})$ is at most $V = 2m+1$. Thus, Assumption 1 holds for this class.

Assumption 3 also holds. Specifically, for any fixed τ, if $\{\boldsymbol{A}_j : 1 \le j \le n\} \cup \{\boldsymbol{0}\}$ are in general position, then the set $S(f, \boldsymbol{Z})$ is in general position almost surely since $\boldsymbol{Z}$ are Gaussian. If $\mathcal{F}$ were countable, this would be enough to ensure Assumption 3 holds. Unfortunately, $\mathcal{F}$ is not countable, however, we can replace it with a countably infinite subset before applying the theorem. Specifically, since $\boldsymbol{x}(\tau, \boldsymbol{Z})$ is right-continuous, for any τ, there exists a rational τ_0 such that $\boldsymbol{x}(\tau, \boldsymbol{Z}) = \boldsymbol{x}(\tau_0, \boldsymbol{Z})$. Hence,

$$\sup_{f \in \mathcal{F}} \left| (\boldsymbol{Z} - \boldsymbol{\mu})^\top \boldsymbol{x}(f, \boldsymbol{Z}) - B(f, \boldsymbol{Z}, \boldsymbol{\lambda}(f, \boldsymbol{Z})) \right| = \sup_{f \in \mathcal{F}_0} \left| (\boldsymbol{Z} - \boldsymbol{\mu})^\top \boldsymbol{x}(f, \boldsymbol{Z}) - B(f, \boldsymbol{Z}, \boldsymbol{\lambda}(f, \boldsymbol{Z})) \right|,$$

where $\mathcal{F}_0 = \{(z, W) \mapsto \frac{W^{-1}}{W^{-1} + \tau} z : \tau \ge 0, \tau \text{ is rational}\}$. Then since $\mathcal{F}_0 \subseteq \mathcal{F}$, $\mathcal{F}_0$ satisfies both Assumptions 1 and 3 and we can apply Theorem 2 to $\mathcal{F}_0$.

Finally, observe that for the above policy class, our correction $B(f, \boldsymbol{Z}, \boldsymbol{\lambda}(f, \boldsymbol{Z}))$ simplifies to the "custom" Bayes correction B^{Bayes} of Gupta and Rusmevichientong (2021).

4.2. Rates of Convergence

In terms of rates, as discussed in (Gupta and Rusmevichientong, 2021), in typical applications $\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{f}^{OR}, \boldsymbol{Z}) = O_p(n)$. In that sense, Corollary 1 shows that the relative regret of optimizing our estimator is at most

$$O_p\left(\left(\frac{m}{n} + V^{1/3} n^{-1/3} \right) \sqrt{\frac{\log n}{\nu_{\min}}} \right).$$

Since $V > m$ in most settings, this bound shows the relative regret vanishes so long as $V \ll \frac{n}{\log^{3/2} n}$. Moreover, the dependence in n (i.e. $\tilde{O}_p(n^{-1/3})$ matches (Gupta and Rusmevichientong, 2021) up to logarithmic factors.

References

Bertsimas, Dimitris and John N Tsitsiklis (1997). *Introduction to linear optimization.* Vol. 6. Athena Scientific Belmont, MA.

Gupta, Vishal, Michael Huang, and Paat Rusmevichientong (2022a). “Debiasing in-sample policy performance for small-data, large-scale optimization”. In: *Operations Research*. Forthcoming.

– (2022b). “Learning Policy Performance Under Weakly-Coupled Settings”. In: *Working Paper*.

Gupta, Vishal and Paat Rusmevichientong (2021). “Small-data, large-scale linear optimization with uncertain objectives”. In: *Management Science* 67.1, pp. 220–241.

Pollard, David (1990). “Empirical Processes: Theory and Applications”. In: Ims.

Wainwright, Martin J (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Vol. 48. Cambridge University Press.