Debiasing In-Sample Policy Performance for Small-Data, Large-Scale Optimization

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

Motivated by the poor performance of cross-validation in settings where data are scarce, we propose a novel estimator of the out-of-sample performance of a policy in data-driven optimization. Our approach exploits the optimization problem's sensitivity analysis to estimate the gradient of the optimal objective value with respect to the amount of noise in the data and uses the estimated gradient to debias the policy's in-sample performance. Unlike cross-validation techniques, our approach avoids sacrificing data for a test set, utilizes all data when training and, hence, is well-suited to settings where data are scarce. We prove bounds on the bias and variance of our estimator for optimization problems with uncertain linear objectives but known, potentially non-convex, feasible regions. For more specialized optimization problems where the feasible region is "weakly-coupled" in a certain sense, we prove stronger results. Specifically, we provide explicit high-probability bounds on the error of our estimator that hold uniformly over a policy class and depends on the problem's dimension and policy class's complexity. Our bounds show that under mild conditions, the error of our estimator vanishes as the dimension of the optimization problem grows, even if the amount of available data remains small and constant. Said differently, we prove our estimator performs well in the small-data, large-scale regime. Finally, we numerically compare our proposed method to state-of-the-art approaches through a case-study on dispatching emergency medical response services using real data. Our method provides more accurate estimates of out-of-sample performance and learns better-performing policies.

Key words: Large-scale, data-driven optimization. Small-data, large-scale regime. Cross-validation.

1. Introduction

The crux of data-driven decision-making is using past data to identify decisions that will have good out-of-sample performance on future, unseen data. Indeed, estimating out-of-sample performance is key to both policy evaluation (assessing the quality of a given policy), and to policy learning (identifying the best policy from a potentially large set of candidates). Estimating out-of-sample performance, however, is non-trivial. Naive estimates that leverage the same data to train a policy and to evaluate its performance often suffer a systematic, optimistic bias, referred to as "in-sample bias" in machine learning and the "optimizer's curse" in optimization (Smith and Winkler 2006).

Consequently, cross-validation and sample-splitting techniques have emerged as the gold-standard approach to estimating out-of-sample performance. Despite the multitude of cross-validation methods, at a high-level, these methods all proceed by setting aside a portion of the data as "testing" data *not* to be used when training the policy, and then evaluating the policy on these testing data. The policy's performance on testing data then serves as an estimate of its performance on future, unseen data, thereby circumventing the aforementioned in-sample bias. Cross-validation is ubiquitous in machine learning and statistics with provably good performance in large sample settings (Bousquet and Elisseeff 2001, Kearns and Ron 1999).

Unfortunately, when data are scarce, cross-validation can perform poorly. Gupta and Rusmevichientong (2021) prove that for the small-data, large-scale regime — when the number of uncertain parameters in an optimization problem is large but the amount of relevant data per parameter is small — each of hold-out, 5-fold, 10-fold and leave-one-out cross validation can have poor performance when used for policy learning, even for very simple optimization problems. Shao (1993) observes a similar failure for leave-one-out cross-validation in a high-dimensional linear regression setting. The key issue in both cases is that when relevant data are *scarce*, estimates of uncertain parameters are necessarily imprecise, and omitting even a small amount of data when training a policy dramatically degrades its performance. Hence, the performance of a policy trained *with a portion of the data* on the test set is not indicative of the performance of the policy trained *with all the data* on future unseen data. We elucidate this phenomenon with a stylized example in Section 1.2 below.

Worse, this phenomenon is not merely an intellectual curiosity. Optimization problems plagued by numerous low-precision estimates are quite common in modern, large-scale operations. For example, optimization models for personalized pricing necessarily include parameters for each distinct customer type, and these parameters can be estimated only imprecisely since relevant data for each type are limited. Similar issues appear in large-scale supply-chain design, promotion optimization, and dispatching emergency response services; see Section 2 for further discussion.

In this paper, we propose a new method for estimating out-of-sample performance without sacrificing data for a test set. The key idea is to debias the in-sample performance of the policy trained on all the data. Specifically, we focus on the optimization problem

$$\boldsymbol{x}^* \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}\subseteq[0,1]^n} \quad \boldsymbol{\mu}^\top \boldsymbol{x} \tag{1.1}$$

where $\mathcal{X}$ is a known, potentially non-convex feasible region contained within $[0,1]^n$, and $\boldsymbol{\mu}\in\mathbb{R}^n$ is an unknown vector of parameters. We assume access to a vector $\boldsymbol{Z}$ of noisy, unbiased predictions of $\boldsymbol{\mu}$ (based on historical data) and are interested in constructing a policy $\boldsymbol{x}(\boldsymbol{Z})$ with good out-of-sample performance $\boldsymbol{\mu}^\top\boldsymbol{x}(\boldsymbol{Z})$. (For clarity, the in-sample performance of $\boldsymbol{x}(\boldsymbol{Z})$ is $\boldsymbol{Z}^\top\boldsymbol{x}(\boldsymbol{Z})$.) Note that for many applications of interest, $\boldsymbol{\mu}^\top\boldsymbol{x}^* = O(n)$ as $n\to\infty$; i.e., the full-information solution grows at least linearly as the dimension grows. Hence, the unknown out-of-sample performance $\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z})$ must also be at least $O_p(n)$ as $n \to \infty$.[1] See Section 2 for examples.

Despite its simplicity, Problem (1.1) subsumes a wide class of optimization problems because $\mathcal{X}$ can be non-convex and/or discrete. This class includes mixed-binary linear optimization problems such as facility location, network design, and promotion maximization. By transforming decision variables, even some non-linear optimization problems such as personalized pricing can be rewritten as Problem (1.1); see Section 2. In this sense, Problem (1.1) is fairly general.

Our estimator applies to affine plug-in policies which are formally defined in Section 2. Loosely, affine plug-in policies are those obtained by solving Problem (1.1) after "plugging-in" some estimator $\boldsymbol{r}(\boldsymbol{Z})$ in place of $\boldsymbol{\mu}$, and $\boldsymbol{r}(\boldsymbol{Z})$ depends affinely on $\boldsymbol{Z}$. Our analysis of this class subsumes many policies used in practice and previously studied (mostly in the large sample setting) in the literature including Sample Average Approximation (SAA), estimate-then-optimize policies based on regression, the Bayes-Inspired policies of Gupta and Rusmevichientong (2021), and the SPO+ policy of Elmachtoub and Grigas (2021). Moreover, this policy class also includes policies motivated by linear smoothers (Buja et al. 1989) including k-nearest neighbors, local-polynomial regression, and smoothing splines. Thus, our estimator provides a theoretically rigorous approach to assessing the quality of optimization policies based on many modern machine learning techniques.

We debias $\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z})$ by exploiting the structure of Problem (1.1) with the plug-in $\boldsymbol{r}(\boldsymbol{Z})$. Specifically, by leveraging this problem's sensitivity analysis, we approximately compute the gradient of its objective value with respect to the variance of $\boldsymbol{Z}$, and use the estimated gradient to debias the in-sample performance. We term this correction the *Variance Gradient Correction* (VGC). Because our method strongly exploits optimization structure, the VGC is Lipschitz continuous in the plug-in values $\boldsymbol{r}(\boldsymbol{Z})$. This continuity is not enjoyed by other techniques such as those in Gupta and Rusmevichientong (2021).

Although the VGC's continuity may seem like mere a mathematical nicety, empirical evidence suggests it improves empirical performance. Similar empirical phenomena – where an estimator that varies smoothly in the data often outperforms similar estimators that change discontinuously – are rife in machine learning. Compare k-nn regression with Gaussian kernel smoothing (Friedman et al. 2001), CART trees with bagged trees (Breiman 1996), or best subset-regression with lasso regression (Hastie et al. 2020a). Theoretically, we exploit this smoothness heavily to establish bounds that hold uniformly over the policy class.

[1] Following Van der Vaart (2000), we say a sequence of random variables $X_n = O_p(a_n)$ if the sequence X_n/a_n is stochastically bounded, i.e., for every $\epsilon > 0$, there exists finite $M > 0$ and finite $N > 0$ such that $\mathbb{P}\{X_n/a_n \geq M\} < \epsilon$, for all $n > N$.

Specifically, we show that, when $\boldsymbol{Z}$ is approximately Gaussian, the bias of our estimator for out-of-sample performance is $\tilde{O}(h)$ as $h \to 0$, where h is a user-defined parameter that controls the accuracy of our gradient estimate (Theorem 3.2). Characterizing the variance is more delicate. We introduce the concept of *Average Solution Instability*, and prove that if the instability of the policy vanishes at rate $O(n^{-\alpha})$ for $\alpha \geq 0$, then the variance of our estimator is roughly $O(n^{3-\alpha}/h)$. Collectively, these results suggest interpreting h as a parameter controlling the bias-variance tradeoff of our estimator. Moreover, when $\alpha > 1$, the variance of our estimator is $o(n^2)$. Since, as mentioned, the unknown out-of-sample performance often grows at least linearly in n, i.e., $\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}) = O_p(n)$, our variance bound shows that when $\alpha > 1$ and n is large, the stochastic fluctuations of our estimator are negligible relative to the out-of-sample performance. In other words, our estimator is quite accurate in these settings.

Our notion of Average Solution Instability is formally defined in Section 3.3. Loosely, it measures the expected change in the j^{th} component of the policy after replacing the k^{th} data point with an i.i.d. copy, where j and k are chosen uniformly at random from $\{1, \ldots, n\}$. This notion of stability is similar to hypothesis stability (Bousquet and Elisseeff 2001), but, to the best of our knowledge, is distinct. Moreover, insofar as we expect that a small perturbation of the data is unlikely to have a large change on the solution for most real-world, large-scale optimization problems, we expect Average Solution Instability to be small and our estimator to have low variance.

We then prove stronger high-probability tail bounds on the error of our estimator for two special classes of "weakly-coupled" instances of Problem (1.1): weakly-coupled-by-variables and weakly-coupled-by-constraints. In Section 4.1, we consider problems that are weakly-coupled-by-variables, i.e., problems that decouple into many, disjoint subproblems once a small number of decision variables are fixed. In Section 4.2 we consider problems that are weakly-coupled-by-constraints, i.e., problems that decouple into many, disjoint subproblems once a small number of constraints are removed. For each problem class, we go beyond bounding the variance to provide an explicit tail bound on the relative error of our estimator that holds uniformly over the policy class. We show that for problems weakly-coupled-by-variables the relative error scales like $\tilde{O}(C_{\mathsf{PI}} \frac{\text{polylog}(1/\epsilon)}{\sqrt[3]{n}})$ where C_{PI} is a constant measuring the complexity of the policy class; see Theorem 4.3. Similarly, we show the relative error for problems weakly-coupled-by-constraints scales like $\tilde{O}(C_{\mathsf{PI}} \frac{\text{polylog}(1/\epsilon)}{\sqrt[4]{n}})$, where C_{PI} measures both the complexity of the policy class and number of constraints of the problem; see Theorem 4.7. Importantly, since these bounds hold uniformly, our debiased in-sample performance can be used both for policy evaluation and policy learning, even when data are scarce, so long as n (the dimension of the problem) is sufficiently large. Said differently, our estimator of out-of-sample performance is particularly well-suited to small-data, large-scale optimization.

Admittedly, weakly-coupled problems as described above do not cover all instances of Problem (1.1) and the appropriateness of modeling $\boldsymbol{Z}$ as approximately Gaussian is application specific. Nonetheless, our results and their proofs highly suggest our estimator will have strong performance whenever the underlying optimization problem is well-behaved enough for certain uniform laws of large numbers to pertain.

Finally, to complement these theoretical results, we perform a numerical case study of dispatching emergency medical services with real data from cardiac arrest incidents in Ontario, Canada. With respect to policy evaluation and learning, we show that our debiased in-sample performance outperforms both traditional cross-validation methods and the Stein correction of Gupta and Rusmevichientong (2021). In particular, while the bias of cross-validation is non-vanishing as the problem size grows for a fixed amount of data, the bias of our VGC converges to zero. Similarly, while both the Stein correction and our VGC have similar asymptotic performance, the smoothness of VGC empirically leads to lower bias and variance for moderate and sized instances.

1.1. Our Contributions

We summarize our contributions as follows:

1. We propose an estimator of out-of-sample performance for Problem (1.1) by debasing in-sample performance through a novel *Variance Gradient Correction* (VGC). Our VGC applies to a general class of affine plug-in policies that subsumes many policies used in practice. Most importantly, unlike cross-validation, VGC does *not* sacrifice data when training, and, hence, is particularly well-suited to settings where data are scarce.
2. We prove that under some assumptions on the data-generating process, for general instances of Problem (1.1), the bias of our estimator is at most $\tilde{O}(h)$ as $h \to 0$, where h is a user-defined parameter. For policy classes that satisfy a certain Average Solution Instability condition, we also prove that that its variance scales like $o\left(\frac{n^2}{h}\right)$ as $n \to \infty$.
3. We prove stronger results for instances of Problem (1.1) in which the feasible region is only weakly-coupled. When the feasible region is weakly-coupled by variables, we prove that, with probability at least $1 - \epsilon$, debiasing in-sample performance with our VGC recovers the true out-of-sample performance up to relative error that is at most $\tilde{O}\left(C_{\mathsf{PI}} \frac{\log(1/\epsilon)}{n^{1/3}}\right)$ as $n \to \infty$, uniformly over the policy class, where C_{PI} is a constant that measures the complexity of the plug-in policy class (Theorem 4.3). Similarly, for certain linear optimization problems that are weakly coupled by constraints, we prove that, with probability at least $1 - \epsilon$, debiasing in-sample performance with VGC estimates the true out-of-sample performance uniformly over the policy class with relative error that is at most $\tilde{O}\left(C_{\mathsf{PI}} \frac{\sqrt{\log(1/\epsilon)}}{n^{1/4}}\right)$ where C_{PI} is a constant measuring the complexity of the plug-in policy class and the number of of constraints of the

problem. (Theorem 4.7). We stress that since both these bounds hold uniformly, our debiased in-sample performance can not only be used for policy evaluation, but also policy learning, even when data are scarce, so long as n (the size of the problem) is sufficiently large.

4. Finally, we present a numerical case study based on real data from dispatching emergency response services to highlight the strengths and weaknesses of our approach relative to cross-validation and the Stein correction of Gupta and Rusmevichientong (2021). Overall, we find that since our VGC exploits the optimization structure of Problem (1.1), it outperforms the benchmarks when the number of uncertain parameters is sufficiently large. Additionally, in settings where the signal to noise ratio is low, VGC more effectively balances the bias-variance trade-off than cross-validation which can be quite sensitive to the number of folds used.

1.2. A Motivating Example: Poor Performance of Cross-Validation with Limited Data

Before proceeding, we present an example that highlights the shortcomings of cross-validation and the benefits of our method when data are limited. Consider a special case of Problem (1.1)

$$\max_{\boldsymbol{x}\in\{0,1\}^n} \sum_{j=1}^{n} \mu_j x_j \tag{1.2}$$

where the true parameters $\boldsymbol{\mu} \in \{-1,1\}^n$ are unknown, but we observe S samples $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_S$ where $\boldsymbol{Y}_i \in \mathbb{R}^n$ and $\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{\mu}, 2\boldsymbol{I})$ for all i and $\boldsymbol{I}$ is the identity matrix. A standard data-driven policy in this setting is Sample Average Approximation (SAA), also called empirical risk minimization, which prescribes the policy

$$\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z}) \in \arg\max_{\boldsymbol{x}\in\{0,1\}^n} \sum_{j=1}^{n} Z_j x_j \qquad \text{where} \qquad Z_j = \frac{1}{S}\sum_{i=1}^{S} Y_{ij}.$$

The key question, of course, is "What is SAA's out-of-sample behavior $\boldsymbol{\mu}^\top \boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z})$?"

To study this question, the left panel of Fig. 1 shows different estimators for the out-of-sample performance of SAA $\mathbb{E}[\boldsymbol{\mu}^\top \boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z})]$ when $S = 3$ in Problem (1.2). The right panel shows the expected relative error (with respect to the oracle) of these estimators as the number of samples S grows. To account for the noise level of the samples, we plot the estimation error with respect to the signal-to-noise ratio (SNR) of Z_j[2]. For reference, Hastie et al. (2020b) argues that SNR greater than 1 is "rare" when working with "noisy, observational data," and an SNR of 0.25 is more "typical."

[2] Following Hastie et al. (2020b), we define $SNR = Var(\mu_\pi)/Var(Z_\pi) = \frac{S}{2n}\sum_{j=1}^{n}\left(\mu_j - \frac{1}{n}\sum_{i=1}^{n}\mu_i\right)^2$ where π is an index drawn uniformly random from 1 to n.

	$\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z})$	$\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z}^{-i})$
In-Sample	18.36	22.33
Cross-Val	−1.86	−9.98
Our Method	2.95	−1.89
Oracle	2.97	−1.87

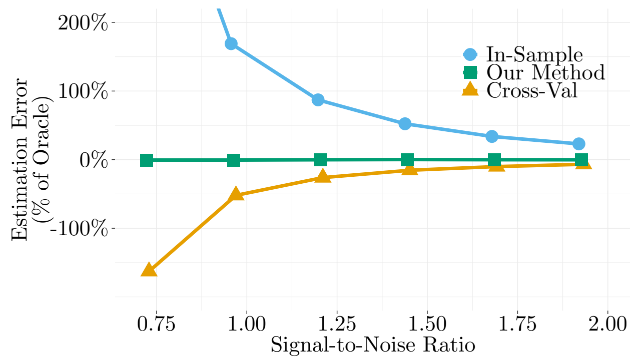

Figure 1 Expected Estimates of Out-of-Sample Performance by Policy for Problem (1.2). In the left table, we take $n = 100$, $S = 3$ and $\mu_j = 1$ if $j \le 14$ and $\mu_j = -1$ otherwise. We estimate the expected out-of-sample perf. across 1,000,000 simulations. Std. errors are less than 0.005. In the right graph, we plot the bias of the estimates with respect to the expected out-of-sample performance as we increase the signal-to-noise ratio. The In-Sample error not shown at 0.75 SNR exceeds 500%.

The first row of Fig. 1 presents the in-sample performance, i.e, the objective of the SAA problem. As expected, we see in-sample performance significantly over-estimates the out-of-sample performance. The right panel of Fig. 1 suggests this effect persists across SNRs and the relative error is at least 23% for SNRs less than 2.

The second row of the left panel of Fig. 1 shows the leave-one-out cross-validation error, which aims to correct the over-optimistic bias and computes $\frac{1}{S}\sum_{i=1}^{S} \boldsymbol{Y}_i^\top \boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z}^{-i})$, where $\boldsymbol{Z}^{-i} = \frac{1}{S-1}\sum_{j\neq i} \boldsymbol{Y}_j$. Cross-validation is also fairly inaccurate, suggesting SAA performs worse than the trivial, non-data-driven policy $\boldsymbol{x} = \boldsymbol{0}$, which has an out-of-sample performance of 0. In the right panel, this incorrect implication occurs for SNRs less than about 0.875.

Why does cross-validation perform so poorly? *By construction* cross-validation omits some data in training, and hence, does *not* estimate the out-of-sample performance of $\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z})$, but rather, that of $\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z}^{-1})$.[3] From the second column of Fig. 1, we see the cross-validation estimate *does* nearly match the true (oracle) performance of $\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z}^{-1})$. When data are scarce, sacrificing even a small amount of data in training can dramatically degrade a policy. As seen in the right panel of Fig. 1, this phenomenon is non-negligible (at least 10% relative error) for signal-to-noise ratios less than or equal to 1.75. Thus, the performance of $\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z}^{-1})$ may not always be a good proxy of the performance of $\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z})$.

How then might we resolve the issue? The third row of the left panel of Fig. 1 presents our estimator based on debiasing the in-sample performance of $\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z})$ with our VGC. Our estimate is essentially unbiased (see also Theorem 3.2 below). The right panel of Fig. 1 confirms this excellent behavior across a range of signal-to-noise ratios. Finally, although this example focuses on the

[3] For clarity, this is the same as the performance as $\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z}^{-2})$ because the data are i.i.d.

bias of our estimator, our results in Section 4 are stronger and bound the (random) error of our estimator directly, rather than its expectation.

1.3. Relationship to Prior Work

Cross-validation is the gold-standard for estimating out-of-sample performance in the large-sample regime with i.i.d. data; see Bousquet and Elisseeff (2001), Kearns and Ron (1999) for some fundamental results. As discussed above, when estimating the performance of a fixed-policy, these approaches entail sacrificing some data in the training step to set aside for validation, and, hence, may be ill-suited to data-scarce settings. Similar issues arise in a variety of other sample-splitting methods, including "honest-trees" Wager and Athey (2018) and most forms of doubly-robust estimation Dudík et al. (2011). By contrast, our VGC based approach to debiasing the in-sample performance effectively uses all the data when training, making it somewhat better suited to data-scarce settings and small-data, large-scale optimization.

Our work also contributes to the growing literature on "optimization-aware" estimation. These works employ a variety of names including operational statistics (Liyanage and Shanthikumar 2005), learning-enabled optimization (Deng and Sen 2018), decision-focused learning (Wilder et al. 2019a), end-to-end learning (Wilder et al. 2019b) and task-based learning (Donti et al. 2017). Fundamentally, this area of research seeks estimators that optimize the out-of-sample performance of a policy in a downstream optimization problem rather than the prediction error of the estimate. Closest to our work is the "Smart 'Predict then Optimize'" framework studied in Elmachtoub and Grigas (2021), Elmachtoub et al. (2020), and El Balghiti et al. (2019). These works also study Problem (1.1), but in a slightly different data setting, and propose policy selection methods for affine and tree-based policies, respectively. Also related is Ito et al. (2018) which develops an unbiased estimate of the sample average approximation (SAA) policy for Problem (1.1), but does not consider higher level moments, policy evaluation for other policies, or policy learning.

To the best of our knowledge, most works on optimization-aware estimation focus on analyzing the particular policy proposed in the respective work. An exception is El Balghiti et al. (2019), which establishes generalization guarantees uniformly over a policy class. Perhaps more importantly, all of these works focus on settings where the amount of data grows large. By contrast, our work develops a general-purpose estimator that is valid for all policies in a class, and, most importantly, centers on the small-data, large-scale regime where data are scarce.

There has been considerably less work in the small-data, large-scale regime, most notably Gupta and Kallus (2021) and Gupta and Rusmevichientong (2021). Of these, Gupta and Rusmevichientong (2021), henceforth GR 2021, is most closely related to our work. Loosely, GR 2021 study a class of weakly coupled linear optimization problems and propose an estimator of the out-of-sample performance based on Stein's Lemma. By leveraging a careful duality argument, the authors prove that the estimation error of their procedure vanishes in both the large-sample and small-data, large-scale regime.

Our work differs in two important respects: First, our estimator applies to a more general class of problems and more general policy classes. Indeed, we focus on Problem (1.1) with specialized results for weakly-coupled instances. Our weakly-coupled by constraints variant in Section 4.2 mirrors the setting of GR 2021, and our weakly-coupled by variables variant in Section 4.1 is more general, allowing us to model, for example, discrete optimization problems. Moreover, our affine plug-in policy class significantly generalizes the "Bayes-Inspired" policy class of GR 2021 by incorporating covariate information.

The second important difference from GR 2021 relates to exploiting optimization structure in Problem (1.1). GR 2021 fundamentally relies on Stein's lemma, a result which applies to general functions and does not specifically leverage optimization structure. By contrast, our method directly leverages the structure of Problem (1.1) through its sensitivity analysis and Danskin's theorem. By leveraging optimization structure, our VGC is, by construction, continuous in the policy class. Evidence from Section 5 suggests this smoothness yields an empirical advantage of our method.

Finally, our work also contributes to a growing literature on debiasing estimates in high-dimensional statistics, most notably for LASSO regression (Javanmard and Montanari 2018, Zhang and Zhang 2014) and M-estimation (Javanmard and Montanari 2014). Like these works, VGC involves estimating a gradient of the underlying system and using this gradient information to form a correction. Unlike these works, however, our gradient estimation strongly leverages ideas from sensitivity analysis in optimization. Moreover, the proofs of our performance guarantees involve substantively different mathematical techniques.

2. Model

As mentioned, our focus is on data-driven instances of Problem (1.1) where the feasible region $\mathcal{X}$ is known, but the parameters $\boldsymbol{\mu}$ are unknown. Despite its simplicity, several applications can be modeled in this form after a suitable transformation of variables.

Example 2.1 (Promotion Optimization) Promotion optimization is an increasingly well-studied application area (Cohen et al. 2017, Baardman et al. 2019). Our formulation mirrors a formulation from the ride-sharing company Lyft around incentive allocation (Schmoys and Wang 2019), but also resembles the online advertising portfolio optimization problem (Rusmevichientong (2006), Pani and Sahin (2017),GR 2021).

The decision-maker (platform) seeks to allocate J different types of coupons (promotions) to K different customer (passenger) types. Coupons are costly, and there is a finite budget C available.

Let μ_{jk} be the reward (induced spending) and c_{jk} be the cost of assigning coupon type j to customer type k. Using x_{jk} to denote the fraction of customers of type k who receive coupons of type j, we can formulate the following linear optimization problem of the form of Problem (1.1).

$$\max_{\boldsymbol{x} \geq 0} \left\{ \sum_{k=1}^{K} \sum_{j=1}^{J} \mu_{jk} x_{jk} \ : \ \sum_{j=1}^{J} x_{jk} \leq 1 \quad \text{for each } k = 1, \ldots, K, \quad \sum_{k=1}^{K} \sum_{j=1}^{J} c_{jk} x_{jk} \leq C \right\} .$$

In typical instances, the cost c_{jk} are likely known (a "\$10 off" coupon costs \$10), whereas the reward μ_{jk} must be estimated from historical data. In settings with many types of coupons and customers, we might further expect that the reward estimates may be imprecise.

Some reflection suggests many linear optimization problems including shortest-path with uncertain edge costs, or even binary linear optimization problems like multi-choice knapsack with uncertain rewards can be cast as above.

We next observe that some two-stage linear optimization problems can also be framed as Problem (1.1).

Example 2.2 (Drone-Assisted Emergency Medical Response) In recent years, emergency response systems have begun utilizing drones as part of their operations, specifically for rapid delivery of automatic electronic defibrillators (AEDs) for out-of-hospital cardiac arrests (OHCA) (Sanfridsson et al. 2019, Cheskes et al. 2020). The intention is that a drone might reach a patient in a remote region before a dispatched ambulance, and (untrained) bystanders can use the AED to assist the patient until the ambulance arrives. Consequently, researchers have begun studying both how to design a drone-assisted emergency response network (where to locate depots) (Boutilier and Chan 2019) and how to create optimal dispatch rules (to which locations should we allocate a drone and from which depot) (Chu et al. 2021). Combining these two problems yields a two-stage optimization problem, similar to facility location, aimed at minimizing the response time.

Namely, let μ_{kl} be the response time of drone routed from a source l for to a patient location k, $l = 1, \ldots, L$ and $k = 1, \ldots, K$. Let $y_l \in \{0, 1\}$ be a binary decision variable encoding if we build a drone depot at location l, and let x_{kl} be a binary decision variable encoding if, after building the network, we should dispatch a drone from location l to patient requests at location k. We let x_{k0} be the choice not to route a drone (sending only an ambulance) to location k and μ_{k0} be the corresponding ambulance travel time. Suppose we can build at most B depots. Then, we have the following optimization problem.

$$\begin{aligned} \min_{\boldsymbol{y} \in \{0,1\}^L, \ \boldsymbol{x} \in \{0,1\}^{K \times L}} \quad & \sum_{k=1}^{K} \sum_{l=0}^{L} \mu_{kl} x_{kl}, \\ \text{s.t.} \quad & \sum_{l=1}^{L} y_l \leq B, \quad x_{kl} \leq y_l, \quad \sum_{l=0}^{L} x_{kl} = 1, \qquad \forall k = 1, \ldots, K, \ \ l = 1, \ldots, L. \end{aligned}$$

Insofar as some drone response times are difficult to predict (depending on the weather, local environment, ability of bystanders to locate and use the drone's payload), we expect in typical instances that estimates μ_{kl} may be imprecise.

Interestingly, some non-linear problems can be transformed into the form of Problem (1.1).

Example 2.3 (Personalized Pricing) Personalized pricing strategies seek to assign a tailored price to each of many customer types reflecting their heterogeneous willingness-to-pay (Cohen et al. 2021, Javanmard et al. 2020, Aouad et al. 2019). One simple formulation posits distinct demand models $D_j(p) = m_j \phi_j(p) + b_j$ in each customer segment j, for some decreasing function $\phi_j(p)$. This yields the revenue maximization problem

$$\max_{\boldsymbol{p} \geq \boldsymbol{0}} \quad \sum_{j=1}^{n} m_j p_j \phi_j(p_j) + b_j p,$$

where p_j is the price for the j^{th} segment. We can cast this nonlinear objective in the form Problem (1.1) by transforming variables,

$$\max_{\boldsymbol{p} \geq \boldsymbol{0}, \boldsymbol{x}} \left\{ \sum_{j=1}^{n} m_j x_j + b_j p_j \ : \ x_j = p_j \phi_j(p_j) \quad \text{for each } j = 1, \ldots, n \right\}$$

where the resulting feasible region is now non-convex. In typical settings, we expect the parameters m_j, b_j are unknown, and estimated via machine learning methods (Aouad et al. 2019). When there are many customer types, these estimates may be imprecise for rarely occurring types.

Finally, we mention as an aside that some dynamic programs like the economic lot-sizing problem in inventory management can be cast in the above form through a careful representation of the dynamic program; see Elmachtoub and Grigas (2021) for details.

2.1. Data

Following GR 2021, we do not assume an explicit data generation procedure. Rather, we assume that as a result of analyzing whatever raw data are available, we obtain noisy, unbiased predictions $\boldsymbol{Z}$ such that $\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{\mu}$ with known precision $\mathbb{E}[(Z_j - \mu_j)^2] = 1/\nu_j$ for $j = 1, \ldots, n$. These predictions might arise as sample averages as in Section 1.2, or as the outputs of some pre-processing regression procedure. We further assume that for each $j = 1, \ldots, n$, we observe a non-random covariate of feature data $\boldsymbol{W}_j \in \mathbb{R}^p$, which may (or may not) be informative for the unknown μ_j.

We believe this set-up reasonably reflects many applications. In the case of drone-assisted emergency response (Example 2.2), $\boldsymbol{W}_j$ encodes features that are predictive of EMS response times such as physical road distance between the patient and the responding ambulance, time of day, day of week, and weather conditions (Chu et al. 2021), while Z_{k0} may be an average of historical response times to location k.

An advantage of modeling $\boldsymbol{Z}$ in lieu of the data generation process is that the precisions ν_j implicitly describe the amount of relevant data available for each μ_j. Let $\nu_{\min} \equiv \min_j \nu_j$ and $\nu_{\max} \equiv \max_j \nu_j$. Then, loosely speaking, the large-sample regime describes instances where $\nu_{\min}$ is large, i.e., where data are plentiful and we can estimate $\boldsymbol{\mu}$ easily. By contrast, the small-data, large-scale regime describes instances in which n is large (large-scale), but there are limited relevant data, and, hence, $\nu_{\max}$ is small.

To simplify our exposition, we will also assume:

Assumption 2.4 (Independent Gaussian Corruptions) *For each $j = 1, \ldots, n$, Z_j has Gaussian distribution with $Z_j \sim \mathcal{N}(\mu_j, 1/\nu_j)$ where ν_j is the* known *precision of Z_j. Moreover, $Z_1, \ldots, Z_n$ are independent.*

Assumption 2.4 is common. GR 2021 employ a similar assumption; Javanmard and Montanari (2018) strongly leverages a Gaussian design assumption when debiasing lasso estimates, and Ito et al. (2018) also assumes Gaussian errors in their debiasing technique. In each case, the idea is if a technique enjoys *provably* good performance under Gaussian corruptions, it will likely have good *practical* performance when data are approximately Gaussian. Indeed, if $\boldsymbol{Z}$ is obtained by maximum likelihood estimation, ordinary linear regression, simple averaging as in Section 1.2, or Gaussian process regression, then the resulting estimates will be approximately Gaussian. We adopt the same perspective in our work.

Note, the independence assumption in Assumption 2.4 is without loss of generality as illustrated in the following example.

Example 2.5 (Correlated Predictions) Suppose we are given an instance of Problem (1.1) and predictions $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a *known*, positive semidefinite matrix. Consider a Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$ and the transformed predictions $\overline{\boldsymbol{Z}} \equiv \boldsymbol{L}^{-1}\boldsymbol{Z}$. Notice $\overline{\boldsymbol{Z}} \sim \mathcal{N}(\overline{\boldsymbol{\mu}}, \boldsymbol{I})$ where $\overline{\boldsymbol{\mu}} \equiv \boldsymbol{L}^{-1}\boldsymbol{\mu}$. We then recast Problem (1.1) as the equivalent problem

$$\min_{\overline{\boldsymbol{x}} \in \overline{\mathcal{X}}} \quad \overline{\boldsymbol{\mu}}^\top \overline{\boldsymbol{x}}, \qquad \text{where} \qquad \overline{\mathcal{X}} \equiv \{\boldsymbol{L}^\top \boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}\}.$$

Our new problem is of the required form with transformed predictions independent across j.

Most importantly, Assumption 2.4 is *not* crucial to many of our results. Violating the Gaussian assumption only affects the bias of our estimator (see Theorem 3.2). Our analysis bounding the variance and tails of the stochastic errors utilize empirical process theory, and can easily be adapted for non-Gaussian corruptions. Moreover, although the bias of our estimator is non-negligible when $\boldsymbol{Z}$ is non-Gaussian, we can bound this bias in terms of the Wasserstein distance between $\boldsymbol{Z}$ and a multivariate Gaussian, suggesting our method has good performance as long as corruptions are *approximately* Gaussian and this distance is small (see Lemma B.4 in Appendix for the bound).

Finally, similar results also hold when $\boldsymbol{\nu}$ is, itself, estimated noisily with the addition of a small bias term related to its estimate's accuracy (see Lemma B.3 in Appendix).

2.2. Affine Plug-in Policy Classes

A data-driven policy for Problem (1.1) is a mapping $\boldsymbol{Z} \mapsto \boldsymbol{x}(\boldsymbol{Z}) \in \mathcal{X}$ that determines a feasible decision $\boldsymbol{x}(\boldsymbol{Z})$ from the observed data $\boldsymbol{Z}$. We focus on classes of affine plug-in policies. Intuitively, a plug-in policy first proxies the unknown $\boldsymbol{\mu}$ by some estimate, $\boldsymbol{r}(\boldsymbol{Z})$, and then solves Problem (1.1) after "plugging-in" this estimate for $\boldsymbol{\mu}$.

Definition 2.6 (Affine Plug-in Policy Classes) *For $j = 1, \ldots, n$, let $r_j(z, \boldsymbol{\theta}) = a_j(\boldsymbol{\theta})z + b_j(\boldsymbol{\theta})$ be an affine function of z where $a_j(\boldsymbol{\theta})$ and $b_j(\boldsymbol{\theta})$ are arbitrary functions of the parameter $\boldsymbol{\theta} \in \Theta$. Let $\boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{\theta}) = (r_1(Z_1, \boldsymbol{\theta}), r_2(Z_2, \boldsymbol{\theta}), \ldots, r_n(Z_n, \boldsymbol{\theta}))^\top \in \mathbb{R}^n$. The plug-in policy with respect to $\boldsymbol{r}(\cdot, \boldsymbol{\theta})$ is given by*

$$\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{\theta})^\top \boldsymbol{x}, \tag{2.1}$$

where ties are broken arbitrarily. Furthermore, we let $\mathcal{X}_\Theta(\boldsymbol{Z}) \equiv \{\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) \in \mathcal{X} : \boldsymbol{\theta} \in \Theta\} \subseteq \mathcal{X}$ denote the corresponding class of plug-in policies over Θ.

When $\boldsymbol{\theta}$ is fixed and clear from context, we suppress its dependence, writing $\boldsymbol{x}(\boldsymbol{Z})$ and $\boldsymbol{r}(\boldsymbol{Z})$. Moreover, for a fixed $\boldsymbol{\theta}$, $r_j(Z_j, \boldsymbol{\theta})$ only depends on the data (linearly) through the j^{th} component.

Plug-in policies are attractive in data-driven optimization because computing $\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta})$ involves solving a problem of the same form as Problem (1.1) . Thus, if a specialized algorithm exists for solving Problem (1.1) – e.g., as with many network optimization problems – the same algorithm can be used to compute the policy. This property does not necessarily hold for other classes of policies such as regularization based policies (GR 2021).

Moreover, many policies used in practice are of the form $\boldsymbol{x}(\boldsymbol{Z}, \hat{\boldsymbol{\theta}}(\boldsymbol{Z}))$ for some $\hat{\boldsymbol{\theta}}(\boldsymbol{Z})$. (See examples below.) Such policies are *not* affine plug-in policies; $r_j(Z_j, \hat{\boldsymbol{\theta}}(\boldsymbol{Z}))$ may depend nonlinearly on all the data $\boldsymbol{Z}$. Nonetheless, our analysis will bound the error of our estimator applied to such policies. Namely, in Section 4, we provide error bounds on our estimator that hold uniformly over $\mathcal{X}_\Theta(\boldsymbol{Z})$. Since these bounds hold uniformly, such bounds also hold for all policies of the form $\boldsymbol{x}(\boldsymbol{Z}, \hat{\boldsymbol{\theta}}(\boldsymbol{Z}))$.

For clarity, we make no claim about the optimality of affine plug-in policies for Problem (1.1); for a particular application, there may exist non-affine policies with superior performance. Our focus on affine plug-ins is motivated by their ubiquity and computational tractability.

We next present examples:

- **Sample Average Approximation (SAA).** The Sample Average Approximation (SAA) is a canonical data-driven policy for Problem (1.1). It is defined by

$$\boldsymbol{x}^{\mathsf{SAA}}(\boldsymbol{Z}) \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{Z}^\top \boldsymbol{x}. \tag{2.2}$$

SAA is thus an affine plug-in policy where the function $r_j(z,\boldsymbol{\theta}) = z$.

- **Plug-ins for Regression Models.** Consider the linear model $r_j(\boldsymbol{Z},\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{W}_j$, which does not depend on $\boldsymbol{Z}$, and the affine plug-in policy

$$\boldsymbol{x}^{\mathsf{LM}}(\boldsymbol{Z},\boldsymbol{\theta}) \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \sum_{j=1}^{n} \boldsymbol{W}_j^\top \boldsymbol{\theta}\cdot x_j. \tag{2.3}$$

As mentioned, many policies in the literature are of the form $\boldsymbol{x}^{\mathsf{LM}}(\boldsymbol{Z},\hat{\boldsymbol{\theta}}(\boldsymbol{Z}))$ for a particular $\hat{\boldsymbol{\theta}}(\boldsymbol{Z})$. For example, letting $\boldsymbol{\theta}^{\mathsf{OLS}}(\boldsymbol{Z}) \in \arg\min_{\boldsymbol{\theta}} \sum_{j=1}^n (Z_j - \boldsymbol{\theta}^\top \boldsymbol{W}_j)^2$ be the ordinary least-squares fit yields the estimate-then-optimize policy $\boldsymbol{x}^{\mathsf{LM}}(\boldsymbol{Z},\boldsymbol{\theta}^{\mathsf{OLS}}(\boldsymbol{Z}))$. Similarly, by appropriately padding the covariate with zeros, we can write the "optimization-aware" SPO and SPO+ methods of Elmachtoub and Grigas (2021) over linear hypothesis classes in the form $\boldsymbol{x}^{\mathsf{LM}}(\boldsymbol{Z},\boldsymbol{\theta}^{\mathsf{SPO}}(\boldsymbol{Z}))$ and $\boldsymbol{\theta}^{\mathsf{SPO+}}(\boldsymbol{Z}))$ where $\boldsymbol{\theta}^{\mathsf{SPO}}(\boldsymbol{Z})$ and $\boldsymbol{\theta}^{\mathsf{SPO+}}(\boldsymbol{Z})$ are obtained by minimizing the so-called SPO and SPO+ losses, respectively. Other methods, e.g., (Wilder et al. 2019a), can be rewritten similarly. As mentioned, our analysis will bound the error when debiasing these polices as well.

Of course, we are not limited to a linear model for $r_j(z,\boldsymbol{\theta})$. We could alternatively use a non-linear specification $r_j(z,\boldsymbol{\theta}) = f(\boldsymbol{W}_j,\boldsymbol{\theta})$ for some, given, nonlinear regression f with parameters $\boldsymbol{\theta}$. This specification of $\boldsymbol{r}(\boldsymbol{Z},\boldsymbol{\theta})$ *still* gives rise to a class of affine plug-in policies. Again, many policies in the literature, including estimate-then-optimize policies and SPO+ over non-linear hypothesis classes can be written in the form $\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}(\boldsymbol{Z}))$ for some particular mapping $\boldsymbol{\theta}(\boldsymbol{Z}) \in \Theta$.

- **Mixed-Effects Policies.** When $\boldsymbol{W}_j$ is not informative for μ_j, plug-ins for regression models perform poorly because no choice of $\boldsymbol{\theta}$ yields a good estimate of $\boldsymbol{\mu}$. By contrast, if $\nu_{\min}$ is large, SAA performs quite well. Mixed-effects policies interpolate between these choices. Define

$$\boldsymbol{x}^{\mathsf{ME}}\big(\boldsymbol{Z},(\tau,\boldsymbol{\beta})\big) \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \sum_{j=1}^{n} \left(\frac{\nu_j}{\nu_j+\tau} Z_j + \frac{\tau}{\nu_j+\tau}\boldsymbol{W}_j^\top \boldsymbol{\beta}\right) x_j, \tag{2.4}$$

where we have focused on a linear model for simplicity and made the dependence on $\boldsymbol{\theta} = (\tau,\boldsymbol{\beta})$ explicit for clarity. Mixed-effects policies are strongly motivated by Bayesian analysis (Gelman et al. 2014). These policies generalize the Bayes-Inspired policy class considered in GR 2021. Again, we observe that $\boldsymbol{x}^{\mathsf{ME}}\big(\boldsymbol{Z},(\tau,\boldsymbol{\beta})\big)$ is an affine-plug in policy. Moreover, we can also consider shrinking towards a nonlinear regression model as in (Ignatiadis and Wager 2019).

- **Plug-Ins for Linear Smoothers** Linear smoothers generalize linear regression models (Buja et al. 1989). Linear smoothers estimate the unknown $\boldsymbol{\mu}$ by $\boldsymbol{S}(\boldsymbol{W},\boldsymbol{\theta})\boldsymbol{Z}$ where $\boldsymbol{S}(\boldsymbol{W},\boldsymbol{\theta}) \in \mathbb{R}^{n\times n}$ is a smoother matrix that depends on the covariates $\boldsymbol{W}$ and some user-defined parameters $\boldsymbol{\theta}$, but does not depend on the data $\boldsymbol{Z}$. Local Polynomial Regression, k-nearest neighbors, cubic smoothing splines, and penalized kernel regression can all be written as a linear smoother. In k-nearest neighbors, e.g., the parameter θ determines the neighborhood size "k". The corresponding affine, plug-in policy is simply
$$\boldsymbol{x}^{\mathsf{LS}}(\boldsymbol{Z},\boldsymbol{\theta}) \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{Z}^\top \boldsymbol{S}(\boldsymbol{W},\boldsymbol{\theta})^\top \boldsymbol{x}.$$
Our framework thus allows us to estimate the out-of-sample performance of optimization policies based on these machine learning methods.

The choice of which affine plug-in policy class to use is largely application dependent. Our bounds in Section 4 provide some preliminary guidance, suggesting a tradeoff between the expressiveness of the policy class and the error of our estimator.

3. Variance Gradient Correction

We make the following assumption on problem parameters for the remainder of the paper:

Assumption 3.1 (Assumptions on Parameters) *There exists a constant $C_\mu > 1$ such that $\|\boldsymbol{\mu}\|_\infty \le C_\mu$, and constants $0 < \nu_{\min} < 1 < \nu_{\max} < \infty$ such that $\nu_{\min} \le \nu_j \le \nu_{\max}$ for all j. Moreover, we assume that $n \ge 3$.*

The assumptions for C_μ and $\nu_{\min}, \nu_{\max}$ are without loss of generality. These assumptions and the assumption on n allow us to simplify the presentation of some results by absorbing lower order terms into leading constants.

The in-sample performance of a policy $\boldsymbol{x}(\boldsymbol{Z})$ is $\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z})$. Let $\boldsymbol{\xi} = \boldsymbol{Z} - \boldsymbol{\mu}$. We call the difference between in-sample and out-of-sample performance, corresponding to $(\boldsymbol{Z}-\boldsymbol{\mu})^\top \boldsymbol{x}(\boldsymbol{Z}) = \boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z})$, the *in-sample optimism.* The expected in-sample optimism $\mathbb{E}[\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z})]$ is the in-sample bias.

Our method estimates the in-sample optimism of an affine, plug-in policy $\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})$. To this end, denote the plug-in objective value by

$$V(\boldsymbol{Z},\boldsymbol{\theta}) \;\equiv\; r(\boldsymbol{Z},\boldsymbol{\theta})^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) \;=\; \min_{\boldsymbol{x}\in\mathcal{X}} \; r(\boldsymbol{Z},\boldsymbol{\theta})^\top \boldsymbol{x}. \tag{3.1}$$

Because it is the minimum of linear functions, $\boldsymbol{Z} \mapsto V(\boldsymbol{Z},\boldsymbol{\theta})$ is always concave. We then estimate the in-sample optimism by the *Variance Gradient Correction* (VGC) defined by

$$D(\boldsymbol{Z},(\boldsymbol{\theta},h)) \;\equiv\; \sum_{j=1}^{n} D_j(\boldsymbol{Z},(\boldsymbol{\theta},h)), \tag{3.2}$$

where for $j = 1, 2, \dots, n$,

$$D_j(\boldsymbol{Z}, (\boldsymbol{\theta}, h)) \equiv \begin{cases} \mathbb{E}\left[\frac{1}{h\sqrt{\nu_j} a_j(\boldsymbol{\theta})} \Big(V(\boldsymbol{Z} + \delta_j \boldsymbol{e}_j) - V(\boldsymbol{Z}) \Big) \Big| \boldsymbol{Z} \right], & \text{if } a_j \neq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{3.3}$$

and $\delta_1, \dots, \delta_n$ are independent Gaussian random variables such that $\delta_j \sim \mathcal{N}\left(0, h^2 + \frac{2h}{\sqrt{\nu_j}}\right)$ for all j. To reduce notation, we define $\bar{\Theta} \equiv \Theta \times [h_{\min}, h_{\max}]$ for $0 < h_{\min} \le h_{\max}$ and write $D(\boldsymbol{Z}, \boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \bar{\Theta}$. We utilize the defined notation for results that require separating h and $\boldsymbol{\theta}$.

The VGC is defined as a (conditional) expectation over the auxiliary random variables δ_j. In practice, we can approximate this expectation to arbitrary precision by simulating δ_j and averaging; see Appendix B.5 for more efficient implementations.

Given the VGC, we estimate the out-of-sample performance by

$$\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) \approx \boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{\theta}) - D(\boldsymbol{Z}, \boldsymbol{\theta}). \tag{3.4}$$

In Section 3.1, we motivate the VGC. We then establish some of its key properties, namely that it is almost an unbiased estimator for the in-sample optimism, its variance is often vanishing as $n \to \infty$, and it is smooth in the policy class.

3.1. Motivating the Variance Gradient Correction (VGC)

Throughout this section, $\boldsymbol{\theta}$ is fixed so we drop it from the notation. Our heuristic derivation of $D(\boldsymbol{Z})$ proceeds in three steps.

Step 1: Re-expressing the In-Sample Optimism via Danskin's Theorem. Fix some j. If $a_j = 0$, then from the plug-in policy problem (Problem (2.1)) we see that $\boldsymbol{x}(\boldsymbol{Z})$ is independent of Z_j and the corresponding term in the in-sample bias is mean-zero, i.e., $\mathbb{E}[\xi_j \boldsymbol{x}(\boldsymbol{Z})] = 0 = D_j(\boldsymbol{Z})$. In other words, we do not correct such terms.

When $a_j \neq 0$, consider the function

$$\lambda \mapsto V(\boldsymbol{Z} + \lambda \xi_j \boldsymbol{e}_j). \tag{3.5}$$

This function is an example of a parametric optimization problem. Danskin's Theorem (Bertsekas 1997, Section B.5) characterizes its derivative with respect to λ.[4] Specifically, for any $\lambda \in \mathbb{R}$ such that $\boldsymbol{x}(\boldsymbol{Z} + \lambda \xi_j \boldsymbol{e}_j)$ is the unique optimizer to Problem (2.1), we have

$$\frac{\partial}{\partial \lambda} V(\boldsymbol{Z} + \lambda \xi_j \boldsymbol{e}_j) = a_j \xi_j x_j(\boldsymbol{Z} + \lambda \xi_j \boldsymbol{e}_j).$$

[4] See Theorem B.1 in Appendix B for a statement of Danskin's Theorem.

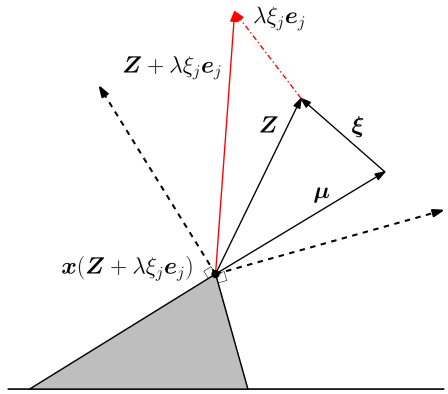

Figure 2 When $\mathcal{X}$ is polyhedral, $\boldsymbol{x}_j(\boldsymbol{Z}+\lambda\xi_j\boldsymbol{e}_j)$ must occur at a vertex if it is unique. Hence, small perturbations to λ do not change the solution (see figure), and the derivative of $V(\boldsymbol{Z}+\lambda\xi_j\boldsymbol{e}_j)$ is entirely determined by the derivative of $\boldsymbol{r}(\boldsymbol{Z}+\lambda\xi_j\boldsymbol{e}_j)$. Similar intuition holds for non-polyhedral $\mathcal{X}$.

When $\boldsymbol{x}(\boldsymbol{Z}+\lambda\xi_j\boldsymbol{e}_j)$ is not the unique optimizer, $a_j\xi_j x_j(\boldsymbol{Z}+\lambda\xi_j\boldsymbol{e}_j)$ is a subgradient, see Fig. 2 for intuition.

Notice that $\frac{\partial}{\partial\lambda}V(\boldsymbol{Z}+\lambda\xi_j\boldsymbol{e}_j)$ is the derivative of the plug-in value when we make the j^{th} component of $\boldsymbol{Z}$ more variable, i.e., variance increases by a factor $(1+\partial\lambda)^2$, where $\partial\lambda$ represents an infinitely small perturbation to λ. This observation motivates our nomenclature "Variance Gradient Correction."

Evaluating the above derivative at $\lambda=0$, dividing by a_j, and summing over j such that $a_j\neq 0$, allows us to re-express the in-sample bias whenever $\boldsymbol{x}(\boldsymbol{Z})$ is unique as

$$\sum_{j=1}^{n}\xi_j x_j(\boldsymbol{Z}) \;=\; \sum_{j:a_j\neq 0}\frac{1}{a_j}\cdot\frac{\partial}{\partial\lambda}V(\boldsymbol{Z}+\lambda\xi_j\boldsymbol{e}_j)\bigg|_{\lambda=0}.$$

Unfortunately, it is not clear how to evaluate these derivatives from the data. This leads to the second step in our derivation.

Step 2: Approximating the Derivative via Randomized Finite Differencing. As a first attempt, we approximate the above derivatives with first-order, forward finite-differences (LeVeque 2007, Chapter 1). Intuitively, we expect that for a sufficiently small step size $h>0$,

$$\frac{\partial}{\partial\lambda}V(\boldsymbol{Z}+\lambda\xi_j\boldsymbol{e}_j)\bigg|_{\lambda=0} \;=\; \frac{1}{h\sqrt{\nu}_j}\Big(V(\boldsymbol{Z}+h\sqrt{\nu_j}\xi_j\boldsymbol{e}_j)-V(\boldsymbol{Z})\Big) \;+\; o_p(1) \quad \text{as } h\to 0, \tag{3.6}$$

which suggests that

$$\sum_{j=1}^{n}\xi_j x_j(\boldsymbol{Z}) \;=\; \sum_{j:a_j\neq 0}\left[\frac{1}{h\sqrt{\nu_j}a_j}\Big(V(\boldsymbol{Z}+h\sqrt{\nu_j}\xi_j\boldsymbol{e}_j)-V(\boldsymbol{Z})\Big)\right] \;+\; o_p(n) \quad \text{as } h\to 0. \tag{3.7}$$

Unfortunately, the right side of Eq. (3.7) is not computable from the data, because we do not observe μ_j, and, hence, do not observe $\xi_j=Z_j-\mu_j$.

To circumvent this challenge, recall that ξ_j is Gaussian and independent across j, and let δ_j be the independent Gaussian random variables defined in the definition of the VGC (Eq. (3.3)). A direct computation shows, $\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j\boldsymbol{e}_j \sim_d \boldsymbol{Z} + \delta_j\boldsymbol{e}_j$, because both $Z + h\sqrt{\nu_j}\xi_j\boldsymbol{e}_j$ and $Z_j + \delta_j\boldsymbol{e}_j$ are Gaussians with matching mean and covariances.[5] Hence, $V(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j\boldsymbol{e}_j) \sim_d V(\boldsymbol{Z} + \delta_j\boldsymbol{e}_j)$.

Inspired by this relation, we replace the unknown $V(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j\boldsymbol{e}_j)$ by $V(\boldsymbol{Z} + \delta_j\boldsymbol{e}_j)$ in our first-order, finite difference approximation, yielding a *randomized* finite difference:

$$\underbrace{\sum_{j=1}^{n}\xi_j x_j(\boldsymbol{\xi}+\boldsymbol{\mu};\boldsymbol{\theta})}_{\text{In-Sample Optimism}} \approx \underbrace{\sum_{j:a_j\neq 0}\left[\frac{1}{h\sqrt{\nu_j}a_j(\boldsymbol{\theta})}\Big(V(\boldsymbol{\mu}+\boldsymbol{\xi}+\delta_j\boldsymbol{e}_j) - V(\boldsymbol{\mu}+\boldsymbol{\xi})\Big)\right]}_{\text{Randomized Finite Difference}}. \tag{3.8}$$

Step 3: De-Randomizing the Correction. Finally, in the spirit of Rao-Blackwellization, we then de-randomize this correction by taking conditional expectations over $\boldsymbol{\delta}$. This de-randomization reduces the variability of our estimator and yields the VGC (Eq. (3.3)).

Higher Order Finite Difference Approximations: Our heuristic motivation above employs a first-order finite difference approximation, and our theoretical analysis below focuses on this setting for simplicity. However, it is possible to use higher order approximations, which in turn reduce the bias. Theoretical analysis of such higher order approximations is tedious, but not substantively different from the first-order case. Hence, it is omitted. In our experiments, we use a particular second-order approximation described in Appendix B.5.

3.2. Bias of Variance Gradient Correction

Our first main result shows that one can make the heuristic derivation of the previous section rigorous when quantifying the bias of the VGC.

Theorem 3.2 (Bias of the Variance Gradient Correction) *Under Assumptions 2.4 and 3.1, for any $0 < h < 1/e$ and any affine, plug-in policy $\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})$, there exists a constant C (depending on $\nu_{\min}$) such that*

$$0 \le \mathbb{E}\left[\sum_{j=1}^{n}\xi_j x_j(\boldsymbol{\mu}+\boldsymbol{\xi},\boldsymbol{\theta}) - \sum_{j=1}^{n}D_j(\boldsymbol{\mu}+\boldsymbol{\xi},\boldsymbol{\theta})\right] \le C\cdot hn\log\left(\frac{1}{h}\right)$$

Recall we expect that, in typical instances, the full-information performance of Problem (1.1) is $O(n)$ as $n\to\infty$. Thus, this theorem asserts that as long as h is small, say $h = h_n = o(1)$ as $n\to\infty$, the bias of VGC is negligible relative to the true out-of-sample performance. In this sense, VGC is asymptotically unbiased for large n.

The proof of the theorem is in Appendix B.1 and proceeds similarly to our heuristic derivation but uses the following monotonicity property to precisely quantify the "little oh" terms.

[5] Here, $\sim_d$ denotes equality in distribution.

Lemma 3.3 (Monotonicity of Affine Plug-in Policies) *For any $\boldsymbol{z}$ and j, the function $t \mapsto \boldsymbol{x}(\boldsymbol{z} + t\boldsymbol{e}_j)$ is non-increasing if $a_j \geq 0$, and the function is non-decreasing if $a_j < 0$.*

Intuitively, the lemma holds because $\boldsymbol{z} \mapsto V(\boldsymbol{z})$ is a concave function (it is the minimum of affine functions). In particular, $t \mapsto V(\boldsymbol{z} + t)$ is also concave, and by Danskin's Theorem, $\frac{d}{dt}V(\boldsymbol{z}+t) = a_j x_j(\boldsymbol{z}+t)$ whenever $\boldsymbol{x}(\boldsymbol{z}+t)$ is unique. Informally, the lemma then follows since the derivative of a concave function is non-increasing. Appendix B.1 provides a formal proof accounting for points of non-differentiability.

Before proceeding, we remark that Theorem 3.2 holds with small modifications under mild violations of the independent Gaussian assumption (Assumption 2.4). Specifically, in cases where ν_j are not known but are estimated, the bias of the VGC constructed with the estimated ν_j increases by a small term depending on the accuracy of the precisions. See Lemma B.3 in the appendix for formal statements and proof.

3.3. The Variance of the VGC

As mentioned in the contributions, the parameter h controls the trade offs between bias and variance in our estimator. Unfortunately, while Theorem 3.2 gives a direct analysis of the bias under mild assumptions, a precise analysis of the variance (or tail behavior) of the VGC is more delicate. In this section we provide a loose, but intuitive bound on the variance of VGC that illustrates the types of problems for which our estimator should perform well. The main message of this section is that the VGC concentrates at its expectation so long as the policy $\boldsymbol{x}(\boldsymbol{Z})$ is "stable" in the sense that perturbing one element of $\boldsymbol{Z}$ does not cause $\boldsymbol{x}(\boldsymbol{Z})$ solution to change too much.

The main challenge in showing $D(\boldsymbol{Z})$ concentrates at its expectation is that $D(\boldsymbol{Z})$ is a sum of *dependent* random variables $D_j(\boldsymbol{Z})$. Worse, this dependence subtly hinges on the structure of Problem (1.1) and the plug-in policy problem (Problem (2.1)) and hence is not amenable to techniques based on mixing or bounding the correlations between terms. We require a different approach.

As a first step towards analyzing $D(\boldsymbol{Z})$, we upper bound its variance by a related, fully-randomized estimator.

Lemma 3.4 (Fully-Randomized VGC) *Suppose that the solution $\boldsymbol{x}(\boldsymbol{Z})$ to Problem (2.1) is almost surely unique. For each j such that $a_j \neq 0$, let*

$$D_j^{\mathsf{R}}(\boldsymbol{Z}) \equiv \frac{\delta_j}{h\sqrt{\nu_j}a_j} x_j(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j) \tag{3.9}$$

where $\tilde{U}_j \sim$ Uniform$[0,1]$ and δ_j is defined in Eq. (3.3)*. Let $D^{\mathsf{R}}(\boldsymbol{Z}) = \sum_{j:a_j \neq 0} D_j^{\mathsf{R}}(\boldsymbol{Z})$ denote the* fully-randomized VGC. *Then, for any j such that $a_j \neq 0$,*

$$D_j(\boldsymbol{Z}) = \mathbb{E}\left[D_j^{\mathsf{R}}(\boldsymbol{Z}) \mid Z\right] \quad \text{and} \quad \mathbb{V}ar(D(\boldsymbol{Z})) \leq \mathbb{V}ar(D^{\mathsf{R}}(\boldsymbol{Z})).$$

Proof: Again, by Danskin's Theorem and the fundamental theorem of calculus,

$$V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j)-V(\boldsymbol{Z}) = \int_0^{\delta_j} a_j x_j(\boldsymbol{Z}+t\boldsymbol{e}_j)dt = \int_0^1 a_j\delta_j\boldsymbol{x}_j(\boldsymbol{Z}+t\delta_j\boldsymbol{e}_j)dt = \mathbb{E}\left[a_j\delta_j\boldsymbol{x}(\boldsymbol{Z}+\delta_j\tilde{U}\boldsymbol{e}_j)\mid \boldsymbol{Z},\delta_j\right].$$

Scaling both sides by $\frac{1}{a_j h\sqrt{\nu_j}}$ and taking expectations over δ_j proves the first statement. The second then follows from Jensen's inequality. □

We next propose upper bounding $\mathbb{V}\mathrm{ar}(D^{\mathsf{R}}(\boldsymbol{Z}))$ with the Efron-Stein Inequality. In particular, let $\overline{\boldsymbol{Z}}$, $\overline{\boldsymbol{\delta}}$ and $\overline{\boldsymbol{U}}$ be i.i.d. copies of $\boldsymbol{Z}$, $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{U}}$ respectively, and let $\boldsymbol{Z}^k$ denote the vector $\boldsymbol{Z}$ but with the k^{th} component replaced by $\overline{Z}_k$. Define $\boldsymbol{\delta}^k$ and $\tilde{\boldsymbol{U}}^k$ similarly. Let $D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta},\tilde{\boldsymbol{U}})$ be the fully-randomized VGC with dependence on all constituent random variables made explicit. Then, by the Efron-Stein Inequality

$$\begin{aligned}\mathbb{V}\mathrm{ar}(D^{\mathsf{R}}(\boldsymbol{Z})) \;\le\; & \frac{1}{2}\sum_{k=1}^n \mathbb{E}\left[(D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta},\tilde{\boldsymbol{U}})-D^{\mathsf{R}}(\boldsymbol{Z}^k,\boldsymbol{\delta},\tilde{\boldsymbol{U}}))^2\right] && \text{(3.10a)}\\ &+\frac{1}{2}\sum_{k=1}^n \mathbb{E}\left[(D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta},\tilde{\boldsymbol{U}})-D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta}^k,\tilde{\boldsymbol{U}}))^2\right] && \text{(3.10b)}\\ &+\frac{1}{2}\sum_{k=1}^n \mathbb{E}\left[(D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta},\tilde{\boldsymbol{U}})-D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta},\tilde{\boldsymbol{U}}^k))^2\right]. && \text{(3.10c)}\end{aligned}$$

Recall that in the typical case, $\boldsymbol{\mu}^\top\boldsymbol{x}(\boldsymbol{Z})=O_p(n)$. Hence in what follows, we will focus on developing conditions for which the upper bound in Eqs. (3.10a) to (3.10c) is $o(n^2)$. Indeed, such a bound would suggest $D(\boldsymbol{Z})-\mathbb{E}[D(\boldsymbol{Z})]=o_p(n)$, i.e., the stochastic fluctuations in the VGC are negligible relative to the magnitude of the out-of-sample error for n sufficiently large. With this perspective, it is not difficult to argue that both Eqs. (3.10b) and (3.10c) both contribute at most $O\left(\frac{n}{h}\right)$ (see proof of Theorem 3.5). Thus, we focus on Eq. (3.10a).

Consider the k^{th} element of the sum. Write,

$$\begin{aligned}\left|D^{\mathsf{R}}(\boldsymbol{Z})-D^{\mathsf{R}}(\boldsymbol{Z}^k)\right| \;&\le\; \frac{1}{ha_k\nu_{\min}}\left|\sum_{j=1}^n \delta_j(x_j(\boldsymbol{Z}+\delta_j\tilde{U}_j\boldsymbol{e}_j)-x_j(\boldsymbol{Z}^k+\delta_j\tilde{U}_j\boldsymbol{e}_j))\right|\\ &\le\; \frac{\|\boldsymbol{\delta}\|_2\sqrt{n}}{ha_k\nu_{\min}}\cdot\left(\frac{1}{n}\sum_{j=1}^n\left|x_j(\boldsymbol{Z}+\delta_j\tilde{U}_j\boldsymbol{e}_j)-x_j(\boldsymbol{Z}^k+\delta_j\tilde{U}_j\boldsymbol{e}_j)\right|^2\right)^{1/2}, \qquad \text{(3.11)}\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz's inequality. Since each δ_j is Gaussian, we expect $\|\boldsymbol{\delta}\|_2^2$ to concentrate sharply at its mean, i.e., $\|\boldsymbol{\delta}\|_2^2=O_p(hn)$. Thus by squaring Eq. (3.11), taking expectations and substituting into Eq. (3.10a), we roughly expect

$$\begin{aligned}\mathbb{V}\mathrm{ar}(D^{\mathsf{R}}(\boldsymbol{Z})) \;\le\; & \underbrace{O\left(\frac{n}{h}\right)}_{\text{Eqs. (3.10b) and (3.10c)}} + \mathbb{E}\left[\frac{\|\boldsymbol{\delta}\|_2^2n^2}{h^2\nu_{\min}^2a_{\min}^2}\cdot\frac{1}{n^2}\sum_{k=1}^n\sum_{j=1}^n\left|x_j(\boldsymbol{Z}+\delta_j\tilde{U}\boldsymbol{e}_j)-x_j(\boldsymbol{Z}^k+\delta_j\tilde{U}\boldsymbol{e}_j)\right|^2\right]\\ \approx\; & O\left(\frac{n}{h}\right)+O\left(\frac{n^3}{h}\right)\cdot\underbrace{\frac{1}{n^2}\sum_{k=1}^n\sum_{j=1}^n\mathbb{E}\left[\left|x_j(\boldsymbol{Z}+\delta_j\tilde{U}_j\boldsymbol{e}_j)-x_j(\boldsymbol{Z}^k+\delta_j\tilde{U}_j\boldsymbol{e}_j)\right|^2\right]}_{\text{Avg. Solution Instability}},\end{aligned}$$

where $a_{\min} \equiv \min_{j:a_j \neq 0} |a_j|$.

We call the indicated term the *Average Solution Instability*. In the worst case, it is at most 1 since $\mathcal{X} \subseteq [0,1]^n$. If, however, it were $O(n^{-\alpha})$ for some $\alpha > 1$, then the $\mathbb{V}\mathrm{ar}(D^{\mathsf{R}}(\boldsymbol{Z})) = o(n^2)$ as desired.

How do we intuitively interpret Average Solution Instability? Roughly, in the limit as $h \to 0$, we might expect that $\boldsymbol{x}(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j) \approx \boldsymbol{x}(\boldsymbol{Z})$ because $\delta_j = O_p(\sqrt{h})$. Then the Average Solution Instability is essentially the expected change in the solution in a randomly chosen component j when we replace the data for a randomly chosen component k with an i.i.d. copy. This interpretation suggests Average Solution Instability should be small so long as a small perturbation to the k^{th} component doesn't change the entire solution vector $\boldsymbol{x}(\boldsymbol{Z})$ by a large amount, i.e., if small perturbations lead to small, local changes in the solutions. Intuitively, many large-scale optimization problems exhibit such phenomenon (see, e.g., Gamarnik (2013)), so we broadly expect the VGC to have low variance.

The above heuristic argument can be made formal as in the following theorem.

Theorem 3.5 (Variance of the VGC) *Suppose that the solution $\boldsymbol{x}(\boldsymbol{Z})$ to Problem (2.1) is almost surely unique, that there exists a constant C_1 (not depending on n) such that $\mathbb{E}\left[\frac{1}{n^2}\sum_{k=1}^n \sum_{j=1}^n \left(x_j(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j) - x_j(\boldsymbol{Z}^k + \delta_j \tilde{U}_j \boldsymbol{e}_j)\right)^2\right] \leq C_1 n^{-\alpha}$ and that Assumption 3.1 holds. Then, there exists a constant C_2 (depending on $\nu_{\min}$ and $a_{\min} \equiv \min_j a_j$) such that for any $0 < h < 1/e$*

$$\mathbb{V}ar(D(\boldsymbol{Z})) = \frac{C_2}{h} \max(n^{3-\alpha}, n).$$

In particular, in the typical case where the full-information solution to Problem (1.1) is $O(n)$, the stochastic fluctuations in the VGC are negligible relative to the out-of-sample performance if $\alpha > 1$.

The proof of Theorem 3.5 is in Appendix B.4.

We remark that Theorem 3.5 provides a *sufficient* condition for the variance of the VGC to be negligible asymptotically and to show that h controls the bias-variance tradeoff, however, the bound is not tight. In Section 4 we provide a tighter analysis given more stringent assumptions on Problems (1.1) and (2.1), which then also provides us guidance on how to select h to approximately balance the bias-variance tradeoff.

3.4. Smoothness and Boundedness of the VGC

One of the key advantages of our VGC is that it is smooth in the policy class, provided $\boldsymbol{\theta} \mapsto r(\cdot, \boldsymbol{\theta})$ is "well-behaved." Other corrections, like the Stein Correction of GR 2021, do not enjoy such smoothness. In Section 5, we argue this smoothness improves the empirical performance of our method. We formalize "well-behaved" in the next assumption:

Assumption 3.6 (Plug-in Function is Smooth) *We assume the functions $a_j(\boldsymbol{\theta})$, $b_j(\boldsymbol{\theta})$ are each L-Lipschitz continuous for all $j=1,\ldots,n$. Moreover, we assume there exists $a_{\max}, b_{\max} < \infty$ such that*

$$\sup_{\boldsymbol{\theta}\in\Theta} |a_j(\boldsymbol{\theta})| \;\leq\; a_{\max} \quad \textit{and} \quad \sup_{\boldsymbol{\theta}\in\Theta} |b_j(\boldsymbol{\theta})| \;\leq\; b_{\max} \quad \forall j=1,\ldots,n.$$

Finally, we assume there exists $a_{\min}$ such that

$$0 \;<\; a_{\min} \;\leq\; \inf\{|a_j(\boldsymbol{\theta})| : a_j(\boldsymbol{\theta}) \neq 0,\; j=1,\ldots,n,\; \boldsymbol{\theta}\in\Theta\}$$

In words, Assumption 3.6 requires the functions $a_j(\boldsymbol{\theta})$ and $b_j(\boldsymbol{\theta})$ to be Lipschitz smooth, bounded, and that the non-zero components of $a_j(\boldsymbol{\theta})$ be bounded away from 0.

Bias and Variance of VGC for Plug-In Linear Regression Models. Recall our Plug-in Linear Model class from Section 2.2. Since $a_j(\boldsymbol{\theta})=0$ for all j, $D(\boldsymbol{Z},\boldsymbol{\theta})=0$ for all $\boldsymbol{Z}$ and (non data-driven) $\boldsymbol{\theta}$ for this class. Said differently, the in-sample performance of a policy is, itself, our estimate of the out-of-sample performance, and, both Theorems 3.2 and 3.5 can both be strengthened; the bias of our estimator and variance of the correction are both zero. More generally, $D(\boldsymbol{Z},\boldsymbol{\theta})=0$ whenever the plug-in functionals $r_j(z,\boldsymbol{\theta})$ do not depend on z for all j.

We stress however that this analysis does not immediately guarantee that the in-sample performance of policies of the form $\boldsymbol{x}^{\mathsf{LM}}(\boldsymbol{Z},\boldsymbol{\theta}(\boldsymbol{Z}))$ is a good estimate of out-of-sample performance, because $\boldsymbol{\theta}(\boldsymbol{Z})$ depends on $\boldsymbol{Z}$. In Section 4 we provide sufficient conditions to ensure that in-sample performance is, indeed, a good estimate of out-of-sample performance. Moreover, when $\boldsymbol{r}(\boldsymbol{Z},\boldsymbol{\theta})$ does depend on $\boldsymbol{Z}$, e.g., as with our Mixed Effects Regression class, $D(\boldsymbol{Z},\boldsymbol{\theta})$ is generally non-zero.

Lemma 3.7 (Smoothness of Variance Gradient Correction) *Under Assumptions 3.1 and 3.6, the following hold:*

i) There exists a constant C_1 (depending on $a_{\min}, a_{\max}, b_{\max}$, and $\nu_{\min}$) such that for any $\boldsymbol{z}\in\mathbb{R}^n$ and any $0<h<1/e$, the function $\boldsymbol{\theta}\mapsto D(z,(\boldsymbol{\theta},h))$ is Lipschitz continuous with parameter $\frac{C_1 n^2 L}{h\sqrt{\nu_{\min}}}(\|\boldsymbol{z}\|_\infty+1)$. Moreover, there exists a constant C_2 (depending on C_μ and C_1) such that for any $R>1$, with probability at least $1-e^{-R}$, the (random) function $\boldsymbol{\theta}\mapsto D(\boldsymbol{Z},(\boldsymbol{\theta},h))$ is Lipschitz continuous with parameter $\frac{C_2 L}{h}\sqrt{\frac{R}{\nu_{\min}}}\cdot n^2\sqrt{\log n}$.

ii) Consider $D(z,(\boldsymbol{\theta},h))$ where $h\in[h_{\min},h_{\max}]$ and $0<h_{\max}-h_{\min}<1$. There exists an absolute constant C_3 such that for any $\boldsymbol{z}\in\mathbb{R}^n$ and $\boldsymbol{\theta}\in\Theta$, the following holds,

$$\left|D(z,(\boldsymbol{\theta},h)) - D(z,(\boldsymbol{\theta},\overline{h}))\right| \leq \frac{C_3 n}{h_{\min}\nu_{\min}^{3/4}}\sqrt{\left|h-\overline{h}\right|}.$$

See Appendix B.2 for a proof. Intuitively, the result follows because $\boldsymbol{\theta} \mapsto V(\boldsymbol{z}, \boldsymbol{\theta})$ is Lipschitz by Danskin's theorem and $D(\boldsymbol{z}, \boldsymbol{\theta})$ is a linear combination of such functions. The second part follows from a high-probability bound on $\|\boldsymbol{Z}\|_\infty$.

In addition to being smooth, the VGC is also bounded as a direct result of taking the conditional expectation over the perturbation parameters δ_j.

Lemma 3.8 (VGC is Bounded) *Suppose that Assumptions 3.1 and 3.6 hold. For any $\boldsymbol{z}$, and any $j = 1, \ldots, n$,*

$$|D_j(\boldsymbol{z})| \le \frac{\sqrt{3}}{\nu_{\min}^{3/4}\sqrt{h}}.$$

The proof can be found in Appendix B.2. The result follows from observing that the j^{th} component of the VGC is the difference of the optimal objectives values of two optimization problems whose cost vector differs by $O(|\delta_j|)$ in one component. Thus, the two optimal objective values can only differ by $O(|\delta_j|)$ which is at most a constant once we take the conditional expectation.

4. Estimating Out-of-sample Performance for Weakly-Coupled Problems

In this section we provide high-probability tail bounds on the error of our estimator for out-of-sample performance that hold uniformly over a given policy class. Such bounds justify using estimator for policy learning, i.e., identifying the best policy within the class. They are also substantively stronger than the variance analysis of Theorem 3.5 as they provide exponential bounds on the tail behavior, rather than bounding the second moment. Additionally, we show the uniform results hold even when $\boldsymbol{\theta} \in \bar{\Theta}$ is chosen in a data-driven manner (which recall also includes h).

From the definition of the VGC out-of-sample estimator (Eq. (3.4)), the error of our estimator of out-of-sample performance for $\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}(\boldsymbol{Z}))$ is

$$\underbrace{\left|\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}(\boldsymbol{Z})) - \left(\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}(\boldsymbol{Z})) - D(\boldsymbol{Z}, \boldsymbol{\theta}(\boldsymbol{Z}))\right)\right|}_{\text{Error Estimating Out of Sample Perf.}} \le \sup_{\boldsymbol{\theta} \in \bar{\Theta}} \underbrace{\left|\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) - D(\boldsymbol{Z}, \boldsymbol{\theta})\right|}_{\text{Error Estimating In-Sample Optimism}}$$

$$\le \sup_{\boldsymbol{\theta} \in \bar{\Theta}} \left|\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) - \mathbb{E}\left[\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta})\right]\right| \tag{4.1a}$$

$$+ \sup_{\boldsymbol{\theta} \in \bar{\Theta}} \left|D(\boldsymbol{Z}, \boldsymbol{\theta}) - \mathbb{E}\left[D(\boldsymbol{Z}, \boldsymbol{\theta})\right]\right| \tag{4.1b}$$

$$+ \sup_{\boldsymbol{\theta} \in \bar{\Theta}} \left|\mathbb{E}\left[\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) - \boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) - D(\boldsymbol{Z}, \boldsymbol{\theta})\right]\right|. \tag{4.1c}$$

We bounded Eq. (4.1c) in Theorem 3.2. Our goal will be to find sufficient conditions to show the remaining terms are also $o_p(n)$ uniformly over $\boldsymbol{\theta} \in \Theta$. Then, in the typical case where the out-of-sample performance is $O_p(n)$, the error of our estimator will be negligible relative to the true out-of-sample performance. Our strategy will be to leverage empirical process theory since the argument of each suprema is a sum of random variables. Importantly, this empirical process analysis does *not* strictly require the independent Gaussian assumption (Assumption 2.4). The challenge of course is that the constraints of Problem (2.1) introduce a complicated dependence between the terms.

Inspired by the average stability condition of Theorem 3.5, we focus on classes of "weakly-coupled" optimization problems. We consider two such classes of problems, those weakly-coupled by variables in Section 4.1 and those weakly-coupled by constraints Section 4.2. We provide formal definitions below.

4.1. Problems Weakly-Coupled by Variables

We say an instance of Problem (1.1) is weakly-coupled by variables if fixing a small number of variables causes the problem to separate into many, decoupled subproblems. Generically, such problems can be written as

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \left(\boldsymbol{\mu}^0\right)^\top \boldsymbol{x}^0 + \sum_{k=1}^{K} \left(\boldsymbol{\mu}^k\right)^\top \boldsymbol{x}^k \\ \text{s.t.} \quad & \boldsymbol{x}^0 \in \mathcal{X}^0, \quad \boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}^0), \quad \forall k = 1, \ldots, K. \end{aligned} \tag{4.2}$$

Here, $\boldsymbol{x}^0$ represents the coupling variables and $k = 1, \ldots, K$ represent distinct subproblems. Notice that once $\boldsymbol{x}^0$ is fixed, each subproblem can be solved separately. Intuitively, if $\dim(\boldsymbol{x}^0)$ is small relative to n, the subproblems of Eq. (4.2) are only "weakly" coupled. Some reflection shows both Examples 2.2 and 2.3 from Section 2 are weakly-coupled by variables.

Let $S_k \subseteq \{1, \ldots, n\}$ be the indices corresponding to $\boldsymbol{x}^k$ for $k = 0, \ldots, K$, and $S_{\max} = \max_{k \geq 0} |S_k|$. The sets $S_0, \ldots, S_K$ form a disjoint partition of $\{1, \ldots, n\}$. Without loss of generality, reorder the indices so that the S_k occur "in order,"; i.e., $(j : j \in S_0), \ldots, (j : j \in S_K)$ is a consecutive sequence.

Given the weakly-coupled structure of Eq. (4.2), we define a generalization of $\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta})$: For each $\boldsymbol{x}^0 \in \mathcal{X}^0$ and $\boldsymbol{\theta} \in \bar{\Theta}$, let

$$\boldsymbol{x}^k(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0) \in \operatorname*{arg\,min}_{\boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}^0)} \boldsymbol{r}^k(\boldsymbol{Z}, \boldsymbol{\theta})^\top \boldsymbol{x}^k, \quad k = 1, \ldots, K, \tag{4.3}$$

where $\boldsymbol{r}^k(\boldsymbol{Z}, \boldsymbol{\theta}) = (r_j(\boldsymbol{Z}, \boldsymbol{\theta}) : j \in S_k)$. Intuitively, the vector

$$\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0) \equiv \left((\boldsymbol{x}^0)^\top, \boldsymbol{x}^1(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0)^\top, \ldots, \boldsymbol{x}^K(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0)^\top\right)^\top$$

satisfies the Average Instability Condition of Theorem 3.5 so long as $S_{\max}$ is not too large since the j^{th} component of the solution changes when perturbing the k^{th} data point if and only if j and k belong to the same subproblem. This event happens with probability at most $S_{\max}/n^2$.

The key to making this intuition formal and obtaining exponential tails for the error of the out-of-sample estimator (Eq. (4.1)) is that

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\bar{\Theta}}\left|\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{E}\left[\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})\right]\right| &\le \sup_{\boldsymbol{\theta}\in\bar{\Theta},\boldsymbol{x}^0\in\mathcal{X}^0}\left|\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)-\mathbb{E}\left[\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\right]\right| \quad \text{and,}\\ \sup_{\boldsymbol{\theta}\in\bar{\Theta}}\left|D(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{E}\left[D(\boldsymbol{Z},\boldsymbol{\theta})\right]\right| &\le \sup_{\boldsymbol{\theta}\in\bar{\Theta},\boldsymbol{x}^0\in\mathcal{X}^0}\left|D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)-\mathbb{E}\left[D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\right]\right|,\end{aligned}$$

where both $\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ and $D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ can, for a fixed $\boldsymbol{\theta},\boldsymbol{x}^0$, be seen as sums of K *independent* random variables. To obtain uniform bounds, we then need to control only the metric entropy of the resulting (lifted) stochastic processes indexed by $(\boldsymbol{\theta},\boldsymbol{x}^0)$.

We propose a simple assumption on the policy class to control this metric entropy. We believe this assumption is easier to verify than other assumptions used in the literature (e.g., bounded linear subgraph dimension or bounded Natarajan dimension), but admittedly slightly more stringent.

Assumption 4.1 (Lifted Affine Plug-in Policy) *Given an affine plug-in policy class defined by* $\boldsymbol{r}(\cdot,\boldsymbol{\theta})$ *for* $\boldsymbol{\theta}\in\Theta$, *we say this class satisfies the lifted affine plug-in policy assumption for problems weakly-coupled by variables (Eq.* (4.2)*) if there exists mapping* $\phi(\cdot)$ *and mappings* $g_k(\cdot)$ *for* $k=1,\dots,K$ *such that*

$$\boldsymbol{x}^k(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\in \operatorname*{arg\,min}_{\boldsymbol{x}^k\in\mathcal{X}^k(\boldsymbol{x}^0)} \phi(\boldsymbol{\theta})^\top g_k(\boldsymbol{Z}^k,\boldsymbol{x}^k,\boldsymbol{x}^0) \quad k=1,\dots,K, \quad \forall \boldsymbol{x}^0\in\mathcal{X}^0.$$

We stress that the mapping $\phi(\cdot)$ is common to all K subproblems and all $\boldsymbol{x}^0\in\mathcal{X}^0$, and both $\phi(\cdot)$ and $g_k(\cdot)$ can be arbitrarily nonlinear. Moreover, $g_k(\cdot)$ may implicitly depend on the precisions $\boldsymbol{\nu}$ and covariates $\boldsymbol{W}$ as these are fixed constants. With the exception of policies from linear smoothers, each of our examples from Section 2.2, satisfies Assumption 4.1. For example, for plug-ins for linear regression models, we can simply take $\phi(\boldsymbol{\theta})=\boldsymbol{\theta}$ and $g_k(\boldsymbol{Z}^k,\boldsymbol{x}^k,\boldsymbol{x}^0)=\sum_{j\in S_k}\boldsymbol{W}_j^\top x_j$.

When $\boldsymbol{x}^k(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ is not the unique minimizer to the problem weakly-coupled variables defined in Eq. (4.3), we require that ties are broken consistently. Let $\mathrm{Ext}(\mathrm{Conv}(\mathcal{X}^k(\boldsymbol{x}^0)))$ denote the set of extreme points of $\mathrm{Conv}(\mathcal{X}^k(\boldsymbol{x}^0))$ and let $\mathcal{X}_{\max}=\max_{k\ge 0}\mathrm{Ext}(\mathrm{Conv}(\mathcal{X}^k(\boldsymbol{x}^0)))$. Note, if $\boldsymbol{x}^k(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ is unique, it is an extreme point.

Assumption 4.2 (Consistent Tie-Breaking) *We assume there exists functions* $\sigma_{k\boldsymbol{x}^0}: 2^{\mathcal{X}^k(\boldsymbol{x}^0)}\mapsto \mathrm{Ext}(\mathrm{Conv}(\mathcal{X}^k(\boldsymbol{x}^0)))$ *such that*

$$\boldsymbol{x}^k(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)=\sigma_{k\boldsymbol{x}^0}\left(\operatorname*{arg\,min}_{\boldsymbol{x}^k\in\mathcal{X}^k(\boldsymbol{x}^0)} \phi(\boldsymbol{\theta})^\top g_k(\boldsymbol{Z}^k,\boldsymbol{x}^k,\boldsymbol{x}^0)\right) \quad k=1,\dots,K, \quad \forall \boldsymbol{x}^0\in\mathcal{X}^0.$$

Consistent tie-breaking requires that if $(\boldsymbol{\theta}_1,\boldsymbol{x}_1^0)$ and $(\boldsymbol{\theta}_2,\boldsymbol{x}_2^0)$ induce the same minimizers in Eq. (4.3) for some $\boldsymbol{Z}$, then $\boldsymbol{x}^k(\boldsymbol{Z},\boldsymbol{\theta}_1,\boldsymbol{x}_1^0)=\boldsymbol{x}^k(\boldsymbol{Z},\boldsymbol{\theta}_2,\boldsymbol{x}_2^0)$, and this point is an extreme point of $\mathcal{X}^k(\boldsymbol{x}^0)$.

Assumptions 4.1 and 4.2 allow us to bound the cardinality of the set $\{(\boldsymbol{x}^0, \boldsymbol{x}^1(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0),\ldots,\boldsymbol{x}^K(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)) \;:\; \boldsymbol{x}^0 \in \mathcal{X}^0, \boldsymbol{\theta} \in \Theta\}$ by adapting a geometric argument counting regions in a hyperplane arrangements from Gupta and Kallus (2021) (Lemma C.7). The cardinality of the set characterizes the metric entropy of the policy class.

Finally, for this section, we say a constant C is *dimension-independent* if C does *not* depend on $\{K, S_{\max}, h, \mathcal{X}^0, \mathcal{X}_{\max}, \dim(\phi)\}$, but may depend on $\{\nu_{\min}, C_\mu, L\}$. We now present the main result of this section:

Theorem 4.3 (Policy Learning for Problems Weakly-Coupled by Variables) *Suppose Assumptions 2.4, 3.1, 3.6, 4.1 and 4.2 all hold. Let* $\mathcal{X}_{\max} \geq \left|\mathrm{Ext}(\mathrm{Conv}(\mathcal{X}^k(\boldsymbol{x}^0)))\right|$ *for all* $k = 1,\ldots,K$ *and* $\boldsymbol{x}^0 \in \mathcal{X}^0$*, and assume* $\mathcal{X}_{\max} < \infty$*. Then, for* $0 < h_{\min} \leq h_{\max} \leq 1$*, there exists a dimension-independent constant* C *such that, for any* $R > 1$*, with probability at least* $1 - 2\exp(-R)$*,*

$$
\begin{aligned}
\sup_{\theta \in \bar{\Theta}} \left|\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - D(\boldsymbol{Z},\boldsymbol{\theta},h)\right| \;\leq\; & C K S_{\max} \cdot h_{\max} \log\left(\frac{1}{h_{\min}}\right) \\
& + \; C S_{\max} R \sqrt{\frac{K \log(1 + |\mathcal{X}^0|)}{h_{\min}}} \left(\log(K) \sqrt{\log(S_{\max}) \log\left(h_{\min}^{-1} \cdot N\left(\sqrt{\frac{h_{\min}}{Kn^2}}, \Theta\right)\right)} \right. \\
& \left. + \; \sqrt{\log(K) \dim(\phi) \log(1 + \mathcal{X}_{\max})} \right).
\end{aligned}
$$

where $N(\epsilon, \Theta)$ *is the* ϵ*-covering number of the set* Θ*.*

In the typical case where Θ does not depend on n or K and

$$
\max(S_{\max}, |\mathcal{X}_0|, \dim(\phi)) = \tilde{O}(1) \text{ as } K \to \infty, \tag{4.4}
$$

we can approximately minimize the above bound by selecting $h \equiv h_k = O(K^{-1/3})$ and noting the relevant covering number grows at most logarithmically in K. This choice of h approximately balances the deterministic and stochastic contributions, and the bound reduces to $\tilde{O}(C_{\mathsf{PI}} K^{2/3})$ for some C_{PI} (depending on $|\mathcal{X}_0|$, $\dim(\phi)$, $\dim(\boldsymbol{\theta})$) that measures the complexity of the policy class. Many applications satisfy the conditions in Eq. (4.4), including the drone-assisted emergency medical response application (Example 2.2).

To illustrate, recall Example 2.2. Here, $\boldsymbol{y}$ represents the binding variables "$\boldsymbol{x}^0$", and we see $|\mathcal{X}^0| = \binom{L}{B}$. Moreover, $S_{\max} = L$, since $\boldsymbol{x}$ decouples across k. Inspecting the constraints, $\mathcal{X}_{\max} \leq B$, since for each k, we choose exactly one depot from which to serve location k, and there are most B available depots. Finally, for a fixed policy class, $\dim(\phi)$ is constant and the log covering number above grows at most logarithmically in K. Most importantly, we expect L (the number of possible depots) and B (the budget) to be fairly small relative to K since regulations and infrastructure limit placement of depots, but there are many possible locations for cardiac events. Here typical instances of Example 2.2 satisfy Eq. (4.4). We return to Example 2.2 in Section 5 where we study the performance of our method numerically.

4.2. Problems with Weakly-Coupled by Constraints

An instance of Problem (1.1) is weakly-coupled by constraints if, after removing a small number of binding constraints, the problem decouples into many separate subproblems. Data-driven linear optimization problems of this form have been well studied by Li and Ye (2019) and GR 2021. In order to facilitate comparisons to the existing literature, we study the specific problem

$$\max_{\boldsymbol{x}\in\mathcal{X}} \ \sum_{j=1}^{n}\mu_j x_j, \qquad \mathcal{X}=\left\{\boldsymbol{x}\in[0,1]^n : \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{A}_j x_j \le \boldsymbol{b}\right\}, \tag{4.5}$$

and corresponding plug-in policies

$$\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} \ \sum_{j=1}^{n} r_j(Z_j,\boldsymbol{\theta})x_j. \tag{4.6}$$

Here, $\boldsymbol{A}_j\in\mathbb{R}^m$ with $m\ge 1$. In particular, we consider a maximization instead of a minimization.

We next introduce a dual representation of Problem (4.6). Specifically, scaling the objective of Problem (4.6) by $\frac{1}{n}$ and dualizing the binding constraints yields

$$\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta}) \in \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^m_+} \left\{\boldsymbol{b}^\top\boldsymbol{\lambda} + \max_{\boldsymbol{x}\in[0,1]^n} \ \frac{1}{n}\sum_{j=1}^{n}(r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda})x_j\right\} \tag{4.7}$$

For a fixed $\boldsymbol{\lambda}$, the inner maximization of Eq. (4.7) can be solved explicitly, yielding

$$\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta}) \in \arg\min_{\boldsymbol{\lambda}\in\mathbb{R}^m_+} \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{Z},\boldsymbol{\theta}), \quad \text{where} \quad \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{Z},\boldsymbol{\theta}) \equiv \boldsymbol{b}^\top\boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^{n}\left(r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right)^+.$$

By strong duality, $V(\boldsymbol{Z},\boldsymbol{\theta}) = \boldsymbol{r}(\boldsymbol{Z},\boldsymbol{\theta})^\top\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) = n\mathcal{L}(\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta}),\boldsymbol{Z},\boldsymbol{\theta})$.

This dual representation highlights the weakly-coupled structure. Indeed, the dependence across terms in the sum in $\mathcal{L}(\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta}),\boldsymbol{Z},\boldsymbol{\theta})$ arises because $\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})$ depends on the entire vector $\boldsymbol{Z}$. However, this dependence has to be "channeled" through the m dimensional vector $\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})$, and, hence, when m is small relative to n, cannot create too much dependence between the summands. Indeed, we will show that if m is small relative to n, then $\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})$ concentrates at its expectation, i.e., a constant, as $n\to\infty$, and, hence, the summands become independent asymptotically. This insight is key to the analysis.

To formalize these ideas, we make assumptions similar to those in GR 2021 and Li and Ye (2019):

Assumption 4.4 (s_0-Strict Feasibility) *There exists an $s_0>0$ and $\boldsymbol{x}_0\in\mathcal{X}$ such that $\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{A}_j x_j^0 + s_0\boldsymbol{e} \le \boldsymbol{b}$.*

Assumption 4.5 (Regularity of Matrix $\boldsymbol{A}$) *There exists a constant C_A such that $\|\boldsymbol{A}_j\|_\infty \le C_A$ for all $1 \le j \le n$. Moreover, there exists a constant $\beta > 0$ such that the minimal eigenvalue of $\frac{1}{n}\sum_{j=1}^n \boldsymbol{A}_j \boldsymbol{A}_j^\top$ is at least β.*

The strict feasibility assumption can often be satisfied by perturbing $\boldsymbol{A}$ or $\boldsymbol{b}$ and ensures the dual optimal values $\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})$ are bounded with high probability. The regularity assumptions on $\boldsymbol{A}$ ensure the function $\boldsymbol{\lambda} \mapsto \mathbb{E}[\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{Z},\boldsymbol{\theta})]$ is strongly convex, a key property in our proof (see below). Such a property holds, e.g., if the columns $\boldsymbol{A}_j$ are drawn randomly from some distribution.

Like Section 4.1, in order to obtain uniform bounds we must also control the metric entropy of the different stochastic error terms in out-of-sample error Eq. (4.1). Generalizing GR 2021, we make the following assumption:

Assumption 4.6 (VC Policy Class) *There exists a function $\rho(\cdot)$ such that*

$$r_j(z_j,\boldsymbol{\theta}) = \rho((z_j,\nu_j,\boldsymbol{W}_j,\boldsymbol{A}_j),\boldsymbol{\theta}), \quad j = 1,\ldots,n$$

and a constant V such that the class of functions

$$\left\{(z,\nu,\boldsymbol{W},\boldsymbol{A}) \mapsto \rho((z,\nu,\boldsymbol{W},\boldsymbol{A}),\boldsymbol{\theta}) - \boldsymbol{A}^\top\boldsymbol{\lambda} \ :\ \boldsymbol{\theta} \in \Theta,\ \boldsymbol{\lambda} \in \mathbb{R}_+^m\right\}$$

has a pseudo-dimension at most V. Without loss of generality, we further assume $V \ge \max(m,2)$.

The size of the constant V captures the complexity of the policy class which typically depends upon the dimension of $\boldsymbol{\theta}$ as well as the number of binding constraints m. As an illustration, recall the plug-in for linear regression models policy class from Section 2. By (Pollard 1990, Lemma 4.4), this policy class satisfies Assumption 4.6 with $V = \dim(\boldsymbol{\theta}) + m$.

Finally, we say a constant C is *dimension-independent* if C does *not* depend on $\{n,h,m,V,\dim(\boldsymbol{\theta})\}$, but may depend on $\{\nu_{\min}, C_A, C_\mu, \beta, s_0, a_{\min}, a_{\max}, b_{\max}, L\}$.

The main result of this section is then:

Theorem 4.7 (Estimation Error for Problems Weakly Coupled Constraints) *Under Assumptions 2.4, 3.1, 3.6 and 4.4 to 4.6, for $0 < h_{\min} \le h_{\max} \le 1$ there exists dimension-independent constants C and n_0, such that for any $R > 1$ and all $n \ge n_0 e^R$, we have that with probability $1 - C\exp(-R)$,*

$$\sup_{\boldsymbol{\theta}\in\bar{\Theta}} \left| \sum_{j=1}^n \xi_j x_j(\boldsymbol{Z},\boldsymbol{\theta}) - D_j(\boldsymbol{Z},\boldsymbol{\theta}) \right| \ \le\ Cn \cdot h_{\max} \log\left(\frac{1}{h_{\min}}\right) \\ + C \cdot V^3 \log^3 V \cdot R \cdot \sqrt{n \log\left(n \cdot N(n^{-3/2},\Theta)\right)} \cdot \frac{\log^5 n}{h_{\min}}$$

To build intuition, consider instances where Θ does not depend on n. Then, $V = O(1)$ and the covering number grows at most logarithmically as $n \to \infty$. We can then minimize the above bound (up to logarithmic terms) by taking taking $h \equiv h_n = O(n^{-1/4})$, yielding a bound of order $\tilde{O}(n^{3/4})$. In particular, in the typical instance where the full-information optimum (c.f. Problem (4.5)) is $O(n)$, the relative error of our estimate is $\tilde{O}(n^{-1/4})$ which is vanishing as $n \to \infty$.

Remark 4.8 The rate above ($\tilde{O}(n^{-1/4})$) is slightly slower than the rate of convergence of the Stein correction in GR 2021 ($\tilde{O}(n^{-1/3})$). We attribute this difference to our choice of a first order finite difference when constructing the VGC. A heuristic argument strongly suggests that had we instead used a second order forward finite difference scheme as in Appendix B.5, we would recover the rate $\tilde{O}(n^{-1/3})$. Moreover, our numerical experiments (with the second order scheme) in Section 5 shows our second-order VGC to be very competitive.

Proof Intuition: Approximate Strong-Convexity and Dual Stability

To build intuition, recall that to show the convergence of VGC, it suffices to bound the Average Solution Instability defined in Theorem 3.5. By complementary slackness, $x_j(\boldsymbol{Z}) = \mathbb{I}\left\{r_j(Z_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z})\right\}$ except possibly for m fractional components. Hence, by rounding the fractional components, we have, for $j \neq k$,

$$\begin{aligned}
&x_j(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j) - x_j(\boldsymbol{Z}^k + \delta_j \tilde{U}_j \boldsymbol{e}_j) \\
&\quad\leq\ \mathbb{I}\left\{r_j(Z_j + \delta_j \tilde{U}_j \boldsymbol{e}_j) \geq \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j)\right\} - \mathbb{I}\left\{r_j(Z_j + \delta_j \tilde{U}_j \boldsymbol{e}_j) > \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z}^k + \delta_j \tilde{U}_j \boldsymbol{e}_j)\right\} \\
&\quad\leq\ \mathbb{I}\left\{r_j(Z_j + \delta_j \tilde{U}_j \boldsymbol{e}_j) \in \left\langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j),\ \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z}^k + \delta_j \tilde{U}_j \boldsymbol{e}_j)\right\rangle\right\},
\end{aligned}$$

where we use $\langle l, u\rangle$ to denote the interval $[\min(l,u),\ \max(l,u)]$. By symmetry, the same bound holds for $x_j(\boldsymbol{Z}^k + \delta_j \tilde{U}_j \boldsymbol{e}_j) - x_j(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j)$. Since summands where $j = k$ each contribute at most 1 to the Average Solution Instability, we thus have that

$$\begin{aligned}
&\frac{1}{n^2}\sum_{k=1}^n \sum_{j=1}^n \left(x_j(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j) - \boldsymbol{x}_j(\boldsymbol{Z}^k + \delta_j \tilde{U}_j \boldsymbol{e}_j)\right)^2 \\
&\quad\leq\ \frac{1}{n^2}\sum_{k=1}^n \sum_{j=1}^n \mathbb{I}\left\{r_j(Z_j + \delta_j \tilde{U}_j \boldsymbol{e}_j) \in \left\langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j),\ \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z}^k + \delta_j \tilde{U}_j \boldsymbol{e}_j)\right\rangle\right\} + O_p\left(\frac{1}{n}\right).
\end{aligned}$$

The principal driver of the Solution Instability is the first double sum; in a worst-case, it might be $O_p(1)$. It will be small if $\|\boldsymbol{\lambda}(\boldsymbol{Z} + \delta_j \tilde{U}_j \boldsymbol{e}_j) - \boldsymbol{\lambda}(\boldsymbol{Z}^k + \delta_j \tilde{U}_j \boldsymbol{e}_j)\|$ is small for most j and k. Said differently, problems like Problem (4.6) that are weakly-coupled by constraints will have small Solution Instability if the dual solutions $\boldsymbol{\lambda}(\cdot)$ are, themselves, stable, i.e., if the dual solution does not change very much when we perturb one data point. Our analysis thus focuses on establishing this dual solution stability.

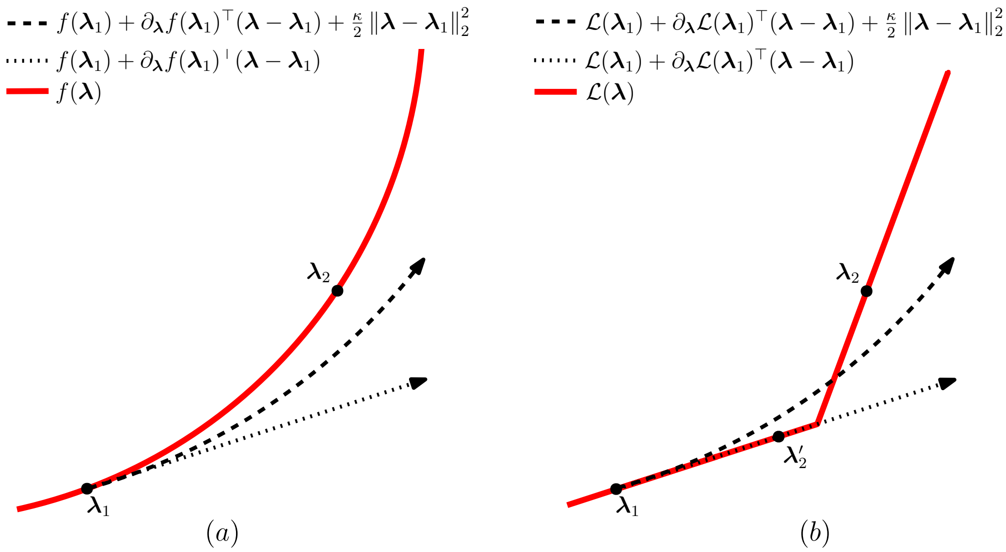

Figure 3 **Approximate Strong Convexity of $\mathcal{L}(\boldsymbol{\lambda})$.** Figure (a) shows a strongly convex function $f(\boldsymbol{\lambda})$ and visualizes the strong convexity condition Eq. (4.8). Figure (b) shows that because $\mathcal{L}(\boldsymbol{\lambda})$ is piecewise linear, it does not satisfy Eq. (4.8) for points on same line segment ($\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2'$). However, when $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$ are sufficiently far apart, they are on different line segments and Eq. (4.8) is satisfied.

Unfortunately, solutions of linear optimization problems need not be stable – a small perturbation to the cost vector might cause the optimal solution to switch between extreme points, inducing a large change. By contrast, solutions of convex optimization problems with strongly-convex objectives *are* stable (see, e.g., Shalev-Shwartz et al. (2010)). The next key step of our proof is to show that although $\boldsymbol{\lambda} \mapsto \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{Z})$ is not strongly-convex, it is still "approximately" strongly-convex with high probability in a certain sense.

To be more precise, recall that a function $f(\boldsymbol{\lambda})$ is κ-strongly-convex if

$$f(\boldsymbol{\lambda}_2) - f(\boldsymbol{\lambda}_1) \geq \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_1)^\top(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1) + \frac{\kappa}{2}\|\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1\|_2^2, \qquad \forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathrm{Dom}(f). \tag{4.8}$$

where $\kappa > 0$ and $\nabla_{\boldsymbol{\lambda}}$ denotes the subgradient. The left panel of Fig. 3 depicts this condition graphically. For any two points $\boldsymbol{\lambda}_2$ and $\boldsymbol{\lambda}_1$, the first-order Taylor series underestimates the function value, and one can "squeeze in" a quadratic correction $\frac{\kappa}{2}\|\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1\|_2^2$ in the gap.

The function $\boldsymbol{\lambda} \mapsto \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{Z})$ does not satisfy this condition, as seen in the right panel for points $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2'$. This function is piecewise-linear, and, for two points on the same line-segment, the first-order Taylor series is exact. However, for points on different line segments, such as $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, the first-order Taylor series is a strict underestimation, and we can squeeze in a quadratic correction. Said differently, Eq. (4.8) does not hold for all $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, but holds for most $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$. In this sense, $\boldsymbol{\lambda} \mapsto \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{Z})$ is "approximately" strongly-convex.

To make a formal statement, it is more convenient to use a different, equivalent definition of strong-convexity. Equation (4.8) is equivalent to the condition

$$\left(\nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_1) - \nabla_{\boldsymbol{\lambda}} f(\boldsymbol{\lambda}_2)\right)^\top (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2) \geq \kappa \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 \quad \forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathrm{Dom}(f). \tag{4.9}$$

Lemma D.9 then shows that $\boldsymbol{\lambda} \mapsto \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{Z})$ is approximately strongly convex in the sense that, with high probability, there exists a $C > 0$ such that

$$\begin{aligned}\left(\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_1, \boldsymbol{Z}) - \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_2, \boldsymbol{Z})\right)^\top (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2) \ &\geq C \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - \frac{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2}}{C\sqrt{n}}, \\ \text{for all } \|\boldsymbol{\lambda}_1\|_1 \leq \lambda_{\max},\ \|\boldsymbol{\lambda}_2\|_1 &\leq \lambda_{\max}, \text{ and } \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \geq \frac{4}{n}\end{aligned} \tag{4.10}$$

where $\lambda_{\max}$ is a dimension independent constant satisfying $\mathbb{E}\|\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta})\|_1 \leq \lambda_{\max}$.

Equation (4.10) mirrors Eq. (4.9). As $n \to \infty$, Eq. (4.10) reduces to the analogue of Eq. (4.9) for $\|\boldsymbol{\lambda}_1\|_1, \|\boldsymbol{\lambda}_2\|_1 \leq \lambda_{\max}$. Moreover, for $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \gg \frac{1}{n}$, the first term on the left of Eq. (4.10) above dominates the second, so that the right hand side is essentially a quadratic in $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2$. These relations motivate our terminology "approximately strongly-convex."

Using this notion of approximate strong-convexity, we show in Lemma D.12 that there exists a set $\mathcal{E}_n \subseteq \mathbb{R}^n$ such that $\boldsymbol{Z} \in \mathcal{E}_n$ with high probability, and, more importantly, for any $\boldsymbol{z} \in \mathcal{E}_n$, the dual solutions are stable, i.e.,

$$\|\boldsymbol{\lambda}(\boldsymbol{z}, \boldsymbol{\theta}) - \boldsymbol{\lambda}(\overline{\boldsymbol{z}}, \boldsymbol{\theta})\|_2 \ \leq \frac{C}{n} \sum_{j=1}^n \mathbb{I}\left\{z_j \neq \overline{z}_j\right\} \quad \forall \overline{\boldsymbol{z}} \text{ s.t. } \|\boldsymbol{\lambda}(\overline{\boldsymbol{z}})\|_1 \leq \lambda_{\max}. \tag{4.11}$$

Equipped with this dual-stability condition, we can bound the average solution instability and the variance of $D(\boldsymbol{Z})$ as in Theorem 3.5.

However, since the above stability condition holds with high probability instead of in expectation, we can actually use a modification of McDiarmid's inequality (see Theorem A.5) to prove the following, stronger tail bound:

Lemma 4.9 (Pointwise Convergence of VGC) *Fix some $\boldsymbol{\theta} \in \Theta$. Under the assumptions of Theorem 4.7, there exists a dimension independent constants C, n_0 such that, for any $R > 1$ and $n \geq n_0 e^R$, we have with probability at least $1 - 4\exp(-R)$,*

$$|D(\boldsymbol{Z}, \boldsymbol{\theta}) - \mathbb{E}\left[D(\boldsymbol{Z}, \boldsymbol{\theta})\right]| \ \leq \ CV^3 \log^2(V) \frac{\log^4(n)\sqrt{n}}{h} \sqrt{R}.$$

We then complete the proof of Theorem 4.7 by i) invoking a covering argument over Θ to extend this tail bound to a uniform concentration for the VGC and ii) again leveraging dual stability to show the in-sample optimism (Eq. (4.1a)) concentrates similarly. See Appendix D.5 for the details.

Comparison to Proof Technique to GR 2021 and Li and Ye (2019)

GR 2021 and Li and Ye (2019) also analyze the behavior of $\boldsymbol{\lambda}(\boldsymbol{Z})$ in the limit as $n \to \infty$. The key to their analysis is observing that the function $\boldsymbol{\lambda} \mapsto \mathbb{E}[\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{Z})]$ is strongly-convex. Using this property, they prove that

$$\|\boldsymbol{\lambda}(\boldsymbol{Z}) - \boldsymbol{\lambda}^*\|_2 = \tilde{O}_p\left(\frac{1}{\sqrt{n}}\right) \tag{4.12}$$

for some constant $\boldsymbol{\lambda}^*$ that does not depend on the realization of $\boldsymbol{Z}$.

Our analysis via approximate strong-convexity takes a different perspective. Specifically, instead of studying the function $\boldsymbol{\lambda} \mapsto \mathbb{E}[\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{Z})]$, we study the (random) function $\boldsymbol{\lambda} \mapsto \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{Z})$. While more complex, this analysis permits a tighter characterization of the behavior of the dual variables. In particular, leveraging Eq. (4.11), one can prove a statement similar to Eq. (4.12) (see Lemma D.17), however, to the best of our knowledge, one cannot easily prove Eq. (4.11) given the strong-convexity of $\boldsymbol{\lambda} \mapsto \mathbb{E}[\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{Z})]$ or Eq. (4.12). A simple triangle inequality from Eq. (4.12) would suggest the much slower rate $\|\boldsymbol{\lambda}(\boldsymbol{Z}) - \boldsymbol{\lambda}(\bar{\boldsymbol{Z}})\| = O_p\left(\frac{1}{\sqrt{n}}\right)$.

It is an open question whether this tighter analysis might yield improved results for the online linear programming setting studied in Li and Ye (2019).

5. Numerical Case Study: Drone-Assisted Emergency Medical Response

We reconsider Example 2.2 using real data from Ontario, Canada. Our data analysis and set-up largely mirror Boutilier and Chan (2019), however, unlike that work, our optimization formulation explicitly models response time uncertainty.

Data and Setup. Recall, our formulation decides the location of drone depots (y_l) and dispatch rules (x_{kl}) where a dispatch rule determines whether to send a drone from depot l to location k when requested. Our objective is to minimize the expected response time to out-of-hospital cardiac arrest (OHCA) events. We consider $L = 31$ potential drone depot locations at existing paramedic, fire, and police stations in the Ontario area.

To study the effect of problem dimension on our estimator, we vary the number of OHCA events via simulation similarly to Boutilier and Chan (2019). Specifically, we estimate a (spatial) probability density over Ontario for the occurrence of OHCA events using a kernel density estimator trained on 8 years of historical OHCA events. We then simulate K (ranging from 50 to 3,200) events according to this density giving the locations k used in our formulation.

In our case-study, μ_{kl} represents the excess time a drone-assisted response takes over an ambulance-only response. (This objective is typically negative). We learn these constants by first training a k-nearest neighbors regression (kNN) for the historical ambulance response times to nearby OHCAs. (For a sense of scale, the maximum ambulance response time is less than $1500s = 25$

min.) We estimate a drone response time based on the (straight-line) distance between k and l assuming an average speed of 27.8 m/s and 20s for take-off and landing (assumptions used in Boutilier and Chan (2019)). We then set μ_{kl} to the difference of the drone time minus the ambulance time. These values are fixed constants throughout our simulations and range from -3100 seconds to 1200 seconds.

We take Z_{kl} be normally corrupted predictions of μ_{kl} where the precisions ν_{kl} are determined by bootstrapping. Specifically, we take many bootstrap samples of our original historical dataset and refit the k-nearest neighbor regression and recompute an estimate of ambulance and drone response times. The precision ν_{kl} is taken to be the reciprocal of the variance of these bootstrap replicates. Precisions range from $\nu_{\min} = 4 \times 10^{-7}$ to $\nu_{\max} = 2 \times 10^{-4}$.

Policy Class. To determine dispatch rules for our case study, we consider the following policies:

$$\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{W}, (\tau, \mu_0)) \in \arg\min_{\boldsymbol{x} \in \mathcal{X}} \sum_{k=1}^{K} \sum_{j=1}^{L} \left(\frac{\nu_{jk}}{\nu_{jk} + \tau} Z_{jk} + \frac{\tau}{\nu_{jk} + \tau} (W_{jk} - \mu_0) \right) x_{jk},$$

where W_{jk} is the computed drone travel time between facility j and OHCA k. Our policy class consists of policies where $\tau \in [0, 100\nu_{\min}]$ and $\mu_0 \in [0, 1500]$. Similar to the Mixed-Effects Policies from Section 2.2, each policy is a weighted average between the SAA policy and a deterministic policy that dispatches to any location whose drone travel-time is at most μ_0.

For the first three experiments, we generate out-of-sample estimates using our VGC, the Stein-correction of GR 2021, and cross-validation using hold-out (2-fold) cross-validation. We assume that we are given two samples $\boldsymbol{Z}^1, \boldsymbol{Z}^2$ so that $Z_{jk} = \frac{1}{2}(Z^1_{jk} + Z^2_{jk})$. We set $h = n^{-1/6}$ for both the VGC and the Stein-Correction based on the recommendation of GR 2021. In the last experiment, we are given one hundred samples $\boldsymbol{Z}^i$ for $i = 1, \ldots, 100$ where $Z_{jk} = \frac{1}{100} \sum_{i=1}^{100} Z^i_{jk}$ and generate out-of-sample estimates for 2, 5, 10, 20, 50, and 100 fold cross-validation. For ease of comparison, we present all results as a percentage relative to full-information optimal performance.

5.1. Results

In our first experiment, we evaluate the bias and square root variance of each method for the out-of-sample performance of the SAA policy ($\tau = 0$) as K, the number of OHCA events grows (see Fig. 4). As predicted by our theoretical analysis, the quality of the out-of-sample estimates improve as we increase the problem size for both the VGC and the Stein Correction. However, cross-validation incurs a non-vanishing bias because it only leverages half the data in training.

As a second experiment, in Fig. 5, we can observe the quality of the estimators over multiple policies in the policy class. We highlight the smoothness of the VGC as τ varies. Since, for large K, the true performance is quite smooth, the worst-case error of VGC is generally smaller than that of

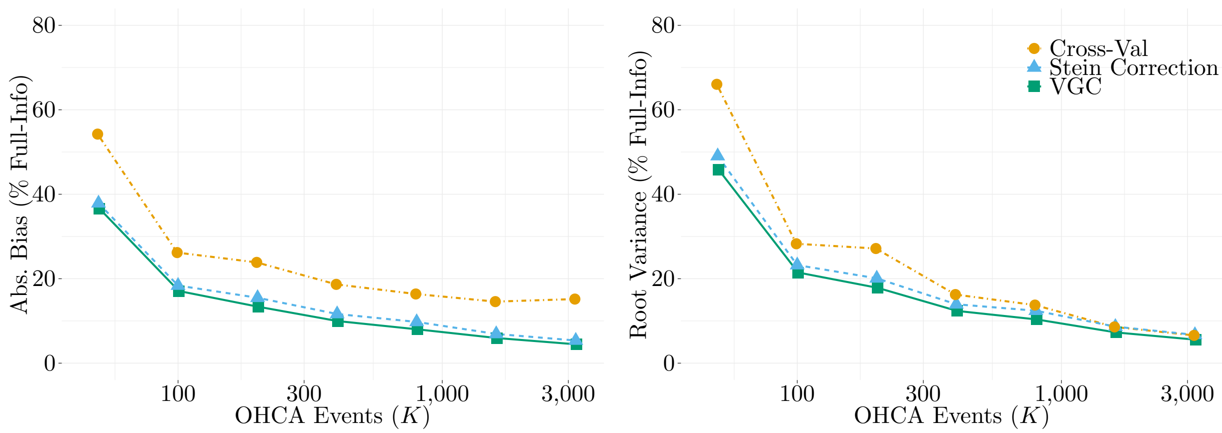

Figure 4 **Bias and variance as $K \to \infty$.** The two graphs plot the bias and variance of the different out-of-sample performance estimators for the Sample Average Approximation (SAA) policy. The bias and variance were estimated across 100 simulations for each K. Although variance vanishes for all methods as K increases, cross-validation exhibits a non-vanishing bias and is uniformly worse for all K.

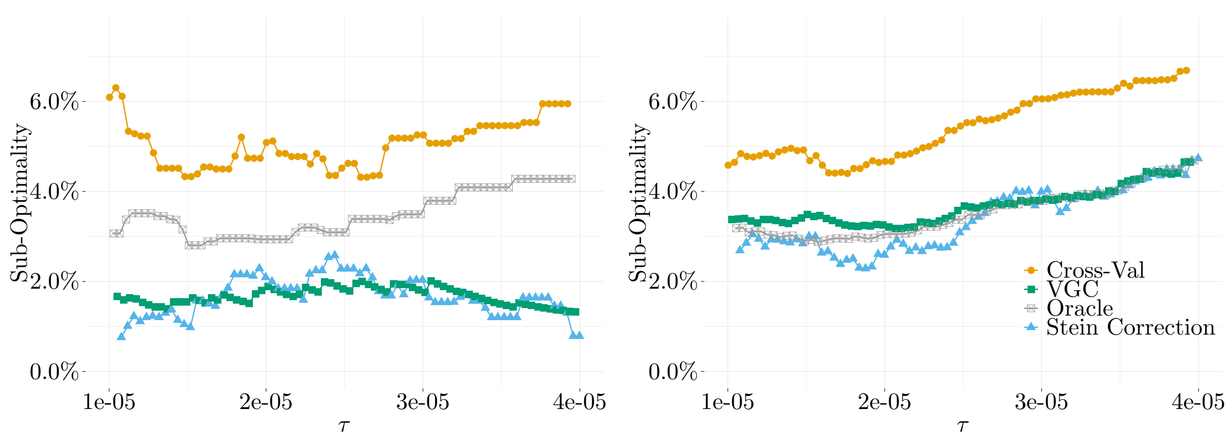

Figure 5 **Estimating Performance across Policy Class.** The first graph shows the estimates of out-of-sample performance across the policy class for the parameter $\tau \in [25\nu_{\min}, 100\nu_{\min}]$ and $\mu_0 = 1000$ for one sample path when $K = 400$. The second graph is similar, but for $K = 3200$. Both plots highlight the smoothness of VGC relative to the Stein-Correction.

the Stein Correction. We also note that while it appears both Stein and VGC systematically overestimate performance, this is an artifact of the particular sample path chosen. By contrast, cross-validation does systematically underestimate performance, because it estimates the performance of a policy trained with half the data, which is necessarily worse.

In our third experiment, we highlight the differences in the policy selected by the VGC estimator and the policy selected by cross-validation. In Fig. 6, we plot the routing decisions of each policy and color code them by the true (oracle) amount of time saved. We highlight two regions (labeled (a) and (b)) on the map where drones arrive after the ambulance. We see that in those regions, the

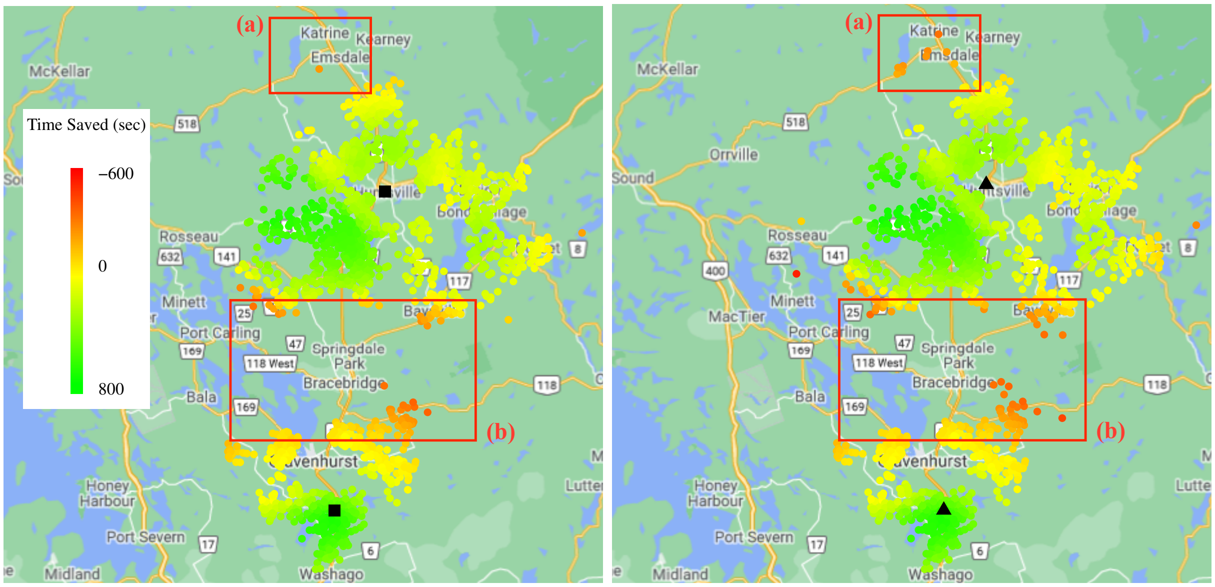

Figure 6 **Comparing Policy Decisions.** Left (resp. right) panel shows routing decisions for the policy selected by VGC (resp. cross-validation). Color indicates time-saved relative to an ambulance-only policy (green is good, red is bad) computed relative to the ground truth. Although routing is largely similar, Regions (a) and (b) highlight some differences where the cross-validation policy makes poorer routing decisions (more orange dots). The larger black points are drone depots.

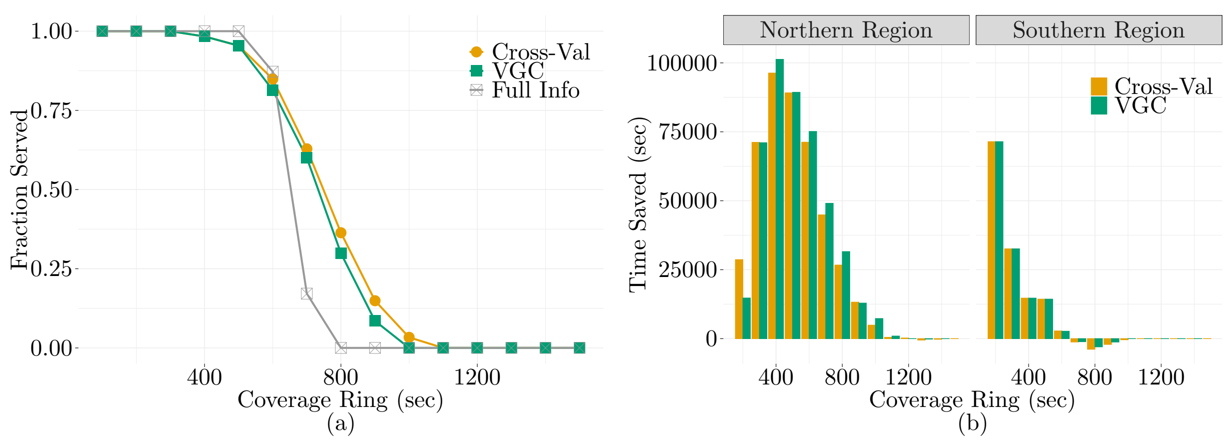

Figure 7 **Estimating Performance across Policy Class.** Each data point in the graph represents the performance metric of each selected policy for a ring shaped region corresponding to distance in time from a drone depot. Graph (a) shows the fraction of patients served in each region for the patients serviced by the Southern depot in Fig. 6. Graph (b) plots the time saved by each policy. The plots highlight the performance difference in routing decisions between the two policies.

cross-validation policy routes to more patients/regions where the drone arrives after the ambulance, thus potentially wasting drone resources and availability. We see in the regions outside of (a) and (b) that routing decisions of the two polices are similar and result in the drone arriving before the

the ambulance. Additionally, we note that the drone depots in the southern region of the map are the same for both policies while the drone depots in the northern region are different.

In Fig. 7, we plot key performance metrics for regions organized by their distance from a drone depot for these two policies. Specifically, we group OHCA events into "coverage rings" based on the travel time from the depot to the event. Each ring is of "width" 100s. For example, the 800 seconds coverage ring corresponds to all OHCA events that are between 701 and 800 seconds away from the drone depot. In Panel (a) of Fig. 7, we restrict attention to the the southern region Fig. 6 where both policies have selected the same drone depot so that we can focus on routing decisions. We plot the fraction of patients served for each coverage ring. We see that the policy chosen by VGC is more conservative with routing in comparison to the policy chosen by cross-validation and more closely aligns the full information benchmark.

In panel (b) of Fig. 7, we compare the time saved between the two policies. We organize the regions into the North and South corresponding to the servicing depots. In the northern region, we see that the VGC policy saves more time in coverage rings further away from the drone depot by sacrificing time saved in closer coverage rings. This difference partially corresponds with the more conservative routing decisions of the VGC, but also can be attributed to the choice of drone depot. We see the VGC policy chooses a depot in less densely populated region that is more centralized overall, while the cross-validation chooses a depot closer to more densely populated regions in terms of OHCA occurrences. In total, we see that the VGC policy saves 1.43% more time in comparison to the cross-validation policy. However, if we breakdown the time saved with respect to minimum distance from a depot, we see that for patients within 600 seconds of an existing drone depot, the VGC performs less than 1% worse in comparison to the cross-validation policy. However, if we consider patients more than 600 seconds away from existing drone depot, the VGC policy saves 13.8% more time in comparison to cross-validation. We interpret this to mean that both the VGC and cross-validation policies make similar performing depot decisions, but VGC makes significantly better routing decisions, particularly at long distances. Since these long distances are precisely where the imprecision is most crucial, we argue this is a relevant advantage.

Finally, we also compare how higher fold cross-validation performs with respect to VGC. In (a) of Fig. 8, we first show how the cross-validation estimators for three different policies performs in estimation error relative to VGC as we vary the number of folds. For all the policies (where increasing τ corresponds to lower variance policies), the plot shows the root mean-squared error (MSE) of cross-validation is uniformly larger than VGC over the folds considered. Furthermore, the plot highlights a drawback of cross-validation, which is that it is not even clear how to select the optimal number of folds in order to minimize the bias variance trade-off of the cross-validation

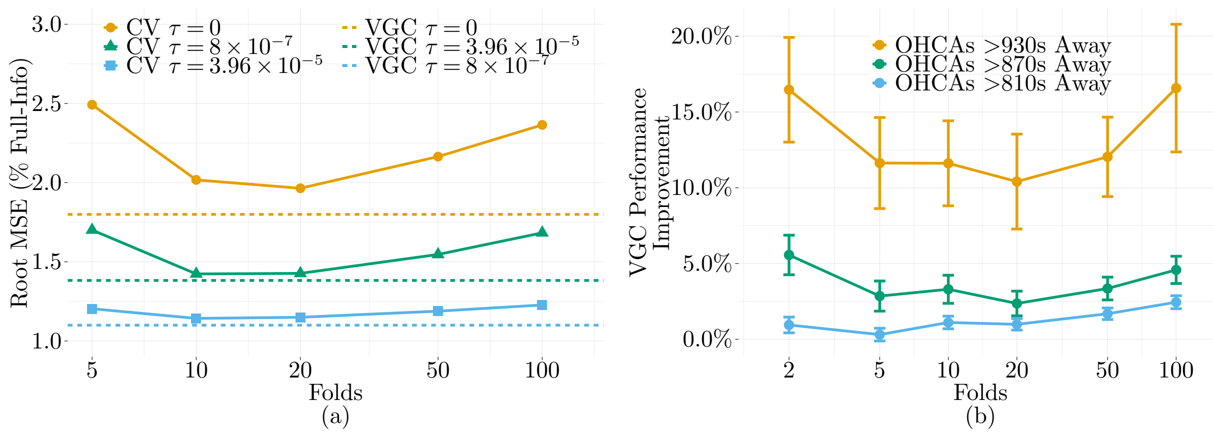

Figure 8 **Varying Cross-Validation Folds.** We plot the policy evaluation and learning performance of cross-validation with different folds across 500 simulations. In each simulation there are 100 samples of $\boldsymbol{Z}^i$. In (a), we plot the mean squared error of cross-validation for three different policies and compare them with the respective VGC estimates represented by the dotted lines. In (b), we plot the percent improvement VGC has over cross-validation, so larger bars indicate lower cross-validation performance.

estimator. In comparison, the VGC with minimal guidance on the choice of the h parameter out-performs common choices of folds for cross-validation such as 5-fold cross-validation and leave-one-out cross-validation. In (b) of Fig. 8, we show how the performance of policy evaluation translates to policy learning in regions further away from the depot, where drone decisions are most crucial. As expected, VGC out-performs cross-validation across all choices of folds and the fold that performs most similarly to VGC, 20-fold, also corresponds to the cross-validation fold with the lowest MSE. This observation suggests with an "ideal" number of folds, cross-validation can perform well in this example, but identifying the right number of folds is non-trivial.

6. Conclusion

Motivated by the poor performance of cross-validation in data-driven optimization problems where data are scarce, we propose a new estimator of the out-of-sample performance of an affine plug-in policy. Unlike cross-validation, our estimator avoids sacrificing data and uses all the data when training, making it well-suited to settings with scarce data. We prove that our estimator is nearly unbiased, and for "stable" optimization problems – problems whose optimal solutions do not change too much when the coefficient of a single random component changes – the estimator's variance vanishes. Our notion of stability leads us to consider two special classes of weakly-coupled optimization problems: weakly-coupled-by-variables and weakly-coupled-by-constraints. For each class of problems, we prove an even stronger result and provide high-probability bounds on the error of our estimator that holds uniformly over a policy class. Additionally, in our analysis of optimization problems weakly-coupled-by-constraints, we provide new insight on the stability of the dual solutions. This new insight may provide further insight in problems that leverage the dual solution such as online linear programming.

Our work offers many exciting directions for future research. Our solution approach to the weakly-coupled problems exploits the decomposability of the underlying optimization problems. We believe that such an approach can be generalized to other settings. Finally, our analysis strongly leverages the linearity of the affine plug-in policy class; it is an open question if similar debiasing techniques might be developed to handle nonlinear objective functions as well.

Acknowledgments

The first two author were partially supported by the National Science Foundation under Grant No. CMMI-1661732. All three authors thank Prof. Tim Chan and Prof. Justin Boutilier for sharing simulation results and details pertaining to their paper (Boutilier and Chan 2019), and the two anonymous reviewers who provided feedback on an earlier revision.

References

Aouad, Ali, Adam N Elmachtoub, Kris J Ferreira, Ryan McNellis. 2019. Market segmentation trees. *arXiv preprint arXiv:1906.01174* .

Baardman, Lennart, Maxime C. Cohen, Kiran Panchamgam, Georgia Perakis, Danny Segev. 2019. Scheduling promotion vehicles to boost profits. *Management Science* **65**(1) 50–70.

Bertsekas, Dimitri P. 1997. Nonlinear programming. *Journal of the Operational Research Society* **48**(3) 334–334.

Bousquet, Olivier, André Elisseeff. 2001. Algorithmic stability and generalization performance. *Advances in Neural Information Processing Systems*. 196–202.

Boutilier, Justin J, Timothy CY Chan. 2019. Response time optimization for drone-delivered automated external defibrillators. *arXiv preprint arXiv:1908.00149* .

Breiman, Leo. 1996. Bagging predictors. *Machine Learning* **24**(2) 123–140.

Buja, Andreas, Trevor Hastie, Robert Tibshirani. 1989. Linear smoothers and additive models. *The Annals of Statistics* 453–510.

Cheskes, Sheldon, Shelley L McLeod, Michael Nolan, Paul Snobelen, Christian Vaillancourt, Steven C Brooks, Katie N Dainty, Timothy CY Chan, Ian R Drennan. 2020. Improving access to automated external defibrillators in rural and remote settings: a drone delivery feasibility study. *Journal of the American Heart Association* **9**(14) e016687.

Chu, Jamal, KH Benjamin Leung, Paul Snobelen, Gordon Nevils, Ian R Drennan, Sheldon Cheskes, Timothy CY Chan. 2021. Machine learning-based dispatch of drone-delivered defibrillators for out-of-hospital cardiac arrest. *Resuscitation* **162** 120–127.

Cohen, Maxime C, Jeremy J Kalas, Georgia Perakis. 2021. Promotion optimization for multiple items in supermarkets. *Management Science* **67**(4) 2340–2364.

Cohen, Maxime C., Ngai-Hang Zachary Leung, Kiran Panchamgam, Georgia Perakis, Anthony Smith. 2017. The impact of linear optimization on promotion planning. *Operations Research* **65**(2) 446–468.

Combes, Richard. 2015. An extension of mcdiarmid's inequality. *arXiv preprint arXiv:1511.05240* .

Deng, Yunxiao, Suvrajeet Sen. 2018. Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming. *INFORMS Journal on Optimization (submitted)* .

Donti, Priya, Brandon Amos, J Zico Kolter. 2017. Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems*. 5484–5494.

Dudík, Miroslav, John Langford, Lihong Li. 2011. Doubly robust policy evaluation and learning. *Proceedings of the 28th International Conference on International Conference on Machine Learning*. 1097–1104.

El Balghiti, Othman, Adam N Elmachtoub, Paul Grigas, Ambuj Tewari. 2019. Generalization bounds in the predict-then-optimize framework. *Advances in Neural Information Processing Systems*. 14412–14421.

Elmachtoub, Adam, Jason Cheuk Nam Liang, Ryan McNellis. 2020. Decision trees for decision-making under the predict-then-optimize framework. *International Conference on Machine Learning*. PMLR, 2858–2867.

Elmachtoub, Adam N, Paul Grigas. 2021. Smart "predict, then optimize". *Management Science* .

Friedman, J., T. Hastie, R. Tibshirani. 2001. *The Elements of Statistical Learning*. Springer, Berlin.

Gamarnik, David. 2013. Correlation decay method for decision, optimization, and inference in large-scale networks. *Theory Driven by Influential Applications*. INFORMS, 108–121.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Rubin. 2014. *Bayesian Data Analysis*, vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Gupta, Vishal, Nathan Kallus. 2021. Data pooling in stochastic optimization. *Management Science* URL `https://doi.org/10.1287/mnsc.2020.3933`.

Gupta, Vishal, Paat Rusmevichientong. 2021. Small-data, large-scale linear optimization with uncertain objectives. *Management Science* **67**(1) 220–241.

Hastie, Trevor, Robert Tibshirani, Ryan Tibshirani. 2020a. Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science* **35**(4) 579 – 592. doi: 10.1214/19-STS733. URL `https://doi.org/10.1214/19-STS733`.

Hastie, Trevor, Robert Tibshirani, Ryan Tibshirani. 2020b. Best subset, forward stepwise or lasso? analysis and recommendations based on extensive comparisons. *Statistical Science* **35**(4) 579–592.

Ignatiadis, Nikolaos, Stefan Wager. 2019. Covariate-powered empirical bayes estimation. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett, eds., *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.

Ito, Shinji, Akihiro Yabe, Ryohei Fujimaki. 2018. Unbiased objective estimation in predictive optimization. Jennifer Dy, Andreas Krause, eds., *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. Stockholm Sweden, 2176–2185.

Javanmard, Adel, Andrea Montanari. 2014. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15**(1) 2869–2909.

Javanmard, Adel, Andrea Montanari. 2018. Debiasing the lasso: Optimal sample size for gaussian designs. *Annals of Statistics* **46**(6A) 2593–2622.

Javanmard, Adel, Hamid Nazerzadeh, Simeng Shao. 2020. Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity. *2020 IEEE International Symposium on Information Theory (ISIT)*. 2652–2657. doi:10.1109/ISIT44484.2020.9174296.

Kearns, Michael, Dana Ron. 1999. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation* **11**(6) 1427–1453.

LeVeque, Randall J. 2007. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM.

Li, Xiaocheng, Yinyu Ye. 2019. Online linear programming: Dual convergence, new algorithms, and regret bounds. *arXiv preprint arXiv:1909.05499* .

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

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

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

Rusmevichientong, David P. Williamson, Paat. 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.

Sanfridsson, J, J Sparrevik, J Hollenberg, P Nordberg, T Djärv, M Ringh, L Svensson, S Forsberg, A Nord, Magnus Andersson-Hagiwara, et al. 2019. Drone delivery of an automated external defibrillator–a mixed method simulation study of bystander experience. *Scandinavian journal of trauma, resuscitation and emergency medicine* **27**(1) 1–9.

Schmoys, David, Shujing Wang. 2019. How to solve a linear optimization problem on incentive allocation? URL `https://eng.lyft.com/how-to-solve-a-linear-optimization-problem-on-incentive-allocation-5a8fb5d04db1`.

Shalev-Shwartz, Shai, Ohad Shamir, Nathan Srebro, Karthik Sridharan. 2010. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research* **11** 2635–2670.

Shao, Jun. 1993. Linear model selection by cross-validation. *Journal of the American statistical Association* **88**(422) 486–494.

Smith, James E, Robert L Winkler. 2006. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science* **52**(3) 311–322.

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

Wager, Stefan, Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**(523) 1228–1242.

Wainwright, Martin J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.

Wilder, Bryan, Bistra Dilkina, Milind Tambe. 2019a. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33. 1658–1665.

Wilder, Bryan, Eric Ewing, Bistra Dilkina, Milind Tambe. 2019b. End to end learning and optimization on graphs. *Advances in Neural Information Processing Systems*. 4672–4683.

Zhang, Cun-Hui, Stephanie S. Zhang. 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* 217–242.

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: Debiasing In-Sample Policy Performance for Small-Data, Large-Scale Optimization

Appendix A: Background Results on Empirical Processes

In this appendix we collect some results on the suprema of empirical processes that we will require in our proofs. All results are either known or easily derived from known results. Our summary is necessarily brief and we refer the reader to Pollard (1990) for a self-contained exposition.

Let $\Psi(t) = \frac{1}{5}\exp(t^2)$. For any real-valued random variable Z, we define the Ψ-norm $\|Z\|_\Psi$ to be $\|Z\|_\Psi \equiv \inf\{C > 0 : \mathbb{E}[\Psi(|Z|/C)] \le 1\}$. Random variables with finite Ψ-norm are sub-Gaussian random variables. We first recall a classical result on the suprema of sub-Gaussian processes over finite sets.

Theorem A.1 (Suprema of Stochastic Processes over Finite Sets) *Let*

$$\boldsymbol{f}(\boldsymbol{\theta}) = (f_1(\boldsymbol{\theta}), \ldots, f_K(\boldsymbol{\theta})) \in \mathbb{R}^K$$

be a vector of K independent stochastic processes indexed by $\boldsymbol{\theta} \in \Theta$. Let $\boldsymbol{F} \in \mathbb{R}^K_+$ be a random vector such that $|f_k(\boldsymbol{\theta})| \le F_k$ for all $\boldsymbol{\theta} \in \Theta$, $k = 1, \ldots, K$, and suppose there exists a constant $M < \infty$ such that $|\{\boldsymbol{f}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}| \le M$ almost surely. Then, for any $R > 1$, there exists an absolute constant C such that with probability $1 - e^{-R}$

$$\sup_{\boldsymbol{\theta}\in\Theta} \left| \sum_{k=1}^K f_k(\boldsymbol{\theta}) - \mathbb{E}\left[\sum_{k=1}^K f_k(\boldsymbol{\theta})\right] \right| \le C \cdot R \cdot \| \|\boldsymbol{F}\|_2 \|_\Psi \sqrt{\log M}.$$

Proof: The result follows from the discussion leading up to Eq. (7.4) of Pollard (1990) after noting that the entropy integral ($J_n(\omega)$ in the notation of Pollard (1990)) is at most $9\|\boldsymbol{F}\|_2\sqrt{\log M}$ given the conditions of the theorem. □

When considering the suprema over potentially infinite sets, we must characterize the "size" of $\{\boldsymbol{f}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ more carefully. Recall for any set $\mathcal{F}$, the ϵ-packing number of $\mathcal{F}$ is the largest number of points we can select in $\mathcal{F}$ such that no two points are within ℓ_2 distance ϵ. We denote this packing number by $D(\epsilon, \mathcal{F})$. We restrict attention to sets whose packing numbers do not grow too fast.

Definition A.2 (Euclidean Sets) *We say a set $\mathcal{F}$ is* Euclidean *if there exists constants A and W such that*

$$D(\epsilon\delta, \mathcal{F}) \le A\epsilon^{-W} \quad \forall 0 < \epsilon < 1,$$

where $\delta \equiv \sup_{\boldsymbol{f}\in\mathcal{F}} \|\boldsymbol{f}\|$.

Furthermore, note that in the special case that $F_k \leq U$, Theorem A.1 bounds the suprema by a term that scales like $U\sqrt{K}$. This bound can be quite loose since $f_k(\boldsymbol{\theta})$ typically takes values much smaller than U and is only occasionally large. Our next result provides a more refined bound on the suprema when the pointwise variance of the process is relatively small and the relevant (random) set is Euclidean. We stress the parameters A and W below must be deterministic.

Theorem A.3 (Suprema of Stochastic Processes with Small Variance) *Suppose that the set $\{\boldsymbol{f}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\} \subseteq \mathbb{R}^K$ is Euclidean with parameters A and W almost surely. Suppose also that*

i) There exists a constant U such that $\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{f}(\boldsymbol{\theta})\|_\infty \leq U$, almost surely, and

ii) There exists a constant σ^2 such that $\sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\left[\|\boldsymbol{f}(\boldsymbol{\theta})\|_2^2\right] \leq K\sigma^2$.

Then, there exists an absolute constant C such that for any $R > 1$, with probability at least $1 - e^{-R}$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \left| \sum_{k=1}^K f_k(\boldsymbol{\theta}) - \mathbb{E}\left[\sum_{k=1}^K f_k(\boldsymbol{\theta})\right] \right| \leq CR \cdot V(A,W)\sqrt{K}\left(\sigma + \frac{UV(A,W)}{\sqrt{K}}\right),$$

where $V(A,W) \equiv \frac{\log A + W}{\sqrt{\log A}}$.

Remark A.4 Notice that when K is sufficiently large, the term in the parenthesis is dominated by 2σ, and hence the bound does not depend on U. Theorem A.3 is not tight in its dependence on R. See, for example, Talagrand's Inequality for the suprema of the empirical process (Wainwright (2019)). We prefer Theorem A.3 to Talagrand's inequality in what follows because it is somewhat easier to apply and is sufficient for our purposes.

Proof of Theorem A.3. For convenience in what follows, let $\mathcal{F}$ be the (random) set $\{\boldsymbol{f}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. Let $\delta \equiv \sup_{\boldsymbol{f}\in\mathcal{F}} \|\boldsymbol{f}\|_2$.

Our goal will be to apply Theorem A.2 of GR 2021. That theorem shows that there exists an absolute constant C_1 such that

$$\sup_{\boldsymbol{f}\in\mathcal{F}} \left| \sum_{k=1}^K f_k - \mathbb{E}\left[\sum_{k=1}^K f_k\right] \right| \leq C_1 R V(A,W) \|\delta\|_\Psi. \tag{A.1}$$

Thus, the remainder of the proof focuses on bounding $\|\delta\|_\Psi$. As an aside, a naive bound $\|\delta\|_\Psi \leq U\sqrt{K}$, so we know that this value is finite. In what follows, we seek a stronger bound.

Write

$$\begin{aligned}
\delta^2 &= \sup_{\boldsymbol{f}\in\mathcal{F}} \|\boldsymbol{f}\|_2^2 \\
&\leq \sup_{\boldsymbol{f}\in\mathcal{F}} \|\boldsymbol{f}\|_2^2 - \mathbb{E}\left[\|\boldsymbol{f}\|_2^2\right] + \sup_{\boldsymbol{f}\in\mathcal{F}} \mathbb{E}\left[\|\boldsymbol{f}\|_2^2\right] \\
&\leq \sup_{\boldsymbol{f}\in\mathcal{F}} \|\boldsymbol{f}\|_2^2 - \mathbb{E}\left[\|\boldsymbol{f}\|_2^2\right] + K\sigma^2
\end{aligned}$$

Let $C_2 > 0$ be a constant to be determined later. Dividing by C_2 and taking expectations above shows

$$\mathbb{E}\left[e^{\frac{\delta^2}{C_2}}\right] \;\le\; e^{\frac{K\sigma^2}{C_2}} \cdot \mathbb{E}\left[e^{\frac{\bar{Z}}{C_2}}\right], \tag{A.2}$$

where

$$\bar{Z} \;\equiv\; \sup_{\boldsymbol{f}\in\mathcal{F}} \left\{ \|\boldsymbol{f}\|_2^2 - \mathbb{E}\left[\|\boldsymbol{f}\|_2^2\right] \right\} \;=\; \sup_{\boldsymbol{f}\in\mathcal{F}} \left\{ \sum_{k=1}^{K} f_j^2 - \mathbb{E}\left[f_j^2\right] \right\}$$

Importantly, $\bar{Z}$ is again the suprema of an empirical process, namely for the "squared" elements. Pollard (1990) provides bounds on $\bar{Z}$ in terms of the entropy integral of the process.

Specifically, let $\boldsymbol{f}^2$ denote the vector whose j^{th} element is f_j^2. Let $\mathcal{F}^2 = \{\boldsymbol{f}^2 : \boldsymbol{f} \in \mathcal{F}\}$. Then the entropy integral of the squared process is defined to be

$$\overline{J} \;\equiv\; 9\int_0^{\overline{\delta}} \sqrt{\log D(x, \mathcal{F}^2)}\,dx,$$

where $\overline{\delta} \equiv \sup_{\boldsymbol{f}\in\mathcal{F}} \|\boldsymbol{f}^2\|_2$.

Then, in the discussion just prior to Eq. (7.4) of Pollard (1990), it is proven that

$$\mathbb{E}\left[e^{\bar{Z}/\|\bar{J}\|_\Psi}\right] \;\le 25. \tag{A.3}$$

Hence, to bound the right side of Eq. (A.2), we will next bound $\|\bar{J}\|_\Psi$. This in turn will allow us to bound $\|\delta\|_\Psi$ and invoke Theorem A.2 of GR 2021.

To this end, observe that for any $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{F}$, we have

$$\|\boldsymbol{f}^2 - \boldsymbol{g}^2\|^2 \;=\; \sum_{k=1}^{K} (f_j^2 - g_j^2)^2 \;= \sum_{k=1}^{K} (f_j + g_j)^2 (f_j - g_j)^2 \;\le\; 4U^2 \|\boldsymbol{f} - \boldsymbol{g}\|^2.$$

Hence, $D(\epsilon, \mathcal{F}^2) \;\le\; D\left(\frac{\epsilon}{2U}, \mathcal{F}\right)$. Write

$$\begin{aligned}
\overline{J} \;&\equiv\; 9\int_0^{\overline{\delta}} \sqrt{\log D(x, \mathcal{F}^2)}\,dx \\
&\le\; 9\int_0^{\overline{\delta}} \sqrt{\log D\left(\frac{x}{2U}, \mathcal{F}\right)}\,dx. \\
&=\; 2U \cdot 9 \int_0^{\frac{\overline{\delta}}{2U}} \sqrt{\log D\left(x, \mathcal{F}\right)}\,dx.
\end{aligned}$$

where the last equality is a change of variables. We now claim we can upper bound this last expression by replacing the upper limit of integration with δ. Indeed, if $\frac{\overline{\delta}}{2U} \le \delta$, then because the integrand is nonnegative, we only increase the integral. If, $\frac{\overline{\delta}}{2U} > \delta$, then note that

$$\int_\delta^{\frac{\overline{\delta}}{2U}} \sqrt{\log D(x, \mathcal{F})}\,dx = 0,$$

since $D(x,\mathcal{F})=1$ for all $x\geq\delta$. Thus, in either case we can replace the upper limit of integration, yielding

$$\bar{J} \;\leq\; 18U\int_0^\delta \sqrt{\log D(x,\mathcal{F})}dx.$$

Recall the entropy integral of the original process is given by

$$J \equiv 9\int_0^\delta \sqrt{\log D(x,\mathcal{F})}dx.$$

Hence,

$$\overline{J} \leq 2UJ.$$

Moreover, Theorem A.2 of GR 2021 shows that $\|J\|_\Psi \leq C_3\|\delta\|_\Psi V(A,W)$ for some absolute constant C_3. Thus we have successfully bounded

$$\|\bar{J}\|_\Psi \;\leq\; 2UC_3\|\delta\|_\Psi V(A,W).$$

Substituting back into Eq. (A.3) shows that

$$\mathbb{E}\left[e^{\frac{\bar{Z}}{2UC_3\|\delta\|_\Psi V(A,W)}}\right] \;\leq\; 25.$$

Now choose C_2 in Eq. (A.2) to be $C_2=\alpha 2UC_3\|\delta\|_\Psi V(A,W)$ for some $\alpha>0$ to be determined later. Substituting our bound on $\bar{Z}$ into Eq. (A.2) shows

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\delta^2}{\alpha 2UC_3\|\delta\|_\Psi V(A,W)}\right)\right] &\leq\; e^{\frac{K\sigma^2}{\alpha 2UC_3\|\delta\|_\Psi V(A,W)}}\cdot\mathbb{E}\left[e^{\frac{\bar{Z}}{\alpha 2UC_3\|\delta\|_\Psi V(A,W)}}\right]\\ &\leq\; \exp\left(\frac{K\sigma^2}{\alpha 2UC_3\|\delta\|_\Psi V(A,W)}\right)\cdot 25^{1/\alpha},\end{aligned}\tag{A.4}$$

where we have used $\alpha>0$ and Jensen's Inequality to simplify.

We now to choose α large enough that the right side is at most 5. Taking logs, it suffices to choose α such that

$$\begin{aligned}\log(5) \;\geq\; &\frac{1}{\alpha}\left(\log(25)+\frac{K\sigma^2}{2UC_3\|\delta\|_\Psi V(A,W)}\right)\\ \iff \alpha \;\geq\; &2+\frac{K\sigma^2}{2UC_3\log(5)\cdot\|\delta\|_\Psi V(A,W)}.\end{aligned}$$

Substituting into Eq. (A.4) shows

$$\|\delta\|_\Psi^2 \;\leq\; \left(2+\frac{K\sigma^2}{2UC_3\log(5)\cdot\|\delta\|_\Psi V(A,W)}\right)2UC_3\|\delta\|_\Psi V(A,W) \;=\; 4C_3UV(A,W)\|\delta\|_\Psi+\frac{K\sigma^2}{\log(5)}$$

In summary, $\|\delta\|_\Psi$ is at most the largest solution to the quadratic inequality

$$y^2-by-c\leq 0,$$

where

$$b = 4C_3 UV(A,W) \qquad \text{and} \qquad c = \frac{K\sigma^2}{\log(5)}.$$

Bounding the largest root shows

$$\begin{aligned} y &\le \frac{b}{2} + \frac{\sqrt{b^2+4c}}{2} \\ &\le \frac{b}{2} + \frac{b+2\sqrt{c}}{2} && \text{(Triangle-Inequality)} \\ &= b+\sqrt{c}. \end{aligned}$$

Or in other words,

$$\|\delta\|_\Psi \ \le \ 4C_3UV(A,W) + \sigma\sqrt{K},$$

where we upper bounded $(\sqrt{\log(5)})^{-1} \le 1$.

Now simply substitute into Eq. (A.1) and collect constants to complete the proof. $\square$

A.1. Method of Bounded Differences Excluding an Exceptional "Bad" Set

In our analysis, we utilize an extension of McDiarmid's inequality due to Combes (2015). Recall, McDiarmid's inequality shows that for a random vector $\boldsymbol{Z} \in \mathcal{X}$ with independent components and function $f : \mathcal{X} \to \mathbb{R}$ such that

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le \sum_{i=1}^n c_i \mathbb{I}\{x_i \ne y_i\} \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^2, \tag{A.5}$$

for some $\boldsymbol{c} \in \mathbb{R}^n$, we have that

$$\mathbb{P}\{|f(\boldsymbol{Z}) - \mathbb{E}[f(\boldsymbol{Z})]| \ge t\} \ \le \ 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

The next result extends McDiarmid's inequality to a setting where Eq. (A.5) only holds for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Y}^2$ where $\mathcal{Y} \subseteq \mathcal{X}$ is a certain "good" set:

Theorem A.5 (Combes (2015)) *Let $\boldsymbol{Z} \in \mathcal{X}$ be a random vector with independent components and $f : \mathcal{X} \mapsto \mathbb{R}$ be a function such that*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \ \le \ \sum_{j=1}^n c_j \mathbb{I}\{x_j \ne y_j\} \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Y},$$

for some vector $\boldsymbol{c} \in R^n$, where $\mathcal{Y} \subseteq \mathcal{X}$. Let $\bar{c} = \sum_{i=1}^n c_i$, and $p = \mathbb{P}\{X \notin \mathcal{Y}\}$. Then, for any $t > 0$

$$\mathbb{P}(|f(\boldsymbol{Z}) - \mathbb{E}[f(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{Y}]| \ge t + p\bar{c}) \ \le \ 2\left(p + \exp\left\{-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right\}\right)$$

In particular, this implies that for any $\epsilon > 2p$, with probability at least $1-\epsilon$,

$$|f(\boldsymbol{Z}) - \mathbb{E}[f(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{Y}]| \ \le p\bar{c} + \|\boldsymbol{c}\|_2 \sqrt{\log\left(\frac{2}{\epsilon - 2p}\right)}.$$

Remark A.6 *In the special case that $\mathcal{Y} = \mathcal{X}$, then $p = 0$, the theorem recovers McDiarmid's inequality.*

Appendix B: Properties of the Variance Gradient Correction (VGC)

First, we state the relevant portion of Danskin's Theorem for reference. See (Bertsekas 1997, Section B.5) for a proof of a more general result:

THEOREM B.1 **(Derivative Result of Danskin's Theorem)**. *Let $Z \subseteq \mathbb{R}^m$ be a compact set, and let $\phi : \mathbb{R}^n \times Z \mapsto \mathbb{R}$ be continuous and such that $\phi(\cdot, z) : \mathbb{R}^n \mapsto \mathbb{R}$ is convex for each $z \in Z$. Additionally, define*

$$Z(x) = \left\{ \bar{z} \ : \ \phi(x, \bar{z}) = \max_{z \in Z} \phi(x, z) \right\}.$$

Consider the function $f : \mathbb{R}^n \mapsto \mathbb{R}$ given by

$$f(x) = \max_{z \in Z} \phi(x, z).$$

If Z(x) consists of a unique point $\bar{z}$ and $\phi(\cdot, \bar{z})$ is differentiable at x, then f is differentiable at x, and $\nabla f(x) = \nabla_x \phi(x, \bar{z})$, where $\nabla_x \phi(x, \bar{z})$ is the vector with coordinates

$$\frac{\partial \phi(x, \bar{z})}{\partial x_i}, \qquad i = 1, \ldots, n.$$

The remainder of the section contains proofs of the results in Section 3.

B.1. Proof of Theorem 3.2

This section contains the omitted proofs leading to the proof of Theorem 3.2. We first relate finite difference approximations of the subgradients of $V(\boldsymbol{z} + t\boldsymbol{e}_j)$ to their true values.

Lemma B.1 (Subgradients Bound Finite Difference Approximation) *For any $\boldsymbol{z} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, we have*

$$a_j x_j(\boldsymbol{z} + t\boldsymbol{e}_j)t \ \leq V(\boldsymbol{z} + t\boldsymbol{e}_j) - V(\boldsymbol{z}) \ \leq \ a_j x_j(\boldsymbol{z})t.$$

Proof: Let $f(t) = V(\boldsymbol{z} + t\boldsymbol{e}_j)$. Recall that $f(t)$ is concave and $f'(t) = a_j x_j(\boldsymbol{z} + t\boldsymbol{e}_j)$ by Danskin's Theorem. Hence, by the subgradient inequality for concave functions, $f(t) \leq f(0) + f'(0)t$ and $f(0) \leq f(t) - tf'(t)$, and thus, $tf'(t) \leq f(t) - f(0) \leq tf'(0)$. This is equivalent to $a_j x_j(\boldsymbol{z} + t\boldsymbol{e}_j)t \ \leq V(\boldsymbol{z} + t\boldsymbol{e}_j) - V(\boldsymbol{z}) \ \leq \ a_j x_j(\boldsymbol{z})t$, which is the desired result. $\square$

Equipped with Lemma B.1, the proof of Lemma 3.3 is nearly immediate.

Proof of Lemma 3.3: The bounds in Lemma B.1 show that $a_j t(x_j(\boldsymbol{z}) - x_j(\boldsymbol{z} + t\boldsymbol{e}_j)) \geq 0$. If $a_j \geq 0$, it follows that $t(x_j(\boldsymbol{z}) - x_j(\boldsymbol{z} + t\boldsymbol{e}_j)) \geq 0$ for all t, which shows that $t \mapsto x_j(\boldsymbol{z} + t\boldsymbol{e}_j)$ is non-increasing. Similarly, if $a_j < 0$, it follows that $t(x_j(\boldsymbol{z}) - x_j(\boldsymbol{z} + t\boldsymbol{e}_j)) \leq 0$ for all t, which shows that $t \mapsto x_j(\boldsymbol{z} + t\boldsymbol{e}_j)$ is non-decreasing. $\square$

Before proving Theorem 3.2, we establish the following intermediary result on the error of a non-randomized forward step, finite difference.

Lemma B.2 (Forward Step Finite Difference Error) *Fix some j such that $a_j \neq 0$ and $0 < h < 1/e$. Then,*

$$\left| \mathbb{E}\left[\xi_j x_j(\boldsymbol{Z}) - \frac{V(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j \boldsymbol{e}_j) - V(\boldsymbol{Z})}{h\sqrt{\nu_j}\boldsymbol{a}_j(\boldsymbol{\theta})} \right] \right| \leq 4h \log\left(\frac{1}{h\sqrt{\nu_{\min}}} \right)$$

In other words, the forward finite step difference introduces a bias of order $\tilde{O}(h)$.

Proof: From Lemma B.1, we see that the term inside the expectation can be upper-bounded by the non-negative term $\xi_j \left[x_j(\boldsymbol{Z}) - x_j(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j \boldsymbol{e}_j) \right]$. Hence,

$$\mathbb{E}\left[\xi_j x_j(\boldsymbol{Z}) - \frac{V(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j \boldsymbol{e}_j) - V(\boldsymbol{Z})}{h\sqrt{\nu_j}\boldsymbol{a}_j(\boldsymbol{\theta})} \right] \leq \mathbb{E}\left[\xi_j \left(x_j(\boldsymbol{Z}) - x_j(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j \boldsymbol{e}_j) \right) \right].$$

To simplify notation, let $g(t) = \mathbb{E}\left[x_j\left(\boldsymbol{Z}_{-j} + \xi_j \boldsymbol{e}_j \right) | \xi_j = t \right]$ where $\boldsymbol{Z}_{-j}$ is identical to $\boldsymbol{Z}$ but has a 0 at the j^{th} component. Then,

$$\left| \mathbb{E}\left[\xi_j \left(x_j(\boldsymbol{Z}) - x_j(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j \boldsymbol{e}_j) \right) \right] \right| = \left| \int_{-\infty}^{\infty} t \left[g(t) - g(t + h\sqrt{\nu_j} t) \right] \phi_j(t)\, dt \right|$$

where $\phi_j(t)$ is the density for $\mathcal{N}(0, 1/\nu_j)$.

To bound the integral, choose a constant $U > 0$ (which we optimize later) and break the integral into three regions, $(-\infty, -U), (-U, U), (U, \infty)$. This yields the upper bound

$$\begin{aligned} &\left| \int_{-\infty}^{\infty} t \left(g(t) - g(t + th\sqrt{\nu_j}) \right) \phi_j(t)\, dt \right| \\ &\quad \leq \underbrace{\int_{-U}^{U} U \left| g(t) - g(t + th\sqrt{\nu_j}) \right| \phi_j(t)\, dt}_{(a)} + \underbrace{\int_{-\infty}^{-U} |t|\, \phi_j(t)\, dt + \int_{U}^{\infty} |t|\, \phi_j(t)\, dt}_{(b)}. \end{aligned}$$

We first bound (a). As the fist step, we attempt to remove the absolute value. From Lemma B.1, $g(\cdot)$ is a monotone function. We claim that for $|t| < U$,

$$\left| g(t) - g(t + h\sqrt{\nu_j} t) \right| \leq \left| g\left(t - Uh\sqrt{\nu_j} \right) - g\left(t + Uh\sqrt{\nu_j} \right) \right|, \tag{B.1}$$

since $(t - Uh\sqrt{\nu_j}, t + Uh\sqrt{\nu_j})$ always contains the interval $(t, t + h\sqrt{\nu_j} t)$. Let

$$b = \begin{cases} Uh\sqrt{\nu_j} & \text{if } a_j > 0, \\ -Uh\sqrt{\nu_j} & \text{otherwise,} \end{cases}$$

so that $\left| g(t - Uh\sqrt{\nu_j}) - g(t + Uh\sqrt{\nu_j}) \right| = g(t - b) - g(t + b)$. Then,

$$\begin{aligned} \int_{-\infty}^{\infty} \left(g(t - b) - g(t + b) \right) \phi_j(t)\, dt &= \int_{-\infty}^{\infty} g\big(t + b\big) \left(\phi_j(t + 2b) - \phi_j(t) \right) dt, \quad \text{(Change of variables)} \\ &= \int_{-\infty}^{\infty} g\left(t + b\right) \int_{t}^{t+2b} -\nu_j z \phi_j(z)\, dz\, dt, \quad (\text{since } \phi_j'(z) = -\nu_j z \phi_j(z)) \end{aligned}$$

$$
\begin{aligned}
&\leq \nu_j \int_{-\infty}^{\infty} |z|\, \phi_j(z) \int_{z-2b}^{z} g\,(t+b)\; dt\, dz, \quad &&\text{(Fubini's Theorem)}\\
&\leq 2\,|b|\, \nu_j \int_{-\infty}^{\infty} |z|\, \phi_j(z) dz, &&\text{(Since } |g(t+b)| \leq 1 \text{ for all } t)\\
&= 4b\sqrt{\frac{\nu_j}{2\pi}}\\
&\leq 2\sqrt{\frac{2}{\pi}} U h \nu_j .
\end{aligned}
$$

To summarize, we have shown that the term (a) satisfies

$$
\int_{-U}^{U} U \left| g(t) - g(t + th\sqrt{\nu_j}) \right| \phi_j(t)\, dt \;\leq\; 2\sqrt{\frac{2}{\pi}} U^2 h \nu_j
$$

To bound (b), we see that first see that

$$
\int_{-\infty}^{-U} |t|\, \phi_j(t)\, dt + \int_{U}^{\infty} |t|\, \phi_j(t)\, dt = 2\int_{U}^{\infty} t\phi_j(t)\, dt = 2\sqrt{\frac{1}{2\pi\nu_j}} \exp\left\{ \frac{-U^2\nu_j}{2} \right\}
$$

where the first equality holds by symmetry.

Putting the bounds of (a) and (b) together, we have

$$
\left| \int_{-\infty}^{\infty} t\left[g(t) - g(t + th\sqrt{\nu_j}) \right] \phi_j(t)\, dt \right| \leq 2\sqrt{\frac{2}{\pi}} U^2 h \nu_j + 2\sqrt{\frac{1}{2\pi\nu_j}} \exp\left\{ \frac{-U^2\nu_j}{2} \right\}
$$

We approximately balance the two terms by letting $U^2 = \frac{2}{\nu_j} \log\left(\frac{1}{h\sqrt{\nu_j}} \right)$. Substituting and simplifying yields

$$
4\sqrt{\frac{2}{\pi}} h \log\left(\frac{1}{h\sqrt{\nu_j}} \right) + h\sqrt{\frac{2}{\pi}} \;\leq\; 4\sqrt{\frac{2}{\pi}} h \log\left(\frac{1}{h\sqrt{\nu_{\min}}} \right) + h\sqrt{\frac{2}{\pi}}.
$$

To simplify, note that $h < 1/e$ and $\nu_{\min} \leq 1$ implies that $\log\left(\frac{1}{h\sqrt{\nu_j}} \right) \geq 1$. Hence, combining the two terms and simplifying provides a bound of $10\sqrt{\frac{1}{2\pi}} h \log\left(\frac{1}{h\sqrt{\nu_{\min}}} \right)$. Note that $10/\sqrt{2\pi} \leq 4$ to complete the proof. □

We can now prove Theorem 3.2.

Proof of Theorem 3.2: Notice that if $a_j(\boldsymbol{\theta}) = 0$, then the j^{th} term contribute nothing to the bias because $\boldsymbol{x}(\boldsymbol{Z})$ is independent of Z_j, so $\mathbb{E}[\xi_j x_j(\boldsymbol{Z})] = 0 = \mathbb{E}[D_j(\boldsymbol{Z})]$. Hence, we focus on terms j where $a_j(\boldsymbol{\theta}) \neq 0$.

Decompose the j^{th} term as

$$
\begin{aligned}
\mathbb{E}[\xi_j x_j(\boldsymbol{Z}) - D_j(\boldsymbol{Z})] \;=\; & \underbrace{\mathbb{E}\left[\xi_j x_j(\boldsymbol{Z}) - \frac{V(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j \boldsymbol{e}_j) - V(\boldsymbol{Z})}{h\sqrt{\nu_j} a_j(\boldsymbol{\theta})} \right]}_{(a)} \\
& + \underbrace{\mathbb{E}\left[\mathbb{E}\left[\frac{V(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j \boldsymbol{e}_j) - V(\boldsymbol{Z})}{h\sqrt{\nu_j} a_j(\boldsymbol{\theta})} - \frac{V(\boldsymbol{Z} + \delta_j \boldsymbol{e}_j) - V(\boldsymbol{Z})}{h\sqrt{\nu_j} a_j(\boldsymbol{\theta})} \,\middle|\, \boldsymbol{Z} \right] \right]}_{(b)}
\end{aligned}
$$

We first bound (b). Canceling out the $V(\boldsymbol{Z})$ yields

$$\frac{1}{h\sqrt{\nu_j}\boldsymbol{a}_j(\boldsymbol{\theta})}\mathbb{E}\left[V(\boldsymbol{Z}+h\sqrt{\nu_j}\xi_j\boldsymbol{e}_j)-V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j)\right].$$

From our previous discussion, $V(\boldsymbol{Z}+h\sqrt{\nu_j}\xi_j\boldsymbol{e}_j)\sim_d V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j)$, whereby $\mathbb{E}\left[V(\boldsymbol{Z}+h\sqrt{\nu_j}\xi_j\boldsymbol{e}_j)-V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j)\right]=0$.

Lemma B.2 bounds (a) by $4h\log\left(\frac{1}{h\sqrt{\nu_{\min}}}\right)$. Summing over j gives us our intended bound. □

B.2. Properties of VGC

We next establish smoothness properties of the VGC.

Proof of Lemma 3.7. We begin with i). We first claim that $\boldsymbol{\theta}\mapsto V(\boldsymbol{z},\boldsymbol{\theta})$ is Lipschitz continuous with parameter $Ln(1+\|\boldsymbol{z}\|_\infty)$. To this end, write

$$\begin{aligned}
V(\boldsymbol{z},\bar{\boldsymbol{\theta}})-V(\boldsymbol{z},\boldsymbol{\theta}) &= \left(\boldsymbol{r}\left(\boldsymbol{z},\bar{\boldsymbol{\theta}}\right)-\boldsymbol{r}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\right)^\top\boldsymbol{x}(\boldsymbol{z},\bar{\boldsymbol{\theta}})-\underbrace{\boldsymbol{r}\left(\boldsymbol{z},\boldsymbol{\theta}\right)^\top\left(\boldsymbol{x}(\boldsymbol{z},\boldsymbol{\theta})-\boldsymbol{x}(\boldsymbol{z},\bar{\boldsymbol{\theta}})\right)}_{\le 0 \text{ by optimality of } \boldsymbol{x}(\boldsymbol{z},\boldsymbol{\theta})}\\
&\le\left|\left(\boldsymbol{r}\left(\boldsymbol{z},\bar{\boldsymbol{\theta}}\right)-\boldsymbol{r}\left(\boldsymbol{z},\boldsymbol{\theta}\right)\right)^\top\boldsymbol{x}(\boldsymbol{z},\bar{\boldsymbol{\theta}})\right|\\
&\le\|\boldsymbol{r}(\boldsymbol{z},\bar{\boldsymbol{\theta}})-\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1\|\boldsymbol{x}(\boldsymbol{z},\bar{\boldsymbol{\theta}})\|_\infty\\
&\le\|\boldsymbol{r}(\boldsymbol{z},\bar{\boldsymbol{\theta}})-\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1 \qquad\qquad (\text{since } \boldsymbol{x}(\boldsymbol{z},\bar{\boldsymbol{\theta}})\in\mathcal{X}\subseteq[0,1]^n)\\
&\le\sum_{j=1}^n\left|a_j(\bar{\boldsymbol{\theta}})-a_j(\boldsymbol{\theta})\right||z_j|+\left|b_j(\bar{\boldsymbol{\theta}})-b_j(\boldsymbol{\theta})\right|\\
&\le\sum_{j=1}^n\left(L\|\boldsymbol{z}\|_\infty\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|+L\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|\right)\\
&=Ln\cdot(1+\|\boldsymbol{z}\|_\infty)\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|.
\end{aligned}$$

Reversing the roles of $\boldsymbol{\theta}$ and $\bar{\boldsymbol{\theta}}$ yields an analogous bound, and, hence,

$$\left|\boldsymbol{r}\left(\boldsymbol{z},\bar{\boldsymbol{\theta}}\right)^\top\boldsymbol{x}(\boldsymbol{z},\bar{\boldsymbol{\theta}})-\boldsymbol{r}\left(\boldsymbol{z},\boldsymbol{\theta}\right)^\top\boldsymbol{x}(\boldsymbol{z},\boldsymbol{\theta})\right|\le Ln\left(1+\|\boldsymbol{z}\|_\infty\right)\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\right\|$$

This proves the first statement.

Next, we claim for any $\boldsymbol{z}$,

$$\left|\frac{1}{a_j(\bar{\boldsymbol{\theta}})}V(\boldsymbol{z},\bar{\boldsymbol{\theta}})-\frac{1}{a_j(\boldsymbol{\theta})}V(\boldsymbol{z},\boldsymbol{\theta})\right| \;\le\; \frac{2nL}{a_{\min}}\left(\frac{a_{\max}}{a_{\min}}\|\boldsymbol{z}\|_\infty+\frac{a_{\max}+b_{\max}}{a_{\min}}\right)\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|. \tag{B.2}$$

Write

$$\begin{aligned}
\left|\frac{1}{a_j(\bar{\boldsymbol{\theta}})}V(\boldsymbol{z},\bar{\boldsymbol{\theta}})-\frac{1}{a_j(\boldsymbol{\theta})}V(\boldsymbol{z},\boldsymbol{\theta})\right| \;&=\; \left|\frac{a_j(\boldsymbol{\theta})V(\boldsymbol{z},\bar{\boldsymbol{\theta}})-a_j(\bar{\boldsymbol{\theta}})V(\boldsymbol{z},\boldsymbol{\theta})}{a_j(\boldsymbol{\theta})a_j(\bar{\boldsymbol{\theta}})}\right|\\
&\le\; \left|\frac{V(\boldsymbol{z},\bar{\boldsymbol{\theta}})-V(z,\boldsymbol{\theta})}{a_j(\bar{\boldsymbol{\theta}})}\right|+\left|\frac{V(z,\boldsymbol{\theta})(a_j(\boldsymbol{\theta})-a_j(\bar{\boldsymbol{\theta}}))}{a_j(\boldsymbol{\theta})a_j(\bar{\boldsymbol{\theta}})}\right|,\\
&\le\; \frac{Ln(1+\|\boldsymbol{z}\|_\infty)\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|}{a_{\min}}+\frac{|V(z,\boldsymbol{\theta})|\,L\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|}{a_{\min}^2},
\end{aligned}$$

where the first inequality follows by adding and subtracting $a_j(\boldsymbol{\theta})V(\boldsymbol{\theta})$ in the numerator, and the second inequality follows from the Lipschitz continuity of $a_j(\boldsymbol{\theta})$ and $V(\boldsymbol{z},\boldsymbol{\theta})$ (Assumption 3.6). Next note that

$$|V(\boldsymbol{z},\boldsymbol{\theta})| \le \|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1 \|\boldsymbol{x}(\boldsymbol{z},\boldsymbol{\theta})\|_\infty \le \|\boldsymbol{a}(\boldsymbol{\theta})\circ \boldsymbol{z}\|_1 + \|\boldsymbol{b}(\boldsymbol{\theta})\|_1 \le n\|\boldsymbol{z}\|_\infty a_{\max} + nb_{\max}.$$

Substituting above and simplifying proves Eq. (B.2)

We can now prove the lemma. Fix a component j. Then,

$$\begin{aligned} D_j(\boldsymbol{z},\bar{\boldsymbol{\theta}}) - D_j(\boldsymbol{z},\boldsymbol{\theta}) &= \mathbb{E}\left[\frac{1}{h\sqrt{\nu_j}a_j(\bar{\boldsymbol{\theta}})}\left(V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\bar{\boldsymbol{\theta}}) - V(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\right) \mid \boldsymbol{Z}=\boldsymbol{z}\right] \\ &\quad - \mathbb{E}\left[\frac{1}{h\sqrt{\nu_j}a_j(\boldsymbol{\theta})}\left(V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta}) - V(\boldsymbol{Z},\boldsymbol{\theta})\right) \mid \boldsymbol{Z}=\boldsymbol{z}\right] \\ &= \frac{1}{h\sqrt{\nu_j}}\mathbb{E}\left[\frac{1}{a_j(\bar{\boldsymbol{\theta}})}V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\bar{\boldsymbol{\theta}}) - \frac{1}{a_j(\boldsymbol{\theta})}V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta}) \mid \boldsymbol{Z}=\boldsymbol{z}\right] \\ &\quad + \frac{1}{h\sqrt{\nu_j}}\mathbb{E}\left[\frac{1}{a_j(\boldsymbol{\theta})}V(\boldsymbol{Z},\boldsymbol{\theta}) - \frac{1}{a_j(\bar{\boldsymbol{\theta}})}V(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) \mid \boldsymbol{Z}=\boldsymbol{z}\right]. \end{aligned}$$

Hence, by taking absolute values and applying Eq. (B.2) twice we obtain

$$\begin{aligned} \left|D_j(\boldsymbol{z},\bar{\boldsymbol{\theta}}) - D_j(\boldsymbol{z},\boldsymbol{\theta})\right| &\le \frac{2nL}{ha_{\min}}\left(\frac{a_{\max}}{a_{\min}}\mathbb{E}\left[\|\boldsymbol{Z}+\delta_j\boldsymbol{e}_j\|_\infty \mid \boldsymbol{Z}=\boldsymbol{z}\right] + \frac{a_{\max}+b_{\max}}{a_{\min}}\right)\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\| \\ &\quad + \frac{2nL}{ha_{\min}}\left(\frac{a_{\max}}{a_{\min}}\|\boldsymbol{z}\|_\infty + \frac{a_{\max}+b_{\max}}{a_{\min}}\right)\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|, \end{aligned}$$

where we have passed through the conditional expectation. Finally, note that $\mathbb{E}\left[\|\boldsymbol{Z}+\delta_j\boldsymbol{e}_j\|_\infty \mid \boldsymbol{Z}=\boldsymbol{z}\right] \le \|\boldsymbol{z}\|_\infty + \mathbb{E}\left[|\delta_j|\right] \le \|\boldsymbol{z}\|_\infty + \sqrt{h^2+2h/\nu_j}$ by Jensen's inequality. We simplify this last expression by noting for $h<1/e$, $h^2<h$, so that

$$\sqrt{h^2+2h/\nu_j} \le \sqrt{h}\sqrt{1+2/\nu_{\min}} \le 2\sqrt{\frac{h}{\nu_{\min}}},$$

using $\nu_{\min}\le 1$. Thus, $\mathbb{E}\left[\|\boldsymbol{Z}+\delta_j\boldsymbol{e}_j\|_\infty \mid \boldsymbol{Z}=\boldsymbol{z}\right] \le \|\boldsymbol{z}\|_\infty + 2\sqrt{\frac{h}{\nu_{\min}}}$. Substituting above and collecting terms yields

$$\frac{4nL}{ha_{\min}}\left(\frac{a_{\max}}{a_{\min}}\|\boldsymbol{z}\|_\infty + \frac{a_{\max}+b_{\max}}{a_{\min}} + \frac{a_{\max}}{a_{\min}}\sqrt{\frac{h}{\nu_{\min}}}\right)\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|. \tag{B.3}$$

We can simplify this expression by letting

$$C_3 \ge \frac{4}{a_{\min}}\cdot\max\left(\frac{a_{\max}}{a_{\min}},\ \frac{a_{\max}+b_{\max}}{a_{min}},\ \frac{a_{\max}}{a_{\min}}\right).$$

Then Eq. (B.3) is at most

$$\frac{C_3nL}{h}\left(\|\boldsymbol{z}\|_\infty + 1 + \sqrt{\frac{h}{\nu_{\min}}}\right)\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_2 \le \frac{C_3nL}{h}\left(\|\boldsymbol{z}\|_\infty + \frac{2}{\sqrt{\nu}_{\min}}\right)\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_2 \le \frac{2C_3nL}{h\sqrt{\nu_{\min}}}\left(\|\boldsymbol{z}\|_\infty+1\right)\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_2,$$

where we have used the bounds on the precisions (Assumption 3.1) and $h < 1/e$ to simplify. Letting $C_1 = 2C_3$ proves the first part of the theorem.

To complete the proof, we require a high-probability bound on $\|\boldsymbol{Z}\|_\infty$. Since $\boldsymbol{Z} - \boldsymbol{\mu}$ is sub-Gaussian, such bounds are well-known (Wainwright 2019), and we have with probability $1 - e^{-R}$,

$$\|\boldsymbol{Z}\|_\infty \;\le\; C_\mu + \|\boldsymbol{Z} - \boldsymbol{\mu}\|_\infty \le C_\mu + \frac{C_4}{\sqrt{\nu_{\min}}}\sqrt{\log n}\sqrt{R},$$

for some universal constant C_4. Substitute this bound into our earlier Lipschitz bound for an arbitrary $\boldsymbol{z}$, and use the Assumption 3.1, $h < 1/e$, and $R > 1$ to collect terms and simplify. We then sum over the n terms of $D(\boldsymbol{Z}, \boldsymbol{\theta})$ to complete the proof for i).

We now bound ii). Focusing on the j^{th} component of $D(\boldsymbol{Z}, (\boldsymbol{\theta}, h))$ and writing

$$D_j\left(\boldsymbol{Z}, (\boldsymbol{\theta}, h)\right) \equiv D_j\left(\boldsymbol{Z}, h, \delta_j^h, \boldsymbol{\theta}\right) = \mathbb{E}\left[\left.\frac{V(\boldsymbol{Z} + \delta_j^h \boldsymbol{e}_j, \boldsymbol{\theta}) - V(\boldsymbol{Z}, \boldsymbol{\theta})}{a_j(\boldsymbol{\theta}) h \sqrt{\nu_j}}\right| \boldsymbol{Z}\right],$$

we see that

$$\begin{aligned} D_j\left(\boldsymbol{Z}, h, \delta_j^h, \boldsymbol{\theta}\right) - D_j\left(\boldsymbol{Z}, \overline{h}, \delta_j^{\overline{h}}, \boldsymbol{\theta}\right) &= \underbrace{D_j\left(\boldsymbol{Z}, h, \delta_j^h, \boldsymbol{\theta}\right) - D_j\left(\boldsymbol{Z}, \overline{h}, \delta_j^h, \boldsymbol{\theta}\right)}_{(a)} \\ &\quad + \underbrace{D_j\left(\boldsymbol{Z}, \overline{h}, \delta_j^h, \boldsymbol{\theta}\right) - D_j\left(\boldsymbol{Z}, \overline{h}, \delta_j^{\overline{h}}, \boldsymbol{\theta}\right)}_{(b)}. \end{aligned}$$

To bound (a) and (b), we see from the proof of Lemma 3.8 that,

$$\left|\frac{V(\boldsymbol{Z}, \boldsymbol{\theta}) - V(\boldsymbol{Z} + Y\boldsymbol{e}_j, \boldsymbol{\theta})}{a_j(\boldsymbol{\theta})}\right| \le |Y|. \tag{B.4}$$

We first bound (a). We see

$$\begin{aligned} \left|D_j\left(\boldsymbol{Z}, h, \delta_j^h, \boldsymbol{\theta}\right) - D_j\left(\boldsymbol{Z}, \overline{h}, \delta_j^h, \boldsymbol{\theta}\right)\right| &= \left|\frac{\overline{h} - h}{h\overline{h}}\right| \left|\mathbb{E}\left[\left.\frac{V(\boldsymbol{Z} + \delta_j^h \boldsymbol{e}_j, \boldsymbol{\theta}) - V(\boldsymbol{Z}, \boldsymbol{\theta})}{\sqrt{\nu_j}}\right| \boldsymbol{Z}\right]\right| \\ &\le \frac{\left|\overline{h} - h\right|}{h_{\min}^2} \left|\mathbb{E}\left[\left.\frac{|\delta_j^h|}{\sqrt{\nu_j}}\right| \boldsymbol{Z}\right]\right|, \text{ by Eq. (B.4)} \\ &\le \frac{\left|\overline{h} - h\right|}{h_{\min}^2} \frac{1}{\sqrt{\nu_{\min}}} \sqrt{\frac{3h}{\sqrt{\nu_{\min}}}} \le \frac{\sqrt{3}\left|\overline{h} - h\right|}{h_{\min}^2 \nu_{\min}^{3/4}}, \end{aligned}$$

where the second to last inequality applies the inequality $\mathbb{E}\left[|\delta_j^h|\right] = \mathbb{E}\left[\sqrt{|\delta_j^h|^2}\right] \le \sqrt{\mathbb{E}\left[|\delta_j^h|^2\right]} \le \sqrt{\frac{3h}{\sqrt{\nu_{\min}}}}$. We then bound (b). We see

$$\begin{aligned} \left|D_j\left(\boldsymbol{Z}, \overline{h}, \delta_j^h, \boldsymbol{\theta}\right) - D_j\left(\boldsymbol{Z}, \overline{h}, \delta_j^{\overline{h}}, \boldsymbol{\theta}\right)\right| &= \left|\mathbb{E}\left[\left.\frac{V(\boldsymbol{Z} + \delta_j^h \boldsymbol{e}_j, \boldsymbol{\theta}) - V(\boldsymbol{Z} + \delta_j^{\overline{h}} \boldsymbol{e}_j, \boldsymbol{\theta})}{a_j(\boldsymbol{\theta}) \overline{h} \sqrt{\nu_j}}\right| \boldsymbol{Z}\right]\right| \\ &= \frac{1}{\overline{h}\sqrt{\nu_j}} \left|\mathbb{E}\left[\left. f\left(\delta_j^h\right) - f\left(\delta_j^{\overline{h}}\right)\right| \boldsymbol{Z}\right]\right| \\ &\le \frac{1}{\overline{h}\sqrt{\nu_j}} W_2\left(\delta_j^h, \delta_j^{\overline{h}}\right), \text{ by Wainwright (2019, pg. 76)} \end{aligned}$$

The Wasserstein distance between two mean-zero Gaussians is known in closed form:

$$W_2\left(\delta_j^h, \delta_j^{\overline{h}}\right) = \left|\sqrt{h^2 + \frac{2h}{\sqrt{\nu_j}}} - \sqrt{\overline{h}^2 + \frac{2\overline{h}}{\sqrt{\nu_j}}}\right| \le \sqrt{\left|h^2 - \overline{h}^2 + \frac{2\left(h-\overline{h}\right)}{\sqrt{\nu_j}}\right|} \le \sqrt{\left(2 + \frac{2}{\sqrt{\nu_j}}\right)\left|h - \overline{h}\right|},$$

where the first inequality comes from the common inequality $\left|\sqrt{a} - \sqrt{b}\right| \le \sqrt{|a-b|}$. Thus,

$$\left|D_j\left(\boldsymbol{Z}, \overline{h}, \delta_j^h, \boldsymbol{\theta}\right) - D_j\left(\boldsymbol{Z}, \overline{h}, \delta_j^{\overline{h}}, \boldsymbol{\theta}\right)\right| \le \frac{1}{h_{\min}\sqrt{\nu_{\min}}}\sqrt{\left(2 + \frac{2}{\sqrt{\nu_{\min}}}\right)\left|h - \overline{h}\right|}.$$

Collecting constants of the bounds of (a) and (b), we obtain our result. □

We now show the that the components of VGC is bounded.

Proof of 3.8: We see

$$\begin{aligned}
&\frac{V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta}) - V(\boldsymbol{Z},\boldsymbol{\theta})}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}} \\
&\quad = \frac{1}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}}\Big(\underbrace{\boldsymbol{r}(\boldsymbol{Z},\boldsymbol{\theta})^\top\left(\boldsymbol{x}(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta}) - \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})\right)}_{\le 0 \text{ by optimality of } \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})} + a_j(\boldsymbol{\theta})\delta_j x_j(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta})\Big) \\
&\quad \le \frac{\delta_j x_j(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta})}{h\sqrt{\nu_j}} \le \frac{|\delta_j|}{h\sqrt{\nu_{\min}}}.
\end{aligned}$$

By an analogous argument,

$$\frac{V(\boldsymbol{Z},\boldsymbol{\theta}) - V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta})}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}} \le \frac{|\delta_j|}{h\sqrt{\nu_{\min}}}.$$

Taking the conditional expectation, we see

$$|D_j(\boldsymbol{z})| \le \mathbb{E}\left[\left|\frac{V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta}) - V(\boldsymbol{Z},\boldsymbol{\theta})}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}}\right|\,\middle|\,\boldsymbol{Z}\right] \le \mathbb{E}\left[\frac{|\delta_j|}{h\sqrt{\nu_{\min}}}\right] \le \frac{\sqrt{3}}{\nu_{\min}^{3/4}\sqrt{h}}.$$

where the first inequality holds by Jensen's inequality and the last inequality holds by Jensen's inequality as $\mathbb{E}\left[|\delta_j|\right] \le \mathbb{E}\left[\sqrt{\delta_j^2}\right] \le \sqrt{\mathbb{E}\left[\delta_j^2\right]} = \sqrt{h^2 + 2h/\sqrt{\nu_j}} \le \sqrt{\frac{3h}{\sqrt{\nu_{\min}}}}$. □

B.3. Bias Under Violations of Assumption 2.4

In cases where the precisions ν_j are not known but estimated by a quantity $\tilde{\nu}_j$, we can construct the VGC in the same fashion, but replacing instances of ν_j with $\tilde{\nu}_j$, giving us,

$$\sum_{j:a_j\neq 0}\frac{1}{h\sqrt{\tilde{\nu}_j}a_j(\boldsymbol{\theta})}\mathbb{E}\left[\left(V(\boldsymbol{\mu}+\boldsymbol{\xi}+\tilde{\delta}_j\boldsymbol{e}_j) - V(\boldsymbol{\mu}+\boldsymbol{\xi})\right)\middle|\boldsymbol{Z}\right]$$

where $\tilde{\delta}_j \sim \mathcal{N}(0, h^2 + 2h/\sqrt{\tilde{\nu}_j})$. The bias of this VGC is similar to Theorem 3.2, except that it picks up an additional bias term due to the approximation error incurred from $\tilde{\nu}_j$, which we quantify in the following lemma.

Lemma B.3 (Bias of VGC with Estimated Precisions) *Suppose Assumption 3.1 holds. Let $\tilde{\nu}_j$ be an estimate of ν_j and let $\tilde{\delta}_j \sim \mathcal{N}(0, h^2 + 2h/\sqrt{\tilde{\nu}_j})$ and assume $\nu_{\min} \le \min_j \tilde{\nu}_j$. For any $0 < h < 1/e$, there exists a constant C dependent on $\nu_{\min}$ such that*

$$\left| \mathbb{E}\left[\sum_{j=1}^n \xi_j x_j(\boldsymbol{Z}) - \sum_{j:a_j \neq 0}^n \mathbb{E}\left[\frac{V(\boldsymbol{Z} + \tilde{\delta}_j \boldsymbol{e}_j) - V(\boldsymbol{Z})}{a_j(\boldsymbol{\theta}) h \sqrt{\tilde{\nu}_j}} \middle| \boldsymbol{Z} \right] \right] \right| \le C \cdot nh \log\left(\frac{1}{h}\right) + \frac{C}{\sqrt{h}} \sum_{j:a_j \neq 0} \left(\left| \nu_j^{1/2} - \tilde{\nu}_j^{1/2} \right| + \sqrt{\left| \nu_j^{1/2} - \tilde{\nu}_j^{1/2} \right|} \right)$$

Proof of Lemma B.3 Move the inner conditional expectation outwards and consider a sample path with a fixed $\boldsymbol{Z}$. Let $D_j(t) \equiv \frac{V(\boldsymbol{Z}+t\boldsymbol{e}_j) - V(\boldsymbol{Z})}{a_j h \sqrt{\tilde{\nu}_j}}$ if $a_j \neq 0$ and 0 otherwise, so that $\mathbb{E}\left[D_j(\tilde{\delta}_j) \mid \boldsymbol{Z} \right]$ is the j^{th} component of the VGC with the estimated precisions and $\sqrt{\frac{\tilde{\nu}_j}{\nu_j}} \mathbb{E}\left[D_j(\delta_j) \mid \boldsymbol{Z} \right]$ is the j^{th} component of the VGC with the correct ν_j. Fix some j^{th} where $a_j \neq 0$. Note,

$$\xi_j x_j(\boldsymbol{Z}) - D_j(\tilde{\delta}_j) = \left(\xi_j x_j(\boldsymbol{Z}) - \frac{V(\boldsymbol{Z} + \delta_j \boldsymbol{e}_j) - V(\boldsymbol{Z})}{a_j h \sqrt{\nu_j}} \right) + \left(\sqrt{\frac{\tilde{\nu}_j}{\nu_j}} D_j(\delta_j) - D_j(\tilde{\delta}_j) \right)$$

The expectation of the first term was bounded in Theorem 3.2, so we focus on the expectation of the second. We see

$$\sqrt{\frac{\tilde{\nu}_j}{\nu_j}} D_j(\delta_j) - D_j(\tilde{\delta}_j) = \underbrace{\left(\sqrt{\frac{\tilde{\nu}_j}{\nu_j}} \right) \left(D_j(\delta_j) - D_j(\tilde{\delta}_j) \right)}_{(a)} + \underbrace{\left(1 - \sqrt{\frac{\tilde{\nu}_j}{\nu_j}} \right) D_j(\tilde{\delta}_j)}_{(b)}.$$

To bound the expectation of (a), we see first see $t \mapsto h\sqrt{\tilde{\nu}_j} D_j(t)$ is $1-$Lipschitz because

$$\frac{\partial}{\partial t} h \sqrt{\tilde{\nu}_j} D_j(t) = \frac{\partial}{\partial t} \frac{V(\boldsymbol{Z} + t\boldsymbol{e}_j) - V(\boldsymbol{Z})}{a_j} = \frac{1}{a_j} a_j x_j(\boldsymbol{Z}) = x_j(\boldsymbol{Z})$$

by Danskin's theorem and because $x_j(\boldsymbol{Z})$ is between 0 and 1. Thus, by (Wainwright 2019, pg. 76)

$$\left| \mathbb{E}\left[h \sqrt{\tilde{\nu}_j} \left(D_j(\delta_j) - D_j(\tilde{\delta}_j) \right) \middle| \boldsymbol{Z} \right] \right| \le W_2\left(\delta_j, \tilde{\delta}_j \right),$$

where $W_2\left(\delta_j, \tilde{\delta}_j \right)$ is the Wasserstein distance between two mean-zero Gaussians δ_j and $\tilde{\delta}_j$, which is known in closed form:

$$W_2\left(\delta_j, \tilde{\delta}_j \right) = \left| \sqrt{h^2 + \frac{2h}{\tilde{\nu}_j^{1/2}}} - \sqrt{h^2 + \frac{2h}{\nu_j^{1/2}}} \right| \le \sqrt{\left| \frac{2h}{\tilde{\nu}_j^{1/2}} - \frac{2h}{\nu_j^{1/2}} \right|} \le \sqrt{2h \left| \frac{\nu_j^{1/2} - \tilde{\nu}_j^{1/2}}{\nu_{\min}} \right|}$$

where $\nu_{\min} \le \min_j \{\min\{\nu_j, \tilde{\nu}_j\}\}$.

To bound the expectation of (b), we see

$$\mathbb{E}\left[\left| D_j(\tilde{\delta}_j) \right| \middle| \boldsymbol{Z} \right] \le \mathbb{E}\left[\frac{\left| a_j \tilde{\delta}_j \right|}{|a_j| \tilde{\nu}_j^{1/2} h} \middle| \boldsymbol{Z} \right] = \frac{1}{\tilde{\nu}_j^{1/2} h} \sqrt{\frac{2}{\pi} \left(h^2 + \frac{2h}{\tilde{\nu}_j^{1/2}} \right)} \le \sqrt{\frac{2}{\pi} \left(\frac{3}{\nu_{\min} h} \right)}$$

where the first equality holds by directly evaluating the expectation and the last inequality holds because $h < 1/e$ and $\nu_{\min} \le 1$.

Putting it all together, we see

$$\begin{aligned}
&\left| \mathbb{E}\left[\sum_{j=1}^n \xi_j x_j(\boldsymbol{Z}) - \sum_{j=1}^n \frac{V(\boldsymbol{Z} + \tilde{\delta}_j \boldsymbol{e}_j) - V(\boldsymbol{Z})}{a_j h \sqrt{\tilde{\nu}_j}} \right] \right| \\
&\le \left| \mathbb{E}\left[\sum_{j=1}^n \xi_j x_j(\boldsymbol{Z}) - \sqrt{\frac{\tilde{\nu}_j}{\nu_j}} D_j(\delta_j) \right] \right| + \left| \sum_{j=1}^n \mathbb{E}\left[\mathbb{E}\left[\sqrt{\frac{\tilde{\nu}_j}{\nu_j}} D_j(\delta_j) - D_j(\tilde{\delta}_j) \middle| \boldsymbol{Z} \right] \right] \right| \\
&\le \sum_{j=1}^n C \cdot h \log\left(\frac{1}{h}\right) + \sum_{j: a_j \neq 0} \frac{1}{h\sqrt{\nu_{\min}}} \sqrt{2h \left| \frac{\nu_j^{1/2} - \tilde{\nu}_j^{1/2}}{\nu_{\min}} \right|} + \left| \frac{\nu_j^{1/2} - \tilde{\nu}_j^{1/2}}{\nu_{\min}} \right| \sqrt{\frac{2}{\pi}\left(\frac{3}{h\nu_{\min}}\right)} \\
&\le \sum_{j=1}^n C \cdot h \log\left(\frac{1}{h}\right) + \sum_{j: a_j \neq 0} \frac{\sqrt{2\left|\nu_j^{1/2} - \tilde{\nu}_j^{1/2}\right|}}{\sqrt{h}\nu_{\min}^{3/2}} + \frac{\left|\nu_j^{1/2} - \tilde{\nu}_j^{1/2}\right| \sqrt{\frac{6}{\pi}}}{\sqrt{h}\nu_{\min}^{3/2}}
\end{aligned}$$

where the first inequality follows from triangle inequality and the second from applying Theorem 3.2 and our bounds on (a) and (b). Collecting constants we obtain our intended result. □

We now highlight when ξ_j are not Gaussian but only sub-Gaussian. Let

Lemma B.4 (Bias VGC with Gaussian assumption violated) *Suppose Assumption 3.1 holds. Let ξ_j be a mean-zero, sub-Gaussian random variable with variance proxy at most σ^2 and admits a density density $\phi(\cdot)$. Additionally, let $\bar{\xi}_j \sim \mathcal{N}(0, 1/\sqrt{\nu_j}$ with density $\bar{\phi}(\cdot)$. Then, there exists a dimension independent constant C, such that*

$$\begin{aligned}
&\left| \mathbb{E}\left[\sum_{j=1}^n \xi_j x_j(\boldsymbol{Z}) - \sum_{j: a_j \neq 0}^n \mathbb{E}\left[\frac{V(\boldsymbol{Z} + \delta_j \boldsymbol{e}_j) - V(\boldsymbol{Z})}{a_j(\boldsymbol{\theta}) h \sqrt{\nu_j}} \middle| \boldsymbol{Z} \right] \right] \right| \\
&\qquad \le n \left(\sigma\sqrt{2\pi} - \log\left(\frac{\left\|\phi - \bar{\phi}\right\|_1}{4} \right) \right) \left\|\phi - \bar{\phi}\right\|_1 + Cnh \log\left(\frac{1}{h}\right) + \sum_{j: a_j \neq 0}^n W_2(\xi_j, \bar{\xi}_j)
\end{aligned}$$

Proof of Lemma B.4 Let $\bar{\boldsymbol{Z}}$ be $\boldsymbol{Z}$, but with the j^{th} component replaced by $\bar{Z}_j = \mu_j + \bar{\xi}_j$ and let

$$D_j(t) = \frac{1}{a_j h \sqrt{\nu_j}} \mathbb{E}\left[V\left(\boldsymbol{Z} + (\delta_j + t - \xi_j)\, \boldsymbol{e}_j\right) - V\left(\boldsymbol{Z} + (t - \xi_j)\, \boldsymbol{e}_j\right) \middle| \boldsymbol{Z} \right].$$

We see

$$\begin{aligned}
\left| \mathbb{E}\left[\xi_j x_j(\boldsymbol{Z}) - D_j(\xi_j) \right] \right| \le& \left| \mathbb{E}\left[\mathbb{E}\left[\xi_j x_j(\boldsymbol{Z}) - \bar{\xi}_j x_j(\bar{\boldsymbol{Z}}) \middle| \boldsymbol{Z}^{-j} \right] \right] \right| + \left| \mathbb{E}\left[\bar{\xi}_j x_j(\bar{\boldsymbol{Z}}) - D_j(\bar{\xi}_j) \right] \right| \\
&+ \left| \mathbb{E}\left[\mathbb{E}\left[D_j(\bar{\xi}_j) - D_j(\xi_j) \middle| \boldsymbol{Z}^{-j} \right] \right] \right|
\end{aligned}$$

By Lemma C.2 GR 2021, we see

$$\left| \mathbb{E}\left[\xi_j x_j(\boldsymbol{Z}) - \bar{\xi}_j x_j(\bar{\boldsymbol{Z}}) \middle| \boldsymbol{Z}^{-j} \right] \right| \le T \left\|\phi - \bar{\phi}\right\|_1 + 4\exp\left(-\frac{T^2}{2\sigma^2}\right) \left(T + \sigma\sqrt{2\pi}\right).$$

To optimize T, we first upperbound the latter term as follows

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

where the last inequality used the fact that the quadratic $-\frac{T^2}{2\sigma^2}+2T$ is maximized at $T^*=4\sigma^2$. Substituting the upperbound, we see

$$T\left\|\phi-\bar{\phi}\right\|_1+4\exp\left(-\frac{T^2}{2\sigma^2}\right)\left(T+\sigma\sqrt{2\pi}\right)\le T\left\|\phi-\bar{\phi}\right\|_1+4\exp\left(\sigma\sqrt{2\pi}-1-T\right)$$

We see the right hand side is minimized at $T^*=\sigma\sqrt{2\pi}-1-\log\left(\frac{\|\phi-\bar{\phi}\|_1}{4}\right)$. Thus, we see

$$\left|\mathbb{E}\left[\xi_j x_j(\boldsymbol{Z})-\bar{\xi}_j x_j(\bar{\boldsymbol{Z}})\,\middle|\,\boldsymbol{Z}^{-j}\right]\right|\le\left(\sigma\sqrt{2\pi}-\log\left(\frac{\left\|\phi-\bar{\phi}\right\|_1}{4}\right)\right)\left\|\phi-\bar{\phi}\right\|_1$$

By Theorem 3.2, we see

$$\left|\mathbb{E}\left[\bar{\xi}_j x_j(\bar{\boldsymbol{Z}})-D_j(\bar{\boldsymbol{Z}})\,\middle|\,\boldsymbol{Z}^{-j}\right]\right|\le Ch\log\left(\frac{1}{h}\right).$$

Finally, since $t\mapsto h\sqrt{\nu_j}D_j(t)$ is $1-$Lipschitz from Lemma B.3, we see that

$$\left|\mathbb{E}\left[D_j(\bar{\boldsymbol{Z}})-D_j(\boldsymbol{Z})\,\middle|\,\boldsymbol{Z}^{-j}\right]\right|\le W_2(\xi_j,\bar{\xi}_j).$$

Putting it all together, we see

$$\left|\mathbb{E}\left[\xi_j x_j(\boldsymbol{Z})-D_j(\xi_j)\right]\right|\le\left(\sigma\sqrt{2\pi}-\log\left(\frac{\left\|\phi-\bar{\phi}\right\|_1}{4}\right)\right)\left\|\phi-\bar{\phi}\right\|_1+Ch\log\left(\frac{1}{h}\right)+W_2(\xi_j,\bar{\xi}_j).$$

Summing over the j terms, we obtain our result □

B.4. Proof of Theorem 3.5.

Before proving the theorem, we require the following lemma.

Lemma B.5 (A χ^2-Tail Bound) *Consider $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_n)^\top$ where δ_j is defined as in the definition of the VGC (Eq. (3.3)) and $0<h<1/e$. Suppose Assumption 3.1 holds. Then,*

$$\mathbb{E}\left[\|\boldsymbol{\delta}\|_2^2\mathbb{I}\left\{\|\boldsymbol{\delta}\|_2^2>\frac{18hn}{\sqrt{\nu_{\min}}}\right\}\right]\;\le\;\frac{36hn}{\sqrt{\nu_{\min}}}e^{-n}.$$

Proof of Lemma B.5. Let $Y_1, \ldots, Y_n$ be independent standard normals.

Observe that since $h < 1/e$, the variance of δ_j is at most $h^2 + \frac{h}{\sqrt{\nu_j}} \leq 2h/\sqrt{\nu_{\min}}$. Then, for $t > 1$

$$
\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{\delta}\|_2^2 > \frac{2hn}{\sqrt{\nu_{\min}}}(1+t)\right) &\leq \mathbb{P}\left(\frac{2h}{\sqrt{\nu_{\min}}}\sum_{j=1}^n Y_j^2 > \frac{2hn}{\sqrt{\nu_{\min}}}(1+t)\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n Y_j^2 > 1+t\right) \\
&\leq e^{-nt/8}, \qquad \text{(B.5)}
\end{aligned}
$$

where the last inequality follows from (Wainwright 2019, pg. 29).

Next, by the tail formula for expectation,

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\delta}\|_2^2 \mathbb{I}\left\{\|\boldsymbol{\delta}\|_2^2 > \frac{18hn}{\sqrt{\nu_{\min}}}\right\}\right] &= \int_0^\infty \mathbb{P}\left(\|\boldsymbol{\delta}\|_2^2 \mathbb{I}\left\{\|\boldsymbol{\delta}\|_2^2 > \frac{18hn}{\sqrt{\nu_{\min}}}\right\} > t\right) dt \\
&= \int_0^{\frac{18hn}{\sqrt{\nu_{\min}}}} \mathbb{P}\left(\|\boldsymbol{\delta}\|_2^2 > \frac{18hn}{\sqrt{\nu_{\min}}}\right) dt + \int_{\frac{18hn}{\sqrt{\nu_{\min}}}}^\infty \mathbb{P}\left(\|\boldsymbol{\delta}\|_2^2 > t\right) dt \\
&\leq \frac{18hn}{\sqrt{\nu_{\min}}} e^{-n} + \int_{\frac{18hn}{\sqrt{\nu_{\min}}}}^\infty \mathbb{P}\left(\|\boldsymbol{\delta}\|_2^2 > t\right) dt \qquad \text{(Applying Eq. (B.5))} \\
&\leq \frac{18hn}{\sqrt{\nu_{\min}}} e^{-n} + \frac{2hn}{\sqrt{\nu_{\min}}} \int_8^\infty \mathbb{P}\left(\|\boldsymbol{\delta}\|_2^2 > \frac{2hn}{\sqrt{\nu_{\min}}}(1+s)\right) ds \\
&\leq \frac{9hn}{\sqrt{\nu_{\min}}} e^{-n} + \frac{2hn}{\sqrt{\nu_{\min}}} \int_8^\infty e^{-ns/8} ds \qquad \text{(Applying Eq. (B.5))} \\
&= \frac{18hn}{\sqrt{\nu_{\min}}} e^{-n} + \frac{16h}{\sqrt{\nu_{\min}}} e^{-n}
\end{aligned}
$$

Rounding up and combining proves the theorem. $\square$

We can now prove the theorem.

Proof of Theorem 3.5. Proceeding as in the main body, we bound each of the three terms of the out-of-sample estimator error (Eq. (3.10)). Before beginning, note that under Assumption 3.1, $\mathbb{V}\text{ar}(\delta_j) \leq \frac{3h}{\sqrt{\nu_{\min}}}$. We use this upper bound frequently.

We start with Eq. (3.10b). Consider the k^{th} non-zero element of the sum. By definition of D^{R},

$$
\begin{aligned}
\left|D^{\mathsf{R}}(\boldsymbol{Z}, \boldsymbol{\delta}, \tilde{\boldsymbol{U}}) - D^{\mathsf{R}}(\boldsymbol{Z}, \boldsymbol{\delta}^k, \tilde{\boldsymbol{U}})\right| &= \left|\frac{\delta_k}{h\sqrt{\nu_k}a_k} x_k(\boldsymbol{Z} + \delta_k \tilde{U}_k \boldsymbol{e}_k) - \frac{\overline{\delta}_k}{h\sqrt{\nu_k}a_k} x_k(\boldsymbol{Z} + \overline{\delta}_k \tilde{U}_k \boldsymbol{e}_k)\right| \\
&\leq \frac{1}{h\sqrt{\nu_k}a_k}\left(|\delta_k| + \left|\overline{\delta}_k\right|\right)
\end{aligned}
$$

Hence, squaring and taking expectations,

$$
\begin{aligned}
\mathbb{E}\left[\left(D^{\mathsf{R}}(\boldsymbol{Z}, \boldsymbol{\delta}, \tilde{\boldsymbol{U}}) - D^{\mathsf{R}}(\boldsymbol{Z}, \boldsymbol{\delta}^k, \tilde{\boldsymbol{U}})\right)^2\right] &\leq \frac{2}{h^2 \nu_k a_k^2}\left(\mathbb{E}\left[\delta_k^2\right] + \mathbb{E}\left[\overline{\delta}_k^2\right]\right) \\
&\leq \frac{12}{h \nu_{\min}^{3/2} a_{\min}}.
\end{aligned}
$$

Summing over k shows

$$Eq.\ (3.10b)\ \leq\ \frac{6n}{h\nu_{\min}^{3/2}a_{\min}}.$$

We now bound Eq. (3.10c). Again, consider the k^{th} non-zero element. By definition,

$$\begin{aligned}\left|D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta},\tilde{\boldsymbol{U}}) - D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta},\tilde{\boldsymbol{U}}^k)\right| &= \frac{|\delta_k|}{h\sqrt{\nu_k}a_k}\left|x_k(\boldsymbol{Z}+\delta_k U_k\boldsymbol{e}_k) - x_k(\boldsymbol{Z}+\delta_k\overline{U}_k\boldsymbol{e}_k)\right| \\ &\leq \frac{2\,|\delta_k|}{h\sqrt{\nu_{\min}}a_{\min}}\end{aligned}$$

Hence,

$$\mathbb{E}\left[\left(D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta},\tilde{\boldsymbol{U}}) - D^{\mathsf{R}}(\boldsymbol{Z},\boldsymbol{\delta},\tilde{\boldsymbol{U}}^k)\right)^2\right] \ \leq\ \frac{4}{h^2\nu_{\min}a_{\min}^2}\mathbb{E}\left[\delta_k^2\right]\ \leq\ \frac{12}{h\nu_{\min}^{3/2}a_{\min}^2}$$

Summing over k shows

$$Eq.\ (3.10c)\ \leq\ \frac{6n}{h\nu_{\min}^{3/2}a_{\min}^2}.$$

Finally, we bound Eq. (3.10a). For convenience, let $\boldsymbol{W}_k\in\mathbb{R}^n$ be the random vector with components $W_{kj} = x_j(\boldsymbol{Z}+\delta_j\tilde{U}\boldsymbol{e}_j) - x_j(\boldsymbol{Z}^k+\delta_j\tilde{U}\boldsymbol{e}_j)$. Then, proceeding as in the main text, we have

$$\begin{aligned}\left(D^{\mathsf{R}}(\boldsymbol{Z}) - D^{\mathsf{R}}(\boldsymbol{Z}^k)\right)^2 &\leq \frac{\|\boldsymbol{\delta}\|_2^2}{h^2a_{\min}^2\nu_{\min}}\cdot\sum_{j=1}^n\left(x_j(\boldsymbol{Z}+\delta_j\tilde{U}\boldsymbol{e}_j) - x_j(\boldsymbol{Z}^k+\delta_j\tilde{U}\boldsymbol{e}_j)\right)^2 \\ &\leq \frac{\|\boldsymbol{\delta}\|_2^2}{h^2a_{\min}^2\nu_{\min}}\cdot\|\boldsymbol{W}_k\|_2^2\end{aligned}$$

Notice that $\mathbb{E}\left[\|\boldsymbol{\delta}\|_2^2\right] = O\left(nh/\nu_{\min}\right)$. We upper bound this expression by splitting on cases where $\|\boldsymbol{\delta}\|_2^2 > \frac{18hn}{\sqrt{\nu_{\min}}}$ or not. Note this quantity is much larger than the mean, so we expect contributions when $\|\boldsymbol{\delta}\|_2^2$ is large to be small.

Splitting the expression yields

$$\begin{aligned}\left(D^{\mathsf{R}}(\boldsymbol{Z}) - D^{\mathsf{R}}(\boldsymbol{Z}^k)\right)^2 &\leq \frac{\|\boldsymbol{\delta}\|_2^2}{h^2a_{\min}^2\nu_{\min}}\cdot\|\boldsymbol{W}_k\|_2^2\mathbb{I}\left\{\|\boldsymbol{\delta}\|_2^2 > \frac{18hn}{\sqrt{\nu_{\min}}}\right\} + \frac{\|\boldsymbol{\delta}\|_2^2}{h^2a_{\min}^2\nu_{\min}}\cdot\|\boldsymbol{W}_k\|_2^2\mathbb{I}\left\{\|\boldsymbol{\delta}\|_2^2 \leq \frac{18hn}{\sqrt{\nu_{\min}}}\right\} \\ &\leq \frac{n}{h^2a_{\min}^2\nu_{\min}}\|\boldsymbol{\delta}\|_2^2\mathbb{I}\left\{\|\boldsymbol{\delta}\|_2^2 > \frac{18hn}{\sqrt{\nu_{\min}}}\right\} + \frac{18n}{ha_{\min}^2\nu_{\min}^{3/2}}\cdot\|\boldsymbol{W}_k\|_2^2\end{aligned}$$

Next take an expectation and apply Lemma B.5 to obtain

$$\mathbb{E}\left[\left(D^{\mathsf{R}}(\boldsymbol{Z}) - D^{\mathsf{R}}(\boldsymbol{Z}^k)\right)^2\right]\ \leq\ \frac{36n^2}{ha_{\min}^2\nu_{\min}^{3/2}}e^{-n} + \frac{18n}{ha_{\min}^2\nu_{\min}^{3/2}}\cdot\mathbb{E}\left[\|\boldsymbol{W}_k\|_2^2\right]$$

Finally summing over k shows

$$Eq.\ (3.10a)\ \leq\ \frac{36n^3}{ha_{\min}^2\nu_{\min}^{3/2}}e^{-n} + \frac{18n^3}{ha_{\min}^2\nu_{\min}^{3/2}}\cdot\frac{1}{n^2}\sum_{k=1}^n\mathbb{E}\left[\|\boldsymbol{W}_k\|_2^2\right]$$

Finally, we combine all three terms in Eq. (3.10) yielding

$$\begin{aligned}\mathbb{V}\text{ar}(D(\boldsymbol{Z})) &\leq \frac{6n}{h\nu_{\min}^{3/2}a_{\min}} + \frac{6n}{h\nu_{\min}^{3/2}a_{\min}^2} + \frac{36n^3}{2ha_{\min}^2\nu_{\min}^{3/2}}e^{-n} + \frac{18n^3}{2ha_{\min}^2\nu_{\min}^{3/2}}\cdot\frac{1}{n^2}\sum_{k=1}^n\mathbb{E}\left[\|\boldsymbol{W}_k\|_2^2\right] \\ &\leq \frac{C_2}{h}\max(n^{3-\alpha}, n),\end{aligned}$$

by collecting the dominant terms. □

B.5. Implementation Details

As mentioned, in our experiments we utilize a second-order forward finite difference approximation. Namely, instead of approximating the derivative using a first-order approximation as in Eq. (3.6), we approximate

$$\frac{\partial}{\partial \lambda} V(\boldsymbol{Z} + \lambda \xi_j \boldsymbol{e}_j)\bigg|_{\lambda=0} \;=\; \frac{1}{2h\sqrt{\nu_j}a_j}\big(4V(\boldsymbol{Z} + h\sqrt{\nu_j}\xi_j \boldsymbol{e}_j) - V(\boldsymbol{Z} + 2h\sqrt{\nu_j}\xi_j \boldsymbol{e}_j) - 3V(\boldsymbol{Z})\big) + O(h^2).$$

The coefficients in this expansion can be derived directly from a Taylor Series. We then use randomization to replace the unknown $h\xi_j$ and $2h\xi_j$ term as before. The j^{th} element of our estimator becomes

$$D_j(\boldsymbol{Z}) \equiv \mathbb{E}\left[\frac{1}{2h\sqrt{\nu_j}a_j}\Big(4V(\boldsymbol{Z} + \delta_j^h \boldsymbol{e}_j) - V(\boldsymbol{Z} + \delta_j^{2h}\boldsymbol{e}_j) - 3V(\boldsymbol{Z})\Big) \bigg| \boldsymbol{Z}\right].$$

where $\delta_j^h \sim \mathcal{N}\left(0, h^2 + \frac{2h}{\sqrt{\nu_j}}\right)$ and $\delta_j^{2h} \sim \mathcal{N}\left(0, 4h^2 + \frac{4h}{\sqrt{\nu_j}}\right)$.

As mentioned, one can always compute the above conditional expectation by Monte Carlo simulation. In special cases, a more computationally efficient method is to utilize a parametric programming algorithm to determine the values of $\boldsymbol{x}(\boldsymbol{Z} + t\boldsymbol{e}_j)$ as t ranges over $\mathbb{R}$. Importantly, for many classes of optimization problems, including, e.g., linear optimization and mixed-binary linear optimization, $\boldsymbol{x}(\boldsymbol{Z} + t\boldsymbol{e}_j)$ is piecewise constant on the intervals (c_i, c_{i+1}), taking value $\boldsymbol{x}^i$, for $i = 1, \ldots, I$, with $c_0 = -\infty$ and $c_I = \infty$. In this case,

$$\mathbb{E}\left[V(\boldsymbol{Z} + \delta_j \boldsymbol{e}_j) \middle| \boldsymbol{Z}\right] = \sum_{i=0}^{I-1} r(\boldsymbol{Z})^\top \boldsymbol{x}^i \cdot \int_{c_i}^{c_{i+1}} \phi_{\delta_j}(t)\, dt + \sqrt{\frac{\sigma^2}{2\pi}} \exp\left(\frac{(c_j - Z_j)^2}{2\sigma^2}\right) \tag{B.6}$$

where $\phi_{\delta_j}(\cdot)$ is the pdf of δ_j. These integrals can then be evaluated in closed-form in terms of the standard normal CDF. A similar argument holds for $\mathbb{E}\left[V(\boldsymbol{Z} + \delta_j^{2h}\boldsymbol{e}_j) \middle| \boldsymbol{Z}\right]$. We follow this strategy in our case study in Section 5.

Appendix C: Problems that are Weakly Coupled by Variables

C.1. Convergence of In-Sample Optimism

In this section we provide a high-probability bound on

$$\sup_{\boldsymbol{x}^0 \in \mathcal{X}^0} \sup_{\boldsymbol{\theta} \in \Theta} \left| \boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0) - \mathbb{E}\left[\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0)\right]\right|.$$

As a first step, we bound the cardinality of

$$\mathcal{X}^{\Theta, \mathcal{X}^0}(\boldsymbol{Z}) \equiv \left\{ \left(\boldsymbol{x}^0, \boldsymbol{x}^1(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0), \ldots, \boldsymbol{x}^K(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0)\right) \,:\, \boldsymbol{\theta} \in \Theta,\ \boldsymbol{x}^0 \in \mathcal{X}^0 \right\} \subseteq \mathbb{R}^n.$$

Lemma C.1 (Cardinality of Lifted, Decoupled Policy Class) *Under the assumptions of Theorem 4.3, there exists an absolute constant C such that*

$$\log \left| \mathcal{X}^{\Theta, \mathcal{X}^0}(\boldsymbol{Z}) \right| \le \dim(\phi) \cdot \log\left(CK \left|\mathcal{X}^0\right| \mathcal{X}_{\max}^2\right)$$

Proof: We adapt a hyperplane arrangement argument from Gupta and Kallus (2021). We summarize the pertinent details briefly. For any $\boldsymbol{x}^0 \in \mathcal{X}^0$, $\boldsymbol{x}^k(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0)$ is fully determined by the relative ordering of the values

$$\left\{\phi(\boldsymbol{\theta})^\top g_k(\boldsymbol{Z}^k, \boldsymbol{x}^k, \boldsymbol{x}^0) \ : \ \boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}^0)\right\}.$$

This observation motivates us to consider the hyperplanes in $\mathbb{R}^{\dim(\phi)}$

$$H_{k,\boldsymbol{Z},\boldsymbol{x}^0}(\boldsymbol{x}^k, \bar{\boldsymbol{x}}^k) = \left\{\phi(\boldsymbol{\theta}) \ : \ \phi(\boldsymbol{\theta})^\top \left(g_k(\boldsymbol{Z}^k, \boldsymbol{x}^k, \boldsymbol{x}^0) - g_k(\boldsymbol{Z}^k, \bar{\boldsymbol{x}}^k, \boldsymbol{x}^0)\right) = 0\right\} \tag{C.1}$$

for all $\boldsymbol{x}^k, \bar{\boldsymbol{x}}^k \in \mathrm{Ext}\left(\mathcal{X}^k(\boldsymbol{x}^0)\right)$, $k = 1, \ldots, K$, and $\boldsymbol{x}^0 \in \mathcal{X}^0$.

On one side of $H_{k,\boldsymbol{Z},\boldsymbol{x}^0}(\boldsymbol{x}^k, \bar{\boldsymbol{x}}^k)$, $\boldsymbol{x}^k$ is preferred to $\bar{\boldsymbol{x}}^k$ in the policy problem Eq. (4.3), on the other side $\bar{\boldsymbol{x}}^k$ is preferred, and on the hyperplane we are indifferent. Thus, if we consider drawing all such hyperplanes in $\mathbb{R}^{\dim(\phi)}$, then the vector $\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0)$ is constant for all $\phi(\boldsymbol{\theta})$ in the relative interior of each induced region. Hence, $\left|\mathcal{X}^{\Theta, \mathcal{X}^0}(\boldsymbol{Z})\right|$ is at most the number of such regions. Gupta and Kallus (2021) prove that the number of such regions is at most $(1 + 2m)^{\dim(\phi)}$ where m is number of hyperplanes in Eq. (C.1). By assumption, $m \le K \mathcal{X}_{\max}^2 \left|\mathcal{X}^0\right|$, and, hence, $\left|\mathcal{X}^\Theta(\boldsymbol{Z}, \boldsymbol{x}^0)\right| \le \left(1 + 2K \left|\mathcal{X}^0\right| \mathcal{X}_{max}^2\right)^{\dim(\phi)}$. Collecting constants yields the bound. $\square$

We can now prove our high-probability tail bound.

Lemma C.2 (Bounding In-sample Optimism) *Under the assumptions of Theorem 4.3, there exists a constant C (depending on $\nu_{\min}$) such that, for any $R > 1$, with probability $1 - e^{-R}$*

$$\sup_{\boldsymbol{\theta} \in \Theta, \boldsymbol{x}^0 \in \mathcal{X}^0} \left|\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0) - \mathbb{E}\left[\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0)\right]\right| \ \le \ C S_{\max} R \sqrt{K \dim(\phi) \log(K \left|\mathcal{X}^0\right| \mathcal{X}_{\max})}$$

Proof: By triangle inequality,

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in \Theta, \boldsymbol{x}^0 \in \mathcal{X}^0} \left|\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0) - \mathbb{E}\left[\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0)\right]\right| \\
&\le \sup_{\boldsymbol{x}^0 \in \mathcal{X}^0} \left|(\boldsymbol{\xi}^0)^\top \boldsymbol{x}^0 - \mathbb{E}\left[(\boldsymbol{\xi}^0)^\top \boldsymbol{x}^0\right]\right| + \sup_{\boldsymbol{x}^0 \in \mathcal{X}^0, \boldsymbol{\theta} \in \Theta} \left|\sum_{k=1}^K \sum_{j \in S_k} \xi_j x_j(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0) - \mathbb{E}\left[\xi_j x_j(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{x}^0)\right]\right|
\end{aligned}$$

Consider the first term. For a fixed $\boldsymbol{x}_0$, this is a sum of sub-Gaussian random variables each with parameter at most $1/\sqrt{\nu_{\min}}$. We apply Hoeffding's inequality to obtain a pointwise bound for a fixed $\boldsymbol{x}_0$ and then take a a union bound over $\mathcal{X}^0$. This yields for some absolute constant c,

$$\mathbb{P}\left\{\sup_{\boldsymbol{x}^0 \in \mathcal{X}^0} \left|\sum_{j \in S_0} \xi_j x_j^0 - \mathbb{E}\left[\xi_j x_j^0\right]\right| > t\right\} \le 2|\mathcal{X}^0| \exp\left(-\frac{c\nu_{\min} t^2}{|S_0|}\right),$$

Rearranging shows that, with probability at least $1 - \exp\{-R\}$,

$$\sup_{\boldsymbol{x}^0 \in \mathcal{X}^0} \left|\sum_{j \in S_0} \xi_j x_j^0 - \mathbb{E}\left[\xi_j x_j^0\right]\right| \le \sqrt{\frac{|S_0| R}{c\nu_{\min}} \log\left(2|\mathcal{X}^0|\right)}$$

For the second component, we apply Theorem A.1. We bounded the cardinality $\mathcal{X}^{\Theta,\mathcal{X}^0}(\boldsymbol{Z})$ in Lemma C.1. Consider the vector

$$\left(\sum_{j\in S_1} \xi_j x_j(\boldsymbol{Z};\boldsymbol{\theta},\boldsymbol{x}^0), \ldots, \sum_{j\in S_K} \xi_j x_j(\boldsymbol{Z};\boldsymbol{\theta},\boldsymbol{x}^0) \right).$$

This vector has independent components. We next construct an envelope $\boldsymbol{F}(\boldsymbol{Z})$ for it and bound the Ψ-norm of $\boldsymbol{F}(\boldsymbol{Z})$.

Since $0 \le x_j(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0) \le 1$, we take

$$\boldsymbol{F}(\boldsymbol{Z}) = \left(\sum_{j\in S_1} |\xi_j|, \ldots, \sum_{j\in S_K} |\xi_j| \right).$$

Note that

$$\left(\sum_{j\in S_k} |\xi_j| \right)^2 \overset{(a)}{\le} |S_k| \sum_{j\in S_k} \left(\frac{\zeta_j}{\sqrt{\nu_j}} \right)^2 \le \frac{S_{\max}}{\nu_{\min}} \sum_{j\in S_k} \zeta_j^2$$

where $\zeta_j \sim \mathcal{N}(0,1)$. Inequality (a) uses $\|\boldsymbol{t}\|_1 \le \sqrt{d}\,\|\boldsymbol{t}\|_2$ for $\boldsymbol{t} \in \mathbb{R}^d$. Plugging into $\|\|\boldsymbol{F}(\boldsymbol{\xi})\|_2\|_\Psi$ we have

$$\|\|\boldsymbol{F}(\boldsymbol{Z})\|_2\|_\Psi \;\le\; \sqrt{\frac{|S_{\max}|}{\nu_{\min}}} \cdot \left\| \sqrt{\sum_{k=1}^{K} \sum_{j\in S_k} \zeta_j^2} \right\|_\Psi \;\overset{(b)}{\le}\; \sqrt{\frac{|S_{\max}|}{\nu_{\min}} \cdot 2|S_{\max}|K}.$$

Inequality (b) follows from Lemma A.1 iv) of GR 2021.

Applying Theorem A.1, combining the bounds of our two components, and collecting constants, we obtain our result. $\square$

C.2. Convergence of VGC

Define $V(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ and $D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ analogously to $V(\boldsymbol{Z},\boldsymbol{\theta})$ and $D(\boldsymbol{Z},\boldsymbol{\theta})$ with $\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})$ replaced by $\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ throughout. The next step in our proof provides a high-probability bound on

$$\sup_{\boldsymbol{\theta}\in\Theta,\boldsymbol{x}^0\in\mathcal{X}^0} \left| D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0) - \mathbb{E}\left[D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0) \right] \right|.$$

We first establish a pointwise bound for a fixed $\boldsymbol{\theta},\boldsymbol{x}^0$.

Lemma C.3 (Pointwise Convergence of VGC for a fixed $\boldsymbol{x}^0$) *Under the assumptions of Theorem 4.3, for fixed $\boldsymbol{\theta},\boldsymbol{x}^0$, there exists a constant C (that depends on $\nu_{\min}$) such that, for any $R > 1$, we have with probability $1 - 2\exp(-R)$,*

$$|D(\boldsymbol{Z},\boldsymbol{\theta}) - \mathbb{E}[D(\boldsymbol{Z},\boldsymbol{\theta})]| \le C\,|S_{\max}| \sqrt{\frac{KR}{h}}$$

Proof: By definition,

$$D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)=\sum_{\substack{j\in S_0\\ a_j(\boldsymbol{\theta})\neq 0}} D_j(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\;+\;\sum_{k=1}^{K}\mathbb{E}\left[\sum_{\substack{j\in S_k\\ a_j(\boldsymbol{\theta})\neq 0}}\frac{V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta},\boldsymbol{x}^0)-V(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}}\,\middle|\,\boldsymbol{Z}\right]. \tag{C.2}$$

Consider the first term. Since $\boldsymbol{x}^0$ is fixed (deterministic),

$$D_j(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\;=\;\frac{1}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}}\mathbb{E}\left[V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta},\boldsymbol{x}^0)-V(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\,\middle|\,\boldsymbol{Z}\right]\;=\;\frac{1}{h\sqrt{\nu_j}}\mathbb{E}\left[\delta_j x_j^0\right]\;=0.$$

This equality holds almost surely. Hence, it suffices to focus on the second term in Eq. (C.2).

Importantly, the second term is a sum of K *independent* random variables for a fixed $\boldsymbol{x}^0$. We next claim that each of these random variables is bounded. For any j such that $a_j(\boldsymbol{\theta})\neq 0$, let S_i be the subproblem such that $j\in S_i$. Then, write

$$\begin{aligned}
&\frac{V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta},\boldsymbol{x}^0)-V(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}}\\
&\quad=\frac{1}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}}\Big(\underbrace{\boldsymbol{r}^k(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)^\top\left(\boldsymbol{x}^k(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta},\boldsymbol{x}^0)-\boldsymbol{x}^k(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\right)}_{\leq 0\text{ by optimality of }\boldsymbol{x}^k(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)}+a_j(\boldsymbol{\theta})\delta_j x_j(\boldsymbol{Z}+\delta_j,\boldsymbol{e}_j,\boldsymbol{\theta},\boldsymbol{x}^0)\Big)\\
&\quad\leq\frac{\delta_j x_j(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta},\boldsymbol{x}^0)}{h\sqrt{\nu_j}}\\
&\quad\leq\frac{|\delta_j|}{h\sqrt{\nu_j}}.
\end{aligned}$$

By an analogous argument,

$$\frac{V(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)-V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta},\boldsymbol{x}^0)}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}}\leq\frac{|\delta_j|}{h\sqrt{\nu_j}}.$$

Hence,

$$\left|D_j(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\right|\;\leq\;\mathbb{E}\left[\left|\frac{V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j,\boldsymbol{\theta},\boldsymbol{x}^0)-V(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)}{a_j(\boldsymbol{\theta})h\sqrt{\nu_j}}\right|\,\middle|\,\boldsymbol{Z}\right]\;\leq\frac{1}{h\sqrt{\nu_j}}\mathbb{E}\left[\delta_j\right]\;\leq\;\frac{1}{h\sqrt{\nu_j}}\cdot\frac{\sqrt{h}}{\nu_j^{1/4}}\;\leq\frac{1}{\sqrt{h}\nu_{\min}^{3/4}}. \tag{C.3}$$

Applying Hoeffding's inequality to Eq. (C.2) and collecting constants shows

$$\mathbb{P}\left\{\left|D(\boldsymbol{Z};\boldsymbol{\theta})-\mathbb{E}\left[D(\boldsymbol{Z};\boldsymbol{\theta})\right]\right|\geq t\right\}\leq 2\exp\left(-\frac{h\cdot t^2}{C_0K\left|S_{\max}\right|^2}\right)$$

for some constant C_0 (depending on $\nu_{\min}$). Thus, with probability $1-\epsilon$, we see

$$\left|D(\boldsymbol{Z};\boldsymbol{\theta})-\mathbb{E}\left[D(\boldsymbol{Z};\boldsymbol{\theta})\right]\right|\leq\sqrt{\frac{C_0\cdot K\left|S_{\max}\right|^2}{h}\log\left(\frac{2}{\epsilon}\right)}$$

Combining constants completes our proof. $\square$

We now bound

$$\sup_{\boldsymbol{\theta}\in\Theta} \left| D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0) - \mathbb{E}\left[D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\right]\right|$$

for a fixed $\boldsymbol{x}^0$. The key idea is to use the Lipschitz continuity of $\boldsymbol{\theta}\mapsto D(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ to cover the set Θ.

Lemma C.4 (Uniform Convergence of VGC for a fixed $\boldsymbol{x}^0$) *Under the assumptions of Theorem 4.3 and for $\mathcal{H}\equiv[h_{\min},h_{\max}]$, there exists a constant C (that depends on $\nu_{\min}$, C_μ, L) such that for any $R>1$, we have with probability $1-2e^{-R}$,*

$$\sup_{\boldsymbol{\theta}\in\bar{\Theta}} |D(\boldsymbol{Z},\boldsymbol{\theta}) - \mathbb{E}[D(\boldsymbol{Z},\boldsymbol{\theta})]| \le C S_{\max}\sqrt{\frac{KR}{h_{\min}}}\sqrt{\log n\cdot \log N\left(\sqrt{\frac{h_{\min}}{Kn^2}},\Theta\right) N\left(\frac{h_{\min}}{K},\mathcal{H}\right)}.$$

Proof. Using the full notation $D(\boldsymbol{Z},(\boldsymbol{\theta},h))$, we first write the supremum to be over $\boldsymbol{\theta}\in\Theta$ and $h\in\mathcal{H}$. Let Θ_0 be a ε_1-covering of Θ and let $\overline{\mathcal{H}}$ be a ε_2-covering of $\mathcal{H}\equiv[h_{\min},h_{\max}]$. Then,

$$\sup_{\substack{\boldsymbol{\theta}\in\Theta\\ h\in\mathcal{H}}} |D(\boldsymbol{Z},(\boldsymbol{\theta},h)) - \mathbb{E}[D(\boldsymbol{Z},(\boldsymbol{\theta},h))]| \le \sup_{\substack{\boldsymbol{\theta}\in\Theta_0\\ h\in\overline{\mathcal{H}}}} |D(\boldsymbol{Z},(\boldsymbol{\theta},h)) - \mathbb{E}[D(\boldsymbol{Z},(\boldsymbol{\theta},h))]| \tag{C.4a}$$

$$+ \sup_{\substack{\boldsymbol{\theta},\overline{\boldsymbol{\theta}}:\|\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\|\le\varepsilon_1\\ h\in\mathcal{H}}} \left|D(\boldsymbol{Z},(\boldsymbol{\theta},h)) - D(\boldsymbol{Z},(\overline{\boldsymbol{\theta}},h))\right| + \sup_{\substack{\boldsymbol{\theta},\overline{\boldsymbol{\theta}}:\|\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\|\le\varepsilon_1\\ h\in\mathcal{H}}} \left|\mathbb{E}\left[D(\boldsymbol{Z},(\boldsymbol{\theta},h)) - D(\boldsymbol{Z},(\overline{\boldsymbol{\theta}},h))\right]\right| \tag{C.4b}$$

$$+ \sup_{\substack{\overline{\boldsymbol{\theta}}\in\Theta_0\\ h,\overline{h}:\|h-\overline{h}\|\le\varepsilon_2}} \left|D(\boldsymbol{Z},(\overline{\boldsymbol{\theta}},h)) - D(\boldsymbol{Z},(\overline{\boldsymbol{\theta}},\overline{h}))\right| + \sup_{\substack{\overline{\boldsymbol{\theta}}\in\Theta_0\\ h,\overline{h}:\|h-\overline{h}\|\le\varepsilon_2}} \left|\mathbb{E}\left[D(\boldsymbol{Z},(\overline{\boldsymbol{\theta}},\overline{h})) - D(\boldsymbol{Z},(\overline{\boldsymbol{\theta}},h))\right]\right| \tag{C.4c}$$

We bound Eq. (C.4b) and Eq. (C.4c) using Lemma 3.7. For Eq. (C.4b), there exists a constant C_1 such that with probability at least $1-e^{-R}$

$$\left|D(\boldsymbol{Z},(\boldsymbol{\theta},h)) - D(\boldsymbol{Z},(\overline{\boldsymbol{\theta}},h))\right| \le \frac{C_1\varepsilon_1 n^2}{h}\sqrt{R\log n}.$$

Similarly, there exists C_2, C_3 and C_4 (depending on $\nu_{\min}, L, C_\mu, a_{\min}, a_{\max}, b_{\max}$) such that

$$\left|\mathbb{E}\left[D(\boldsymbol{Z},(\boldsymbol{\theta},h)) - D(\boldsymbol{Z},(\overline{\boldsymbol{\theta}},h))\right]\right| \le \frac{C_2\varepsilon_1 n^2}{h}\left(\mathbb{E}\left[\|\boldsymbol{z}\|_\infty\right]+1\right) \le \frac{C_3\varepsilon_1 n^2}{h}\left(\sqrt{\log n}+C_\mu\right) \le \frac{C_4\varepsilon_1 n^2}{h}\sqrt{\log n},$$

where the second inequality uses a standard bound on the maximum of n sub-Gaussian random variables, and we have used Assumption 3.1 to simplify. Combining and taking the supremum over h shows

$$Eq.\ \text{(C.4b)} \le \frac{C_5\varepsilon_1 n^2}{h_{\min}}\sqrt{R\log n},$$

for some constant C_5.

For Eq. (C.4c), there exists a constant C_6 such that

$$\left|D(\boldsymbol{Z},(\boldsymbol{\theta},\overline{h})) - D(\boldsymbol{Z},(\boldsymbol{\theta},h))\right| \le \frac{C_6\varepsilon_2^{1/2} n}{h_{\min}}.$$

Since the same holds for the expectation of the same quantity, we see

$$Eq.\ (\mathrm{C.4c}) \ \leq\ \frac{2C_6\varepsilon_2^{1/2}n}{h_{\min}},$$

Similarly, using Lemma C.3 and applying a union bound over elements in $\overline{\Theta}$ and $\overline{\mathcal{H}}$ shows with probability at least $1-e^{-R}$,

$$Eq.\ (\mathrm{C.4a})\ \leq C_7 S_{\max}\sqrt{\frac{KR}{h_{\min}}\log\left(N(\varepsilon_1,\Theta)N(\varepsilon_2,\mathcal{H})\right)},$$

for some constant C_7. Choosing $\varepsilon_1 = \frac{S_{\max}\sqrt{Kh_{\min}}}{n^2}$, $\varepsilon_2 = \frac{S_{\max}^2 K h_{\min}}{n^2}$, we see

$$Eq.\ (\mathrm{C.4a}) + Eq.\ (\mathrm{C.4b}) + Eq.\ (\mathrm{C.4c}) \leq C_8 S_{\max}\sqrt{\frac{KR}{h_{\min}}\log\left(N(\varepsilon_1,\Theta)N(\varepsilon_2,\mathcal{H})\right)},$$

Finally, we obtain our result by simplifying the above bound slightly since $n = \sum_{k=0}^{K}|S_k| \leq KS_{\max}$, and hence,

$$\begin{aligned}\varepsilon_1 &= \frac{KS_{\max}\sqrt{Kh_{\min}}}{Kn^2} \ \geq\ \sqrt{\frac{h_{\min}}{Kn^2}}\\ \varepsilon_2 &= \frac{S_{\max}^2K^2h_{\min}}{Kn^2}\ \geq\ \frac{h_{\min}}{K}.\end{aligned}$$

Substituting the lower-bounds, we obtain our intended result.

□

We can now prove Theorem 4.3.

Proof of Theorem 4.3: We proceed to bound each term on the right side of Eq. (4.1).

To bound Eq. (4.1a), observe by definition of our lifted policy class and Lemma C.2, we have, with probability at least $1-e^{-R}$, that

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\bar{\Theta}}\left|\boldsymbol{\xi}^\top\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - \mathbb{E}\left[\boldsymbol{\xi}^\top\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})\right]\right| \ &\leq\ \sup_{\boldsymbol{\theta}\in\bar{\Theta},\boldsymbol{x}^0\in\mathcal{X}^0}\left|\boldsymbol{\xi}^\top\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0) - \mathbb{E}\left[\boldsymbol{\xi}^\top\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\right]\right|\\ &\leq\ CS_{\max}R\sqrt{K\dim(\phi)\log(K\left|\mathcal{X}^0\right|\mathcal{X}_{\max})}.\end{aligned}$$

To bound Eq. (4.1b), let $\mathcal{H}\equiv[h_{\min},h_{\max}]$. Then, by applying the union bound to Lemma C.4 with $R\leftarrow R+\log(1+|\mathcal{X}^0|)$ we have that with probability at least $1-2e^{-R}$,

$$\begin{aligned}\sup_{\substack{\boldsymbol{\theta}\in\Theta\\ \boldsymbol{x}^0\in\mathcal{X}^0}}\left|D(\boldsymbol{Z},\boldsymbol{\theta}) - \mathbb{E}\left[D(\boldsymbol{Z},\boldsymbol{\theta})\right]\right| \ &\leq\ C_1S_{\max}\sqrt{\frac{K(R+\log(|1+\mathcal{X}^0|)}{h_{\min}}}\sqrt{\log n\cdot\log\left(N\left(\varepsilon_1,\Theta\right)N\left(\varepsilon_2,\mathcal{H}\right)\right)}\\ &\leq\ C_2S_{\max}\sqrt{\frac{KR\log(|1+\mathcal{X}^0|)}{h_{\min}}}\sqrt{\log n\cdot\log\left(N\left(\varepsilon_1,\Theta\right)N\left(\varepsilon_2,\mathcal{H}\right)\right)},\end{aligned}$$

for some constants C_1 and C_2 and where $\varepsilon_1 = \sqrt{\frac{h_{\min}}{Kn^2}}$ and $\varepsilon_2 = \frac{h_{\min}}{K}$.

Finally, to bound Eq. (4.1c), use Theorem 3.2 and take the supremum over $h\in\mathcal{H}$ to obtain

$$Eq.\ (4.1\mathrm{c})\ \leq\ C_5h_{\max}KS_{\max}\log(1/h_{\min}).$$

Substituting these three bounds into Eq. (4.1) and collecting constants proves the theorem. □

Appendix D: Problems that are Weakly Coupled by Constraints

D.1. Properties of the Dual Optimization Problem

Throughout the section, we use the notation $\langle \ell, u\rangle$ to denote the interval $[\min(\ell,u), \max(\ell,u)]$.

Recall our dual formulation

$$\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta}) \in \arg\min_{\boldsymbol{\lambda}\geq \mathbf{0}} \; \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{z},\boldsymbol{\theta}), \quad \text{where} \;\; \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{z},\boldsymbol{\theta}) \;\equiv\; \boldsymbol{b}^\top\boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^n \left[r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right]^+.$$

Since $\mathcal{L}(\boldsymbol{\lambda})$ is non-differentiable, its (partial) subgradient is not-unique. We identify a particular subgradient by

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{z},\boldsymbol{\theta}) \;=\; \boldsymbol{b} - \frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{r_j(z_j,\boldsymbol{\theta}) > \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right\}\boldsymbol{A}_j.$$

The following identity that characterizes the remainder term in a first order Taylor-series expansion of $\mathcal{L}(\boldsymbol{\lambda})$ with this subgradient.

Lemma D.1 (A Taylor Series for $\mathcal{L}(\boldsymbol{\lambda})$) *For any $\boldsymbol{\lambda}\in\mathbb{R}^m_+$, $\boldsymbol{z}\in\mathbb{R}^n$, $\boldsymbol{\theta}\in\Theta$,*

$$\mathcal{L}(\boldsymbol{\lambda}_2,\boldsymbol{z},\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\lambda}_1,\boldsymbol{z},\boldsymbol{\theta}) = \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}_1,\boldsymbol{z},\boldsymbol{\theta})^\top(\boldsymbol{\lambda}_2-\boldsymbol{\lambda}_1) + \frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{r_j(z_j,\boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\rangle\right\}\left|r_j(\boldsymbol{z},\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right|.$$

Proof of Lemma D.1: Since $\boldsymbol{z}$ and $\boldsymbol{\theta}$ are fixed, drop them from the notation. Using the definitions of $\mathcal{L}$ and $\nabla_{\boldsymbol{\lambda}}\mathcal{L}$, we see it is sufficient to prove that for each j,

$$[r_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2]^+ - [r_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1]^+ + \mathbb{I}\left\{r_j > \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right\}\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_2-\boldsymbol{\lambda}_1) \;=\; \mathbb{I}\left\{r_j \in \langle \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\rangle\right\}\left|r_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right| \tag{D.1}$$

First notice that if $\boldsymbol{A}_j^\top\boldsymbol{\lambda}_1 = \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2$, then both sides of Eq. (D.1) equal zero. Further, if $r_j \notin \langle \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\rangle$, then both sides are again zero. Thus, we need only considering the case where $r_j \in \langle \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\rangle$. We can confirm the identity directly by considering the cases where $\boldsymbol{A}^\top\boldsymbol{\lambda}_1 < \boldsymbol{A}^\top\boldsymbol{\lambda}_2$ and $\boldsymbol{A}^\top\boldsymbol{\lambda}_1 > \boldsymbol{A}^\top\boldsymbol{\lambda}_2$ separately. □

The following result is proven in Lemma D.3 of GR 2021. We reproduce it here for completeness.

Lemma D.2 (Dual Solutions Bounded by Plug-in) *If $\mathcal{X}$ is s_0-strictly feasible, then*

i) $\|\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta})\|_1 \leq \frac{2}{ns_0}\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1$

ii) $\mathbb{E}\left[\|\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta})\|_1\right] \leq \frac{2}{ns_0}\mathbb{E}\left[\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1\right]$

Proof of Lemma D.2: By optimality, $\mathcal{L}(\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta}),\boldsymbol{z},\boldsymbol{\theta}) \leq \mathcal{L}(\mathbf{0},\boldsymbol{z},\boldsymbol{\theta}) \leq \frac{1}{n}\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1$. Since $\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta}) \geq \mathbf{0}$, it follows that $\|\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta})\|_1 = \boldsymbol{e}^\top\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta})$. Thus,

$$\begin{aligned}\|\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta})\|_1 \;\leq\; & \max_{\boldsymbol{\lambda}\geq 0} \boldsymbol{e}^\top\boldsymbol{\lambda} \\ & \text{s.t.} \quad \boldsymbol{b}^\top\boldsymbol{\lambda} + \frac{1}{n}\sum_{j=1}^n \left(r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right)^+ \;\leq\; \frac{1}{n}\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1.\end{aligned}$$

We upper bound this optimization by relaxing the constraint with penalty $1/s_0 > 0$ to see that

$$
\begin{aligned}
\|\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta})\|_1 &\le \max_{\boldsymbol{\lambda}\ge 0} \quad \boldsymbol{e}^\top\boldsymbol{\lambda} + \frac{1}{s_0}\left(\frac{1}{n}\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1 - \boldsymbol{b}^\top\boldsymbol{\lambda} - \frac{1}{n}\sum_{j=1}^n \left(r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right)^+\right)\\
&= \max_{\boldsymbol{\lambda}\ge 0} \quad \boldsymbol{e}^\top\boldsymbol{\lambda} + \frac{1}{s_0}\left(\frac{1}{n}\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1 - \boldsymbol{b}^\top\boldsymbol{\lambda} - \frac{1}{n}\sum_{j=1}^n \max_{x_j\in[0,1]} x_j\left(r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right)\right)\\
&\le \max_{\boldsymbol{\lambda}\ge 0} \quad \boldsymbol{e}^\top\boldsymbol{\lambda} + \frac{1}{s_0}\left(\frac{1}{n}\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1 - \boldsymbol{b}^\top\boldsymbol{\lambda} - \frac{1}{n}\sum_{j=1}^n x_j^0\left(r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right)\right)\\
&= \max_{\boldsymbol{\lambda}\ge 0} \quad \left(\boldsymbol{e} - \frac{1}{s_0}\boldsymbol{b} + \frac{1}{ns_0}\boldsymbol{A}\boldsymbol{x}^0\right)^\top\boldsymbol{\lambda} + \frac{1}{ns_0}\left(\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1 - \boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})^\top\boldsymbol{x}^0\right).
\end{aligned}
$$

By s_0-strict feasibility, $\frac{1}{n}\boldsymbol{A}\boldsymbol{x}^0 + s_0\boldsymbol{e} \le \boldsymbol{b} \longleftrightarrow \boldsymbol{e} - \frac{1}{s_0}\boldsymbol{b} + \frac{1}{ns_0}\boldsymbol{A}\boldsymbol{x}^0 \le 0$. Hence, $\boldsymbol{\lambda} = \boldsymbol{0}$ is optimal for this optimization problem. Thus, for all $\boldsymbol{\theta}\in\Theta$,

$$
\|\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta})\|_1 \;\le\; \frac{1}{ns_0}\left(\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1 - \boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})^\top\boldsymbol{x}^0\right) \;\le\; \frac{2}{ns_0}\|\boldsymbol{r}(\boldsymbol{z},\boldsymbol{\theta})\|_1.
$$

This proves i). Applying the expectation to both sides completes the proof. □

D.2. Constructing the Good Set

To construct the set of $\boldsymbol{Z}$ where approximate strong convexity holds or the "good" set, we first define the following constants:

$$
\lambda_{\max} \equiv \frac{2}{s_0}\left(a_{\max}\left(C_\mu + \frac{4}{\sqrt{\nu_{\min}}}\right) + b_{\max}\right), \tag{D.2a}
$$

$$
\phi_{\min} \equiv \frac{\sqrt{\nu_{\min}}}{a_{\max}\sqrt{2\pi}}\exp\left(-\frac{\nu_{\max}(a_{\max}C_\mu + b_{\max} + C_A\lambda_{\max})^2}{2a_{\min}^2}\right). \tag{D.2b}
$$

$$
\Lambda_n = \left\{(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)\in\mathbb{R}_+^m\times\mathbb{R}_+^m \;:\; \|\boldsymbol{\lambda}_1\|_1 \le \lambda_{\max},\; \|\boldsymbol{\lambda}_2\|_1 \le \lambda_{\max},\; \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \ge 4/n\right\},\text{ and} \tag{D.2c}
$$

$$
T_n = \left\{(\boldsymbol{\lambda},\boldsymbol{\theta},\Gamma)\in\mathbb{R}_+^m\times\Theta\times R \;:\; \|\boldsymbol{\lambda}\|_1 \le \lambda_{\max},\; \Gamma \ge \frac{1}{n}\right\}. \tag{D.2d}
$$

These values depend on the constants defined in Assumption 4.4 and Assumption 4.5.

We now define the "good" set,

$$
\begin{aligned}
\mathcal{E}_n \;\equiv\; \Bigg\{\boldsymbol{z} \;:\; &\left(\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}_1,\boldsymbol{z},\boldsymbol{\theta}) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}_2,\boldsymbol{z},\boldsymbol{\theta})\right)^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\\
&\qquad \ge\; \phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2}V^2\log(V)\frac{\log^2 n}{\sqrt{n}} \quad \forall(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)\in\Lambda_n,\;\forall\boldsymbol{\theta}\in\Theta,\\
&\frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{\left|r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right| \le \Gamma\right\} \;\le\; \Gamma\sqrt{\nu_{\max}} + \Gamma^{1/2}V\log(V)\frac{\log^2 n}{\sqrt{n}}, \quad \forall(\boldsymbol{\lambda},\boldsymbol{\theta},\Gamma)\in T_n\\
&\|\boldsymbol{z}\|_1 \le nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}},\\
&\|\boldsymbol{z}\|_\infty \le \log n\Bigg\},
\end{aligned}
$$

For clarity, we stress that $\phi_{\min} > 0$ and $\lambda_{\max} > 0$ are dimension independent constants.

We show in the next section that $\mathbb{P}(\boldsymbol{Z} \notin \mathcal{E}_n) = \tilde{O}(1/n)$. Thus, the event $\{\boldsymbol{Z} \in \mathcal{E}_n\}$ happens with high-probability, and we will perform our subsequent probabilistic analysis conditional on this "good" set.

D.3. Bounding the "Bad" Set

The purpose of this section is to bound $\mathbb{P}(\boldsymbol{Z} \notin \mathcal{E}_n)$. Since, $\mathcal{E}_n$ consists of four conditions, we treat each separately. The last two conditions on $\|\boldsymbol{Z}\|_1$ and $\|\boldsymbol{Z}\|_\infty$ can be analyzed using standard techniques for sub-Gaussian random variables.

Lemma D.3 (Bounding $\|\boldsymbol{Z}\|_1$) *Under Assumption 3.1,*

$$\mathbb{P}\left(\|\boldsymbol{Z}\|_1 > nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}\right) \le e^{-n/32}.$$

Proof of Lemma D.3: Note that $\mathbb{E}[|Z_j - \mu_j|] \le \frac{1}{\sqrt{\nu_{\min}}}$ by Jensen's inequality. Furthermore, because each Z_j is sub-Gaussian with variance proxy $\frac{1}{\nu_{\min}}$, we have by Lemma A.1 of GR 2021 that $\| |Z_j - \mu_j| \|_\Psi = \|Z_j - \mu_j\|_\Psi \le \frac{2}{\sqrt{\nu_{\min}}}$. Thus, $|Z_j - \mu_j| - \mathbb{E}[|Z_j - \mu_j|]$ is a mean-zero sub-Gaussian random variable with variance proxy at most $\frac{16}{\nu_{\min}}$. Finally, observe $\|\boldsymbol{Z}\|_1 \le nC_\mu + \sum_{j=1}^n |Z_j - \mu_j|$. Hence,

$$\begin{aligned}
\mathbb{P}\left(\|\boldsymbol{Z}\|_1 > nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}\right) &\le \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n |Z_j - \mu_j| > \frac{2}{\sqrt{\nu_{\min}}}\right) \\
&\le \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n |Z_j - \mu_j| - \mathbb{E}[|Z_j - \mu_j|] > \frac{1}{\sqrt{\nu_{\min}}}\right) \\
&\le e^{-\frac{n}{32}},
\end{aligned}$$

by the usual bound on the sum of independent sub-Gaussian random variables.

□

Lemma D.4 (Bounding $\|\boldsymbol{Z}\|_\infty$) *Under Assumption 3.1, there exists a dimension independent constant n_0 such that for all $n \ge n_0$,*

$$\mathbb{P}(\|\boldsymbol{Z}\|_\infty > \log n) \le \frac{1}{n^2}.$$

Proof of Lemma D.4: By Wainwright (2019), $\mathbb{E}[\|\boldsymbol{Z}\|_\infty\|] \le C_1\sqrt{\log n}$, for some dimension independent constant C_1. Moreover, by (Wainwright 2019, Example 2.29), $\|\boldsymbol{Z}\|_\infty - \mathbb{E}[\|\boldsymbol{Z}\|_\infty]$ is sub-Gaussian with variance proxy at most $1/\nu_{\min}$. Hence,

$$\mathbb{P}(\|\boldsymbol{Z}\|_\infty > \log n) \le \mathbb{P}\left(\|\boldsymbol{Z}\|_\infty - \mathbb{E}[\|\boldsymbol{Z}\|_\infty] > \log n - C_1\sqrt{\log n}\right).$$

For n sufficiently large, $\log n - C_1\sqrt{\log n} \geq \frac{1}{2}\log n$. Hence, for n sufficiently large, this last probability is at most

$$\mathbb{P}\left(\|\boldsymbol{Z}\|_\infty - \mathbb{E}\left[\|\boldsymbol{Z}\|_\infty\right] > \frac{1}{2}\log n \right) \;\leq\; \exp\left(-\frac{\nu_{\min}\log^2 n}{8} \right)$$
$$\leq\; n^{-\frac{\nu_{\min}\log n}{8}}.$$

For n sufficiently large, the exponent is at most -2, proving the lemma. □

We next establish that the inequality bounding the behavior over T_n hold with high probability. As a preparation, we first bound the supremum of a particular stochastic process over this set.

Lemma D.5 (Suprema over T_n) *Recall the definition of T_n in Eq.* (D.2d). *Under Assumptions 3.1, 3.6 and 4.6, there exist dimension independent constants C and n_0 such for all $n \geq n_0$, we have that for any $R > 1$, with probability at least $1 - e^{-R}$,*

$$\sup_{(\boldsymbol{\lambda},\boldsymbol{\theta},\Gamma)\in T_n} \left| \sum_{j=1}^n \mathbb{I}\left\{ \left| r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq \Gamma \right\} - \mathbb{P}\left(\left| r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq \Gamma \right) \right| \Gamma^{-1/2} \;\leq\; CRV\log V\sqrt{n}.$$

Proof of Lemma D.5: Our goal will be to apply Theorem A.3. As a first step, we claim that for a fixed $\boldsymbol{\lambda}, \boldsymbol{\theta}, \Gamma$, there exists a dimension independent constant C_1 such that

$$\mathbb{P}\left(\left| r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \;\leq\; \Gamma \right) \;\leq \sqrt{\nu_j} C_1 \Gamma.$$

To prove the claim, notice that this quantity is the probability that a Gaussian random variables lives in an interval of length 2Γ. Upper bounding the density of the Gaussian by its value at its mean shows the probability is at most $\sqrt{\frac{\nu_j}{2\pi}}$. Thus,

$$\mathbb{P}\left(\left| r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \;\leq\; \Gamma \right) \;\leq 2\Gamma\sqrt{\frac{\nu_j}{2\pi}} \;\leq \Gamma\sqrt{\nu_j}. \tag{D.3}$$

This upperbound further implies that there exists a dimension independent constant C_2 such that the parameter "σ^2" in Theorem A.3 is at most C_2, because the indicator squared equals the indicator. We also take the parameter "U" to be $\sqrt{n}$ since $\Gamma \;\geq \frac{1}{n}$.

Thus, to apply the theorem it remains to show the set

$$\mathcal{F} \equiv \left\{ \left(\mathbb{I}\left\{ \left| r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq \Gamma \right\} \right)_{j=1}^n \;:\; (\boldsymbol{\lambda},\boldsymbol{\theta},\Gamma) \in T_n \right\}$$

is Euclidean and compute its parameters.

Consider the set

$$\mathcal{F}_1 \equiv \left\{ \left(\mathbb{I}\left\{ r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \leq \Gamma \right\} \right)_{j=1}^n \;:\; (\boldsymbol{\lambda},\boldsymbol{\theta},\Gamma) \in T_n \right\}.$$

By Assumption 4.6, this set has VC-dimension at most V, and, hence, also has pseudo-dimension at most V. The same is true of the set

$$\mathcal{F}_2 \equiv \left\{ \left(\mathbb{I}\left\{ \boldsymbol{A}_j^\top \boldsymbol{\lambda} - r_j(Z_j, \boldsymbol{\theta}) \leq \Gamma \right\} \right)_{j=1}^n \;:\; (\boldsymbol{\lambda}, \boldsymbol{\theta}, \Gamma) \in T_n \right\}.$$

Since $\mathcal{F} = \mathcal{F}_1 \wedge \mathcal{F}_2$, by (Pollard 1990, Lemma 5.1) there exists an absolute constant C_2 such that $\mathcal{F}$ has pseudo-dimension at most $C_2 V$. By Theorem A.3 of GR 2021, $\mathcal{F}$ is Euclidean with parameters $A = (C_2 V)^{6 C_2 V}$ and $W = 4 C_2 V$. The relevant complexity parameter "$V(A, W)$" is then at most

$$\frac{6 C_2 V \log(C_2 V) + 4 C_2 V}{\sqrt{6 C_2 V \log(C_2 V)}} \;\leq\; C_3 \sqrt{V \log V},$$

for some dimension independent constant C_3.

Theorem A.3 now bounds the suprema by $C_4 R V \log(V) \sqrt{n}$, completing the proof. $\square$

Equipped with Lemma D.5, we can now show the relevant condition holds with high-probability.

Lemma D.6 (Bounding Away from Degeneracy) *Recall the definition of T_n in Eq.* (D.2d). *Under Assumptions 3.1, 3.6 and 4.6 there exists a dimension independent constant n_0 such that for all $n \geq n_0$, with probability at least $1 - 1/n$ we have that*

$$\frac{1}{n} \sum_{j=1}^n \mathbb{I}\left\{ \left| r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq \Gamma \right\} \;\leq\; \Gamma \sqrt{\nu_{\max}} \;+\; \Gamma^{1/2} V \log(V) \frac{\log^2 n}{\sqrt{n}}, \quad \forall (\boldsymbol{\lambda}, \boldsymbol{\theta}, \Gamma) \in T_n.$$

Proof of Lemma D.6 Apply Lemma D.5 with $R = \log n$ to conclude that with probability at least $1 - 1/n$, for all $(\boldsymbol{\lambda}, \boldsymbol{\theta}, \Gamma) \in T_n$ simultaneously, we have

$$\left| \sum_{j=1}^n \mathbb{I}\left\{ \left| r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq \Gamma \right\} - \mathbb{P}\left(\left| r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq \Gamma \right) \right| \;\leq\; C \Gamma^{1/2} V \log(V) \log(n) \sqrt{n}.$$

Then observe that as in the proof of Eq. (D.3), $\mathbb{P}\left(\left| r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \right| \leq \Gamma \right) \;\leq \Gamma \sqrt{\nu_{\max}}$. Finally, for n sufficiently large, $C \leq \log n$. Rearranging then completes the proof. $\square$

Remark D.7 *We describe the condition in Lemma D.6 as "Bounding Away from Degeneracy" because $r_j(\boldsymbol{Z}_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}$ is the reduced cost of the j^{th} component at the dual solution $\boldsymbol{\lambda}$. Hence, the lemma asserts that there are not too many reduced costs that are less than $1/n$.*

It remains to establish that the approximate strong convexity condition over Λ_n holds with high probability. As preparation, we again bound the suprema of a particular stochastic process.

Lemma D.8 (Suprema over Λ_n) *Under Assumptions 3.1, 3.6 and 4.6, there exists a dimension independent constant C such that for any $R > 1$, with probability at least $1 - e^{-R}$, we have*

$$\sup_{(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n, \boldsymbol{\theta} \in \Theta} \left| \frac{1}{n} \sum_{j=1}^n \left(\mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle \right\} - \mathbb{P}\left(r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle \right) \right) \frac{\left| \boldsymbol{A}_j^\top (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2) \right|}{\| \boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2 \|_2^{3/2}} \right|$$
$$\leq\; C R V^2 \log(V) \sqrt{n}.$$

Proof of Lemma D.8: Our strategy will be to apply Theorem A.3. To this end, we first claim that there exists a dimension independent constant $\phi_{\max}$ such that for any fixed $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n$ and $\boldsymbol{\theta} \in \Theta$

$$\mathbb{P}\left(r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle\right) \;\le\; \phi_{\max} \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_1.$$

To prove the claim, notice that this is the probability that a Gaussian random variable lives in an interval of length at most $\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right| \le C_A \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_1$. Upper bounding the Gaussian density by the square root of its precision proves the claim.

We next argue that this claim implies that there exists a dimension independent constant C_1 such that the parameter "σ^2" in Theorem A.3 is at most C_1. Indeed, an indicator squared is still the same indicator. Scaling by

$$\frac{\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right|^2}{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^3} \;\le\; \frac{C_A^2 m}{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2},$$

and then averaging over j proves that σ^2 is at most $C_1 m$.

We can take the parameter "U" to be $C_A\sqrt{n}$ since

$$\frac{\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right|}{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2}} \;\le\; C_A\sqrt{m}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{-1/2} \le C_A\sqrt{mn},$$

because $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n$.

Thus, to apply Theorem A.3 we need only show that the set

$$\mathcal{F} \equiv \left\{ \left(\mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle \right\} \frac{\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right|}{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2}} \right)_{j=1}^n \;:\; (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n,\; \boldsymbol{\theta} \in \Theta \right\}$$

is Euclidean and determine its parameters. To this end, first consider the sets

$$\begin{aligned}
\mathcal{F}_1 &\equiv \left\{ \left(\mathbb{I}\left\{r_j(Z_j, \boldsymbol{\theta}) \ge \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right\}\right)_{j=1}^n \;:\; \boldsymbol{\lambda}_1 \in \mathbb{R}_+^m,\; \boldsymbol{\theta} \in \Theta \right\} \\
\mathcal{F}_2 &\equiv \left\{ \left(\mathbb{I}\left\{r_j(Z_j, \boldsymbol{\theta}) \le \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right\}\right)_{j=1}^n \;:\; \boldsymbol{\lambda}_2 \in \mathbb{R}_+^m,\; \boldsymbol{\theta} \in \Theta \right\} \\
\mathcal{F}_3 &\equiv \left\{ \left(\mathbb{I}\left\{r_j(Z_j, \boldsymbol{\theta}) \ge \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right\}\right)_{j=1}^n \;:\; \boldsymbol{\lambda}_2 \in \mathbb{R}_+^m,\; \boldsymbol{\theta} \in \Theta \right\} \\
\mathcal{F}_4 &\equiv \left\{ \left(\mathbb{I}\left\{r_j(Z_j, \boldsymbol{\theta}) \le \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right\}\right)_{j=1}^n \;:\; \boldsymbol{\lambda}_1 \in \mathbb{R}_+^m,\; \boldsymbol{\theta} \in \Theta \right\}
\end{aligned}$$

By Assumption 4.6, each of these sets has VC-dimension at most V (they are indicator sets for functions with pseudo-dimension at most V). Now define the set

$$\mathcal{F}_5 \equiv \left\{ \left(\mathbb{I}\left\{r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \rangle\right\}\right)_{j=1}^n \;:\; (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n \right\},$$

and notice that

$$\mathcal{F}_5 \subseteq (\mathcal{F}_1 \wedge \mathcal{F}_2) \vee (\mathcal{F}_3 \wedge \mathcal{F}_4).$$

Hence, by (Pollard 1990, Lemma 5.1), there exists an absolute constant $C_2 > 1$ such that $\mathcal{F}_5$ has pseudodimension at most $C_2 V$. By Theorem A.3 of GR 2021, $\mathcal{F}_5$ is thus Euclidean with parameters $A = (C_2 V)^{6C_2 V}$ and $W = 4C_2 V$.

Now consider the set

$$\mathcal{F}_6 = \left\{ \left(\frac{\boldsymbol{A}_j^\top (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)}{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2}} \right)_{j=1}^n \ : \ (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_1) \in \Lambda_n \right\},$$

and notice

$$\mathcal{F}_6 \subseteq \{ \left(\boldsymbol{A}_j^\top \boldsymbol{\lambda} \right)_{j=1}^n \ : \boldsymbol{\lambda} \in \mathbb{R}^m \}.$$

This latter set belongs to a vector space of dimension at most m, and hence has pseudo-dimension at most $m \leq V$. Thus, by Theorem A.3 of GR 2021, it is Euclidean with parameters at most $A = V^{6V}$ and $W = 4V$.

To conclude, notice that $\mathcal{F}$ is the pointwise product of $\mathcal{F}_5$ and $\mathcal{F}_6$. Hence, by (Pollard 1990, Lemma 5.3), we have that $\mathcal{F}$ is Euclidean with parameters $A = (C_3 V)^{C_3 V} \cdot C_3^{C_3 V}$ and $W = C_3 V$ for some absolute constant C_3. In particular, the relevant complexity parameter "$V(A, W)$" for $\mathcal{F}$ is at most $C_4 \sqrt{V \log(V)}$ for some dimension independent parameter C_4.

Applying Theorem A.3 now shows that suprema of the lemma is at most $C_5 R (\sqrt{V \log V})^2 m \sqrt{mn}$, for some dimension independent C_5. Since $m \leq V$, this completes the lemma. □

Equipped with Lemma D.8, we can prove the approximate strong convexity condition holds with high probability.

Lemma D.9 (Approximate Strong Convexity with High Probability) *Under Assumptions 3.1, 3.6 and 4.6, there exists a dimension independent constant n_0 such that for all $n \geq n_0$, we have with probability at least $1 - \frac{1}{n}$ that the following inequality holds simultaneously for all $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n$ and $\boldsymbol{\theta} \in \Theta$:*

$$\left(\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_1, \boldsymbol{z}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_2, \boldsymbol{z}, \boldsymbol{\theta}) \right)^\top (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2) \ \geq \ \phi_{\min} \beta \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2} V^2 \log(V) \frac{\log^2 n}{\sqrt{n}}.$$

Proof of Lemma D.9: By choosing $R = \log n$, Lemma D.8 shows that there exists a dimension independent constant C_1 with probability at least $1 - 1/n$

$$\begin{aligned}
&\sup_{(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n, \boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{j=1}^n \left(\mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle \right\} - \mathbb{P}\left(r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle \right) \right) \frac{\left| \boldsymbol{A}_j^\top (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2) \right|}{\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|^{3/2}} \\
&\quad \geq \ -C V^2 \log(V) \sqrt{n} \log n.
\end{aligned}$$

This inequality implies that for any $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n$ and $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle\right\} \left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right| \\
&\qquad \geq \frac{1}{n}\sum_{j=1}^n \mathbb{P}\left(r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle\right) \left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right| - C\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|^{3/2} V^2 \log(V) \frac{\log n}{\sqrt{n}}.
\end{aligned}$$

Thus our first goal will be to bound the summation on the right side. Isolate the j^{th} term. The probability $\mathbb{P}\left(r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle\right)$ is the probability that a Gaussian random variable lives in an interval of length at most $\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right|$. Moreover, the endpoints of this interval are most $\left|\boldsymbol{A}_j^\top \boldsymbol{\lambda}_i\right| \leq C_A \boldsymbol{\lambda}_{\max}$ for $i = 1, 2$, since $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n$. It follows that these endpoints are no further than $a_j(\boldsymbol{\theta})\mu_j + b_j(\boldsymbol{\theta}) + C_A \lambda_{\max}$ from the mean of the relevant Gaussian. Thus, we can lower bound the density of the Gaussian on this interval. This reasoning proves

$$\begin{aligned}
&\mathbb{P}\left(r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle\right) \left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right| \\
&\geq \left(\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right)^2 \cdot \frac{\sqrt{\nu_j}}{a_j(\boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{\nu_j(a_j(\boldsymbol{\theta})\mu_j + b_j(\boldsymbol{\theta}) + C_A\lambda_{\max})^2}{2a_j(\boldsymbol{\theta})^2}\right) \\
&\geq \phi_{\min}\left(\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right)^2.
\end{aligned}$$

Averaging over j shows

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^n \mathbb{P}\left(r_j(Z_j, \boldsymbol{\theta}) \in \langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \rangle\right) \left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right| &\geq \phi_{\min}(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)^\top \frac{1}{n}\sum_{j=1}^n \boldsymbol{A}_j \boldsymbol{A}_j^\top (\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2) \\
&\geq \phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2,
\end{aligned}$$

by Assumption 4.5.

Substitute above, and notice if $n_0 = e^C$, then $\log n \geq C$ to complete the proof. □

Finally, a simple union bound gives

Lemma D.10 ($\boldsymbol{Z} \in \mathcal{E}_n$ with High Probability) *Under Assumptions 3.1, 3.6 and 4.6 there exists a dimension independent constant n_0 such that for all $n \geq n_0$, $\mathbb{P}(\boldsymbol{Z} \in \mathcal{E}_n) \geq 1 - \frac{4}{n}$.*

Proof. Combine Lemmas D.3, D.4, D.6 and D.9 and apply a union bound.

D.4. Properties of the Good Set

In this section, we argue that for data realizations $\boldsymbol{z} \in \mathcal{E}_n$, our optimization problems satisfy a number of properties, and, in particular, the dual solutions and VGC satisfy a bounded differences condition. We start by showing that small perturbations to the data $\boldsymbol{z}$ still yield dual solutions that are bounded. Note, any $\boldsymbol{z} \in \mathcal{E}_n$ satisfies the assumptions of the next lemma.

Lemma D.11 (Bounded Duals) *Suppose Assumptions 3.1 and 3.6 hold and* $\|\boldsymbol{t}\|_\infty \le \frac{3\sqrt{n}}{\sqrt{\nu_{\min}}}$ *and* $\boldsymbol{z}$ *satisfies* $\|\boldsymbol{z}\|_1 \le nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}$*. Then, for all* $j = 1,\ldots,n$*,*

$$\sup_{\boldsymbol{\theta}\in\Theta} \|\boldsymbol{\lambda}(\boldsymbol{z}+t_j\boldsymbol{e}_j,\boldsymbol{\theta})\|_1 \;\le\; \lambda_{\max}.$$

Proof of Lemma D.11: Write

$$\begin{aligned}
\|\boldsymbol{\lambda}(\boldsymbol{z}+t_j\boldsymbol{e}_j,\boldsymbol{\theta})\|_1 \;&\le\; \frac{2}{ns_0}\|\boldsymbol{r}(\boldsymbol{z}+t_j\boldsymbol{e}_j,\boldsymbol{\theta})\|_1 && \text{(Lemma D.2)}\\
&\le\; \frac{2}{ns_0}\left(a_{\max}\|\boldsymbol{z}\|_1 + a_{\max}\,|t_j| + b_{\max}n\right) && \text{(Definition of } \boldsymbol{r}(\cdot,\boldsymbol{\theta}))\\
&\le\; \frac{2}{s_0}\left(a_{\max}\left(C_\mu + \frac{2}{\sqrt{\nu_{\min}}} + \frac{3}{\sqrt{n\nu_{\min}}}\right) + b_{\max}\right) && \text{(by assumptions on } \|\boldsymbol{z}\|_1 \text{ and } \|\boldsymbol{t}\|_\infty)\\
&\le\; \lambda_{\max},
\end{aligned}$$

since $3/\sqrt{n} \le \sqrt{3} \le 2$. Taking the supremum of both sides over $\boldsymbol{\theta}\in\Theta$ completes the proof. □

We next establish a bounded differences condition for the dual solution $\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta})$.

Lemma D.12 (Bounded Differences for the Dual) *Suppose Assumptions 3.1 and 3.6 hold and that* $\boldsymbol{z}\in\mathcal{E}_n$ *and* $\|\boldsymbol{\lambda}(\overline{\boldsymbol{z}},\boldsymbol{\theta})\| \le \lambda_{\max}$*. Then, there exists a dimension independent constant* C *such that*

$$\|\boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta}) - \boldsymbol{\lambda}(\bar{\boldsymbol{z}},\boldsymbol{\theta})\|_2 \;\le\; CV^3\log^2(V)\frac{\log^4 n}{n}\sum_{j=1}^n \mathbb{I}\left\{z_j \ne \bar{z}_j\right\}.$$

Proof of Lemma D.12: To declutter the notation, define

$$\begin{aligned}
f_1(\boldsymbol{\lambda}) &\equiv \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{z},\boldsymbol{\theta}), \quad \boldsymbol{\lambda}_1 \equiv \boldsymbol{\lambda}(\boldsymbol{z},\boldsymbol{\theta}),\\
f_2(\boldsymbol{\lambda}) &\equiv \mathcal{L}(\boldsymbol{\lambda},\overline{\boldsymbol{z}},\boldsymbol{\theta}), \quad \boldsymbol{\lambda}_2 \equiv \boldsymbol{\lambda}(\overline{\boldsymbol{z}},\boldsymbol{\theta}).
\end{aligned}$$

Furthermore, let $I_j = \langle \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\rangle$.

Notice if $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \le 4/n$, the inequality is immediate for $C = 4$ since $m \ge 1$. Hence, we assume throughout that $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 > 4/n$.

Using Lemma D.1 we have that

$$\begin{aligned}
f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) \;&=\; \nabla f_1(\boldsymbol{\lambda}_1)^\top(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1) + \frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{r_j(z_j)\in I_j\right\}\left|r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right|\\
&\ge\; \frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{r_j(z_j)\in I_j\right\}\left|r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right|, \qquad \text{(D.4)}
\end{aligned}$$

where the inequality uses $f_1(\cdot)$ is convex and $\boldsymbol{\lambda}_1$ is an optimizer. Analogously, we have that

$$f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2) \;\ge\; \frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{r_j(\bar{z}_j)\in I_j\right\}\left|r_j(\bar{z}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|. \qquad \text{(D.5)}$$

Adding Eqs. (D.4) and (D.5) yields

$$\begin{aligned}
&f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) + f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2)\\
&\quad\geq \frac{1}{n}\sum_{j=1}^{n}\left(\mathbb{I}\left\{r_j(\bar{z}_j) \in I_j\right\}\left|r_j(\bar{z}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| + \mathbb{I}\left\{r_j(z_j)\in I_j\right\}\left|r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right|\right)
\end{aligned}$$

Isolate the j^{th} term on the right. To lower bound this term, note that when $z_j \neq \bar{z}_j$,

$$\begin{aligned}
&\left|\mathbb{I}\left\{r_j(\bar{z}_j)\in I_j\right\}\left|r_j(\bar{z}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| - \mathbb{I}\left\{r_j(z_j)\in I_j\right\}\left|r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|\right|\\
&\quad\leq \mathbb{I}\left\{r_j(\bar{z}_j)\in I_j\right\}\left|r_j(\bar{z}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| + \mathbb{I}\left\{r_j(z_j)\in I_j\right\}\left|r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|\\
&\quad\overset{\text{(a)}}{\leq} 2\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right|\\
&\quad\leq 2C_A\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2,
\end{aligned}$$

where inequality (a) follows because each indicator is non-zero only when the corresponding r is in the interval $\langle \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\rangle$. Hence, substituting above and rearranging yields

$$\begin{aligned}
&f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) + f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2) + 2C_A\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \cdot \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left\{z_j \neq \bar{z}_j\right\}\\
&\quad\geq \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left\{r_j(z_j)\in I_j\right\}\left(\left|r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right| + \left|r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|\right)\\
&\quad= \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left\{r_j(z_j)\in I_j\right\}\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\right|, \qquad \text{(D.6)}\\
&\quad= \left(\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}_1, \boldsymbol{z}, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}_2, \boldsymbol{z}, \boldsymbol{\theta})\right)^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2)\\
&\quad\geq \phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - V^2\log(V)\cdot\frac{\log^2 n}{\sqrt{n}}\cdot\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2}, \qquad (\text{since } \boldsymbol{z}\in\mathcal{E}_n, (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)\in\Lambda_n))
\end{aligned}$$

where Eq. (D.6) follows because when the indicator is non-zero, $r_j(z_j)$ is between $\boldsymbol{A}_j^\top\boldsymbol{\lambda}_1$ and $\boldsymbol{A}_j^\top\boldsymbol{\lambda}_2$.

To summarize the argument so far, we have shown that

$$\begin{aligned}
f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) + f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2) \; + \; 2C_A\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2\frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\left\{z_j \neq \bar{z}_j\right\} \qquad \text{(D.7)}\\
\geq \; \phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - V^2\log(V)\frac{\log^2 n}{\sqrt{n}}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2}.
\end{aligned}$$

The next step of the proof upper bounds the left side. By definition of $f_1(\cdot), f_2(\cdot)$,

$$\begin{aligned}
&f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) + f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2)\\
&\quad= \frac{1}{n}\sum_{j=1}^{n}\left(\left[r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right]^+ - \left[r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right]^+ + \left[r_j(\bar{z}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right]^+ - \left[r_j(\bar{z}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right]^+\right).
\end{aligned}$$

The j^{th} term is non-zero only if $z_j \neq \bar{z}_j$. In that case,

$$\begin{aligned}
&\left[r_j(z_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right]^+ - \left[r_j(z_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right]^+ + \left[r_j(\bar{z}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right]^+ - \left[r_j(\bar{z}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right]^+ \\
&\quad \leq\ 2\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1)\right| \\
&\quad \leq\ 2C_A\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1\|_2 \\
&\quad \leq\ 2C_A\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1\|_2.
\end{aligned}$$

Summing over j for which $z_j \neq \bar{z}_j$ and substituting into the left side of Eq. (D.7) yields,

$$4C_A\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1\|_2 \cdot \frac{1}{n}\sum_{j=1}^n \mathbb{I}\{z_j \neq \bar{z}_j\}\ \geq\ \phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - V^2\log(V)\frac{\log^2 n}{\sqrt{n}}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2}. \tag{D.8}$$

To simplify this expression, recall that by assumption

$$\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \geq \frac{4}{n} \implies \sqrt{n}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{1/2} \geq 1.$$

Hence we can inflate the left side of Eq. (D.8) by multiplying by $\sqrt{n}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{1/2}$ and then rearranging to obtain

$$\begin{aligned}
\phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 &\leq\ 4C_A\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|^{3/2} \cdot \frac{1}{\sqrt{n}}\sum_{j=1}^n \mathbb{I}\{z_j \neq \bar{z}_j\} + V^2\log(V)\frac{\log^2 n}{\sqrt{n}}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2} \\
&\leq\ C_1V^2\log(V)\frac{\log^2 n}{\sqrt{n}}\sum_{j=1}^n \mathbb{I}\{z_j \neq \bar{z}_j\}\,\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2},
\end{aligned}$$

for some dimension independent constant C_1. Dividing both sides by $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2}$ and combining constants yields

$$\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{1/2}\ \leq\ C_2V^2\log(V)\frac{\log^2 n}{\sqrt{n}}\sum_{j=1}^n \mathbb{I}\{z_j \neq \bar{z}_j\}, \tag{D.9}$$

for some dimension independent constant C_2. Multiply Eq. (D.9) by $V\log(V)\frac{\log^2 n}{\sqrt{n}}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2$ to see that

$$V\log(V)\frac{\log^2 n}{\sqrt{n}}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2} \leq C_2V^3\log^2(V)\frac{\log^4 n}{n}\sum_{j=1}^n \mathbb{I}\{z_j \neq \bar{z}_j\}\ \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2.$$

Substitute this upper-bound to Eq. (D.8), yielding

$$4C_A\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_1\|_2 \cdot \frac{1}{n}\sum_{j=1}^n \mathbb{I}\{z_j \neq \bar{z}_j\}\ \geq\ \phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - C_2V^3\log^2(V)\frac{\log^4 n}{n}\sum_{j=1}^n \mathbb{I}\{z_j \neq \bar{z}_j\}\ \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2.$$

Now divide by $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2$ and rearrange to complete the proof. □

We now use this result to show the VGC is also Lipschitz in the Hamming distance. The key to the following proof is that that strong-duality shows $V(\boldsymbol{z}, \boldsymbol{\theta}) = n\mathcal{L}(\boldsymbol{\lambda}(\boldsymbol{z}, \boldsymbol{\theta}), \boldsymbol{z}, \boldsymbol{\theta})$.

Lemma D.13 (Bounded Differences for VGC) *Let $\boldsymbol{z}, \overline{\boldsymbol{z}} \in \mathcal{E}_n$. Suppose Assumptions 3.1 and 3.6 hold. Then, there exists a dimension independent constant C, such that for any n such that $\frac{\log n}{nh} \leq 1$, we have that*

$$|D(\boldsymbol{z},\boldsymbol{\theta}) - D(\overline{\boldsymbol{z}},\boldsymbol{\theta})| \leq \frac{C}{h} \; V^3 \log^2(V) \log^4(n) \; \sum_{i=1}^{n} \mathbb{I}\{z_i \neq \overline{z}_i\}$$

Proof of Lemma D.13: Notice if $\boldsymbol{z} = \bar{\boldsymbol{z}}$ the lemma is trivially true. Hence, throughout, we assume $\boldsymbol{z} \neq \bar{\boldsymbol{z}}$. Since $\boldsymbol{\theta}$ is fixed throughout, we also drop it from the notation.

As a first step, we will prove the two inequalities

$$V(\boldsymbol{z}+\delta_j \boldsymbol{e}_j) - V(\boldsymbol{z}) \;\geq\; \left(r_j(\boldsymbol{z}) + a_j\delta_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{z}+\delta_j\boldsymbol{e}_j)\right)^+ - \left(r_j(\boldsymbol{z}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{z}+\delta_j\boldsymbol{e}_j)\right)^+, \quad \text{(D.10a)}$$

$$V(\boldsymbol{z}+\delta_j \boldsymbol{e}_j) - V(\boldsymbol{z}) \;\leq\; \left(r_j(\boldsymbol{z}) + a_j\delta_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{z})\right)^+ - \left(r_j(\boldsymbol{z}_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{z})\right)^+. \quad \text{(D.10b)}$$

To prove the first inequality, write

$$\begin{aligned} V(\boldsymbol{z}+\delta_j\boldsymbol{e}_j) - V(\boldsymbol{z}) &= n\mathcal{L}\left(\boldsymbol{\lambda}\left(\boldsymbol{z}+\delta_j\boldsymbol{e}_j\right), \boldsymbol{z}+\delta_j\boldsymbol{e}_j\right) - n\mathcal{L}\left(\boldsymbol{\lambda}\left(\boldsymbol{z}\right), \boldsymbol{z}\right) \quad \text{(D.11)} \\ &\leq n\mathcal{L}\left(\boldsymbol{\lambda}\left(\boldsymbol{z}\right), \boldsymbol{z}+\delta_j\boldsymbol{e}_j\right) - n\mathcal{L}\left(\boldsymbol{\lambda}\left(\boldsymbol{z}\right), \boldsymbol{z}\right) \\ &= \left(r_j(z_j) + a_j\delta_j - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{z})\right)^+ - \left(r_j(z_j) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{z})\right)^+ \end{aligned}$$

where the inequality holds by the sub-optimality of $\boldsymbol{\lambda}(\boldsymbol{z})$ for $\mathcal{L}\left(\boldsymbol{\lambda}, \boldsymbol{z}+\delta_j\boldsymbol{e}_j\right)$, and the last equality holds since all terms except the j^{th} in the summation of the Lagrangian cancel out. A similar argument using the sub-optimality of $\boldsymbol{\lambda}(\boldsymbol{z}+\delta_j\boldsymbol{e}_j)$ for $\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{z}+\delta_j\boldsymbol{e}_j)$ proves the lower bound.

The next step of the proof establishes that there exists a dimension independent constant C_1 such that

$$\begin{aligned} &\mathbb{E}\left[V(\boldsymbol{z}+\delta_j\boldsymbol{e}_j) - V(\boldsymbol{z}) - \left(V(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j) - V(\overline{\boldsymbol{z}})\right)\right] \\ &\quad\leq\; C_1\left(V^3\log^2(V)\frac{\log^4 n}{n}\sum_{i=1}^{n}\mathbb{I}\{z_i \neq \overline{z}_i\}\right)\mathbb{I}\{z_j = \overline{z}_j\} + C_1\sqrt{h}\;\mathbb{I}\{z_j \neq \overline{z}_j\} \quad \text{(D.12)} \end{aligned}$$

As suggested by the bound, we will consider two cases depending on whether $z_j = \bar{z}_j$.

Case 1: $z_j \neq \bar{z}_j$. Notice the inequalities Eq. (D.10) apply as well when $\boldsymbol{z}$ is replaced by $\bar{\boldsymbol{z}}$. Hence, applying the upper bound for the first term and the lower bound for the second term shows

$$\begin{aligned} &V(\boldsymbol{z}+\delta_j\boldsymbol{e}_j) - V(\boldsymbol{z}) - \left(V(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j) - V(\overline{\boldsymbol{z}})\right) \\ &\quad\leq \left(\left(r_j(z_j) + a_j\delta_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{z})\right)^+ - \left(r_j(z_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{z})\right)^+\right) \\ &\qquad - \left(\left(r_j(\bar{z}_j) + a_j\delta_j - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\right)^+ - \left(r_j(\bar{z}_j) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\right)^+\right) \\ &\quad\leq\; 2\,|a_j|\,|\delta_j|, \end{aligned}$$

because $t \mapsto t^+$ is a 1-Lipschitz function. Take expectations of both sides, using Jensen's inequality and upper bounding the variance of δ_j shows

$$\mathbb{E}\left[V(\boldsymbol{z}+\delta_j\boldsymbol{e}_j)-V(\boldsymbol{z})-(V(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)-V(\overline{\boldsymbol{z}}))\right] \;\le\; 2a_{\max}\mathbb{E}\left[|\delta_j|\right] \;\le 2a_{\max}\frac{\sqrt{3h}}{\nu_{\min}^{1/4}}.$$

Collecting constants proves the inequality when $z_j \neq \bar{z}_j$.

Case 2: $z_j = \bar{z}_j$. Proceeding as in Case 1,

$$\begin{aligned}
&V(\boldsymbol{z}+\delta_j\boldsymbol{e}_j)-V(\boldsymbol{z})-(V(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)-V(\overline{\boldsymbol{z}}))\\
&\quad\le \left(r_j(z_j)+a_j\delta_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{z})\right)^+-\left(r_j(z_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{z})\right)^+\\
&\qquad-\left(\left(r_j(\bar{z}_j)+a_j\delta_j-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\right)^+-\left(r_j(\bar{z}_j)-\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j))\right)^+\right)\\
&\quad\le 2\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}(\boldsymbol{z})-\boldsymbol{\lambda}(\bar{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j))\right|\\
&\quad\le 2C_A\|\boldsymbol{\lambda}(\boldsymbol{z})-\boldsymbol{\lambda}(\bar{\boldsymbol{z}})\|+2C_A\|\boldsymbol{\lambda}(\overline{\boldsymbol{z}})-\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\|,
\end{aligned}$$

where the second inequality follows again because $t \mapsto t^+$ is a contraction, but we group the terms in a different order, and the last inequality follows from the triangle-inequality and the Cauchy-Schwarz inequality. We can bound the first term by invoking Lemma D.12 yielding

$$\begin{aligned}
&V(\boldsymbol{z}+\delta_j\boldsymbol{e}_j)-V(\boldsymbol{z})-(V(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)-V(\overline{\boldsymbol{z}}))\\
&\qquad\le C_2V^3\log^2(V)\frac{\log^4 n}{n}\sum_{i=1}^n\mathbb{I}\{z_i\neq\bar{z}_i\}+2C_A\|\boldsymbol{\lambda}(\overline{\boldsymbol{z}})-\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\|,
\end{aligned}\tag{D.13}$$

for some dimension independent constant C_2. Taking expectations shows that to prove a bound, it will suffice to bound $\mathbb{E}\left[\|\boldsymbol{\lambda}(\overline{\boldsymbol{z}})-\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\|\right]$. To this end, consider splitting the expectation based on whether $|\delta_j|\geq 3\sqrt{\frac{n}{\nu_{\min}}}$.

If $|\delta_j|\leq 3\sqrt{\frac{n}{\nu_{\min}}}$, then by Lemma D.11, $\|\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\|_1 \le \lambda_{\max}$. Hence we can invoke Lemma D.12 again yielding

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\lambda}(\overline{\boldsymbol{z}})-\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\|\mathbb{I}\left\{|\delta_j|\leq 3\sqrt{\frac{n}{\nu_{\min}}}\right\}\right] \;&\le\; C_3V^3\log^2(V)\frac{\log^4 n}{n}\mathbb{P}\left(|\delta_j|\leq 3\sqrt{\frac{n}{\nu_{\min}}}\right)\\
&\le\; C_3V^3\log^2(V)\frac{\log^4 n}{n}.
\end{aligned}$$

Next, assume $|\delta_j|\geq 3\sqrt{\frac{n}{\nu_{\min}}}$. Write

$$\begin{aligned}
\|\boldsymbol{\lambda}(\overline{\boldsymbol{z}})-\boldsymbol{\lambda}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\| \;&\le\; \lambda_{\max}+\frac{2}{ns_0}\|\boldsymbol{r}(\overline{\boldsymbol{z}}+\delta_j\boldsymbol{e}_j)\|_1 && \text{(by Lemma D.2)}\\
&\le\; \lambda_{\max}+\frac{2}{ns_0}\|\boldsymbol{r}(\overline{\boldsymbol{z}})\|_1+\frac{2\,|a_j|}{ns_0}\|\,|\delta_j| && \text{(by def. of } \boldsymbol{r}(\cdot)\text{)}\\
&\le\; \lambda_{\max}+C_4+\frac{2a_{\max}}{ns_0}\,|\delta_j|\,,
\end{aligned}$$

for some dimension independent constant C_4, because $\overline{\boldsymbol{z}} \in \mathcal{E}$ implies that $\|\boldsymbol{r}(\boldsymbol{z})\|/n$ is bounded by a (dimension-independent) constant.

Thus, for some dimension independent constant C_5 we have

$$\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{\lambda}(\overline{\boldsymbol{z}}) - \boldsymbol{\lambda}(\overline{\boldsymbol{z}} + \delta_j \boldsymbol{e}_j)\| \mathbb{I}\left\{|\delta_j| > 3\sqrt{\frac{n}{\nu_{\min}}}\right\}\right] \;&\leq\; \mathbb{E}\left[\left(C_5 + \frac{C_5}{n}|\delta_j|\right)\mathbb{I}\left\{|\delta_j| > 3\sqrt{\frac{n}{\nu_{\min}}}\right\}\right] \\
&\leq\; \mathbb{E}\left[\left(C_5 + \frac{C_5}{n}\right)|\delta_j|\,\mathbb{I}\left\{|\delta_j| > 3\sqrt{\frac{n}{\nu_{\min}}}\right\}\right] \\
&=\; \mathbb{E}\left[2C_5\,|\delta_j|\,\mathbb{I}\left\{|\delta_j| > 3\sqrt{\frac{n}{\nu_{\min}}}\right\}\right],
\end{aligned}$$

where the final inequality uses $3\sqrt{\frac{n}{\nu_{\min}}} \geq 1$ since $n \geq 3$. Integration by parts with the Gaussian density shows there exists a dimension independent constant C_6 such that

$$\mathbb{E}\left[|\delta_j|\,\mathbb{I}\left\{|\delta_j| > 3\sqrt{\frac{n}{\nu_{\min}}}\right\}\right] \;\leq C_6\sqrt{h}e^{-nh/\sqrt{\nu_{\min}}} \;\leq\; C_6\sqrt{h}e^{-nh} \;\leq\; \frac{C_6}{n},$$

because $nh > \log n$ and $h < 1$ by assumption.

Combining the two cases shows that

$$\mathbb{E}\left[\|\boldsymbol{\lambda}(\overline{\boldsymbol{z}}) - \boldsymbol{\lambda}(\overline{\boldsymbol{z}} + \delta_j \boldsymbol{e}_j)\|\right] \;\leq\; C_7 V^3 \log^2(V)\frac{\log^4 n}{n}$$

for some dimension independent constant C_7.

Taking the expectation of Eq. (D.13), substituting this bound and collecting constants proves

$$V(\boldsymbol{z} + \delta_j \boldsymbol{e}_j) - V(\boldsymbol{z}) - \left(V(\overline{\boldsymbol{z}} + \delta_j \boldsymbol{e}_j) - V(\overline{\boldsymbol{z}})\right) \;\leq\; C_8 V^3 \log^2(V)\frac{\log^4 n}{n}\sum_{j=1}^n \mathbb{I}\{z_j \neq \overline{z}_j\}$$

for some constant C_8. Combining with Case 1 establishes Eq. (D.12).

Now, by symmetry, Eq. (D.12) holds with the roles of $\boldsymbol{z}$ and $\overline{\boldsymbol{z}}$ reversed. Hence, Eq. (D.12) also holds after taking the absolute values of both sides.

We can now write,

$$\begin{aligned}
|D(\boldsymbol{z}) - D(\overline{\boldsymbol{z}})| \;&\leq\; \sum_{j=1}^n \frac{1}{h\sqrt{\nu_j}\,|a_j|}\left|\mathbb{E}\left[V(\boldsymbol{z} + \delta_j \boldsymbol{e}_j) - V(\boldsymbol{z}) - \left(V(\overline{\boldsymbol{z}} + \delta_j \boldsymbol{e}_j) - V(\overline{\boldsymbol{z}})\right)\right]\right| \\
&\leq\; \frac{C_1}{h\sqrt{\nu_{\min}}a_{\min}} V^3 \log^2(V)\frac{\log^4 n}{n}\sum_{i=1}^n \mathbb{I}\{z_i \neq \overline{z}_i\}\sum_{j=1}^n \mathbb{I}\{z_j = \overline{z}_j\} + \frac{C_1}{\sqrt{h\nu_{\min}}a_{\min}}\sum_{j=1}^n \mathbb{I}\{z_j \neq \overline{z}_j\} \\
&\leq\; \frac{C_9}{h}\, V^3 \log^2(V)\log^4(n)\,\sum_{i=1}^n \mathbb{I}\{z_i \neq \overline{z}_i\},
\end{aligned}$$

for some constant C_9. This completes the proof. □

Finally, we show that $\boldsymbol{\theta} \mapsto \boldsymbol{\lambda}(\boldsymbol{z}, \boldsymbol{\theta})$ is also smooth on $\mathcal{E}_n$, at least locally.

Lemma D.14 (Local Smoothness of Dual Solution in $\boldsymbol{\theta}$) *Suppose $\boldsymbol{z} \in \mathcal{E}_n$ and that Assumptions 3.1, 3.6 and 4.6 hold. Then, there exist dimension independent constants C and n_0 such that for any $n \geq n_0$ and any $\bar{\boldsymbol{\theta}}$ such that $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \leq \frac{1}{n}$, we have that*

$$\left\|\boldsymbol{\lambda}(\boldsymbol{z}, \boldsymbol{\theta}) - \boldsymbol{\lambda}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})\right\|_2 \;\leq\; CV^2 \log V \frac{\log^{5/4} n}{\sqrt{n}}$$

Proof of Lemma D.14: The proof is similar to that of Lemma D.12. To declutter the notation, define

$$f_1(\boldsymbol{\lambda}) \equiv \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{z}, \boldsymbol{\theta}), \quad \boldsymbol{\lambda}_1 \equiv \boldsymbol{\lambda}(\boldsymbol{z}, \boldsymbol{\theta})$$

$$f_2(\boldsymbol{\lambda}) \equiv \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{z}, \bar{\boldsymbol{\theta}}), \quad \boldsymbol{\lambda}_2 \equiv \boldsymbol{\lambda}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})$$

Furthermore, let $I_j = \left\langle \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1, \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2 \right\rangle$.

If $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \leq 4V^2 \log V \frac{\log n}{\sqrt{n}}$, then the lemma holds trivially for $C = 4$. Hence, for the remainder, we assume $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 > 4V^2 \log V \frac{\log n}{\sqrt{n}}$. In particular, by Lemma D.11, this implies $(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \in \Lambda_n$.

Using Lemma D.1, we have that

$$\begin{aligned}
f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) &= \nabla f_1(\boldsymbol{\lambda}_1)^\top (\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1) + \frac{1}{n} \sum_{j=1}^n \mathbb{I}\left\{r_j(z_j, \boldsymbol{\theta}) \in I_j\right\} \left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right| \\
&\geq \frac{1}{n} \sum_{j=1}^n \mathbb{I}\left\{r_j(z_j, \boldsymbol{\theta}) \in I_j\right\} \left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right|
\end{aligned}$$

where $f_1(\cdot)$ is convex and $\boldsymbol{\lambda}_1$ is an optimizer. Similarly, we have that

$$f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2) \geq \frac{1}{n} \sum_{j=1}^n \mathbb{I}\left\{r_j(z_j, \bar{\boldsymbol{\theta}}) \in I_j\right\} \left|r_j(z_j, \bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right|.$$

Adding yields

$$\begin{aligned}
&f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) + f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2) \\
&\quad \geq \; \frac{1}{n} \sum_{j=1}^n \mathbb{I}\left\{r_j(z_j, \boldsymbol{\theta}) \in I_j\right\} \left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right| + \mathbb{I}\left\{r_j(z_j, \bar{\boldsymbol{\theta}}) \in I_j\right\} \left|r_j(z_j, \bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right|.
\end{aligned} \tag{D.14}$$

We would like to combine the j^{th} summand to simplify. To this end, adding and subtracting $\mathbb{I}\left\{r_j(z_j, \boldsymbol{\theta}) \in I_j\right\} \left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right|$ yields,

$$\begin{aligned}
&\mathbb{I}\left\{r_j(z_j, \boldsymbol{\theta}) \in I_j\right\} \left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right| + \mathbb{I}\left\{r_j(z_j, \bar{\boldsymbol{\theta}}) \in I_j\right\} \left|r_j(z_j, \bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right| \\
&\quad = \; \mathbb{I}\left\{r_j(z_j, \boldsymbol{\theta}) \in I_j\right\} \left(\left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right| + \left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right| \right) && \text{(D.15a)} \\
&\qquad + \mathbb{I}\left\{r_j(z_j, \bar{\boldsymbol{\theta}}) \in I_j\right\} \left|r_j(z_j, \bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right| - \mathbb{I}\left\{r_j(z_j, \boldsymbol{\theta}) \in I_j\right\} \left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right| && \text{(D.15b)}
\end{aligned}$$

We simplify Eq. (D.15a) by noting that when the indicator is non-zero,

$$\left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_2\right| + \left|r_j(z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}_1\right| \;=\; \left|\boldsymbol{A}_j^\top (\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1)\right|$$

Hence,

$$\text{Eq. (D.15a)} \;=\; \mathbb{I}\left\{r_j(z_j,\boldsymbol{\theta}) \in I_j\right\}\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1)\right|.$$

We rewrite Eq. (D.15b) as

$$\begin{aligned}
\text{Eq. (D.15b)} \;=\;& \mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \in I_j\right\}\Big(\left|r_j(z_j,\bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| - \left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|\Big) \\
&+ \left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|\Big(\mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \in I_j\right\} - \mathbb{I}\left\{r_j(z_j,\boldsymbol{\theta}) \in I_j\right\}\Big) \\
\overset{(a)}{\geq}\;& -\mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \in I_j\right\}\left|r_j(z_j,\bar{\boldsymbol{\theta}}) - r_j(z_j,\boldsymbol{\theta})\right| \\
&+ \left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|\Big(\mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \in I_j\right\} - \mathbb{I}\left\{r_j(z_j,\boldsymbol{\theta}) \in I_j\right\}\Big), \\
\overset{(b)}{\geq}\;& -\left|r_j(z_j,\bar{\boldsymbol{\theta}}) - r_j(z_j,\boldsymbol{\theta})\right| - \left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|\mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \notin I_j,\; r_j(z_j,\boldsymbol{\theta}) \in I_j\right\}, \\
\overset{(c)}{\geq}\;& -L\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2(|z_j| + 1) - \left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|\mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \notin I_j,\; r_j(z_j,\boldsymbol{\theta}) \in I_j\right\}, \\
\overset{(d)}{\geq}\;& -\frac{L}{n}(|z_j| + 1) - \left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right|\mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \notin I_j,\; r_j(z_j,\boldsymbol{\theta}) \in I_j\right\},
\end{aligned}$$

where inequality (a) is the triangle inequality, inequality (b) rounds the indicators, inequality (c) follows from the Lipschitz assumptions on $a_j(\boldsymbol{\theta})$ and $b_j(\boldsymbol{\theta})$, and inequality (d) uses $\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| \leq \frac{1}{n}$. Finally note that when the last indicator is non-zero,

$$\left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| \;\leq\; \left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1)\right| \;\leq\; 2C_A\sqrt{m}\lambda_{\max}.$$

Substituting this bound above and the resulting lower bound on Eq. (D.15b) into Eq. (D.15) proves

$$\begin{aligned}
&\mathbb{I}\left\{r_j(z_j,\boldsymbol{\theta}) \in I_j\right\}\left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right| + \mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \in I_j\right\}\left|r_j(z_j,\bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| \\
&\quad\geq \mathbb{I}\left\{r_j(z_j,\boldsymbol{\theta}) \in I_j\right\}\left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1)\right| - \frac{L}{n}(|z_j| + 1) - 2C_A\sqrt{m}\lambda_{\max}\mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \notin I_j,\; r_j(z_j,\boldsymbol{\theta}) \in I_j\right\}.
\end{aligned} \tag{D.16}$$

We can further clean up the last indicator by noting that

$$\begin{aligned}
&\mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \notin I_j,\; r_j(z_j,\boldsymbol{\theta}) \in I_j\right\} \\
&\quad\Longrightarrow\; \text{Either } \left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| \leq \left|r_j(\boldsymbol{z},\boldsymbol{\theta}) - r_j(\boldsymbol{z},\bar{\boldsymbol{\theta}})\right| \text{ or } \left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right| \leq \left|r_j(\boldsymbol{z},\boldsymbol{\theta}) - r_j(\boldsymbol{z},\bar{\boldsymbol{\theta}})\right|.
\end{aligned}$$

Moreover, because $\boldsymbol{z} \in \mathcal{E}$, we can use the Lipschitz assumptions on $a_j(\boldsymbol{\theta})$ and $b_j(\boldsymbol{\theta})$ and the fact that $\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| \leq \frac{1}{n}$ to write

$$|r_j(\boldsymbol{z},\boldsymbol{\theta}) - r_j(\boldsymbol{z},\boldsymbol{\theta})| \;\leq\; 2L\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|\log n \;\leq\; \frac{2L\log n}{n}.$$

Thus,

$$\mathbb{I}\left\{r_j(z_j,\bar{\boldsymbol{\theta}}) \notin I_j, r_j(z_j,\boldsymbol{\theta}) \in I_j\right\} \;\leq\; \mathbb{I}\left\{\left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| \leq \frac{2L\log n}{n}\right\} + \mathbb{I}\left\{\left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right| \leq \frac{2L\log n}{n}\right\}.$$

Making this substitution into Eq. (D.16), averaging over j, and substituting this bound into Eq. (D.14) shows

$$
\begin{aligned}
&f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) + f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2) \\
&\qquad \geq \frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{r_j(z_j,\boldsymbol{\theta}) \in I_j\right\} \left|\boldsymbol{A}_j^\top(\boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1)\right| - \frac{L}{n^2}\sum_{j=1}^n (|z_j|+1) \\
&\qquad\quad - 2C_A\sqrt{m}\lambda_{\max}\frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{\left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| \leq \frac{2L\log n}{n}\right\} + \mathbb{I}\left\{\left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right| \leq \frac{2L\log n}{n}\right\} \\
&\qquad = \left(\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}_1,\boldsymbol{z},\boldsymbol{\theta}) - \nabla_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}_2,\boldsymbol{z},\boldsymbol{\theta})\right)^\top(\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2) - \frac{L}{n}(\|\boldsymbol{z}\|_1/n + 1) \\
&\qquad\quad - 2C_A\sqrt{m}\lambda_{\max}\frac{1}{n}\sum_{j=1}^n \mathbb{I}\left\{\left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right| \leq \frac{2L\log n}{n}\right\} + \mathbb{I}\left\{\left|r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right| \leq \frac{2L\log n}{n}\right\} \\
&\qquad \geq \phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - V^2\log(V)\frac{\log^2 n}{\sqrt{n}}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2} - \frac{2L}{n}\left(C_\mu + 2/\sqrt{\nu_{\min}}\right) \\
&\qquad\quad - 8C_A\sqrt{m}\lambda_{\max}\left(L\sqrt{\nu_{\max}}\frac{\log n}{n} + \sqrt{2L}V\log(V)\frac{\log^{5/2} n}{n}\right),
\end{aligned}
$$

because $\frac{2L\log n}{n} \geq \frac{1}{n}$ by Assumption 3.1 and $\boldsymbol{z} \in \mathcal{E}_n$. Using Assumption 3.1 to further simplify, we have thus far shown that for some dimension independent constant C_2,

$$
\begin{aligned}
f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) + f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2) \;\geq\; &\phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - V^2\log(V)\frac{\log^2 n}{\sqrt{n}}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2} \\
&- C_2V^2\log V\frac{\log^{5/2} n}{n}
\end{aligned}
\tag{D.17}
$$

We next proceed to upper bound left side of this inequality. By definition of $f_1(\cdot), f_2(\cdot)$,

$$
\begin{aligned}
&f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) + f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2) \\
&\quad = \frac{1}{n}\sum_{j=1}^n\left(\left[r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right]^+ - \left[r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right]^+ + \left[r_j(z_j,\bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right]^+ - \left[r_j(z_j,\bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right]^+\right).
\end{aligned}
$$

Focusing on the j^{th} term, we see

$$
\begin{aligned}
&\left[r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right]^+ - \left[r_j(z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right]^+ + \left[r_j(z_j,\bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_1\right]^+ - \left[r_j(z_j,\bar{\boldsymbol{\theta}}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}_2\right]^+ \\
&\quad \leq 2\left|r_j(z_j,\boldsymbol{\theta}) - r_j(z_j,\bar{\boldsymbol{\theta}})\right| \\
&\quad \leq 2L\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2(|z_j|+1), \\
&\quad \leq \frac{2L}{n}(|z_j|+1),
\end{aligned}
$$

where the penultimate inequality uses Assumption 3.6. Averaging over j, we see

$$
f_1(\boldsymbol{\lambda}_2) - f_1(\boldsymbol{\lambda}_1) + f_2(\boldsymbol{\lambda}_1) - f_2(\boldsymbol{\lambda}_2) \;\leq\; \frac{2L}{n}\left(\frac{\|\boldsymbol{z}\|_1}{n} + 1\right) \;\leq\; \frac{C_3}{n}
$$

for some constant C_3, since $\boldsymbol{z} \in \mathcal{E}_n$.

Substitute into Eq. (D.17) to see that

$$\frac{C_3}{n} \;\geq\; \phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2} V^2 \log(V)\frac{\log^2 n}{\sqrt{n}} - C_2 V^2 \log V \frac{\log^{5/2} n}{n}$$

Rearranging and collecting constants shows

$$\phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{3/2} V^2 \log(V)\frac{\log^2 n}{\sqrt{n}} \;\leq\; C_4 V^2 \log V \frac{\log^{5/2} n}{n},$$

for some dimension-independent constant C_4.

We can also lower bound the left side by recalling

$$\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 > V^2 \log V \frac{\log n}{\sqrt{n}} \implies \frac{n^{1/4}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^{1/2}}{\log^{1/2}(n)\, V^2 \log(V)} > 1.$$

Hence, inflating the second term on the left yields

$$\phi_{\min}\beta\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 - \frac{\log^{3/2} n}{n^{1/4}}\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 \leq C_5 V^2 \log V \frac{\log^{5/2} n}{n}.$$

For n sufficiently large, the first term on the left is at least twice the second. Rearranging and taking square roots shows

$$\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \;\leq\; C_6 \sqrt{V^2 \log V}\frac{\log^{5/4} n}{\sqrt{n}}.$$

Recalling that $V \geq 2$ proves the theorem.

□

D.5. Pointwise Convergence Results

To prove our theorem, we require the uniform convergence of the in-sample optimism to expectation and the uniform convergence of VGC to its expectation. In this section, we first establish several pointwise convergence results to assist with this task. Our main workhorse will be Theorem A.5 where $\mathcal{E}_n$ defines the good set on which our random variables satisfy a bounded differences condition.

As a first step, we will show that the dual solutions converge (for a fixed $\boldsymbol{\theta}$) to their expectations. In preparation, we first bound the behavior of the dual on the bad set.

Lemma D.15 (Dual Solution Conditional Expectation Bound) *Suppose Assumptions 3.1 and 3.6 both hold. Let*

$$\mathcal{E}_{1,n} \equiv \left\{ \boldsymbol{z} : \|\boldsymbol{z}\|_1 \leq nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}} \right\}.$$

Then, there exists a dimension independent constant C, such that

$$\mathbb{E}\left[\lambda_i(\boldsymbol{Z}, \boldsymbol{\theta})\mathbb{I}\left\{\boldsymbol{Z} \in \mathcal{E}_{1,n}^c\right\}\right] \leq C \exp\left(-\frac{n}{C}\right)$$

Proof of Lemma D.15: We first bound

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{Z}\|_1 \mathbb{I}\left\{\boldsymbol{Z} \in \mathcal{E}_{1,n}^c\right\}\right] &= \int_0^\infty \mathbb{P}\left(\|\boldsymbol{Z}\|_1 \mathbb{I}\left\{\boldsymbol{Z} \in \mathcal{E}_{1,n}^c\right\} \geq t\right) dt \\
&= \int_0^{nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}} \mathbb{P}\left(\|\boldsymbol{Z}\|_1 > nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}\right) dt \;+\; \int_{nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}}^\infty \mathbb{P}\left(\|\boldsymbol{Z}\|_1 \geq t\right) dt \\
&\leq \left(nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}\right) e^{-n/32} \;+\; \int_{nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}}^\infty \mathbb{P}\left(\|\boldsymbol{Z}\|_1 \geq t\right) dt
\end{aligned}
$$

By inspection, there exists a dimension independent constant C_1 such that the first term is at most $C_1 e^{-n/C_1}$.

To analyze the second term, recall $\|\boldsymbol{Z}\|_1 \leq nC_\mu + \sum_{j=1}^n |Z_j - \mu_j|$. Hence,

$$
\begin{aligned}
\int_{nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}}^\infty \mathbb{P}\left(|\boldsymbol{Z}|_1 \geq t\right) dt &= \int_{nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}}^\infty \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n |Z_j - \mu_j| \geq \frac{1}{n}t - C_\mu\right) dt \\
&= \int_{nC_\mu + \frac{2n}{\sqrt{\nu_{\min}}}}^\infty \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n |Z_j - \mu_j| - \mathbb{E}\left[|Z_j - \mu_j|\right] \geq \frac{1}{n}t - C_\mu - \frac{1}{\nu_{\min}}\right) dt,
\end{aligned}
$$

since $\mathbb{E}\left[|Z_j - \mu_j|\right] \leq \frac{1}{\sqrt{\nu_{\min}}}$ by Jensen's inequality. Now make the change of variable $s = \frac{t}{n} - C_\mu - \frac{1}{\sqrt{\nu_{\min}}}$ to obtain

$$
n\int_{\nu_{\min}^{-1/2}}^\infty \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n |Z_j - \mu_j| - \mathbb{E}\left[|Z_j - \mu_j|\right] \geq s\right) ds \;\leq\; n\int_{\nu_{\min}^{-1/2}}^\infty e^{-\frac{s^2 \nu_{\min} n}{32}} ds
$$

because $|Z_j - \mu_j| - \mathbb{E}\left[|Z_j - \mu_j|\right]$ is a mean-zero, sub-Gaussian random variable with variance proxy at most $\frac{16}{\nu_{\min}}$. (See Lemma D.3 for clarification.) Making another change of variables proves this last integral is equal to

$$
\frac{4}{\sqrt{\nu_{\min} n}} \int_{\sqrt{n}/4}^\infty e^{-t^2/2} dt, \;\leq\; \frac{16}{n\sqrt{\nu_{\min}}} e^{-\frac{n}{32}}
$$

This value is also at most $C_2 e^{-n/C_2}$ for some constant C_2.

In summary, we have shown that there exists a dimension independent constant C_3 such that

$$
\mathbb{E}\left[\|\boldsymbol{Z}\|_1 \mathbb{I}\left\{\boldsymbol{Z} \in \mathcal{E}_{1,n}^c\right\}\right] \leq C_3 e^{-n/C_3}.
$$

Now to prove the lemma, recall by Lemma D.2,

$$
\lambda_i(\boldsymbol{Z}, \boldsymbol{\theta}) \leq \|\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta})\|_1 \leq \frac{2}{s_0 n} \|\boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{\theta})\|_1 \leq \frac{2}{s_0 n} \left(a_{\max} \|\boldsymbol{Z}\|_1 + b_{\max} n\right) \tag{D.18}
$$

where the second inequality holds by Assumption 3.6. Multiplying by $\mathbb{I}\left\{\boldsymbol{Z} \in \mathcal{E}_{1,n}^c\right\}$ and taking expectations hows,

$$
\mathbb{E}\left[\lambda_i(\boldsymbol{Z}, \boldsymbol{\theta}) \mathbb{I}\left\{\boldsymbol{Z} \in \mathcal{E}_{1,n}^c\right\}\right] \;\leq\; \frac{C_4}{n} e^{-n/C_4} + C_4 \mathbb{P}\left(\boldsymbol{Z} \in \mathcal{E}_{1,n}^c\right) \;\leq\; \frac{C_4}{n} e^{-n/C_4} + C_4 e^{-n/32},
$$

by Lemma D.3. Collecting constants proves the lemma. □

We now use Theorem A.5 to prove that the dual solution concentrates at its expectation for any fixed $\boldsymbol{\theta} \in \Theta$.

Lemma D.16 (Pointwise Convergence Dual Solution) *Fix some $\boldsymbol{\theta} \in \Theta$ and $i = 1, \dots, m$. Under Assumptions 3.1, 3.6 and 4.6 There exists dimension independent constants C and n_0, such that for all $n \geq n_0 e^R$, the following holds with probability $1 - \exp(-R)$*

$$|\lambda_i(\boldsymbol{Z}, \boldsymbol{\theta}) - \mathbb{E}[\lambda_i(\boldsymbol{Z}, \boldsymbol{\theta})]| \leq C V^3 \log^2 V \frac{\log^4 n}{\sqrt{n}} \sqrt{R}.$$

Proof of Lemma D.16: The proof will use the dual stability condition (Lemma D.12) to apply Theorem A.5. Since $\boldsymbol{\theta}$ is fixed throughout, we drop it from the notation.

By triangle inequality,

$$|\lambda_i(\boldsymbol{Z}) - \mathbb{E}[\lambda_i(\boldsymbol{Z})]| \;\leq\; \underbrace{|\lambda_i(\boldsymbol{Z}) - \mathbb{E}[\lambda_i(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{E}_n]|}_{(a)} + \underbrace{|\mathbb{E}[\lambda_i(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{E}_n] - \mathbb{E}[\lambda_i(\boldsymbol{Z})]|}_{(b)}. \tag{D.19}$$

We first bound (b) by a term that is $O(1/n)$. We see that

$$\begin{aligned}
\mathbb{E}[\lambda_i(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{E}_n] - \mathbb{E}[\lambda_i(\boldsymbol{Z})] \;&=\; \Big(\mathbb{E}[\lambda_i(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{E}_n] - \mathbb{E}[\lambda_i(\boldsymbol{Z}) \mid \boldsymbol{Z} \notin \mathcal{E}_n]\Big)\mathbb{P}\{\boldsymbol{Z} \notin \mathcal{E}_n\} \\
&\leq\; \frac{C_1}{n} + \mathbb{E}[\lambda_i(\boldsymbol{Z})\mathbb{I}\{\boldsymbol{Z} \notin \mathcal{E}_n\}],
\end{aligned}$$

where we used Lemmas D.10 and D.11 to bound the first term and C_1 is a dimension independent constant. To bound the second term, define the set

$$\mathcal{E}_0 \equiv \left\{ \boldsymbol{z} : \|\boldsymbol{z}\|_1 \leq n C_\mu + \frac{2n}{\sqrt{\nu_{\min}}} \right\}.$$

Notice, $\mathcal{E}_n \subseteq \mathcal{E}_0$. Then write,

$$\begin{aligned}
\mathbb{E}[\lambda_i(\boldsymbol{Z})\mathbb{I}\{\boldsymbol{Z} \notin \mathcal{E}_n\}] &= \mathbb{E}[\lambda_i(\boldsymbol{Z})(\mathbb{I}\{\boldsymbol{Z} \in \mathcal{E}_0^c\} + \mathbb{I}\{\boldsymbol{Z} \in \mathcal{E}_0 \cap \mathcal{E}_n^c\})] \\
&\leq C_2 \exp\left(-\frac{n}{C_2}\right) + \lambda_{\max}\mathbb{P}\{\boldsymbol{Z} \in \mathcal{E}_n^c\} && \text{(Lemmas D.11 and D.15)} \\
&\leq C_3 \exp\left(-\frac{n}{C_3}\right) + \frac{C_2}{n} && \text{(Lemma D.10)},
\end{aligned}$$

for dimension independent constants C_2 and C_3.

Collecting terms shows that there exists a dimension independent constant C_3 such that for n sufficiently large,

$$\text{Term (b) of Eq. (D.19)} \;=\; |\mathbb{E}[\lambda_i(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{E}_n] - \mathbb{E}[\lambda_i(\boldsymbol{Z})]| \leq \frac{C_3}{n}.$$

We now bound Term (a) by leveraging Theorem A.5. First note that for any $\boldsymbol{Z}, \bar{\boldsymbol{Z}} \in \mathcal{E}_n$, we have

$$|\lambda_i(\boldsymbol{Z}) - \lambda_i(\bar{\boldsymbol{Z}})| = \sqrt{|\lambda_i(\boldsymbol{Z}) - \lambda_i(\bar{\boldsymbol{Z}})|^2} \leq \sqrt{\sum_{i=1}^m |\lambda_i(\boldsymbol{Z}) - \lambda_i(\bar{\boldsymbol{Z}})|^2} = \|\boldsymbol{\lambda}(\boldsymbol{Z}) - \boldsymbol{\lambda}(\bar{\boldsymbol{Z}})\|_2.$$

Thus, by Lemma D.12, we see that

$$\left|\lambda_i(\boldsymbol{Z}) - \lambda_i(\bar{\boldsymbol{Z}})\right| \le C_4 V^3 \log^2 V \cdot \frac{\log^4(n)}{n} \sum_{j=1}^n \mathbb{I}\left\{Z_j \neq \bar{Z}_j\right\},$$

and, hence, $\lambda_i(\cdot)$ satisfies the bounded differences condition on $\mathcal{E}_n$. By Lemma D.10, $\mathbb{P}\{\boldsymbol{Z} \notin \mathcal{E}_n\} \le \frac{C_5}{n}$. By the assumptions, $n > 4C_5 e^R \implies e^{-R} > \frac{2C_5}{n}$. Theorem A.5 then shows that with probability at least $1 - e^{-R}$,

$$\begin{aligned}
\lambda_i(\boldsymbol{Z}) - \mathbb{E}[\lambda_i(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{E}_n] \;&\le\; C_5 C_4 V^3 \log^2(V) \frac{\log^4 n}{n} + C_4 V^3 \log^2(V) \frac{\log^4 n}{\sqrt{n}} \sqrt{\log\left(\frac{2}{e^{-R} - 2C_5/n}\right)} \\
&\le\; C_6 V^3 \log^2(V) \frac{\log^4 n}{\sqrt{n}} \sqrt{\log\left(\frac{2}{e^{-R} - 2C_5/n}\right)} \\
&\le\; C_6 V^3 \log^2(V) \frac{\log^4 n}{\sqrt{n}} \sqrt{\log\left(\frac{4}{e^{-R}}\right)} \\
&\le\; C_7 V^3 \log^2(V) \frac{\log^4 n}{\sqrt{n}} \sqrt{R},
\end{aligned}$$

where the third inequality again uses $n > 4C_5 e^R$, and the remaining inequalities simply collect constants and dominant terms.

To summarize, by substituting the two bounds into the upperbound of $|\lambda_i(\boldsymbol{Z}) - \mathbb{E}[\lambda_i(\boldsymbol{Z})]|$ in Eq. (D.19), we obtain that with probability at least $1 - e^{-R}$

$$|\lambda_i(\boldsymbol{Z}) - \mathbb{E}[\lambda_i(\boldsymbol{Z})]| \le C_8 V^3 \log^2(V) \frac{\log^4 n}{\sqrt{n}} \sqrt{R} + \frac{C_8}{n}.$$

Collecting terms completes the proof.

□

Proof of Lemma 4.9: Since $\boldsymbol{\theta}$ is fixed, we drop it from the notation. By triangle inequality,

$$|D(\boldsymbol{Z}) - \mathbb{E}[D(\boldsymbol{Z})]| \le |D(\boldsymbol{Z}) - \mathbb{E}[D(\boldsymbol{Z}) | \boldsymbol{Z} \in \mathcal{E}_n]| + |\mathbb{E}[D(\boldsymbol{Z}) | \boldsymbol{Z} \in \mathcal{E}_n] - \mathbb{E}[D(\boldsymbol{Z})]|. \tag{D.20}$$

We bound the latter term first. Since $D(\boldsymbol{Z})$ is bounded by Lemma 3.8, we see

$$\begin{aligned}
|\mathbb{E}[D(\boldsymbol{Z}) | \boldsymbol{Z} \in \mathcal{E}_n] - \mathbb{E}[D(\boldsymbol{Z})]| &= \Big|\mathbb{E}[D(\boldsymbol{Z}) | \boldsymbol{Z} \in \mathcal{E}_n] - \mathbb{E}[D(\boldsymbol{Z}) | \boldsymbol{Z} \notin \mathcal{E}_n]\Big| \mathbb{P}\{\boldsymbol{Z} \notin \mathcal{E}_n\} \\
&\le \frac{C_1}{n\sqrt{h}}
\end{aligned}$$

for some dimension independent constant C_1 by using Lemma D.10.

We now bound the first term. We use Theorem A.5. Recall from Lemma D.13 that

$$\left|D(\boldsymbol{Z}) - D(\overline{\boldsymbol{Z}})\right| \le C_2 \log^4 n \cdot V^3 \log^2 V \cdot \frac{1}{h} \sum_{j=1}^n \mathbb{I}\left\{Z_j \neq \overline{Z}_j\right\}$$

for $\boldsymbol{Z}, \overline{\boldsymbol{Z}} \in \mathcal{E}_n$ and from Lemma D.10 $\mathbb{P}\{\boldsymbol{Z} \notin \mathcal{E}_n\} \le \frac{C_1}{n}$. Finally if $n > 4C_1 e^R$, then $2C_1/n < \frac{1}{2}e^{-R}$, and we have by Theorem A.5 that

$$\begin{aligned}|D(\boldsymbol{Z}) - \mathbb{E}[D(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{E}_n]| \;&\le\; C_3 V^3 \log^2(V) \frac{\log^4(n)}{h} + C_3 V^3 \log^2(V) \frac{\log^4(n)\sqrt{n}}{h} \sqrt{\log\left(\frac{2}{e^{-R} - 2C_1/n}\right)} \\ &\le\; C_4 V^3 \log^2(V) \frac{\log^4(n)\sqrt{n}}{h} \sqrt{\log\left(\frac{2}{e^{-R} - 2C_1/n}\right)} \\ &\le\; C_5 V^3 \log^2(V) \frac{\log^4(n)\sqrt{n}}{h} \sqrt{R},\end{aligned}$$

where the last line again uses $n > 4C_1 e^R$.

Returning to the initial upper bound Eq. (D.20), we apply our two bounds to see

$$|D(\boldsymbol{Z}) - \mathbb{E}[D(\boldsymbol{Z})]| \;\le\; C_6 V^3 \log^2(V) \frac{\log^4(n)\sqrt{n}}{h}\sqrt{R} + \frac{C_6}{n\sqrt{h}}$$

By Assumption 3.1, $h < 1 < n$ implies that $\frac{\sqrt{n}}{h} \ge \frac{1}{n\sqrt{h}}$. Hence, collecting dominant terms completes the proof. □

D.6. Uniform Convergence of Dual Solutions

The goal of this section is to extend our previous pointwise results to uniform results over all $\boldsymbol{\theta} \in \Theta$. Let $\overline{\Theta}$ be a minimal $\frac{1}{n}$-covering of Θ. Then, for every $\boldsymbol{\theta} \in \Theta$ there exists $\bar{\boldsymbol{\theta}} \in \overline{\Theta}$ such that $\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2 \le \frac{1}{n}$.

Lemma D.17 (Uniform Convergence Dual Solution) *Under the assumptions of Theorem 4.7, there exists dimension independent constants C and n_0 such that for any $R > 1$ and any $n \ge n_0 e^R$, the following holds with probability $1 - 2e^{-R}$:*

$$\sup_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta})]\|_\infty \le C V^2 \log^2 V \log m \sqrt{R \log N\left(\frac{1}{n}, \Theta\right)} \frac{\log^4 n}{\sqrt{n}}$$

Proof of Lemma D.17: By triangle inequality,

$$\begin{aligned}\sup_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta}) - \mathbb{E}[\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta})]\|_\infty \le & \underbrace{\sup_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta}) - \boldsymbol{\lambda}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\|_\infty}_{(a)} + \underbrace{\sup_{\boldsymbol{\theta} \in \Theta} \|\mathbb{E}\left[\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta}) - \boldsymbol{\lambda}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\right]\|_\infty}_{(b)} \\ & + \underbrace{\sup_{\bar{\boldsymbol{\theta}} \in \bar{\Theta}} \|\boldsymbol{\lambda}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) - \mathbb{E}\left[\boldsymbol{\lambda}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\right]\|_\infty}_{(c)}\end{aligned}$$

We bound each term separately.

First we bound Term (a). If $\boldsymbol{Z} \in \mathcal{E}_n$, then, from Lemma D.14, and bounding the ℓ_∞-norm by the ℓ_2-norm,

$$\sup_{\boldsymbol{\theta} \in \Theta} \|\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta}) - \boldsymbol{\lambda}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\|_\infty \le C_1 V^3 \log^2 V \frac{\log^{5/4} n}{\sqrt{n}}$$

for some dimension independent constant C_1. By Lemma D.10, this occurs with probability at least $1-4/n$.

Next, we bound (b). Telescoping the expectation as before, we have for any $i=1,\ldots,m$ that

$$\begin{aligned}\mathbb{E}\left[\lambda_i(\boldsymbol{Z},\boldsymbol{\theta})-\lambda_i(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\right] &= \mathbb{E}\left[\lambda_i(\boldsymbol{Z},\boldsymbol{\theta})-\lambda_i(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\middle|\boldsymbol{Z}\in\mathcal{E}_n\right]\mathbb{P}\left\{\boldsymbol{Z}\in\mathcal{E}_n\right\}\\ &\quad+\mathbb{E}\left[\lambda_i(\boldsymbol{Z},\boldsymbol{\theta})-\lambda_i(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\middle|\boldsymbol{Z}\notin\mathcal{E}_n\right]\mathbb{P}\left\{\boldsymbol{Z}\notin\mathcal{E}_n\right\}\end{aligned}$$

We can bound the first term using Lemmas D.10 and D.14. To bound the second term, define the set

$$\mathcal{E}_{1,n}\equiv\left\{\boldsymbol{z}\;:\;\|\boldsymbol{z}\|_1\;\le nC_\mu+\frac{2n}{\sqrt{\nu}_{\min}}\right\},$$

and recall that $\mathcal{E}_n\subseteq\mathcal{E}_{1,n}$. Observe that

$$\begin{aligned}\mathbb{E}\left[\lambda_i(\boldsymbol{Z},\boldsymbol{\theta})\middle|\boldsymbol{Z}\notin\mathcal{E}_n\right]\mathbb{P}\left(\boldsymbol{Z}\notin\mathcal{E}_n\right) &= \mathbb{E}\left[\lambda_i(\boldsymbol{Z},\boldsymbol{\theta})\mathbb{I}\left\{\boldsymbol{Z}\notin\mathcal{E}_n\right\}\right]\\ &\le \mathbb{E}\left[\lambda_i(\boldsymbol{Z},\boldsymbol{\theta})\left(\mathbb{I}\left\{\boldsymbol{Z}\notin\mathcal{E}_{1,n}\right\}+\mathbb{I}\left\{\boldsymbol{Z}\in\mathcal{E}_{1,n},\boldsymbol{Z}\notin\mathcal{E}_n\right\}\right)\right]\\ &\le \mathbb{E}\left[\lambda_i(\boldsymbol{Z},\boldsymbol{\theta})\mathbb{I}\left\{\boldsymbol{Z}\notin\mathcal{E}_{1,n}\right\}\right]+\lambda_{\max}\mathbb{P}\left(\boldsymbol{Z}\notin\mathcal{E}_n\right) && \text{(Lemma D.11)}\\ &\le C_2\exp\left(-\frac{n}{C_2}\right)+\frac{C_2\lambda_{\max}}{n} && \text{(Lemmas D.10 and D.15),}\end{aligned}$$

for some dimension independent constant C_2. Combining these observations shows that

$$\begin{aligned}\mathbb{E}\left[\lambda_i(\boldsymbol{Z},\boldsymbol{\theta})-\lambda_i(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\right] &\le C_3V^3\log^2 V\frac{\log^{5/4}n}{\sqrt{n}}+C_3\exp\left(-\frac{n}{C_3}\right)+\frac{C_3\lambda_{\max}}{n}\\ &\le C_4V^3\log^2 V\frac{\log^{5/4}n}{\sqrt{n}}\end{aligned}$$

where C_3 and C_4 are dimension-independent constants. Taking the supremum over $\boldsymbol{\theta}\in\Theta$ and over $i=1,\ldots,m$, bounds Term (b).

Finally, we bound Term (c). We see that

$$\sup_{\bar{\boldsymbol{\theta}}\in\bar{\Theta},0\le i\le m}\left|\lambda_i(\boldsymbol{Z},\bar{\boldsymbol{\theta}})-\mathbb{E}\left[\lambda_i(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\right]\right|\le C_5V^3\log^2(V)\frac{\log^4 n}{\sqrt{n}}\sqrt{R\log\left(m\cdot N\left(\frac{1}{n},\Theta\right)\right)}$$

by applying Lemma D.16 and taking the union bound over the $\left|\overline{\Theta}\right|\le N\left(\frac{1}{n},\Theta\right)$ elements in $\overline{\Theta}$ by the m choices of i.

Taking a union bound over the probabilities that bounds hold on Terms (a) and (c) and adding term (b) shows that there exists a dimension independent constant C such that with probability $1-e^{-R}-4/n$

$$\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{E}\left[\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right]\|_\infty\;\le\;CV^3\log^2(V)\log m\sqrt{R\log N\left(\frac{1}{n},\Theta\right)}\frac{\log^4 n}{\sqrt{n}}.$$

Finally, note that if $n>4e^{-R}$, this last probability is at least $1-2e^{-R}$ to complete the proof. $\square$

D.7. Uniform Convergence of In-Sample Optimism

In this section, we construct a high-probability bound for

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\xi_j x_j(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{E}\left[\xi_j x_j(\boldsymbol{Z},\boldsymbol{\theta})\right]\right|$$

where we recall that $\boldsymbol{Z}=\boldsymbol{\mu}+\boldsymbol{\xi}$. Note for convenience we have scaled this by $\frac{1}{n}$.

Constructing the bound requires decomposing the in-sample optimism into several subcomponents. We outline the subcomponents by providing the proof to Lemma D.18 first. For convenience, in this section *only*, we use the notation $\boldsymbol{\lambda}(\boldsymbol{\theta})\equiv\mathbb{E}\left[\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right]$ as shorthand.

Lemma D.18 (Uniform In-sample Optimism for Coupling Constraints) *Let $N(\varepsilon,\Theta)$ be the $\varepsilon-$covering number of Θ. Under the assumptions of Theorem 4.7, there exists dimension independent constants C and n_0 such that for any $R>1$ and $n\geq n_0 e^R$, the following holds with probability $1-6\exp(-R)$.*

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\xi_j x_j(\boldsymbol{Z};\boldsymbol{\theta})-\mathbb{E}\left[\xi_j x_j(\boldsymbol{Z};\boldsymbol{\theta})\right]\right|\leq CV^3\log^3 V\sqrt{\log N\left(\frac{1}{n},\Theta\right)}\cdot\frac{R\log^4(n)}{\sqrt{n}}$$

Proof of Lemma D.18. By triangle inequality, we see,

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\xi_j x_j(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{E}\left[\xi_j x_j(\boldsymbol{Z},\boldsymbol{\theta})\right]\right|\\
&\leq\underbrace{\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\xi_j\left(x_j(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\}\right)\right|}_{\text{Rounding Error}}\\
&\quad+\underbrace{\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\xi_j\left(\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\}-\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}\right)\right|}_{\text{Dual Approximation Error}}\\
&\quad+\underbrace{\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}-\mathbb{E}\left[\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}\right]\right|}_{\text{ULLN for Dual Approximation}}\\
&\quad+\underbrace{\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\xi_j\left(\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}-\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\}\right)\right]\right|}_{\text{Expected Dual Approximation Error}}\\
&\quad+\underbrace{\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\left[\xi_j\left(\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq\boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\}-x_j(\boldsymbol{Z},\boldsymbol{\theta})\right)\right]\right|}_{\text{Expected Rounding Error}}.
\end{aligned}$$

For Rounding and Expected Rounding Error, we have

$$\begin{aligned}\left|\sum_{j=1}^{n}\xi_j\left(x_j(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\}\right)\right| &\leq \|\boldsymbol{\xi}\|_\infty\left\|x_j(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\}\right\|_1\\ &\leq \|\boldsymbol{\xi}\|_\infty m\end{aligned}$$

where the first inequality follows Holder's inequality and the second inequality holds by complementary slackness. Note,

$$\mathbb{P}\left\{\|\boldsymbol{\xi}\|_\infty\geq t\right\}\leq\sum_{j=1}^{n}\mathbb{P}\left\{|\xi_j|\geq t\right\}\leq\sum_{j=1}^{n}2\exp\left\{-\frac{\nu_{\min}t^2}{2}\right\}.$$

Moreover, $\mathbb{E}\left[\|\boldsymbol{\xi}\|_\infty\right]\leq C_1\sqrt{\log n}$ for some dimension independent constant C_1. Thus, with probability at least $1-e^{-R}$, we have

$$\begin{aligned}\text{Rounding Error}+\text{Expected Rounding Error} &\leq \|\boldsymbol{\xi}\|_\infty\frac{m}{n}+\mathbb{E}\|\boldsymbol{\xi}\|_\infty\frac{m}{n}\\ &= \frac{m}{n}\left(\|\boldsymbol{\xi}\|_\infty+\mathbb{E}\|\boldsymbol{\xi}\|_\infty\right)\\ &\leq \frac{C_2 m}{n}\sqrt{R\log n},\end{aligned}$$

for some dimension independent constant C_2.

We bound the Dual Approximation Error terms in Lemma D.20 with our uniform bounds on the dual solutions from Lemma D.17 below, proving that with probability at least $1-4e^{-R}$,

$$\begin{aligned}&\text{Dual Approximation Error}+\text{Expected Dual Approximation Error} && \text{(D.21)}\\ &\leq C_3R\sqrt{\frac{V}{n}}+C_3V^2\log^2 V\log(m)\cdot\sqrt{R\log N\left(\frac{1}{n},\Theta\right)}\cdot\frac{\log^4(n)}{\sqrt{n}} && \text{(D.22)}\\ &\leq C_4V^2\log^2 V\log(m)\frac{R}{\sqrt{n}}\cdot\sqrt{\log N\left(\frac{1}{n},\Theta\right)}\frac{\log^4(n)}{\sqrt{n}} && \text{(D.23)}\end{aligned}$$

for some dimension independent constants C_3 and C_4.

We bound the ULLN for Dual Approximation term in Lemma D.19 below to prove that

$$\text{ULLN for Dual Approximation}\leq C_5R\sqrt{\frac{V}{n}}$$

with probability $1-\exp(-R)$.

Taking a union bound over all probabilities and summing all bounds yields the result. □

Lemma D.19 (ULLN for Dual Approximation) *Under Assumptions 3.1 and 4.6, there exists a dimension independent constant C such that for any $R>1$, with probability at least $1-e^{-R}$,*

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{j=1}^{n}\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}-\mathbb{E}\left[\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}\right]\right|\leq C\cdot R\sqrt{Vn}$$

Proof of Lemma D.19: We first note that

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{j=1}^{n}\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}-\mathbb{E}\left[\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}\right]\right|\\
&\leq \sup_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\lambda}\in\mathbb{R}^m}\left|\sum_{j=1}^{n}\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right\}-\mathbb{E}\left[\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right\}\right]\right|
\end{aligned}$$

and the summation is a sum of centered independent random variables. We will apply Theorem A.1 to the last expression. Specifically, we consider the envelope $\boldsymbol{F}(\boldsymbol{Z})=(|Z_j|)_{j=1}^n$. Then, we have

$$\|\|\boldsymbol{F}(\boldsymbol{Z})\|_2\|_\Psi \overset{(a)}{\leq} \left\|\frac{\|\boldsymbol{\zeta}\|_2}{\sqrt{\nu_{\min}}}\right\|_\Psi \overset{(b)}{\leq} \sqrt{\frac{2n}{\nu_{\min}}}=C_1\sqrt{n}$$

for some dimension independent constant C_1. We see (a) holds by letting $\zeta_j=\sqrt{\nu_j}\xi_j$ and (b) holds by Lemma A.1 iv) from GR2020.

Next,

$$\left|\left\{\left(\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right\}\right)_{j=1}^n \,:\, \boldsymbol{\theta}\in\Theta,\,\boldsymbol{\lambda}\in\mathbb{R}^m\right\}\right| \leq \left|\left\{\left(\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right\}\right)_{j=1}^n \,:\, \boldsymbol{\theta}\in\Theta,\,\boldsymbol{\lambda}\in\mathbb{R}^m\right\}\right|,$$

and by Assumption 4.6, the latter set has VC-dimension V and hence cardinality at most 2^V.

Thus, we see that with probability $1-e^{-R}$, that

$$\sup_{\boldsymbol{\theta}\in\Theta,\boldsymbol{\lambda}\in\mathbb{R}^m}\left|\sum_{j=1}^{n}\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right\}-\mathbb{E}\left[\xi_j\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}\right\}\right]\right|\leq C_2R\sqrt{Vn}$$

for some absolute constant C_2. □

We next provide bounds for the Dual Approximation Error terms in the proof of Lemma D.18.

Lemma D.20 (Dual Approximation Error) *Assume Assumptions 3.1 and 4.6 hold. Then, there exists dimension independent constants C and n_0 such that for any $R>1$ and $n>n_0e^R$, we have with probability at least $1-4e^{-R}$, the following two inequalities hold simultaneously:*

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{j=1}^{n}\xi_j\left(\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\}-\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}\right)\right|\leq CR\sqrt{Vn},\\
&\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{j=1}^{n}\mathbb{E}\left[\xi_j\left(\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\}-\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta})\geq \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}\right)\right]\right|\\
&\quad\leq CV^2\log^3(V)\cdot\sqrt{R\log N\left(\frac{1}{n},\Theta\right)}\cdot\log^4(n)\sqrt{n}.
\end{aligned}$$

Proof of Lemma D.20: First observe that under the conditions of the theorem, Lemma D.17 implies that for some dimension independent constant C_1, with probability at least $1-2e^{-R}$,

$$\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})-\boldsymbol{\lambda}(\boldsymbol{\theta})\|_2 \;\leq\; \underbrace{C_1V^2\log^2(V)\log(m)\cdot\sqrt{R\log N\left(\frac{1}{n},\Theta\right)}\cdot\frac{\log^4(n)}{\sqrt{n}}}_{\equiv\delta},$$

where we have by bounding the ℓ_2-norm by $\sqrt{m}$ times the ℓ_∞ norm. Define the right side to be the constant δ as indicated.

We will restrict attention to the events where both the above inequality holds and also $\boldsymbol{Z} \in \mathcal{E}_n$. By the union bound and Lemma D.10, this event happens with probability at least $1 - 2e^{-R} - 4/n$. For $n > 4e^R$, this probability is at least $1 - 3e^{-R}$.

Now write

$$\begin{aligned}
&\left| \sum_{j=1}^n \xi_j \left(\mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) \geq \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta}) \right\} - \mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) \geq \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{\theta}) \right\} \right) \right| \qquad \text{(D.24)} \\
&\quad \leq \sum_{j=1}^n \left| \xi_j \left(\mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) \geq \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta}) \right\} - \mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) \geq \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{\theta}) \right\} \right) \right| \\
&\quad \leq \sum_{j=1}^n |\xi_j| \, \mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{\theta}) \in [-C_A\delta, C_A\delta] \right\},
\end{aligned}$$

because $\|\boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta}) - \boldsymbol{\lambda}(\boldsymbol{\theta})\|_2 \leq \delta$. Furthermore, when the indicator is non-zero, we can bound

$$|\xi_j| \;\leq\; \frac{1}{a_{\min}} \left(C_A \lambda_{\max} + C_A \delta + b_{\max} \right) + C_\mu \;\leq\; C_2(1+\delta), \qquad \text{(D.25)}$$

for some dimension independent C_2.

By Lemma D.2, $\|\boldsymbol{\lambda}(\boldsymbol{\theta})\|_1 \leq \frac{2}{ns_0} \mathbb{E}\|\boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{\theta})\|_1 \leq \lambda_{\max}$ and thus we can upper bound,

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in \Theta} \left| \sum_{j=1}^n \xi_j \left(\mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) \geq \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{Z}, \boldsymbol{\theta}) \right\} - \mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) \geq \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{\theta}) \right\} \right) \right| \\
&\leq \underbrace{C_2(1+\delta) \sup_{\boldsymbol{\theta} \in \Theta, \|\boldsymbol{\lambda}\|_1 \leq \lambda_{\max}} \left| \sum_{j=1}^n \mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \in [-C_A\delta, C_A\delta] \right\} - \mathbb{P}\left\{ r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \in [-C_A\delta, C_A\delta] \right\} \right|}_{(i)} \\
&\qquad + \underbrace{C_2(1+\delta) \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{P}\left\{ r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda}(\boldsymbol{\theta}) \in [-C_A\delta, C_A\delta] \right\}}_{(ii)}
\end{aligned}$$

To bound the supremum (i), we will apply Theorem A.1. Not that the vector $\boldsymbol{e}$ is a valid envelope. To bound the cardinality of

$$\mathcal{F} \equiv \left\{ \left(\mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \in [-C_A\delta, C_A\delta] \right\} \right)_{j=1}^n \;:\; \boldsymbol{\lambda} \in \mathbb{R}_+^m, \;\; \boldsymbol{\theta} \in \Theta \right\},$$

consider the two sets

$$\begin{aligned}
\mathcal{F}_1 &\equiv \left\{ \left(\mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \geq -C_A\delta \right\} \right)_{j=1}^n \;:\; \boldsymbol{\lambda} \in \mathbb{R}_+^m, \;\; \boldsymbol{\theta} \in \Theta \right\}, \\
\mathcal{F}_2 &\equiv \left\{ \left(\mathbb{I}\left\{ r_j(Z_j, \boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \leq C_A\delta \right\} \right)_{j=1}^n \;:\; \boldsymbol{\lambda} \in \mathbb{R}_+^m, \;\; \boldsymbol{\theta} \in \Theta \right\}.
\end{aligned}$$

Under Assumption 4.6, both sets have pseudo-dimension at most V. Furthermore, $\mathcal{F} = \mathcal{F}_1 \wedge \mathcal{F}_2$. Hence, by Pollard (1990), there exists an absolute constant C_3 such that the pseudo-dimension of $\mathcal{F}$ is at most $C_3 V$, and hence its cardinality is at most $n^{C_3 V}$.

Thus, applying Theorem A.1 shows that there exists a constant C_4 such that with probability at least $1 - e^{-R}$,

$$\text{Term (i)} \ \le \ C_4(1+\delta)R\sqrt{Vn\log n}.$$

To evaluate Term (ii), we recognize it as the probability as the probability that a Gaussian random variable lives in an interval of length $2C_A\delta$. Upper bounding the Gaussian density by its value at the mean shows

$$\mathbb{P}\left\{r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top \boldsymbol{\lambda} \in [-C_A\delta, C_A\delta]\right\} \le 2C_A\sqrt{\frac{\nu_{\max}}{2\pi}}\delta \ \le C_5\delta. \tag{D.26}$$

Thus,

$$\text{Term } (ii) \ \le C_6(1+\delta)\delta.$$

Combining our bounds, we see that with probability at least $1 - 4e^{-R}$,

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{j=1}^n \xi_j\left(\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta}) \ge \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\} - \mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta}) \ge \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}\right)\right| \\
&\quad \le \ C_7(1+\delta)R\sqrt{Vn} + C_7(1+\delta)\delta \\
&\quad \le \ C_7 R\sqrt{Vn},
\end{aligned}$$

by substituting the value of δ and only retaining the dominant terms. This proves the first result of the lemma.

To prove the second result of the lemma, note that

$$\begin{aligned}
&\left|\sum_{j=1}^n \mathbb{E}\left[\xi_j\left(\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta}) \ge \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\} - \mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta}) \ge \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}\right)\right]\right| \\
&\quad \le \sum_{j=1}^n \mathbb{E}\left[\left|\xi_j\left(\mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta}) \ge \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\right\} - \mathbb{I}\left\{r_j(Z_j,\boldsymbol{\theta}) \ge \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta})\right\}\right)\right|\right] \\
&\quad \le C_2(1+\delta)\sum_{j=1}^n \mathbb{P}\left(r_j(Z_j,\boldsymbol{\theta}) - \boldsymbol{A}_j^\top\boldsymbol{\lambda}(\boldsymbol{\theta}) \in [-C_A\delta, C_A\delta]\right) \\
&\quad \le nC_8(1+\delta)\delta,
\end{aligned}$$

where the second inequality uses the bound on $|\xi_j|$ (Eq. (D.25)) and the last inequality follows from argument leading to Eq. (D.26) above. Substituting the value of δ, using the assumption that $V \ge m$, and retaining only the dominant terms completes the proof. $\square$

D.8. Uniform Convergence of VGC

Lemma D.21 (Uniform VGC for Coupling Constraints) *Let $N(\varepsilon, \Theta)$ be the $\varepsilon-$covering number of Θ and the assumptions under Theorem 4.7 hold. There exists dimension independent constants C and n_0 such that for $n \geq n_0 e^R$ the following holds with probability $1 - Ce^{-R}$,*

$$\sup_{\boldsymbol{\theta} \in \bar{\Theta}} |D(\boldsymbol{Z}, \boldsymbol{\theta}) - \mathbb{E}[D(\boldsymbol{Z}, \boldsymbol{\theta})]| \leq C \cdot V^2 \log^2 V \cdot \sqrt{R \cdot n \log(n \cdot N(n^{-3/2}, \Theta))} \cdot \frac{\log^4 n}{h_{\min}}$$

Proof of Lemma D.21 We follow a similar strategy to Lemma C.4 and again consider the full notation version of the VGC, $D(\boldsymbol{Z}, (\boldsymbol{\theta}, h))$, and take the supremum over $\boldsymbol{\theta} \in \Theta$ and $h \in \mathcal{H}$ where $\mathcal{H} \equiv [h_{\min}, h_{\max}]$. Let Θ_0 be a minimal $n^{-3/2}$-covering of Θ. In particular, for any $\boldsymbol{\theta} \in \Theta$ there exists a $\bar{\boldsymbol{\theta}} \in \Theta_0$ such that $\left\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2 \leq n^{-3/2}$. Similarly, let $\bar{\mathcal{H}}$ be the n^{-1}-covering of $\mathcal{H}$. By telescoping,

$$\begin{aligned}
\sup_{\boldsymbol{\theta} \in \Theta} \left|D(\boldsymbol{Z}, (\boldsymbol{\theta}, \bar{h})) - \mathbb{E}\left[D(\boldsymbol{Z}, (\boldsymbol{\theta}, \bar{h}))\right]\right| \leq & \underbrace{\sup_{\substack{\bar{\boldsymbol{\theta}} \in \Theta_0 \\ \bar{h} \in \overline{\mathcal{H}}}} \left|D(\boldsymbol{Z}, (\bar{\boldsymbol{\theta}}, \bar{h})) - \mathbb{E}\left[D(\boldsymbol{Z}, (\bar{\boldsymbol{\theta}}, \bar{h}))\right]\right|}_{(i)} \\
& + \underbrace{\sup_{\substack{\boldsymbol{\theta}, \bar{\boldsymbol{\theta}} : \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| \leq n^{-3/2} \\ h \in \mathcal{H}}} \left|D(\boldsymbol{Z}, (\boldsymbol{\theta}, h)) - D(\boldsymbol{Z}, (\bar{\boldsymbol{\theta}}, h))\right| + \sup_{\substack{\boldsymbol{\theta}, \bar{\boldsymbol{\theta}} : \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\| \leq n^{-3/2} \\ h \in \mathcal{H}}} \left|\mathbb{E}\left[D(\boldsymbol{Z}, (\bar{\boldsymbol{\theta}}, h))\right] - \mathbb{E}[D(\boldsymbol{Z}, (\boldsymbol{\theta}, h))]\right|}_{(ii)} \\
& + \underbrace{\sup_{\substack{\bar{\boldsymbol{\theta}} \in \Theta_0 \\ h, \bar{h} : \|h - \bar{h}\| \leq n^{-1}}} \left|D(\boldsymbol{Z}, (\bar{\boldsymbol{\theta}}, h)) - D(\boldsymbol{Z}, (\bar{\boldsymbol{\theta}}, \bar{h}))\right| + \sup_{\substack{\bar{\boldsymbol{\theta}} \in \Theta_0 \\ h, \bar{h} : \|h - \bar{h}\| \leq n^{-1}}} \left|\mathbb{E}\left[D(\boldsymbol{Z}, (\bar{\boldsymbol{\theta}}, h))\right] - \mathbb{E}\left[D(\boldsymbol{Z}, (\bar{\boldsymbol{\theta}}, \bar{h}))\right]\right|}_{(iii)}
\end{aligned} \tag{D.27}$$

We bound Term (i) by taking a union bound over the $N(n^{-3/2}, \Theta)$ elements of $\bar{\Theta}$ and $N(n^{-1}, \mathcal{H})$ elements of $\overline{\mathcal{H}}$ in combination with the pointwise bound from Lemma 4.9. This shows that with probability at least $1 - 4e^{-R}$

$$\sup_{\substack{\bar{\boldsymbol{\theta}} \in \Theta_0 \\ \bar{h} \in \overline{\mathcal{H}}}} \left|D(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) - \mathbb{E}\left[D(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\right]\right| \leq C_1 V^2 \log^2(V) \sqrt{R \log\left(N\left(n^{-3/2}, \Theta\right) N(n^{-1}, \mathcal{H})\right)} \frac{\sqrt{n} \log^4(n)}{h_{\min}}.$$

Terms (ii) and (iii) of Eq. (D.27) can be bounded as follows.

First, for (ii), we see by Lemma 3.7 that for $\|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\| \leq n^{-3/2}$ there exists a constant C_1 such that

$$\left|D(\boldsymbol{Z}, (\boldsymbol{\theta}, h)) - D(\boldsymbol{Z}, (\bar{\boldsymbol{\theta}}, h))\right| \leq \frac{C_1 L}{h} \sqrt{\frac{R}{\nu_{\min}}} \cdot n^{1/2} \sqrt{\log n}$$

with probability $1 - \exp(-R)$. Similarly, there exists C_2, C_3 and C_4 (depending on $\nu_{\min}, L, C_\mu, a_{\min}, a_{\max}, b_{\max}$) such that

$$\left|\mathbb{E}\left[D(\boldsymbol{Z}, (\boldsymbol{\theta}, h)) - D(\boldsymbol{Z}, (\overline{\boldsymbol{\theta}}, h))\right]\right| \leq \frac{C_2 n^{1/2}}{h} \left(\mathbb{E}\left[\|\boldsymbol{z}\|_\infty\right] + 1\right) \leq \frac{C_3 n^{1/2}}{h} \left(\sqrt{\log n} + C_\mu\right) \leq \frac{C_4 n^{1/2}}{h} \sqrt{\log n},$$

where the second inequality uses a standard bound on the maximum of n sub-Gaussian random variables, and we have used Assumption 3.1 to simplify. Combining the two terms and taking the supremum over $h \in [h_{\min}, h_{\max}]$, we see there exists a constant C_5 (depending on C_1 and C_4), such that

$$(ii) \leq \frac{C_5\sqrt{Rn\log n}}{h_{\min}}$$

Finally, for (iii) we see by Lemma 3.7 that for $\|h - \overline{h}\| \leq n^{-1}$, there exists an absolute constant C_6 such that,

$$(iii) \leq \frac{C_6\sqrt{n}}{h_{\min}\nu_{\min}}$$

Combining, we see there exists dimension independent constants C_7 and C_8 such that with probability $1 - 5e^{-R}$,

$$\begin{aligned}\sup_{\substack{\boldsymbol{\theta}\in\Theta\\ h\in\mathcal{H}}} |D(\boldsymbol{Z},\boldsymbol{\theta}) - \mathbb{E}[D(\boldsymbol{Z},\boldsymbol{\theta})]| &\leq C_7 V^2 \log^2(V)\sqrt{Rn\log\left(N\left(n^{-3/2},\Theta\right)N(n^{-1},\mathcal{H})\right)}\frac{\log^4(n)}{h_{\min}} + C_7\frac{\sqrt{Rn\log n}}{h_{\min}} \\ &\leq C_8 \cdot V^2\log^2 V \cdot \sqrt{R\cdot n\log\left(n\cdot N(n^{-3/2},\Theta)\right)}\cdot\frac{\log^4 n}{h_{\min}}\end{aligned}$$

where the last inequality holds as $N(n^{-1},\mathcal{H}) \leq n$. This completes the proof. □

To obtain uniform bounds, we characterize the complexity of the policy class through the $n^{-3/2}$ covering number of the parameter space Θ. As an example to demonstrate the size of the covering number, consider the case where Θ is a compact subset of $\mathbb{R}^p$ with a finite diameter Γ. Applying Lemma 4.1 of Pollard (1990) and that the $\epsilon-$packing number bounds the $\epsilon-$covering number, we see that $N(n^{-3/2},\Theta) \leq \left(3n^{3/2}\Gamma\right)^p$. Combining this bound with Eq. (D.27), we obtain the following corollary.

Corollary D.22 (Uniform Convergence for Finite Policy Class) *Let Θ be a compact subset of $\mathbb{R}^p$ with a finite diameter Γ. There exists dimension independent constants C, n_0 such that for $n \geq n_0$ the following holds with probability $1 - C\exp\{-R\}$,*

$$\sup_{\boldsymbol{\theta}\in\Theta}\frac{1}{n}|D(\boldsymbol{Z},\boldsymbol{\theta}) - \mathbb{E}[D(\boldsymbol{Z},\boldsymbol{\theta})]| \leq C\log^4 n\cdot V^2\log V\cdot\frac{1}{h_{\min}}\sqrt{\frac{R}{n}\cdot p\log\left(3n^{3/2}\Gamma\right)}$$

This corollary shows that the complexity of the policy class depends on the number of parameters of the plug-in policies. We see from Section 2.2 that p for many common policy classes does not depend on n, implying that the convergence of the VGC estimator to its expectation follows the rate from Corollary D.22 up to log terms. For example, p for mixed effect policies depends on the dimension of $\boldsymbol{W}_j$ which reflects the information available, such as features, for each μ_j. This is typically fixed even as the number of observations n may increase. This implies that for many policy classes, the estimation error converges to 0 as $n \to \infty$.

D.9. Proof of Theorem 4.7

We can now prove Theorem 4.7.

Proof of Theorem 4.7: We proceed to bound each term on the right side of Eq. (4.1). To bound Eq. (4.1a), we have by Lemma D.18, that with probability at least $1-e^{-R}$, that

$$\sup_{\boldsymbol{\theta}\in\bar{\Theta}}\left|\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{E}\left[\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})\right]\right| \;\leq\; CV^3\log^3 V\sqrt{n\cdot\log N\left(\frac{1}{n},\Theta\right)}\cdot R\log^4(n).$$

To bound Eq. (4.1b), let $\mathcal{H}\equiv[h_{\min},h_{\max}]$. Then, by Lemma D.21, we have that for some dimension independent constant C_1 that with probability at least $1-C_1e^{-R}$,

$$\sup_{\boldsymbol{\theta}\in\bar{\Theta}}\left|D(\boldsymbol{Z},\boldsymbol{\theta})-\mathbb{E}\left[D(\boldsymbol{Z},\boldsymbol{\theta})\right]\right|\leq C_1\cdot V^2\log^2 V\cdot\sqrt{R\cdot n\log\left(n\cdot N(n^{-3/2},\Theta)\right)}\cdot\frac{\log^4 n}{h_{\min}}.$$

Finally, to bound Eq. (4.1c), use Theorem 3.2 and take the supremum over $h\in\mathcal{H}$ to obtain

$$Eq.\ (4.1c)\;\leq\; C_2 h_{\max} n\log(1/h_{\min}).$$

Substituting these three bounds into Eq. (4.1) and collecting constants proves the theorem. □