Decision-Aware Denoising

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

Michael Huang
Zicklin School of Business, Baruch College, City University of New York, New York, NY 10010, michael.huang@baruch.cuny.edu

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

Modern decision-making in urban planning, climate change, and healthcare leverage large geospatial and panel datasets. These data are often extremely noisy, resulting in low-quality downstream decisions. We propose a new "light touch" framework to adapt techniques from machine learning originally designed for denoising these data by smoothing to instead guide decision-making. The key to our method is a novel estimator of out-of-sample policy performance that we call the *one-shot Variance Gradient Correction* (one-shot VGC). Using the one-shot VGC, we tune the machine learning methods to minimize downstream costs (instead of minimizing prediction error or maximizing signal recovery). We uniformly bound the relative error of the one-shot VGC as an estimate of downstream costs by an intuitive measure of solution stability for the problem plus a term that vanishes as problem dimension grows. Solution stability depends both on the policy and the structure of the downstream optimization problem. We bound the solution stability for three classes of policies and problems – i) regularized plug-in policies for convex feasible regions ii) (unregularized) plug-in policies for strongly-convex feasible regions and iii) (unregularized) affine plug-in policies for weakly-coupled (potentially non-convex) problems. In all cases, we show the solution stability vanishes (relative to out-of-sample cost) as the dimension of the optimization problem grows. Finally, we present a case study based on real traffic accident data from New York City on deploying speed humps to reduce pedestrian injury. Our "light-touch" decision-aware approach outperforms traditional decision-blind techniques and highlights that the optimal level of smoothing for a denoising algorithm should depend on the downstream decision-problem.

Key words: Decision-aware learning. Small-data, large-scale regime. Predict-then-Optimize Framework.

---

1. Introduction

Many decision-making problems involve large-scale systems, such as cities in urban design, road networks in transportation, and ecosystems in wildlife conservation. To analyze and make decisions in these settings, practitioners must use data to construct estimates of their many uncertain parameters. Crucial data are often missing or very imprecise as they must be collected over time through a limited number of sensors, satellites, and people. As a result, estimates of the key model parameters can be noisy leading to poor decision-making.

A canonical solution to reducing the noise of estimates is denoising. Denoising describes a broad variety of approaches that seek to estimate an underlying signal that has been corrupted by noise. Across domains, denoising has been successfully applied to extract the heart's electrical signal from electrocardiograms (Chatterjee et al., 2020), separate speech from background noise (Wilson et al., 2008), study the impact of drugs, cancer progression, and cell development at a molecular level (Kavran and Clauset, 2021), and improve weather forecasts (Kim et al., 2020). Importantly, in each of these applications, the goal is either signal recovery or improved prediction; there is no explicit downstream decision task.

By contrast, this paper proposes a new method for denoising when the estimated signal informs a downstream optimization problem by explicitly leveraging the structure of that problem. Such approaches are sometimes called *decision-aware*. By contrast *estimate-then-optimize* or *decision-blind* procedures denoise the parameters using standard statistical and machine learning tools (without knowledge of the downstream optimization) and then plug-in the denoised estimates into the downstream optimization. Recent works have shown that decision-aware methods often out-perform decision-blind methods, particularly in small-data, large-scale settings (Gupta, Huang, and Rusmevichientong, 2022a) and when models are misspecified (Elmachtoub and Grigas, 2022; Elmachtoub et al., 2023).

More specifically, we study the following optimization problem with linear objective:

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

where $\mathcal{X}$ is a known, potentially non-convex, feasible region, and $\boldsymbol{\mu} \in \mathbb{R}^n$ is an unknown vector of parameters (the signal). We observe a corrupted version of μ_j, i.e., $Z_j = \mu_j + \xi_j$ for each j, and features $\boldsymbol{W}_j \in \mathbb{R}^p$ that are fixed and known.

We consider a class of policies that first tries to denoise the Z_j to recover μ_j (perhaps leveraging the $\boldsymbol{W}_j$ and components Z_k for $k \neq j$) and then plugs in this denoised estimate for $\boldsymbol{\mu}$ in Problem (1) or a regularized variant of Problem (1). Such policies are called *plug-in* and *regularized* plug-in policies, respectively (c.f. Definition 2.2).

Given a class of plug-in or regularized plug-in policies, our goal is to find the policy $\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{W})$ in that class with the best out-of-sample performance $\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{W})$ with respect to the objective in Problem (1).

Since $\boldsymbol{\mu}$ is unknown, even estimating the out-of-sample performance of a given policy $\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{W})$ is not trivial. The naive in-sample estimator $\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{W})$ is biased, i.e.,

$\mathbb{E}\left[\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{W})\right] \neq \mathbb{E}\left[\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{W})\right]$. (See Gupta, Huang, and Rusmevichientong, 2022a for discussion).

Instead, we propose a novel estimator we call the *one-shot Variance Gradient Correction* (one-shot VGC) that debiases the in-sample performance $\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{W})$ to obtain an estimate of $\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{W})$. The one-shot VGC approximately computes the gradient of the objective value with respect to $\boldsymbol{Z} - \boldsymbol{\mu}$, similar to the method of Gupta, Huang, and Rusmevichientong (2022a), which we refer to as the *multi-shot VGC*. We provide a detailed comparison of the one-shot VGC and multi-shot VGC in Sections 1.2, 3.3 and 7.1 below.

We prove uniform bounds on the error of our estimator of out-of-sample performance. The key insight is that this error is bounded (almost surely) by a particular measure of solution stability plus an approximation term that is easy to control (c.f. Theorem 4.3). Thus, to bound the error of our estimator, it is sufficient to study solution stability.

This perspective yields a simple analysis for problems where the plug-in policy is continuous. For example, with regularized plug-in policies and convex $\mathcal{X}$, the plug-in policy is Lipschitz continuous, and we show that the error of the one-shot VGC essentially scales like $\tilde{O}_p(n^{1/4})$ for ℓ_1-strongly-convex regularizers and like $\tilde{O}_p(n^{3/4})$ for ℓ_2-strongly-convex regularizers, uniformly over the class (c.f. Theorem 5.2 and Corollary 5.3). Similarly, for (unregularized) plug-in policies and strongly-convex $\mathcal{X}$ (c.f. Theorem 5.4), the plug-in policy is *almost* Lipschitz continuous, and we prove similar rates of convergence on the uniform error (c.f. Theorem 5.6). In typical applications, the full information optimum of Problem (1) is $\Theta(n)$, and hence, these results show that the relative error of the one-shot VGC as an estimator of out-of-sample performance vanishes as $n \to \infty$, uniformly over the policy class. This further implies that choosing a policy by optimizing our debiased estimate incurs vanishing relative regret as $n \to \infty$. The proofs of these results are short and straightforward highlighting the pivotal role of our new notion of solution stability.

The case of non-convex $\mathcal{X}$ and potentially unregularized plug-in policies is more delicate because plug-ins are typically discontinuous. Solution stability in this case depends strongly on the interplay between the plug-in function and the constraints of $\mathcal{X}$. We restrict attention to affine plug-in functions and feasible regions $\mathcal{X}$ that are "weakly-coupled" in the sense of Gupta, Huang, and Rusmevichientong (2022a). As we argue in Section 2.3, many machine learning and denoising methods can be rewritten as linear smoothers or linear filters (Buja, Hastie, and Tibshirani, 1989; Wahba, 1990) and, hence, give rise to affine plug-in policies. In this sense, the restriction to affine plug-ins is arguably mild. Moreover, linear optimization problems with wide constraint matrices and some two-stage stochastic programs are weakly-coupled, so this class subsumes many applications of interest.

We tailor our one-shot VGC correction to this setting. We prove that the error of our estimator essentially scales like $\tilde{O}_p(n^{3/4})$ (c.f. Theorem 6.11 and discussion surrounding Corollary 6.7), so that in the typical case, the relative error of our debiased estimator of out-of-sample performance vanishes as $n \to \infty$. Again, this implies that choosing a policy by optimizing our debiased estimate incurs vanishing relative regret as $n \to \infty$.

The key to our analysis in this setting is a novel proof technique that breaks the dependence structure in weakly-coupled problems by considering a suitably lifted problem. In the lifted space, we can characterize dependence between components of the plug-in policy by the chromatic number of a particular graph constructed from $\mathcal{X}$ and the plug-in function. This analysis allows us to bypass many of the involved duality arguments and approximate strong-convexity arguments used in Gupta, Huang, and Rusmevichientong (2022a) to analyze weakly-coupled programs, and may have independent interest.

Finally, in Section 7.2, we perform a numerical case study on prioritizing requests for speed humps using motor vehicle accident data from New York City. Our decision-aware denoising approach outperforms decision-blind denoising in identifying locations with the most pedestrian injuries caused by motor vehicle accidents. The improvement arises because decision-blind methods "oversmooth" in this example – when restricting to regions with higher pedestrian injuries, less smoothing is needed than when considering all of New York, and our decision-aware method correctly exploits this feature.

1.1. Our Contributions

1. We propose the one-shot Variance Gradient Correction (one-shot VGC) for debiasing in-sample policy performance to estimate out-of-sample policy performance. The one-shot VGC applies to any plug-in or regularized plug-in policy for Problem (1).
2. Under common assumptions on the data-generating process, we bound the variance (Theorem 3.2) and tail behavior (Theorem 3.4) of the one-shot VGC. Furthermore we show the expected error of the one-shot VGC in estimating in-sample bias is bounded by a specific measure of solution stability that depends on both the policy and structure of Problem (1) (Theorem 4.3).

By further bounding this solution stability, we prove that:

3. When Problem (1) is convex, the plug-in function is Lipschitz, and the policy is regularized, we can uniformly bound the estimation error of the one-shot VGC (Theorem 5.2) in terms of properties of the regularizer. In particular, this bound implies the relative error of our procedure vanishes, and one can learn a best-in-class policy as the problem size grows in typical settings.
4. When $\mathcal{X}$ is strongly-convex, the plug-in function is Lipschitz, and the policy is unregularized, we can uniformly bound the estimation error (Theorem 5.6) in terms of properties of the feasible region and the magnitude of the plug-in function. This bound also implies the relative error of our procedure vanishes, and one can learn a best-in-class policy as the problem size grows.
5. When Problem (1) is potentially non-convex but weakly coupled, the plug-in function is affine, and the policy is unregularized, we uniformly bound the estimation error of the one-shot VGC in terms of the chromatic number of a particular graph depending on both $\mathcal{X}$ and the affine policy (Theorem 6.11). We show that even with crude upper bounds on this chromatic number, the relative error of our procedure vanishes, and it is possible to learn a best-in-class policy as the problem size grows in typical settings.
6. We present a case study on deploying speed humps to reduce pedestrian injuries in motor vehicle accidents using data from New York City. We compare our decision-aware denoising approach to a decision-blind estimate-then-optimize approach. We find that the ideal amount of smoothing depends on the downstream loss, and, hence, decision-blind methods cannot always perform optimally. By contrast, our decision-aware approach finds a policy with performance comparable to oracle performance.

1.2. Relationship to Prior Work

Denoising has a long history in both signal processing and statistics. Our review is necessarily incomplete. Methods in signal processing include using Fourier analysis to discard high-frequency noise (Rabiner and Gold, 1975), solving ℓ_1-penalized optimization problems to deblur edges in images (Candes, Romberg, and Tao, 2006), and, more recently, leveraging neural networks and autoencoders to enhance blurry video (Tian et al., 2020). In statistics, denoising is often cast as a type of "smoothing" (Buja, Hastie, and Tibshirani, 1989; Seeger, 2004; Wahba, 1990). Such smoothing methods are particularly common when analyzing geospatial data (Chiles and Delfiner, 2012). Both streams of literature focus on the quality of signal recovery.

By contrast, our work focuses on using denoising to improve decision-making in a downstream optimization problem. Unlike the two previous streams, we are not concerned with recovering the original signal or the error of our denoised estimate from the original signal.

In this sense, our work contributes to the growing literature of "decision-aware learning," also called "operational data analysis" (Liyanage and Shanthikumar, 2005; Grigas, Qi, et al., 2021; Elmachtoub and Grigas, 2022; Wilder, Dilkina, and Tambe, 2019; Hu, Kallus, and Mao, 2022; Feng and Shanthikumar, 2023; Chu et al., 2023). Most research in decision-aware learning studies algorithms in the traditional large-sample limit where one has increasing amounts of data and, hence, increasingly precise estimates of key parameters. In our notation, this would correspond to the limit where $Z_j \to \mu_j$ for each j and denoising is not needed.

We adopt a different perspective. As mentioned previously, many interesting applications of denoising exhibit poor data quality, either due to imprecise sensors or missing covariates. Improving sensor quality or collecting additional covariates can be costly, time-consuming, or may raise privacy concerns. These features have recently spurred the development of methods tailored to "small-data" settings (Besbes and Mouchtaki, 2023; Besbes, Ma, and Mouchtaki, 2022). At the same time, many applications of interest are large-scale – i.e. n is large. Thus, we focus on a small-data, large-scale optimization regime where the precision of each Z_j is bounded, and n is large. This regimes exhibits fundamentally different asymptotics than the large-sample regime (Gupta and Rusmevichientong, 2021), distinguishing our work from most prior work on decision-aware learning.

Our one-shot VGC is based on a perturbation argument and Danskin's Theorem. Similar perturbation arguments are used in Ito, Yabe, and Fujimaki (2018) and Guo, Jordan, and Zhou (2022) in very different contexts (parameter estimation and causal inference, respectively). Both works offer asymptotic analysis in the large-sample limit. By contrast, we prove explicit (finite n) bounds in a small-data, large-scale regime.

Finally, our work builds on the results of Gupta, Huang, and Rusmevichientong (2022a). Both works seek to learn plug-in policies for optimization problems with uncertain linear objective in a small-data, large-scale data setting. Gupta, Huang, and Rusmevichientong (2022a) proposes (what we call) the *multi-shot VGC* to construct an estimator of out-of-sample performance and proves some performance guarantees for weakly-coupled optimization problems. The emphasis of the work is to advocate for such debiasing over cross-validation in small-data, large-scale regimes.

Our work improves upon and complements Gupta, Huang, and Rusmevichientong (2022a) in several ways: First, our analysis of the one-shot VGC applies more generally than that of the multi-shot VGC; it applies to any regularized plug-in policy with a Lipschitz plug-in function when $\mathcal{X}$ is convex, (unregularized) plug-in policies with Lipschitz plug-in functions when $\mathcal{X}$ is strongly-convex, and to general affine plug-in policies when $\mathcal{X}$ is weakly-coupled. By contrast, the analysis in Gupta, Huang, and Rusmevichientong (2022a) applies only to unregularized, ***separable*** affine plug-in functions (c.f. Remark 2.3) and weakly-coupled $\mathcal{X}$. Crucially, most denoising methods do *not* give rise to separable plug-ins. Indeed, the key intution in denoising is to combine information from "nearby" observations to smooth out corruptions. This deficiency was precisely what motivated our study of this problem.

Second, preliminary experiments suggest our one-shot VGC outperforms the multi-shot variant (c.f. Section 7.1). Section 3.3 offers some intuition for this performance difference. Essentially, for separable affine plug-ins, both approaches appear (approximately) unbiased, but our one-shot VGC exhibits less variance. For non-separable affine plug-ins, the multi-shot VGC has both larger variance and a non-vanishing bias.

Third, from a more theoretical perspective, our work introduces two technical tools that facilitate our simpler and more general proofs. First, we use a novel notion of solution stability (c.f. Theorem 4.1) to bound our estimation error with high-probability. By comparison, Gupta, Huang, and Rusmevichientong (2022a) introduced a different notion of stability – average solution stability – in their work. They show small average solution stability is a sufficient condition to bound the variance of their estimator for a fixed policy. However, such pointwise variance bounds are not sufficient for learning, and hence their performance guarantees do not explicitly leverage average solution stability. By contrast, we prove our bounds by directly analyzing our novel solution stability concept leading to more streamlined and intuitive proofs. The second technical tool is our aforementioned "lifting" technique that allows us to bypass the difficult duality and approximate strong-convexity argument of Gupta, Huang, and Rusmevichientong (2022a) used to analyze weakly-coupled problems. Both ideas may be useful in other settings.

2. Model

Let $\boldsymbol{\mu} \in \mathbb{R}^n$ be a fixed but unknown signal. We observe corrupted estimates $\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{\xi}$ where $\boldsymbol{\xi} \in \mathbb{R}^n$ is a mean-zero, random vector with independent components, and ξ_j has (known) precision $\nu_j \geq 0$. (Recall the variance of ξ_j is $1/\nu_j$.) For convenience, let $\nu_{\min} \equiv \min_j \nu_j$. We observe fixed features $\boldsymbol{W}_j \in \mathbb{R}^p$ that potentially contain information relating components of $\boldsymbol{\mu}$, and let $\boldsymbol{W} \in \mathbb{R}^{n \times p}$ be the matrix with j^{th} row $\boldsymbol{W}_j$ and $\boldsymbol{\Sigma}$ be the covariance of $\boldsymbol{\xi}$.

In traditional signal-processing applications, e.g., with geospatial data, $\boldsymbol{Z}$ represents sensor readings at n locations. The random corruptions $\boldsymbol{\xi}$ represent irreducible measurement error, and the features $\boldsymbol{W}_j$ encode information about the j^{th} location (latitude, longitude, elevation, terrain type, rainfall level, etc.). Often, such applications entail many locations (large n), and improving sensor quality is prohibitively expensive (low precisions ν_j). Such applications are perhaps best studied in the so-called *small-data, large-scale regime*, (Gupta and Rusmevichientong, 2021; Gupta, Huang, and Rusmevichientong, 2022a; Gupta and Kallus, 2022), and our analysis takes this perspective.

Our model can also approximate some machine learning applications with panel data (Ignatiadis and Wager, 2019; Chung et al., 2022; Gupta and Kallus, 2023). Here, j refers to individual "units" such as patients, facilities, or regions, with $\boldsymbol{W}_j$ encoding the corresponding unit information. The "observations" $\boldsymbol{Z}$ represent predictions from an upstream, black-box machine learning model with known precisions $\boldsymbol{\nu}$. Because of privacy constraints, missing data and a host of other issues, precisions can be low (see above references for examples). Moreover, typical applications often focus on a specific set of existing units — e.g., a certain set of existing facilities — rather than assessing the performance of a policy on a hypothetical "out-of-sample" unit.

In either interpretation, we might believe two components j and k with "similar" features $\boldsymbol{W}_j$ and $\boldsymbol{W}_k$ have similar underlying μ_j and μ_k. Thus, we expect the observation Z_k to be informative of the underlying signal μ_j.

To highlight an application of Problem (1) with such data, consider the following:

Example 2.1 (Speed Hump Planning in NYC) Certain studies such as Elvik et al. (2009) have shown speed humps can reduce the number of crashes by 50%–60%. They are highly requested traffic calming measures used to reduce crashes, injuries, and fatalities. In New York City, residents can request speed humps; however, Kuntzman (2022) recently found that the wait time for such requests can be as long as 12 *years*. Speed studies that assess the feasibility of deploying a speed hump at a requested location are a salient bottleneck, taking an average two years to complete.

To prioritize speed hump requests, consider the following linear optimization problem that chooses speed humps that would, if built, reduce the most pedestrian injuries:

$$\min_{\boldsymbol{x}} \quad \sum_{k=1}^{K} \mu_k x_k, \qquad \text{s.t.} \quad \sum_{k=1}^{K} x_k \le B, \quad x_k \in \{0,1\}\ \forall k = 1,\dots,K. \tag{2}$$

Here $\{1,2,\dots,K\}$ denotes the set of requests for speed humps, and B is the number of speed humps that can be processed in a year. The binary variable x_k encodes if request k will be processed. We assume that the potential *relative* reduction of pedestrian injuries is the same for all requests. The cost μ_k is the expected number of pedestrian injuries per year at location k. In cities like New York City, estimates Z_k of μ_k can be formed from traffic accident reports, but are often noisy. Features $\boldsymbol{W}_k$ capture information such as longitude and latitude of the accident, street width, speed limit, and number of bike lanes.

As mentioned, many denoising methods share information across similar locations to estimate $\boldsymbol{\mu}$. We next introduce two policy classes that exploit this intuition.

2.1. Policy Classes

We study a class of (potentially regularized) plug-in policies. Let $\boldsymbol{T}: \mathbb{R}^n \times \bar{\Theta} \mapsto \mathbb{R}^n$ be a transformation of the data depending on a user-defined parameter $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$, and let $\phi: [0,1] \mapsto \mathbb{R}_+$ be a strongly convex regularizer.

Definition 2.2 (Regularized Plug-in Policy) *A regularized plug-in policy* $\boldsymbol{x}\left(\boldsymbol{Z},(\rho,\bar{\boldsymbol{\theta}})\right)$ *for Problem (1) takes the form*[1]

$$\boldsymbol{x}(\boldsymbol{Z},(\rho,\bar{\boldsymbol{\theta}})) \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}} \sum_{j=1}^{n} \left(T_j(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) x_j + \rho\phi(x_j) \right). \tag{3}$$

The policy is defined by $(\rho,\bar{\boldsymbol{\theta}})$ – i.e., T and ϕ are fixed – hence we sometimes write $\boldsymbol{\theta} = (\rho,\bar{\boldsymbol{\theta}})$ for brevity. Given $\bar{\Theta}$ and an interval $[\rho_{\min},\rho_{\max}]$, let $\Theta = \mathbb{R}_+ \times \bar{\Theta}$ for brevity, and define the policy class $\mathcal{X}_\Theta(\boldsymbol{Z}) \equiv \left\{\boldsymbol{x}(\boldsymbol{Z},(\rho,\bar{\boldsymbol{\theta}})) : (\rho,\bar{\boldsymbol{\theta}}) \in \Theta\right\}$. We assume throughout that Θ is compact.

[1] When Eq. (3) admits multiple optima, we tie-break via some deterministic rule, e.g., taking the lexicographically smallest solution.

Intuitively, regularized plug-in policies proxy the unknown $\boldsymbol{\mu}$ by a transformation $\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$ of the data $\boldsymbol{Z}$, and plug this proxy into a version of Problem (1) with a regularizer. The use of the regularizer $\phi(\cdot)$ in this proxy problem is common (see, e.g. Grigas, Qi, et al. (2021) and Wilder, Dilkina, and Tambe (2019)), because when $\mathcal{X}$ is convex it ensures policies are (almost everywhere) differentiable with respect to $\bar{\boldsymbol{\theta}}$. However, in settings where $\mathcal{X}$ is non-convex – e.g., when Problem (1) encodes a minimum spanning tree problem – we lose this differentiability, and the value of regularization is unclear. Hence, in what follows, we treat i) regularized policies, i.e., $\rho > 0$ and ii) unregularized policies, i.e., $\rho = 0$ separately in Sections 5 and 6, respectively.

We focus on regularized plug-in policies because they are "light touch" in the sense of Chung et al. (2022), i.e., they are only marginally more complex to implement than an estimate-then-optmize procedure. Specifically, for a fixed $\bar{\boldsymbol{\theta}}$, one need not alter the measurement or prediction procedure used to obtain $\boldsymbol{Z}$. The transformation $\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$ is applied as a post-processing step. Unregularized plug-in policies are also "light touch" with respect to the optimization. If one has a specialized algorithm to solve Problem (1) efficiently, one can apply this algorithm directly to the transformed vector $\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$.

Remark 2.3 (Relation to Separable, Affine Plug-in Policies) Regularized plug-in policies generalize the *separable* affine plug-in policy class studied in Gupta, Huang, and Rusmevichientong (2022a). Namely, separable affine plug-in policies take $\rho = 0$ and $T_j(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) = a_j(\bar{\boldsymbol{\theta}}) Z_j + b_j(\bar{\boldsymbol{\theta}})$, where the functions a_j and b_j may depend on $\boldsymbol{W}$. We focus on general plug-ins as they are strictly necessary to exploit our earlier intuition that Z_k may be informative for μ_j when $\boldsymbol{W}_j$ and $\boldsymbol{W}_k$ are similar. Because $T_j(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$ may depend on $\boldsymbol{Z}_k$ for $k \neq j$, our analysis of general, regularized plug-in policies below must use substantively different techniques than those in Gupta, Huang, and Rusmevichientong (2022a).

2.2. A Decision-Aware Oracle Benchmark

Unlike previous works on denoising, we do *not* assume that $\boldsymbol{\mu} = T(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}^*)$ for some value of $\bar{\boldsymbol{\theta}}^*$. Instead, we take a best-in-class perspective. We define the oracle parameter

$$\boldsymbol{\theta}^{OR} \in \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \Theta} \ \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}), \tag{4}$$

and corresponding oracle policy $\boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}^{OR})$. This oracle is "best-in-class" over Θ, but not easily computable because $\boldsymbol{\mu}$ is unknown.

In what follows, we consider two tasks:

1. (Policy Evaluation) Given a fixed $\boldsymbol{\theta}$, provide an estimate of $\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{T}(\boldsymbol{Z},\boldsymbol{\theta}))$. Our strategy will be to debias the observed performance $\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})$.
2. (Policy Learning) Find a parameter $\hat{\boldsymbol{\theta}}(\boldsymbol{Z})$ (depending on the data) such that $\boldsymbol{\mu}^\top(\boldsymbol{x}(\boldsymbol{Z},\hat{\boldsymbol{\theta}}(\boldsymbol{Z}) - \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}^{OR})))$ is "small." Our strategy will be to show that our debiased estimators from task 1 are accurate uniformly over Θ, and hence optimizing the estimate yields a policy with performance comparable to oracle performance.

2.3. Examples of Denoising Plug-Ins

Before delving into the details of our approach, we describe popular denoising plug-ins inspired by modern machine learning methods. The following examples are all *affine* transformations of $\boldsymbol{Z}$. Affine structure is not necessary for the majority of our paper, but is necessary for our results in Section 6.

Example 2.4 (Kernel Smoothers) Kernel smoothers define similarity via a kernel function $K_{\bar{\boldsymbol{\theta}}}(\boldsymbol{W}_j, \boldsymbol{W}_k)$ with parameters $\bar{\boldsymbol{\theta}}$. For example, the box-kernel is defined by $K_{\bar{\theta}}(\boldsymbol{W}_j, \boldsymbol{W}_k) = \mathbb{I}\left\{\|\boldsymbol{W}_j - \boldsymbol{W}_k\| \le \bar{\theta}\right\}$. Inspired by kernel smoothing, we consider a plug-in policy using Nadaraya-Watson regression as our transformation, i.e.,

$$T_j^{KR}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) = \sum_{i=1}^n \frac{K_{\bar{\boldsymbol{\theta}}}(\boldsymbol{W}_j, \boldsymbol{W}_i)}{\sum_{l=1}^n K_{\bar{\boldsymbol{\theta}}}(\boldsymbol{W}_j, \boldsymbol{W}_l)} Z_i. \tag{5}$$

This expression is affine in $\boldsymbol{Z}$. By varying $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$, we obtain a class of plug-in-policies each with a different amount of smoothing.

Example 2.5 (Local Regression) Local linear regression uses a kernel to upweight similar data points when predicting for a given test point. To estimate μ_j, it solves

$$\boldsymbol{\beta}_j^{\mathrm{LR}}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) \in \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n K_{\bar{\boldsymbol{\theta}}}(\boldsymbol{W}_j, \boldsymbol{W}_i)\,(\boldsymbol{W}_j \boldsymbol{\beta} - Z_j)^2,$$

and outputs $T_j^{LR}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) = \boldsymbol{W}_j^\top \boldsymbol{\beta}_j^{\mathrm{LR}}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})$. Defining the Graham matrix $\boldsymbol{K} = (K_{\bar{\boldsymbol{\theta}}}(\boldsymbol{W}_j, \boldsymbol{W}_k))_{ij} \in \mathbb{R}^{n\times n}$, we can rewrite this estimator as

$$T_j^{LR}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) = \boldsymbol{W}_j^\top \left(\boldsymbol{W}^\top \mathrm{diag}(\boldsymbol{K}_j)\boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \mathrm{diag}(\boldsymbol{K}_j)\boldsymbol{Z},$$

which is again affine in $\boldsymbol{Z}$. By varying $\bar{\boldsymbol{\theta}}$ (and thus the Graham Matrix), we obtain a class of plug-in policies based on local linear regression with different sized neighborhoods. Polynomial local regression methods can also be used in a similar way.

Example 2.6 (Ridge Regression and Random Features) In high dimensional settings when $p \gg n$, regularized regression is often use to improve stability. For example, ridge regression solves

$$\boldsymbol{\beta}^{\mathrm{RR}}(\boldsymbol{Z},\bar{\theta}) \in \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{W}\boldsymbol{\beta} - \boldsymbol{Z}\|_2^2 + \bar{\theta}\|\boldsymbol{\beta}\|_2, \tag{6}$$

and then proxies $\boldsymbol{\mu}$ by

$$\boldsymbol{T}^{RR}(\boldsymbol{Z},\bar{\theta}) = \boldsymbol{W}\boldsymbol{\beta}^{\mathrm{RR}}(\boldsymbol{Z},\bar{\theta}) = \boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{W} + \bar{\theta}\boldsymbol{I}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{Z},$$

which is affine in $\boldsymbol{Z}$. By varying $\bar{\theta}$, we can construct a class of plug-in policies corresponding to different degrees of regularization.

If we first apply a non-linear transformation to each feature $\boldsymbol{W}_j \mapsto \psi(\boldsymbol{W}_j) \in \mathbb{R}^{\bar{p}}$ and then apply ridge regression, the resulting transformation is still in affine in $\boldsymbol{Z}$. Thus, plug-in policies based on kernelized ridge regression also fall within our scope. Recently, randomized feature maps (Gallicchio and Scardapane, 2020; Rahimi and Recht, 2008) have been used to approximate kernelized ridge regression. Such methods also give rise to affine policies.

Example 2.7 (Clustering) Consider a clustering method indexed by a hyperparameter $\bar{\boldsymbol{\theta}}$. (For example, take K-means clustering indexed by the number of clusters K.) Applied to the features $\boldsymbol{W}_j$, $j = 1,\ldots,n$, clustering yields a partition $\mathcal{S}(\bar{\boldsymbol{\theta}}) = \{C_1,\ldots,C_K\}$ of the set $\{1,\ldots,n\}$. We estimate μ_j by the average of its cluster:

$$T_j^{CL}\left(\boldsymbol{Z},\bar{\boldsymbol{\theta}}\right) = \sum_{k=1}^{K} \mathbb{I}\left\{j \in C_k\right\}\left(\frac{1}{|C_k|}\sum_{i \in C_k} Z_i\right).$$

This transformation is affine in $\boldsymbol{Z}$. By varying $\bar{\boldsymbol{\theta}}$ (e.g. varying the number of clusters), we again obtain a class of plug-in policies with different local neighborhoods.

The above list is not exhaustive. Many transformations $\boldsymbol{T}(\cdot)$ can be cast as linear smoothers (see Buja, Hastie, and Tibshirani (1989)). Such predictors yield affine plug-in policies, and are often indexed by a low dimensional parameter $\bar{\boldsymbol{\theta}}$.

Of course, not all policies are affine plug-in policies. For example, when using a deep neural network to predict μ_j, the resulting predictions are generally not affine in $\boldsymbol{Z}$. Outside of Section 6, our results will not require affine structure. However, we stress that non-affine plug-ins can sometimes be approximated by affine plug-ins. For example, Jacot, Gabriel, and Hongler (2018) show that kernel regression using neural tangent kernels approximates infinitely wide fully-connected neural networks, and Arora et al. (2020) provide evidence that this approximation has comparably strong performance in small-data tasks.

3. One-Shot Variance Gradient Correction for Policy Evaluation

To denoise the in-sample performance $\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})$, we must essentially estimate the in-sample optimism since

$$\underbrace{\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})}_{\text{Out-of-Sample Performance}} = \underbrace{\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})}_{\text{In-Sample Performance}} - \underbrace{\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})}_{\text{In-Sample Optimism}} .$$

Loosely, our approach to estimating $\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})$ is as follows: Define

$$\hat{\boldsymbol{x}}(\boldsymbol{t},\rho) \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{t}^\top \boldsymbol{x} + \rho \sum_{j=1}^n \phi(x_j) \quad \text{and} \quad V(\boldsymbol{t},\rho) \equiv \boldsymbol{t}^\top \hat{\boldsymbol{x}}(\boldsymbol{t},\rho) + \rho \sum_{j=1}^n \phi(\hat{x}_j(\boldsymbol{t},\rho)).$$

Fix some small user-defined constant $h > 0$ and random vector $\boldsymbol{\delta}_h \in \mathbb{R}^n$. (We discuss how this choice affects estimation quality in Theorem 4.3 below.) Finally, compute the *one-shot VGC Correction*:

$$D(\boldsymbol{Z},\boldsymbol{\theta}) = D\left(\boldsymbol{Z},(\rho,\bar{\boldsymbol{\theta}})\right) \equiv \frac{1}{h}\mathbb{E}\left[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) + \boldsymbol{\delta}_h,\rho) - V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho) \,\middle|\, \boldsymbol{Z}\right]. \tag{7}$$

Computationally, the expectation can be approximated by simulating $\boldsymbol{\delta}_h$ and averaging.

The one-shot VGC $D(\boldsymbol{Z},\boldsymbol{\theta})$ is our estimate of the in-sample optimism $\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})$. Hence our estimate of out-of-sample performance is $\boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) \approx \boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - D(\boldsymbol{Z},\boldsymbol{\theta})$.

3.1. A Heuristic Derivation

To provide some intuition, we outline a heuristic derivation of our correction similar to that in Gupta, Huang, and Rusmevichientong (2022a). We consider a fixed policy and drop $\rho, \bar{\boldsymbol{\theta}}$ from the notation in this derivation.

Consider the function

$$\lambda \mapsto V(\boldsymbol{T}(\boldsymbol{Z}) + \lambda\boldsymbol{\xi}), \tag{8}$$

where $\boldsymbol{\xi} \equiv \boldsymbol{Z} - \boldsymbol{\mu}$. By Danskin's theorem, whenever $\hat{\boldsymbol{x}}(r(\boldsymbol{Z}) + \lambda\boldsymbol{\xi})$ is the unique optimizer, the derivative of this function is

$$\frac{\partial}{\partial\lambda} V(\boldsymbol{T}(\boldsymbol{Z}) + \lambda\boldsymbol{\xi})\bigg|_{\lambda=0} = \boldsymbol{\xi}^\top \hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z})).$$

(When $\hat{\boldsymbol{x}}(T(\boldsymbol{Z}) + \lambda\boldsymbol{\xi})$ is not the unique optimizer, then $\boldsymbol{\xi}^\top \hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}))$ is a subgradient.) Hence, instead of estimating the rightside, we attempt to estimate the leftside.

Since we cannot evaluate the derivative directly, we approximate it with a first-order, forward finite-difference approximation (LeVeque, 2007, Chapter 1). For sufficiently small $h > 0$, we expect that

$$\frac{\partial}{\partial \lambda} V(\boldsymbol{T}(\boldsymbol{Z}) + \lambda \boldsymbol{\xi})\bigg|_{\lambda=0} \;=\; \frac{1}{h}\Big(V(\boldsymbol{T}(\boldsymbol{Z}) + h\boldsymbol{\xi}) - V(\boldsymbol{T}(\boldsymbol{Z}))\Big) \;+\; o_p(h) \quad \text{as } h \to 0. \tag{9}$$

Finally, since $\boldsymbol{\xi}$ is unknown, we approximate $\boldsymbol{T}(\boldsymbol{Z}) + h\boldsymbol{\xi}$ by $\boldsymbol{T}(\boldsymbol{Z}) + \boldsymbol{\delta}_h$. The resulting estimator $\frac{1}{h}\left(V(\boldsymbol{T}(\boldsymbol{Z}) + \boldsymbol{\delta}_h) - V(\boldsymbol{T}(\boldsymbol{Z}))\right)$ depends on the random variable $\boldsymbol{\delta}_h$. To reduce variance, we derandomize our estimator by taking a conditional expectation over $\boldsymbol{\delta}_h$ (Rao-Blackwellization).

3.2. Properties of the One-Shot VGC

Under minimal assumptions, we can bound the variance of the one-shot VGC $D(\boldsymbol{Z}, \boldsymbol{\theta})$.

Assumption 3.1 (Lipschitz Plug-in) *The function $\boldsymbol{Z} \mapsto \boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$ is $C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}})$-Lipschitz with respect to the ℓ_1 norm, i.e., $\|\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) - \boldsymbol{T}(\boldsymbol{Y}, \bar{\boldsymbol{\theta}})\|_1 \le C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}}) \|\boldsymbol{Z} - \boldsymbol{Y}\|_1$.*

Define $C_{\boldsymbol{T}} = \max_{\bar{\boldsymbol{\theta}} \in \bar{\Theta}} C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}})$.

When $\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) = \boldsymbol{L}(\bar{\boldsymbol{\theta}})\boldsymbol{Z} + \boldsymbol{l}(\bar{\boldsymbol{\theta}})$ is affine, $C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}}) = \max_j \|\boldsymbol{L}_j(\bar{\boldsymbol{\theta}})\|_1$, i.e., the maximal ℓ_1-norm of the columns. For many linear smoothers presented in Section 2.3, the matrix $\boldsymbol{L}(\bar{\boldsymbol{\theta}})$ has non-negative entries with columns that sum to 1. Hence, $C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}}) = 1$ in these examples.

We apply the Efron-Stein Inequality to bound the variance of the one-shot VGC:

Theorem 3.2 (Variance of the VGC) *Under Assumption 3.1, the variance of the one-shot VGC satisfies*

$$\mathbb{V}ar(D(\boldsymbol{\mu} + \boldsymbol{\xi}, \boldsymbol{\theta})) \;\le\; \frac{4C_{\boldsymbol{T}}^2(\bar{\boldsymbol{\theta}})n}{h^2 \nu_{\min}}.$$

This bound on the variance suggests that if $C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}})$ is $O_p(1)$, the stochastic fluctuations of the one-shot VGC are $O_p(n^{-1/2}/h)$. Thus, in typical cases where the full information solution to Problem (1) is $O(n)$, the stochastic contributions are asymptotically negligible if $h\sqrt{n} \to \infty$ as $n \to \infty$.

We can strengthen this result to a high probability tail bound assuming $\boldsymbol{\xi}$ has light tails:

Assumption 3.3 (Independent Sub-Gaussian Corruptions) *For all $j = 1, \ldots, n$, ξ_j is sub-Gaussian with variance proxy at most κ^2.*

Theorem 3.4 (Concentration of the VGC) *Suppose Assumptions 3.1 and 3.3 hold. Then, with probability at least* $1-\epsilon$,

$$|D(\boldsymbol{Z},\boldsymbol{\theta}) - \mathbb{E}[D(\boldsymbol{Z},\boldsymbol{\theta})]| \ \le\ C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}})\frac{\kappa}{h}\sqrt{n\log\left(\frac{2}{\epsilon}\right)}.$$

3.3. Comparison to Multi-Shot VGC

In Gupta, Huang, and Rusmevichientong (2022a), a similar approach is used to motivate what we refer to as the *multi-shot VGC*. Providing a rigorous, theoretical comparison of the methods is difficult because the multi-shot VGC only applies to *separable* affine plug-in policies. To develop some intuition, we compare the two estimators in the special case that $T(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) = \boldsymbol{Z} + \ell(\bar{\boldsymbol{\theta}})$, $\rho = 0$, and ξ_j are independent standard normals for all j.

Under these simplifying assumptions, the multi-shot VGC with a stepsize of $\bar{h}$ is

$$D^{MS}(\boldsymbol{Z}) = \sum_{j=1}^{n}\frac{1}{\bar{h}}\mathbb{E}\left[(V(\boldsymbol{Z}+\delta_j\boldsymbol{e}_j) - V(\boldsymbol{Z}))\mid \boldsymbol{Z}\right], \text{ where } \delta_j \sim \mathcal{N}(0,\bar{h}^2+2\bar{h}),\ j=1,\ldots,n. \tag{10}$$

Define the random variable $\bar{\boldsymbol{\delta}} \in \mathbb{R}^n$ by $\bar{\boldsymbol{\delta}} \sim \delta_j\boldsymbol{e}_j$ with probability $\frac{1}{n}$, and let $h = \bar{h}/n$. Then,

$$D^{MS}(\boldsymbol{Z}) = \frac{1}{h}\mathbb{E}\left[V(\boldsymbol{Z}+\bar{\boldsymbol{\delta}}) - V(\boldsymbol{Z})\mid \boldsymbol{Z}\right].$$

Thus, multishot VGC is a special case of the one-shot VGC with a specific random step.

As we will show in Theorem 4.3 and Proposition 4.4, when $\boldsymbol{\xi}_j$ are independent, standard normals, the random variable $\boldsymbol{\delta}_h$ which minimizes estimation error in our one shot VGC is $\boldsymbol{\delta}_h \sim \mathcal{N}(\boldsymbol{0},(h^2+2h)\boldsymbol{I})$. Hence the relevant comparison is to

$$D(\boldsymbol{Z}) = \frac{1}{h}\mathbb{E}\left[V(\boldsymbol{Z}+\boldsymbol{\delta}_h) - V(\boldsymbol{Z})\mid \boldsymbol{Z}\right]. \tag{11}$$

When $h < 1/n$, we see that both methods take a step size of length $O_p(\sqrt{nh})$. When $h > 1/n$, multi-shot VGC takes a step of size $O_p(nh)$, while the one-shot VGC takes a smaller step of size $O_p(\sqrt{nh})$. In both cases, the multishot VGC only steps in coordinate directions, while the one-shot VGC steps in a uniformly random direction. Because of this difference in step-size and direction, the two corrections are different in general. However, in the special case that the optimization problem fully decouples, i.e, $\hat{x}_j(\boldsymbol{t},0)$ only depends on t_j for all j, the two corrections are essentially equivalent. (See Proposition B.3 in the appendix for a formal statement.)

From a theoretical point of view, the different structure of one-shot and multi-shot VGC give rise to different analyses. Specifically, because it restricts to coordinate directions, bounding the bias of the multishot VGC is straightforward (Gupta, Huang, and Rusmevichientong, 2022a, Theorem 3.2), but analyzing its stochasic behavior is quite difficult and involves a subtle duality argument and a number of assumptions on the policy and optimization. See Gupta, Huang, and Rusmevichientong (2022a, Theorem 3.5). By contrast, as we have shown in Theorems 3.2 and 3.4, by fully randomizing the step direction, the one-shot VGC admits a very simple analysis of its stochastic deviation around its mean, but analyzing its bias is now more difficult and considered in the next section.

From an empirical point of view, we study the multi-shot and one-shot VGC numerically in Section 7.1 and find that the larger steps of the multi-shot VGC can cause it to have larger variance and estimation error than the one-shot VGC. Moreover, when the optimization problem does not decouple, the bias of the multi-shot VGC grows with the degree of coupling. Our one-shot variant thus performs much better.

4. Bounding the Estimation Error by Solution Stability

In this section, we bound the estimation error of our method by a more manageable term related to the stability of our policies:

Definition 4.1 *For any $\boldsymbol{\theta} \in \Theta$ and $h > 0$, we define the* solution stability *of a policy by*

$$\mathsf{SS}(\boldsymbol{\xi}, h, \boldsymbol{\theta}) \ = \ \left|\boldsymbol{\xi}^\top \left(\hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}), \rho) - \hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + h\boldsymbol{\xi}, \rho)\right)\right|.$$

Intuitively, solution-stability measures how much the solution changes given a small perturbation of the data $\boldsymbol{Z}$ *in the direction of the noise* $\boldsymbol{\xi}$. In particular, solution-stability is random, and not directly computable (because it depends on $\boldsymbol{\xi}$). Below, we first show that the error of our estimator is bounded by solution stability (plus a term that is easy to control) and then discuss how to further bound solution-stability.

To that end, write

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta} &\underbrace{\left|\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) - D(\boldsymbol{Z}, \boldsymbol{\theta}) - \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}))\right|}_{\text{Estimation Error}} \\ &\le \sup_{\boldsymbol{\theta}\in\Theta} \underbrace{\left|\boldsymbol{\xi}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) - \frac{1}{h}\left(V(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + h\boldsymbol{\xi}, \rho) - V(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}), \rho)\right)\right|}_{\text{Finite Difference Error}} \qquad \text{(12a)}\end{aligned}$$

$$+\sup_{\boldsymbol{\theta}\in\Theta}\underbrace{\left|\frac{1}{h}\left(V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho)\right)-D(\boldsymbol{Z},\boldsymbol{\theta})\right|}_{\text{Replication Error}}. \tag{12b}$$

Intuitively, the Finite Difference Error measures the error incurred by approximating the derivative by a forward finite step approximation. The Replication Error measures the error introduced by replacing $\boldsymbol{\xi}$ by $\boldsymbol{\delta}$.

We can further bound Finite Difference Error by invoking the concavity of $f(\lambda)=V(\boldsymbol{T}(\boldsymbol{Z},\boldsymbol{\theta})+\lambda\boldsymbol{\xi},\rho)$.[2] By concavity,

$$f(h)\ \le\ f(0)+hf'(0)\implies\frac{V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho)}{h}\ \le\ \boldsymbol{\xi}^\top\hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho),$$
$$f(0)\ \le\ f(h)-hf'(h)\implies\frac{V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho)}{h}\ \ge\ \boldsymbol{\xi}^\top\hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho),$$

where we have again used Danskin's theorem to compute the derivatives. Combining shows

$$\text{Eq. (12a)}\ \le\ \sup_{\boldsymbol{\theta}\in\Theta}\left|\boldsymbol{\xi}^\top\left(\hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho)-\hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)\right)\right|\ =\ \sup_{\boldsymbol{\theta}\in\Theta}\ \mathsf{SS}(\boldsymbol{\xi},h,\boldsymbol{\theta}). \tag{13}$$

As we will argue in Section 4.1 below, Solution Stability depends strongly on the structure of Problem (1) and the choice of policy class.

By contrast, we can bound the Replication Error with minimal additional assumptions. Our results primarily depend on the complexity of the plug-in class which we capture by deterministic covering of the set Θ. Specifically, we assume the following covering exists.

Assumption 4.2 (Plug-in and Regularization Covering) *For any $\varepsilon>0$, there exists a deterministic set $\Theta_0(\varepsilon)$ such that for every $(\bar{\boldsymbol{\theta}},\rho)\in\Theta$ there exists a $(\bar{\boldsymbol{\theta}}_0,\rho_0)\in\Theta_0(\varepsilon)$ such that $\|\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})-\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)\|_1\le\varepsilon C_\Theta(\|\boldsymbol{Z}\|_1+1)$ and $|\rho-\rho_0|\le\varepsilon$ for some universal constant $C\ge 1$.*

Many of the machine learning plug-ins in Section 2.3 satisfy the assumption due to their affine structure as $\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})$ is $C(\|Z\|_1+1)$-Lipschitz with respect to $\bar{\boldsymbol{\theta}}$ for some constant C. We also note the assumption is trivially satisfied if Θ_0 is finite as often happens with clustering plug-ins (Example 2.7).

Combining these two bounds yields the first main result of our paper:

[2] This map is concave because it is the minimum of linear functions of λ.

Theorem 4.3 (Bounding Estimation Error by Stability) *Suppose Assumptions 3.1, 3.3 and 4.2 hold. Then, there exists a constant C (depending on $\phi_{\max}$, $\nu_{\min}$, $\|\boldsymbol{\mu}\|_\infty$) such that the following holds with probability at least $1-2\epsilon$,*

$$\sup_{\boldsymbol{\theta}\in\Theta} \underbrace{\left|\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - D(\boldsymbol{Z},\boldsymbol{\theta},h) - \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})\right|}_{\text{Estimation Error}} \le \underbrace{C\frac{(C_{\boldsymbol{T}}+h)\kappa}{h}\sqrt{n\cdot\log n\cdot\log\left|\Theta_0\left(n^{-1/2}\right)\right|\cdot\log\left(\frac{2}{\epsilon}\right)}}_{\text{Deviation of Replication Error}}$$
$$+\underbrace{\frac{1}{h}\sup_{\boldsymbol{\theta}\in\Theta}\mathbb{W}_1\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\,\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h\big)}_{\text{Expected Replication Error}}$$
$$+\sup_{\boldsymbol{\theta}\in\Theta}\mathsf{SS}(\boldsymbol{\xi},h,\boldsymbol{\theta})$$

where $\mathbb{W}_1$ the Wasserstein 1-distance metric.

For many of the policies presented earlier in Section 2.3, e.g., Examples 2.4 and 2.7, $\left|\Theta_0\left(n^{-1/2}\right)\right|$ is polynomial in n and, hence, the first term in the bound is $\tilde{O}(\sqrt{n}/h)$. The second term of the bound depends on how well we replicate the (unknown) perturbation $h\boldsymbol{\xi}$. In special cases, it possible to exactly replicate $h\boldsymbol{\xi}$ so that this term is 0.

Proposition 4.4 (Exact Distribution Replication with Multivariate Gaussians) *Let $\boldsymbol{Z}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and let $\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})=\boldsymbol{L}(\bar{\boldsymbol{\theta}})\boldsymbol{Z}+\boldsymbol{l}(\bar{\boldsymbol{\theta}})$ be affine. If $\boldsymbol{L}(\bar{\boldsymbol{\theta}})\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\boldsymbol{L}(\bar{\boldsymbol{\theta}})^\top$ is a positive semi-definite matrix, then choosing $\boldsymbol{\delta}_h\sim\mathcal{N}(\boldsymbol{0},h(\boldsymbol{L}(\bar{\boldsymbol{\theta}})\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\boldsymbol{L}(\bar{\boldsymbol{\theta}})^\top)+h^2\boldsymbol{\Sigma})$, yields $\mathbb{W}_1\left(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\,\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h\right)=0$.*

Finally the third term in Theorem 4.3 is the solution stability. As we argue in the next section, we believe that some dependence on solution stability is fundamental and unavoidable. In Sections 5 and 6 we use problem structure to bound this term.

4.1. Is Dependence on Solution Stability Unavoidable?

We build on our heuristic derivation from Section 3.1 to develop some intuition for why solution stability arises in our analysis and why we believe it is a fundamental quantity. Other authors (Yu, 2013; Shalev-Shwartz et al., 2010) have similarly observed the importance of stability in learning applications. We fix $\boldsymbol{\theta}$ and drop it from the notation throughout this subsection.

For a twice differentiable function, a classical Taylor series argument shows that the error in approximating the first derivative by a first-order finite step difference scales like the second derivative, i.e.,

$$\left| \frac{f(h) - f(0)}{h} - f'(0) \right| = \frac{h}{2} |f''(0)| + o(h).$$

Hence, for the function $\lambda \mapsto V(\boldsymbol{T}(\boldsymbol{Z}) + \lambda \boldsymbol{\xi})$ from Eq. (8), we might expect that

$$\left| \frac{V(\boldsymbol{T}(\boldsymbol{Z}) + h\boldsymbol{\xi}) - V(\boldsymbol{T}(\boldsymbol{Z}))}{h} - \boldsymbol{\xi}^\top \hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z})) \right| \;\approx\; \frac{h}{2} \left| \frac{\partial^2}{\partial \lambda^2} V(\boldsymbol{T}(\boldsymbol{Z}) + \lambda \boldsymbol{\xi}) \Big|_{\lambda=0} \right| ,$$

and

$$\frac{\partial^2}{\partial \lambda^2} V(\boldsymbol{T}(\boldsymbol{Z}) + \lambda \boldsymbol{\xi}) \Big|_{\lambda=0} = \frac{\partial}{\partial \lambda} \boldsymbol{\xi}^\top \hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}) + \lambda \boldsymbol{\xi}) \Big|_{\lambda=0} = \boldsymbol{\xi}^\top \boldsymbol{J}(\boldsymbol{T}(\boldsymbol{Z}))\boldsymbol{\xi},$$

where $\boldsymbol{J}(\boldsymbol{T}(\boldsymbol{Z})) \in \mathbb{R}^{n\times n}$ is the Jacobian of $\lambda \mapsto \hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}) + \lambda \boldsymbol{\xi})$. This heuristic analysis suggests that the magnitude of changes in the solution vector $\hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}))$ to perturbations of its input drives the error in our method. More specifically, if we apply a first order finite difference to $\frac{\partial}{\partial \lambda} \boldsymbol{\xi}^\top \hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}) + \lambda \boldsymbol{\xi})$ at $\lambda = 0$, we recover $\boldsymbol{\xi}^\top \left(\hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}) + h\boldsymbol{\xi}) - \hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z})) \right)$, the solution stability of Eq. (13).

This perspective also highlights why solution stability depends on the particular structure of Problem (1) and the policy class. Specifically, the $(i,j)^{\text{th}}$ element of $J(\boldsymbol{T}(\boldsymbol{Z}))$ is $\frac{\partial}{\partial Z_j} \hat{x}_i(\boldsymbol{T}(\boldsymbol{Z}))$, measuring how much the i^{th} component of the solution changes with small perturbations of the j^{th} component of the input. The magnitude of this change depends strongly on the constraints in the feasible set $\mathcal{X}$ of Problem (1) and how $\boldsymbol{T}(\cdot)$ magnifies/attenuates the perturbation. Hence, unlike bounding the replication error, bounding the finite difference error seemingly requires stronger assumptions on Problem (1).

In the next two sections, we present rigorous bounds on the solution stability (and hence, finite difference error) under different structural assumptions on Problem (1). In Section 5, we treat the simpler case of regularized policies and convex $\mathcal{X}$, where we can bound the solution stablity using strong convexity. In Section 6, we treat the more difficult case of unregularized affine policies where we instead require a weakly-coupled assumption on $\mathcal{X}$.

5. Bounds for Continuous Plug-in Policies

5.1. Regularized Policies

We first present bounds for regularized plug-in policies when the feasible set $\mathcal{X}$ is convex. The key insight is that when $\boldsymbol{t} \mapsto V(\boldsymbol{t}, \rho)$ is strongly-convex, plug-in policies are Lipschitz in their argument. Other authors have utilized this Lipschitz continuity in a variety of contexts. We use it to derive a simple upper bound on solution stability.

Lemma 5.1 (Solution Stability of Regularized Policies) *If $\mathcal{X}$ is convex, $\rho > 0$, and $\phi(\cdot)$ is 1-strongly convex with respect to the norm $\|\cdot\|$,*

$$|\mathsf{SS}(\boldsymbol{\xi}, h, \boldsymbol{\theta})| \;\le\; \frac{h}{\rho}\|\boldsymbol{\xi}\|_*^2, \;\; a.s..$$

The bound only depends on the regularizer through the dual norm, but does not otherwise depend on the plug-in $\boldsymbol{T}(\boldsymbol{Z}, \boldsymbol{\theta})$. In this sense, it is general purpose.

Plugging Lemma 5.1 into Theorem 4.3 bounds estimation error uniformly.

Theorem 5.2 (Estimation Error for Regularized Policies) *Suppose Assumptions 3.1, 3.3 and 4.2 hold, $\mathcal{X}$ is convex, and $\Theta = [p_{\min}, p_{\max}] \times \bar{\Theta}$ for $0 < p_{\min} < p_{\max}$. If $\phi(\cdot)$ is 1-strongly convex with respect to $\|\cdot\|$, then there exists a constant C (depending on $\phi_{\max}$, $\nu_{\min}$, $\|\boldsymbol{\mu}\|_\infty$) such that with probability at least $1-\epsilon$*

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - D(\boldsymbol{Z},\boldsymbol{\theta},h) - \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})\right| \;\le\; \frac{h}{\rho_{\min}}\|\boldsymbol{\xi}\|_*^2$$
$$+\, C\frac{(C_{\boldsymbol{T}}+h)\kappa}{h}\sqrt{n\log\left|\Theta_0\left(n^{-1/2}\right)\right|\log\left(\frac{1}{\epsilon}\right)} + \frac{1}{h}\sup_{\boldsymbol{\theta}\in\Theta}\mathbb{W}_1(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) + h\boldsymbol{\xi}, \boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) + \boldsymbol{\delta}_h).$$

Common regularizers include $\phi(x) = \frac{1}{2}x^2$, which is 1-strongly convex with respect to the ℓ_2 norm (so $\|\cdot\|_*$ is the ℓ_2 norm), and $\phi(x) = x\log x$, which is 1-strongly convex with respect to the ℓ_1 norm (so $\|\cdot\|_*$ is the ℓ_∞ norm). In either case, $\|\boldsymbol{\xi}\|_*^2$ can be bounded using standard techniques.

In the special case that the Wasserstein component is zero (c.f. Proposition 4.4), we can optimize the choice of h to obtain explicit rates:

Corollary 5.3 (Regularized Policies with Exact Replication) *Suppose the assumptions of Theorem 5.2 and Proposition 4.4 hold and $\left|\Theta_0(n^{-1/2})\right|$ is polynomial in n.*

i) If $\phi(\cdot)$ is 1-strongly convex with respect to $\|\cdot\|_1$, letting $h = \rho_{\min}^{1/2} n^{1/4}$ shows

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - D(\boldsymbol{Z},\boldsymbol{\theta},h) - \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})\right| \;=\; \tilde{O}_p\left(n^{1/4}\rho_{\min}^{-1/2}\right),$$

ii) If $\phi(\cdot)$ is 1-strongly convex with respect to $\|\cdot\|_2$, letting $h = \rho_{\min}^{1/2} n^{-1/4}$ shows

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - D(\boldsymbol{Z},\boldsymbol{\theta},h) - \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})\right| \;=\; \tilde{O}_p\left(n^{3/4}\rho_{\min}^{-1/2}\right).$$

For $\rho_{\min} \to 0$ sufficiently slowly as n grows, both errors vanish relative to out-of-sample performance in the typical case that the optimal vaue of Problem (1) is $O(n)$.

5.2. Strongly Convex Feasible Regions

We next study unregularized policies in the special case where $\mathcal{X}$ is strongly convex by adapting techniques from El Balghiti et al., 2022. We first summarize key results from that work and then show how they can be used to bound solution stability and, hence, estimation error, in our setting.

For a given norm $\|\cdot\|$, let $B(\boldsymbol{x}_0, r) = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{x}_0\| \le r\}$ be the ball of radius r centered at $\boldsymbol{x}_0$. Recall the following classical definition:

Definition 5.4 (Strongly Convex Feasible Region) *We say the feasible region $\mathcal{X}$ is α-strongly convex with respect to $\|\cdot\|$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$, and any $\lambda \in [0,1]$,*

$$B(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y},\ \frac{\alpha}{2}\lambda(1-\lambda)\|\boldsymbol{x} - \boldsymbol{y}\|^2) \subseteq \mathcal{X}$$

In words, strong convexity of the feasible region ensures that $\mathcal{X}$ is "round." El Balghiti et al. (2022, Theorem 7, Theorem 3a) establishes that when $\mathcal{X}$ is α-strongly convex, the plug-policies are *almost* Lipschitz continuous, namely[3]

$$\|\hat{\boldsymbol{x}}(\boldsymbol{t}, 0) - \hat{\boldsymbol{x}}(\bar{\boldsymbol{t}}, 0)\| \ \le\ \frac{2\|\boldsymbol{t} - \bar{\boldsymbol{t}}\|_*}{\alpha\left(\|\boldsymbol{t}\|_* + \|\bar{\boldsymbol{t}}\|_*\right)}, \qquad \forall \boldsymbol{t}, \bar{\boldsymbol{t}} \in \mathbb{R}^n.$$

These results provide a simple bound on solution-stability:

Lemma 5.5 (Solution Stability for Strongly-Convex Feasible Regions) *Suppose $\mathcal{X}$ is α-strongly convex with respect to $\|\cdot\|$. Then,*

$$|\mathsf{SS}(\boldsymbol{\xi}, h, \boldsymbol{\theta})| \ \le\ \frac{2h\|\boldsymbol{\xi}\|_*^2}{\alpha\|\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\|_*}.$$

Unlike Lemma 5.1, our bound depends not only on the strong-convexity parameter, but also the magnitude of $\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$. Intuitively, if $\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$ is close to $\mathbf{0}$, then a small perturbation might cause the plug-in policy to change wildly.

We can combine this bound with our previous results to provide an estimation bound for unregularized policies and strongly-convex feasible regions:

[3] El Balghiti et al. (2022, Theorem 3a) uses a slightly worse constant in which they bound $\frac{1}{2}\left(\|\boldsymbol{t}\|_* + \|\bar{\boldsymbol{t}}\|_*\right) \ge \min\left(\|\boldsymbol{t}\|_*, \|\bar{\boldsymbol{t}}\|_*\right)$. We use the tighter constant above.

Theorem 5.6 (Estimation Error for Strongly-Convex Feasible Regions) *Suppose Assumptions 3.1, 3.3 and 4.2 hold, $\mathcal{X}$ is α-strongly convex with respect to $\|\cdot\|$ and $\Theta = [0] \times \bar{\Theta}$ (i.e. $\rho = 0$) and $\inf_{\bar{\boldsymbol{\theta}} \in \bar{\Theta}} \|\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\|_* \geq T_{\min}$. Then there exists a constant C (depending on $\phi_{\max}, \nu_{\min}, \|\boldsymbol{\mu}\|_\infty$) such that with probability at least $1-\epsilon$,*

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) - D(\boldsymbol{Z}, \boldsymbol{\theta}, h) - \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) \right| \leq \frac{Ch\|\boldsymbol{\xi}\|_*^2}{\alpha T_{\min}}$$
$$+ C\frac{(C_{\boldsymbol{T}} + h)\kappa}{h}\sqrt{n \log |\Theta_0(n^{-1/2})| \log\left(\frac{1}{\epsilon}\right)} + \frac{1}{h}\sup_{\boldsymbol{\theta} \in \Theta} \mathbb{W}_1(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + h\boldsymbol{\xi}, \boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + \boldsymbol{\delta}_h).$$

Much like Theorem 5.2, when the Wasserstein contribution is zero, we can optimize the choice of h to obtain an explicit rate. The results are identical to Corollary 5.3 with $\rho_{\min}$ replaced by $\alpha T_{\min}$ and hence omitted for brevity.

6. Bounds for Unregularized Plug-in Policies in Weakly-Coupled Problems

For combinatorial optimization problems with non-convex feasible sets, the plug-in policy is typically *not* continuous with respect to its parameter. Hence, bounding solution stability requires us to exploit the structure of the underlying problem more heavily.

In this section, we prove estimation error bounds for non-regularized plug-in policies, i.e. when $\rho = 0$ and $\Theta = \{0\} \times \bar{\Theta}$. We focus on a general class of linear optimization problems known as *weakly-coupled* linear optimization problems, also studied in Gupta, Huang, and Rusmevichientong (2022a).

To describe such problems, let $S_0, \ldots, S_K$ be a disjoint partition of $\{1, \ldots, n\}$ and let $\boldsymbol{x}^k = (x_j : j \in S_k)$ denote vectors of decision variables for $k = 0, \ldots, K$. Without loss of generality, reorder the indices so that the S_k occur "in order," i.e., $(j \in S_0), \ldots, (j \in S_K)$ is a consecutive sequence. Weakly-coupled optimization problems have the following form:

$$\min_{\boldsymbol{x}^0, \ldots, \boldsymbol{x}^K} \quad \left(\boldsymbol{\mu}^0\right)^\top \boldsymbol{x}^0 + \sum_{k=1}^K \left(\boldsymbol{\mu}^k\right)^\top \boldsymbol{x}^k \tag{14a}$$
$$\text{s.t.} \quad \sum_{k=0}^K \sum_{j \in S_k} \boldsymbol{A}_j^0(\boldsymbol{x}^0) x_j \leq \boldsymbol{b}^0(\boldsymbol{x}^0) \tag{14b}$$
$$\boldsymbol{x}^0 \in \mathcal{X}^0 \tag{14c}$$
$$\boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}^0), \quad \forall k = 1, \ldots, K, \tag{14d}$$

where for each fixed $\boldsymbol{x}^0$, $\boldsymbol{A}_j^0(\boldsymbol{x}^0) \in \mathbb{R}^m$ for all j, $\boldsymbol{b}^0(\boldsymbol{x}^0) \in \mathbb{R}^m$, and $\mathcal{X}^k(\boldsymbol{x}^0)$ for $k = 1, \ldots, K$ are convex. The coupling set $\mathcal{X}^0$ is possibly non-convex. We let $\mathcal{X}^{WC}$ denote the feasible set that satisfies Eqs. (14b) to (14d) and assume $\mathcal{X}^{WC} \subseteq [0,1]^n$. For convenience, define $S_{\max} \equiv \max_k |S_k|$ to be the size of the largest subproblem.

The optimization problem is "weakly-coupled" in the sense that by removing the linear constraints (Eq. (14b)) and fixing a choice of $\boldsymbol{x}^0 \in \mathcal{X}^0$ (Eq. (14c)), the optimization problem decouples into K separate convex subproblems of the form:

$$\min_{x^k} \; \left(\boldsymbol{\mu}^k\right)^\top \boldsymbol{x}^k, \quad \text{s.t. } \boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}^0). \tag{15}$$

We assume throughout that after fixing $\boldsymbol{x}^0$, Problem (14) satisfies strong convex duality.

Many applications of linear optimization have the form of Problem (14). For example, two-stage stochastic optimization problems are often weakly coupled by variables. In these problems, $\boldsymbol{x}^0$ are first-stage decision variables (which are sometimes binary), and each $\boldsymbol{x}^k$ represents second-stage decisions for the K different scenarios. Gupta, Huang, and Rusmevichientong (2022a) consider such an instance to model drone assisted emergency medical response. In that application, $m = 0$ and $|\mathcal{X}^0| = O(1)$ as $K \to \infty$. Other applications are weakly-coupled by constraints, where the binding constraints (Eq. (14b)) model resource budgets, e.g., on time or labor. Example 2.1 is such an example, where the binding constraint limits the number of speed humps that can be built. In general, we argue that in typical applications, the number of subproblems $K = O(n)$, the number of coupling constraints $m = O(1)$, and $\log|\mathcal{X}^0| = o(n)$. Although we present results for general instances, the reader may want to focus on these scalings as they interpret our results.

Additional Assumptions

Proving strong estimation error bounds for Problem (14) with general policies appears challenging. We make several simplifying assumptions. First, we specialize our policy class.

Definition 6.1 (Affine Plug-in Policy) *An affine plug-in policy* $\boldsymbol{x}\left(\boldsymbol{Z}, (0, \bar{\boldsymbol{\theta}})\right)$ *is a plug-in policy where* $\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}}) = \boldsymbol{L}(\bar{\boldsymbol{\theta}})\boldsymbol{z} + \boldsymbol{l}(\bar{\boldsymbol{\theta}})$ *and* $\boldsymbol{L}(\cdot)$, $\boldsymbol{l}(\cdot)$ *may implicitly depend* $\boldsymbol{W}$.

As highlighted in Section 2.3, many machine learning methods are $\boldsymbol{W}$-dependent affine transformations of $\boldsymbol{Z}$. The separable affine plug-in policy class studied in Gupta, Huang, and Rusmevichientong (2022a) corresponds to the special case where $\boldsymbol{L}(\bar{\boldsymbol{\theta}})$ is diagonal.

To simplify exposition, we also assume the parameters are appropriately rescaled:

Assumption 6.2 (Scaling of Plug-in Parameters) *Let $\sigma_j(\bar{\boldsymbol{\theta}})$ be the j^{th} largest singular value of $\boldsymbol{L}(\bar{\boldsymbol{\theta}})$. We assume that $\sup_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}} \sigma_1(\bar{\boldsymbol{\theta}}) \leq 1$.*

Assumption 6.2 almost holds without loss of generality because $(\boldsymbol{L}(\bar{\boldsymbol{\theta}}), \boldsymbol{l}(\bar{\boldsymbol{\theta}}))$ can be scaled by small positive constant; such scaling does not alter $\boldsymbol{x}(\boldsymbol{Z}, (0, \bar{\boldsymbol{\theta}}))$. We say "almost" because Assumption 6.2 requires this constant be chosen uniformly over $\bar{\Theta}$. This uniformity holds for many examples of interest. In Example 2.4 and Example 2.5 this property holds whenever the kernel K is supported on a compact set (like the box-kernel). In Example 2.6, it holds so long as $\bar{\theta}$ is bounded away from zero, and it always holds in Example 2.7.

We also require invertibility of the submatrices of $\boldsymbol{L}(\bar{\boldsymbol{\theta}})$ and $\boldsymbol{L}(\bar{\boldsymbol{\theta}}) + h\boldsymbol{I}$ for our analysis:

Assumption 6.3 (Invertibility of Submatrices) *Let $\boldsymbol{L}^k\left(\bar{\boldsymbol{\theta}}\right) \in \mathbb{R}^{|S_k|\times|S_k|}$ be the submatrix of $\boldsymbol{L}\left(\bar{\boldsymbol{\theta}}\right)$ induced by the set S_k. We assume there exists a constant $\sigma_{\min}$ such that*

$$0 < \sigma_{\min} \leq \inf_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}} \sigma_{\min}(\boldsymbol{L}^k(\bar{\boldsymbol{\theta}})), \ \text{ and } 0 < \sigma_{\min} \leq \inf_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}} \sigma_{\min}(\boldsymbol{L}^k(\bar{\boldsymbol{\theta}}) + h\boldsymbol{I}), \quad \forall k = 1, \ldots, K,$$

where $\sigma_{\min}(\cdot)$ denotes the smallest singular value, and h is the step size in the one-shot VGC (cf. Eq. (7)).

Practically, we can satisfy Assumption 6.3 by adding a small perturbation $\gamma\boldsymbol{I}$ to $\boldsymbol{L}(\bar{\boldsymbol{\theta}})$. In Section 7.2, however, we show our approach is effective even Assumption 6.3 does not hold.

Finally, following Gupta and Rusmevichientong (2021) and Gupta, Huang, and Rusmevichientong (2022a), we assume Gaussian corruptions. Recall $\boldsymbol{\Sigma}$ is the covariance of $\boldsymbol{Z}$.

Assumption 6.4 (Gaussian Corruptions) *We assume $\boldsymbol{Z} \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$ and that $\boldsymbol{L}(\bar{\boldsymbol{\theta}})\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{L}(\bar{\boldsymbol{\theta}})^\top$ is positive semidefinite for every $\bar{\boldsymbol{\theta}} \in \bar{\Theta}$.*

We focus on the Gaussian case for simplicity. Using Gupta, Huang, and Rusmevichientong (2022a, Lemma B.4) one might extend our results to the case of approximately Gaussian noise, but we do not pursue this below. The assumption that $\boldsymbol{L}(\bar{\boldsymbol{\theta}})^\top\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{L}(\bar{\boldsymbol{\theta}})^T$ is positive semidefinite allows us to invoke Proposition 4.4 to simplify our analysis.

Assumptions 6.3 and 6.4 are arguably more stringent than Assumption 6.2 or requiring affine policies. In our numerical experiments, we study a setting where these assumptions may not hold and our method still demonstrates strong performance.

We next analyze the special case of block decoupled problems, i.e., when $m=0$ and $|\mathcal{X}^0|=0$ in Section 6.1. We utilize these results to construct a specialized one-shot VGC for general weakly-coupled problems in Section 6.2. We bound the estimation error of this specialized one-shot for weakly-coupled problems in Section 6.3.

Since $\rho=0$ throughout this section, we write $\hat{\boldsymbol{x}}(\boldsymbol{t})$ and drop ρ from the notation.

6.1. Block Decoupled Problems

Our first result bounds the solution stability. The proof uses the chromatic number of a particular graph built from the policy class to bound the dependence between subproblems. For simplicity of exposition, we present a weaker bound using Brook's Theorem to upperbound this chromatic number. (See Section E.4 for the tighter bound and proof.)

Lemma 6.5 (Solution Stability of Block-Decoupled Problems) *Suppose Assumptions 3.1 and 6.2 to 6.4 hold. Consider Problem (14) with $m=0$, and $|\mathcal{X}^0|=0$. If $\Theta=\{0\}\times\bar{\Theta}$, then there exists a constant C (depending on $\nu_{\min},\nu_{\max},\sigma_{\min}$) such that, with probability at least $1-\epsilon$,*

$$\sup_{\boldsymbol{\theta}\in\Theta}\mathsf{SS}(\boldsymbol{\xi},h,\boldsymbol{\theta}) \;\le\; CS_{\max}nh\sqrt{\log\left(\frac{1}{h}\right)}+CS_{\max}^{3/2}\sqrt{nT_{\max}\log\mathcal{X}_{\max}}\cdot\log\left(\frac{S_{\max}T_{\max}}{\epsilon}\right)$$

where $T_{\max}=\max_j\sum_{i=1}^n\max_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}}\mathbb{I}\left\{\boldsymbol{L}_{ij}(\bar{\boldsymbol{\theta}})\neq 0\right\}$, and $\mathcal{X}_{\max}=\max_{\boldsymbol{z}\in\mathbb{R}^n}|\mathcal{X}_\Theta(\boldsymbol{z})|$.

The first term on the right bounds $\sup_{\boldsymbol{\theta}\in\Theta}\mathbb{E}\left[\mathsf{SS}(\boldsymbol{\xi},h,\boldsymbol{\theta})\right]$. It scales with $S_{\max}$, the size of the blocks. In the typical case where $S_{\max}=O(1)$ as $n\to\infty$, such as Example 2.1, the expected solution stability is $\tilde{O}_p(nh)$. The second term bounds $\sup_{\boldsymbol{\theta}\in\Theta}|\mathsf{SS}(\boldsymbol{\xi},\boldsymbol{\theta},h)-\mathbb{E}\left[\mathsf{SS}(\boldsymbol{\xi},\boldsymbol{\theta},h)\right]|$ and depends on the level of independence between policy components $x_j(\boldsymbol{Z},\boldsymbol{\theta})$ as measured by $S_{\max}T_{\max}$, which bounds number of subproblems that can be correlated to one another. The $\sqrt{\log\mathcal{X}_{\max}}$ measures the policy class complexity to allow for a uniform bound. The constants $T_{\max}$ and $\mathcal{X}_{\max}$ depend on our choice of policy class. Loosely, when $\boldsymbol{L}(\bar{\boldsymbol{\theta}})$ is sparse, e.g. when denoising is based on "local" neighborhoods, $T_{\max}=O(1)$. Furthermore, if $\left|\bar{\Theta}\right|$ is small, we expect $\mathcal{X}_{\max}$ is also small. We formalize these intuitions in Section E.1 by revisiting our examples from Section 2.3.

We can now combine Theorem 4.3 and Lemma 6.5 to obtain a bound on the estimation error of the one-shot VGC. Let $L_{\max}=\sup_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}}\sup_{i,j}|\boldsymbol{L}_{ij}(\bar{\boldsymbol{\theta}})|$.

Theorem 6.6 (Estimation Error for Block Decoupled Problems) *Suppose Assumptions 3.1, 4.2 and 6.2 to 6.4 hold. Consider Problem (14) with $m = 0$, and $|\mathcal{X}^0| = 0$. Define $\boldsymbol{\delta}_h$ as in Proposition 4.4. Finally assume $h < C_T$. If $\Theta = \{0\} \times \bar{\Theta}$, then there exists a constant C (depending on $\nu_{\min}, \nu_{\max}, \sigma_{\min}, \|\boldsymbol{\mu}\|_\infty, L_{\max}$) such that with probability at least $1 - \epsilon$,*

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in \Theta} \left| \boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) - D(\boldsymbol{Z}, \boldsymbol{\theta}, h) - \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) \right| \\
&\quad \leq \underbrace{C \left(S_{\max} n^{1/2} h + \frac{T_{\max}}{h} \right)}_{\text{Step-Size Trade-Off}} \sqrt{n \log n \cdot \log\left(\frac{1}{h}\right) \cdot \log \left|\Theta_0\left(n^{-1/2}\right)\right| \cdot \log\left(\frac{2}{\epsilon}\right)} \\
&\qquad + C S_{\max}^{3/2} \sqrt{n T_{\max} \log \mathcal{X}_{\max}} \log\left(\frac{S_{\max} T_{\max}}{\epsilon}\right).
\end{aligned}$$

The VGC step-size trade-off groups terms that depend on the choice of h, and the stochastic error groups terms related to the structure of the optimization problem and plug-in estimator. By optimizing h, we obtain the following corollary.

Corollary 6.7 (Optimized Step-Size) *Under the assumptions of Theorem 6.6, if $h = O\left(n^{-1/4} T_{\max}^{1/2} S_{\max}^{-1/2}\right)$ there exists a constant C (depending on $\nu_{\min}, \nu_{\max}, \sigma_{\min}, \|\boldsymbol{\mu}\|_\infty, L_{\max}$) such that with probability at least $1 - \epsilon$,*

$$\begin{aligned}
&\sup_{\boldsymbol{\theta} \in \Theta} \left| \boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) - D(\boldsymbol{Z}, \boldsymbol{\theta}, h) - \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z}, \boldsymbol{\theta}) \right| \\
&\quad \leq C S_{\max}^{3/4} n^{3/4} \log n \sqrt{T_{\max} \log \mathcal{X}_{\max} \log \left|\Theta_0\left(n^{-1/2}\right)\right|} \cdot \log\left(\frac{S_{\max} T_{\max}}{\epsilon}\right).
\end{aligned}$$

The corollary shows that when all problem and policy parameters are $\tilde{O}(1)$, then the estimation error grows at a rate $\tilde{O}_p(n^{3/4})$. Thus, in typical settings where full-information performance is of order $O(n)$, the one-shot VGC estimator has vanishing relative error.

6.2. A Modified VGC for Weakly-Coupled Problems

Our strategy for the weakly-coupled Problem (14) will be to i) "fix" values of the coupling variables Eq. (14c) and Lagrangian duals for the coupling constraints Eq. (14b) ii) apply our previous results to the resulting decoupled problem iii) take a union bound over all possible values of the "fixed" quantities. To execute this strategy we define a lifted policy class indexed by the "fixed" quantities and then define a modified VGC for this lifted class.

Lifted Affine Policy Class

We first introduce dual notation. Recall the general affine plug-in for the weakly-coupled optimization problem Problem (14) solves

$$\begin{aligned}
\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) \in \arg\min_{\boldsymbol{x}} \quad & \sum_{j=1}^{n} T_j(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) x_j && \text{(16a)}\\
\text{s.t.} \quad & \sum_{k=0}^{K} \sum_{j\in S_k} \boldsymbol{A}_j^0(\boldsymbol{x}^0) x_j \le \boldsymbol{b}^0(\boldsymbol{x}^0) && \text{(16b)}\\
& \boldsymbol{x}^0 \in \mathcal{X}^0 && \text{(16c)}\\
& \boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}_0), \quad \forall k=1,\ldots,K. && \text{(16d)}
\end{aligned}$$

where $\boldsymbol{A}_j^0(\boldsymbol{x}^0) \in \mathbb{R}^m$ for $j=1,\ldots,n$ and $\boldsymbol{b}^0(\boldsymbol{x}^0) \in \mathbb{R}^m$ are vectors that may or may not depend on $\boldsymbol{x}^0$.

Fix some $\boldsymbol{x}^0 \in \mathcal{X}^0$ and consider relaxing Eq. (16b) with dual variables $\boldsymbol{\lambda} \in \mathbb{R}_+^m$ to obtain the block decoupled optimization problem

$$\min_{\boldsymbol{x}} \sum_{k=1}^{K} \langle \boldsymbol{T}^k(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) + \boldsymbol{A}_{S_k}^0(\boldsymbol{x}^0)^\top \boldsymbol{\lambda}, \boldsymbol{x}^k \rangle, \quad \text{s.t. } \boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}^0),\ \forall k=1,\ldots,K \tag{17}$$

where $\boldsymbol{A}_{S_k}^0(\boldsymbol{x}^0)$ is formed by taking columns $j \in S_k$. Let $\tilde{\boldsymbol{x}}^k(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}})$ denote the solution of the k^{th} subproblem and $\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}}) = \left(\boldsymbol{x}^0, \tilde{\boldsymbol{x}}^1(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}})^\top, \ldots, \tilde{\boldsymbol{x}}^K(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}})^\top\right)^\top$. Although the notation is onerous, the intuition is to view Eq. (17) as defining the plug-in policy $\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}})$ for the target optimization problem

$$\min_{\boldsymbol{x}} \sum_{k=1}^{K} \langle \boldsymbol{\mu}^k, \boldsymbol{x}^k \rangle, \quad \text{s.t. } \boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}^0),\ \forall k=1,\ldots,K,$$

where $\left(\boldsymbol{x}^0,\boldsymbol{\lambda},\bar{\boldsymbol{\theta}}\right)$ define the plug-in policy. Our strategy will be to use Theorem 6.6 to debias the in-sample performance of this policy, and then relate it back to the original plug-in policy for our weakly coupled problem. To that end, define the lifted affine policy class

$$\tilde{\mathcal{X}}(\boldsymbol{Z}) = \left\{ \tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}}) \ :\ \boldsymbol{\lambda} \in \mathbb{R}_+^m,\ \boldsymbol{x}^0 \in \mathcal{X}^0,\ \bar{\boldsymbol{\theta}} \in \bar{\Theta} \right\}$$

and the corresponding plug-in function

$$\tilde{V}\left(\boldsymbol{t},\boldsymbol{x}^0\right) = \min\left\{ \boldsymbol{t}^\top \boldsymbol{x} \ :\ \boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}^0),\ \forall k=1,\ldots,K. \right\}$$

Our one-shot VGC applied to this lifted, block-decoupled problem is

$$\tilde{D}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}},h) \equiv \frac{1}{h}\mathbb{E}\left[\tilde{V}(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+(\boldsymbol{A}^0)^\top\boldsymbol{\lambda}+\boldsymbol{\delta}_h,\boldsymbol{x}^0)-\tilde{V}(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+(\boldsymbol{A}^0)^\top\boldsymbol{\lambda},\boldsymbol{x}^0)\Big|\,\boldsymbol{Z}\right],$$

where $\boldsymbol{\delta}_h$ is defined as in Proposition 4.4.

Of course Eq. (17) is not the same as Problem (16). However, one might intuit that if we set $(\boldsymbol{\lambda},\boldsymbol{x}^0)$ to their optimal values, Eq. (17) might closely approximate Problem (16) and the corresponding VGC might approximately debias Problem (16). To that end, for a fixed $\boldsymbol{x}^0$, we define the optimal dual variables

$$\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0) \in \operatorname*{arg\,max}_{\boldsymbol{\lambda}\geq\boldsymbol{0}} \; -\left\langle \boldsymbol{b}^0(\boldsymbol{x}^0),\,\boldsymbol{\lambda}\right\rangle + \sum_{k=0}^{K}\left\langle \boldsymbol{T}^k(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{A}^0_{S_k}(\boldsymbol{x}^0)^\top\boldsymbol{\lambda},\,\tilde{\boldsymbol{x}}^k(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}})\right\rangle \tag{18}$$

and optimal coupling variables $\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})$ as the portion of the optimal solution to Problem (16) corresponding to block S_0. Our modified VGC correction is then

$$D^{WC}(\boldsymbol{Z},\boldsymbol{\theta},h) \equiv \tilde{D}(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\bar{\boldsymbol{\theta}},h). \tag{19}$$

Remark 6.8 (Computing the Modified VGC) Computing the modified VGC is not much harder than solving Problem (16) since one can extract $\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})$ and then solve a convex problem to obtain $\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ and $\tilde{\boldsymbol{x}}\,(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\boldsymbol{\theta})$. In practice, however, we adopt a simpler approach. Let $V(\boldsymbol{t})$ denote the optimal objective value of Problem (16) when the cost coefficients are $\boldsymbol{t}\in\mathbb{R}^n$. We approximate $D^{WC}(\boldsymbol{Z},\boldsymbol{\theta},h)\approx \frac{1}{h}\mathbb{E}\left[V\left(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,0\right)-V\left(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),0\right)\big|\,\boldsymbol{Z}\right]$. As can be seen in the proof, these quantities are asymptotically identical. We treat Eq. (19) in our analysis for simplicity.

6.3. Estimation Error of Weakly-Coupled One-Shot VGC

The key insight to leveraging Theorem 6.6 to analyzing D^{WC} is relating solutions to Problem (16) to our lifted affine policy class. In particular, we will show that

$$\boldsymbol{x}^k(\boldsymbol{Z},\boldsymbol{\theta}) = \tilde{\boldsymbol{x}}^k\left(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\bar{\boldsymbol{\theta}}\right)$$

whenever $\tilde{\boldsymbol{x}}^k\left(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\bar{\boldsymbol{\theta}}\right)$ is the unique solution to its defining optimization problem. We next introduce a technical assumption that allows us to bound the number of blocks with multiple solutions.

Let $\mathcal{D}^k(\boldsymbol{x}^0)$ be the set of unit vectors $\boldsymbol{d}$ such that $\min_{\boldsymbol{x}\in\mathcal{X}^k(\boldsymbol{x}^0)} \boldsymbol{d}^\top \boldsymbol{x}$ has multiple optima and define the set of vectors

$$\mathcal{F}(\boldsymbol{Z},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}}) \equiv \bigcup_{k=1}^{K} \left\{ \begin{pmatrix} T_j(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) \\ \boldsymbol{A}_j^0 \\ -d_j \end{pmatrix} : j \in S_k,\ \boldsymbol{d} \in \mathcal{D}^k(\boldsymbol{x}^0) \right\} \subseteq \mathbb{R}^{m+2}.$$

We assume these vectors are in general position.

Assumption 6.9 (Induced Cost Vectors in General Position) *We have*

$$\mathbb{P}\left(\mathcal{F}(\boldsymbol{Z},\boldsymbol{x}^0,\bar{\boldsymbol{\theta}}) \text{ are in general position for all } \bar{\boldsymbol{\theta}} \in \bar{\Theta} \text{ and } \boldsymbol{x}^0 \in \mathcal{X}^0\right) = 1.$$

We use this assumption to show that the number of subproblems of our lifted affine policy that have multiple solutions is at most m in Lemma F.1.

Additionally, we require assumptions on the boundedness of the optimal dual values $\boldsymbol{\lambda}^0(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)$ to cover the policy class. We adopt a simple assumption similar to Gupta, Huang, and Rusmevichientong (2022a) and Gupta and Rusmevichientong (2021).

Assumption 6.10 ($\bar{s}$-Strict Feasibility of Coupling Constraints) *For each $\boldsymbol{x}^0 \in \mathcal{X}^0$, there exists an $\bar{s} > 0$ and $\bar{\boldsymbol{x}} \in \mathcal{X}^{WC}$ such that $\sum_{j=1}^n \boldsymbol{A}_j^0(\boldsymbol{x}^0)\bar{x}_j + \bar{s}\boldsymbol{e} \le \boldsymbol{b}^0(\boldsymbol{x}^0)$.*

Lemma F.2 bounds $\|\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta})\|_1$ under this assumption.

We now can present our uniform bound on the estimation error of the weakly-coupled one-shot VGC. Let $\|\boldsymbol{A}^0\|_\infty$ be the entry-wise infinity norm for matrices.

Theorem 6.11 (Estimation Error of the Weakly-Coupled VGC) *Assume Assumptions 3.1, 4.2, 6.2 to 6.4, 6.9 and 6.10 hold and let $\Theta = \{0\} \times \bar{\Theta}$. Choose $h = O\left(n^{-1/4} T_{\max}^{1/2} S_{\max}^{-1/2}\right)$ and $\boldsymbol{\delta}_h$ as in Proposition 4.4. Consider Problem (14). There exists a constant C (depending on $\nu_{\min}$, $\nu_{\max}$, $\sigma_{\min}$, $\|\boldsymbol{\mu}\|_\infty$, $L_{\max}$, $\log \bar{s}^{-1}$, $\|\boldsymbol{A}^0\|_\infty$) such that the following holds with probability at least $1-\epsilon-\exp(-n)$,*

$$\sup_{\boldsymbol{\theta}\in\Theta} \underbrace{\left|\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - D^{WC}(\boldsymbol{Z},\boldsymbol{\theta},h) - \boldsymbol{\mu}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})\right|}_{\text{Estimation Error}}$$
$$\le \underbrace{C_4 S_{\max}^{3/4} n^{3/4} \log^{3/2} n \sqrt{m T_{\max} \log \mathcal{X}_{\max} \log \left|\Theta_0\left(n^{-1/2}/2\right)\right|} \cdot \log\left(\frac{S_{\max} T_{\max} |\mathcal{X}^0|}{\epsilon}\right)}_{\text{Block Decoupled Bound}},$$
$$+ \underbrace{C m S_{\max} \sqrt{\log n \cdot \log\left(\frac{1}{\epsilon}\right)}}_{\text{Lifted Policy Estimation Error}}$$

Since we apply the one-shot VGC to the lifted class, our bound for weakly-coupled problems only differs from Theorem 6.6 by the additional lifted policy approximation error. This error bounds the difference between the in-sample optimism of the lifted policy and the original policy. In typical settings where $m = O(1)$ (c.f. Example 2.1), this term is $\tilde{O}(1)$.

We stress that since our weakly-coupled VGC applies to the lifted decoupled problem, it depends on the complexity of the lifted policy class $\tilde{\mathcal{X}}_\Theta(\boldsymbol{Z})$, not just $\mathcal{X}_\Theta(\boldsymbol{Z})$. Since the lifted class simply shifts the plug-in by an affine amount, it can often be bounded using the same techniques as in Section E.1

7. Computational Study

7.1. Comparing One-Shot and Multi-Shot VGC

As discussed, the one-shot and multi-shot VGC are provably equal in special cases, most notably when the optimization problem fully decouples (c.f. Proposition B.3 in the appendix). In this section, we compare these two methods empirically, focusing on settings where the feasible region has coupling constraints. For clarity, the theoretical analysis of multi-shot VGC in Gupta, Huang, and Rusmevichientong, 2022a does not cover non-separable affine plug-in policies, but the algorithm can still be applied (without theoretical guarantees) to this setting. We utilize synthetic data where can individually vary the degree of coupling and size of the problem.

Specifically, we consider on the full information optimization problem:

$$\max_{\boldsymbol{x}\in\mathcal{X}} \sum_{k=1}^{K}\sum_{j=1}^{B} \mu_{kj} x_{kj}, \quad \text{where } \mathcal{X} = \left\{ \boldsymbol{x} \, : \, \sum_{j=1}^{B} x_{kj} \le 1 \, \forall k; \; x_{kj} \in \{0,1\} \, \forall k, j \right\}$$

We observe $W_j = j$, and $Z_{kj} \sim \mathcal{N}(\mu_{kj}, 1/\nu_{kj})$ with

$$(\mu_{kj}, \nu_{kj}) = \begin{cases} (0,2) & \text{w.p. } \frac{1}{2}, \\ (0.5, 10) & \text{w.p. } \frac{1}{4}, \\ (1,6) & \text{w.p. } \frac{1}{4}. \end{cases}$$

The "ground-truth" values of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are generated once and fixed throughout the experiment. The data $\boldsymbol{Z}$ are generated each simulation run, and $B = 4$ unless otherwise specified.

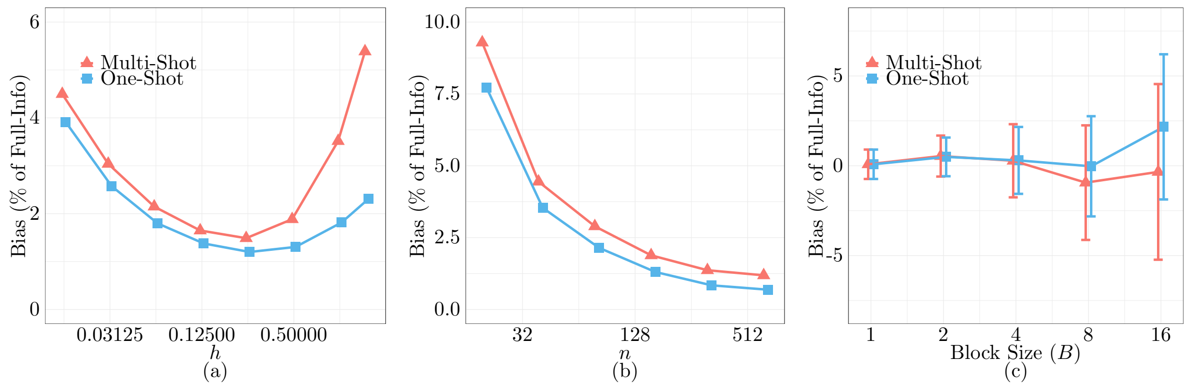

Figure 1 **Multi-Shot vs. One-Shot: Separable Policy**. The graphs show the root mean squared error (RMSE) of the one-shot VGC (Eq. (11)) is better than the RMSE of the multi-shot VGC (Eq. (10)) for the plug-in policy with $\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) = \boldsymbol{Z}$. For figures (a) and (b) the values are computed over 200 trials and use the same $\boldsymbol{\delta}_h$. Plot (a) varies the step size parameter h for $n = 160$ and plot (b) varies the number of decisions n for $h = .5$. Figure (c) plots the estimated bias over block sizes for a fixed $n = 160$ and step-size $h = 2^{-5}$. Error bars are 95% confidence intervals computed using 200 trials.

Separable Plug-In. In Fig. 1, we first consider a separable affine plug-in function $T_j(\boldsymbol{Z}) = Z_j$. We build on our intuition from Section 3.3 that the different step-sizes of one-shot and multi-shot VGC impacts estimation quality. The first two panels show that multi-shot VGC has uniformly worse mean-square error than our one-shot variant across different step-sizes and problems izes. Panel (c) shows that both estimates yield approximately unbiased estimates, even as the degree of coupling grows. Hence, this difference in MSE is due to increased variance.

Non-Separable Plug-in. Fig. 2 considers a non-separable affine plug-in $T_j(\boldsymbol{Z}) = 0.5Z_j + 0.25(Z_{j-1} + Z_{j+1})$, which can be seen as a 1-nearest-neighbor smoothing. In this case, multi-shot VGC not only exhibits higher variance, but is also biased. Panel (a) shows the size of the bias increases as the degree of coupling in the feasible region increases. Panel (b) shows this bias is non-vanishing in n, even for a fixed step-size h. By contrast, bias (scaled by full information) of the one-shot VGC closely corresponds to the predicted rate of $O(Bh)$ from Theorem 6.6. Overall, the plots suggest that the bias for the multi-shot VGC seems scales poorly with B when applied to non-separabe policies.

7.2. Case Study: Prioritizing Speed Hump Requests

Our case study further develops Example 2.1, i.e., prioritizing speed hump requests to improve traffic safety. We use motor vehicle accident data from the New York Police

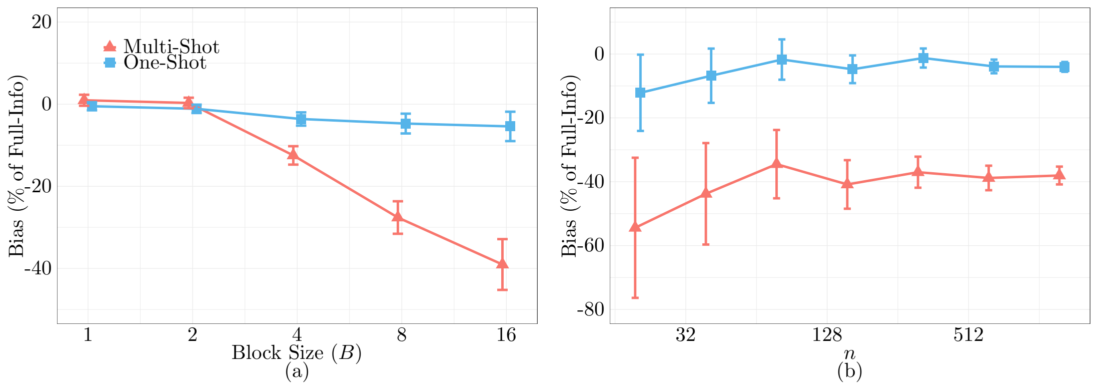

Figure 2 **Multi-Shot vs. One-Shot: Non-Separable Policy**. The graphs plots the estimates of the bias of the one-shot VGC (Eq. (11)) and multi-shot VGC (Eq. (10)) for the non-separable plug-in policy $T_j(\boldsymbol{Z}) = 0.5Z_j + 0.25(Z_{j-1} + Z_{j+1})$. Figure (a) increases the size of the blocks (B) for fixed $n = KB = 160$ and $h = 2^{-5}$. Figure (b) increases the number of decision variables n for $B = 20$ and $h = 2^{-5}$. Error bars are 95% confidence intervals computed using 200 trials.

Department from 2012-2023 (New York, 2023). Our goal is to identify census tracts that have the highest incidence of pedestrian injury due to motor vehicle accidents subject to a budget constraint on the number of tracts that can be serviced in a given year (B).

Notably, our data lack covariate information such as foot and vehicle traffic that are (intuitively) highly predictive of pedestrian injury rates. Hence, obtaining precise estimates is difficult or impossible. While one might argue for collecting and integrating these (and other) useful covariates into the dataset, such integration can be costly, time-consuming, and may conflict with data-privacy regulations. Our approach instead seeks to identify the highest quality decisions possible using only the data at hand.

All code and data for these experiments are available on Github[4].

Data and Setup. We consider selecting the top 5% of the 2,157 total census tracts in New York city to prioritize speed hump requests, i.e., $B = 107$ and $n = 2{,}157$.

We let μ_j represent the expected number of crashes per year with pedestrian, cyclist, or motorist injuries in census tract j. We model $\mu_j = \lambda_j p_j$ where λ_j and p_j are the respective crash rate per year and probability a pedestrian is injured given an accident for census tract j. As ground truth, we take p_j to be the observed pedestrian injury rate for each census tract across the *entire* time period covered by the dataset (2012-2023). These ground-truth values range from 0 to 0.1576 across census tracts. We treat these values as *unknown.*

[4] https://github.com/mh3166/Decision-Aware-Denoising

Similarly, we take the ground-truth λ_j to be the observed average number of crashes per year over the time frame. These values range from 1.4 to 448 across census tracts with an average of 52 crashes per year. However, we treat these value as *known* and fixed.

We set N_j to be the number of accidents observed in 2018. To generate our observations, we let $\hat{p}_j \sim \text{Binomial}(\gamma N_j, p_j)$ for $\gamma \in \{1,2,3,4\}$. Intuitively, this represents estimating the pedestrian injury rate based on γ years of historical data. The predicted number of crashes with injuries is then $Z_j = \lambda_j \hat{p}_j / \gamma$. This setup allows us to generate multiple simulation paths Z_j and also vary the amount of historical data available in forming these estimates.

Policy Class. We consider a non-regularized kernel smoother plug-in policy class (Example 2.4):

$$T_j(\hat{\boldsymbol{p}}, \boldsymbol{W}, \theta) \;=\; \lambda_j \sum_{i=1}^{n} \hat{p}_i \frac{\mathbb{I}\{\|\boldsymbol{W}_j - \boldsymbol{W}_i\| \le \theta\}}{\sum_{l=1}^{n} \mathbb{I}\{\|\boldsymbol{W}_j - \boldsymbol{W}_l\| \le \theta\}},$$

where we let the bandwidth $\theta \ge 0$ vary. Here $\boldsymbol{W}_j$ are the longitude and latitude coordinates of the centroid of census tract j. Our goal is to select θ that maximizes the out-of-sample performance $\sum_{j=1}^{n} \lambda_j p_j x_j(\hat{\boldsymbol{p}}, \boldsymbol{W}, \theta)$.

Experiments. We compare decision-aware methods to decision-blind methods for selecting θ as we vary the amount of crash data observed (γ).

For decision-aware methods, we compare the multi-shot VGC implementation discussed in Section 3.3 and the Stein correction from Gupta and Rusmevichientong (2021). All three approaches estimate the out-of-sample performance of different choices of θ. We set $h = n^{-1/4}$ for the one-shot VGC and simulate $\boldsymbol{\delta}_h$ as multivariate Gaussians similar to Proposition 4.4. The Stein correction and multi-shot VGC also require a choice of h which we set as $h = n^{-1/6}$ as recommended. For full implementation details, see Appendix G.1.

For our decision-blind policy, we use a predict-then-optimize policy that chooses the θ that minimizes (oracle) mean squared error (MSE) of the predicted pedestrians injured. Using oracle MSE favors the decision-blind by avoiding any noise from misestimating MSE. Formally, the MSE policy selects

$$\theta^{MSE} \in \arg\min_{\theta \ge 0} \frac{1}{n} \sum_{j=1}^{n} \left(T_j(\hat{\boldsymbol{p}}, \boldsymbol{W}, \theta) - \lambda_j p_j\right)^2$$

Finally, we also benchmark against the oracle θ that optimizes the true out-of-sample performance for each realization of our data $\hat{\boldsymbol{p}}$. This oracle represents an upperbound on achievable performance.

For ease of comparison, we present performance relative to the full information optimal performance $\max_{\boldsymbol{x}\in\mathcal{X}} \sum_{j=1}^{n} \lambda_j p_j x_j$.

7.3. Case Study Results

Figure 3 plots the estimates of the expected out-of-sample performance of the all the different methods as we increase the amount of data. We highlight several key observations.

First, we show that decision-aware and decision-blind approaches can both achieve performance comparable to the oracle policy as the data increases. However, our proposed decision-aware one-shot VGC approach converges faster, showing that it can also perform well in small-data settings. We also stress that we are using an *oracle* version of MSE, so the MSE performance here optimisitc.

Second, we see that among decision-aware approaches, the one-shot VGC also converges faster to oracle performance. We hypothesize the one-shot VGC converges faster than the Stein correction due to leveraging the optimization structure more directly. The same observation was also made in Gupta, Huang, and Rusmevichientong (2022a) with the multi-shot VGC. For the multi-shot VGC, we see the bias and variance issues discussed in Section 7.1 directly translate into learning worse policies.

Lifted Optimization Problems. Inspired by Theorem 6.11, we next consider a modifications of the one-shot VGC and multi-shot VGC in which we apply these corrections to the lifted policy class introduced in Section 6.2. Fig. EC.1 in the appendix provides a larger plot with all methods for ease of comparison.

As predicted by Theorem 6.11, the “lifted” one-shot VGC performs similarly to the non-lifted variant. However, we see that the “lifted” multi-shot VGC outperforms its non-lifted version, achieving performance similar to the Stein correction. This empirical result combined with Proposition B.3 suggest some of the bias of multi-shot VGC from Section 7.1 might be heuristically addressed by “lifting” coupling constraints. While not pursued here, we conjecture that techniques from Theorem 6.11 and Gupta, Huang, and Rusmevichientong (2022b) might be used to formalize this heuristic.

Advantage of Decision-Aware Approach. In Figure 4, we map the selected census tracts of the decision-aware one-shot VGC policy and the decision-blind MSE policy. The map also plots contours of pedestrian injury rates generated from observed crashes. Many of the census tracts selected by the MSE policy lie on regions where there are multiple

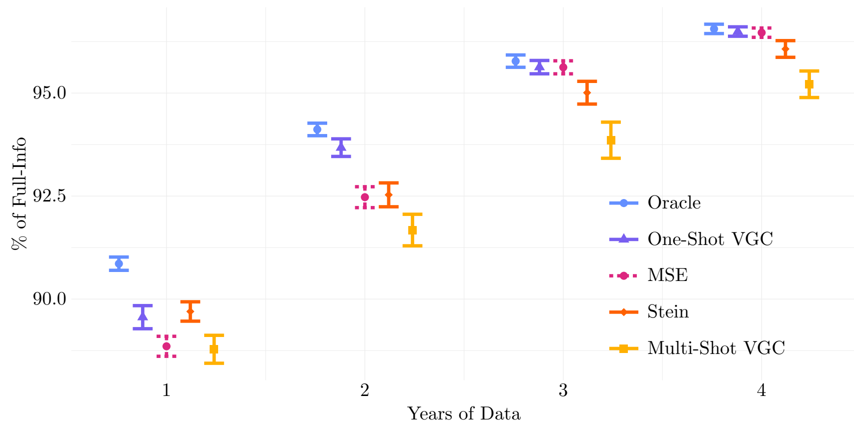

Figure 3 **Performance Results**. We compare the estimated expected out-of-sample performance of our method to various benchmarks and various over 100 trials. The error bars are 95% confidence intervals. The experiments vary the amount of data available but keep the number of decisions fixed.

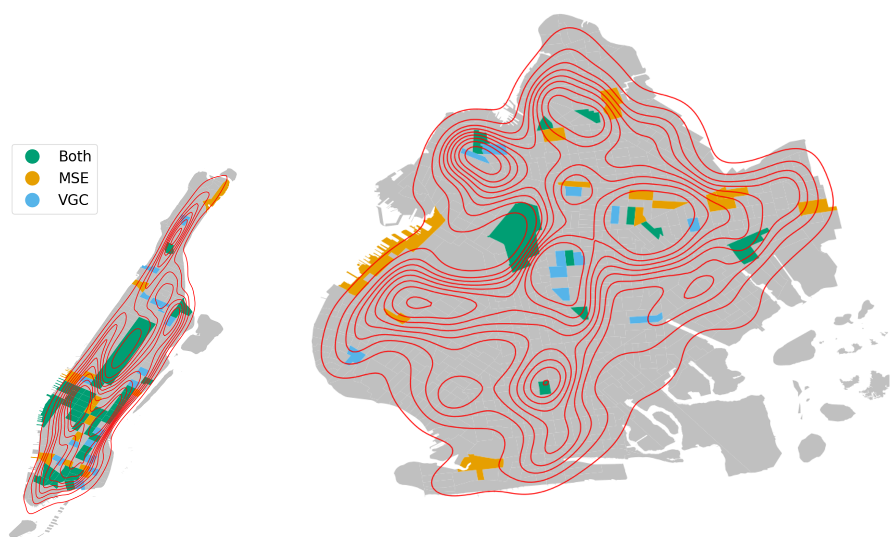

Figure 4 **Map of Census Tracts Selected.** We plot the census tracts selected by the MSE policy and the one-shot VGC policy for the Manhattan and Brooklyn boroughs of New York City. The contour lines plot the density of crashes with pedestrian injuries.

contour lines indicating a changing density of crashes with pedestrian injury. By contrast, the one-shot VGC policy tends to select census tracts in the center of regions with a high pedestrian injury rate.

To better understand this phenomenon, we plot the distribution of θ^{MSE}, θ^{OR} and θ^{VGC} (the θ that optimizes our one-shot VGC estimator) in Fig. 5. Note that θ^{OR} and θ^{MSE} are quite different. By construction, θ^{MSE} is the best possible θ to minimize MSE, not downstream decision loss. In this example, it systematically over-smooths the data. This

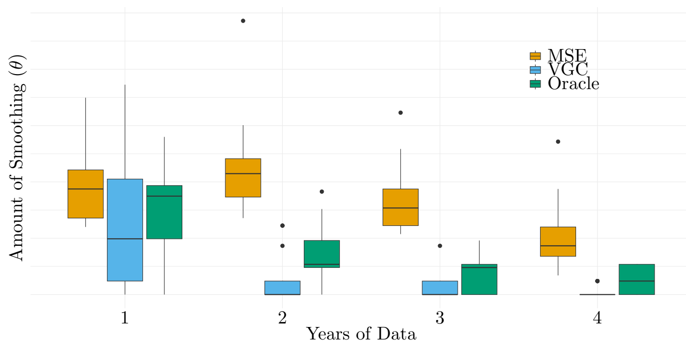

Figure 5 **Amount of Smoothing by Method** The figure compares the values of θ chosen by each method. Across realizations, θ^{MSE} tends to oversmooth relative to θ^{OR}, particularly as the amount of historical data grows. By contrast, the one-shot VGC performs less smoothing.

partially explains why it chooses census tracts that cross contour lines – the larger θ effectively smooths away the differences across the contour. By contrast, the one-shot VGC selects a θ^{VGC} much closer to the oracle value. We see this as strong evidence that i) Decision-aware approaches are necessary for finding the "right" level of smoothing in these types of applications, and ii) our one-shot VGC, despite being light-touch, is able to identify a good level of smoothing.

8. Conclusion

Our paper highlights the benefits of adapting traditional denoising techniques to decision-making settings. We provide rigorous guarantees of our approach for a wide class of regurlarized plug-in and non-regularized plug-in policies. Specifically, we provide a general bound of the error of the approach in terms of the stability of the underlying policies. For regularized plug-in policies and convex optimization, this approach readily yields strong guarantees that show the relative error of the method is vanishing as the problem size increases. We also develop a customized VGC estimator for general affine plug-in policies for weakly-coupled, potentially non-convex optimization problems. Under somewhat stronger assumptions, we also show the relative error in these settings vanishes as the problem size grows.

The fundamental role of stability in these results complements the role of stability in more traditional generalization guarantees in machine learning. Exploring other aspects of stability in data-driven optimization is an open area.

Finally, while our results primarily focus on debiasing policy performance for optimization problems with linear objectives, developing similar debiasing techniques for other structured optimization classes remains an exciting direction for future work.

References

Arora, Sanjeev et al. (2020). "Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks". In: *ICLR 2020*.

Besbes, Omar, Will Ma, and Omar Mouchtaki (2022). "Beyond IID: data-driven decision-making in heterogeneous environments". In: *Advances in Neural Information Processing Systems* 35, pp. 23979–23991.

Besbes, Omar and Omar Mouchtaki (2023). "How big should your data really be? data-driven newsvendor: learning one sample at a time". In: *Management Science*.

Buja, Andreas, Trevor Hastie, and Robert Tibshirani (1989). "Linear smoothers and additive models". In: *The Annals of Statistics*, pp. 453–510.

Candes, Emmanuel J, Justin K Romberg, and Terence Tao (2006). "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 59.8, pp. 1207–1223.

Chan, Grace and Andrew TA Wood (1997). "Algorithm AS 312: An Algorithm for simulating stationary Gaussian random fields". In: *Applied Statistics*, pp. 171–181.

Chatterjee, Shubhojeet et al. (2020). "Review of noise removal techniques in ECG signals". In: *IET Signal Processing* 14.9, pp. 569–590.

Chiles, Jean-Paul and Pierre Delfiner (2012). *Geostatistics: modeling spatial uncertainty*. Vol. 713. John Wiley & Sons.

Chu, Leon Yang et al. (2023). "Solving the price-setting newsvendor problem with parametric operational data analytics (ODA)". In: *Available at SSRN 4400568*.

Chung, Tsai-Hsuan et al. (2022). "Decision-Aware Learning for Optimizing Health Supply Chains". In: *arXiv preprint arXiv:2211.08507*.

Devroye, Luc, Abbas Mehrabian, and Tommy Reddad (2018). "The total variation distance between high-dimensional Gaussians". In: *arXiv preprint arXiv:1810.08693* 6.

El Balghiti, Othman et al. (2022). "Generalization Bounds in the Predict-Then-Optimize Framework". In: *Mathematics of Operations Research* 0.0, null. DOI: `10.1287/moor.2022.1330`. URL: `https://pubsonline.informs.org/doi/abs/10.1287/moor.2022.1330`.

Elmachtoub, Adam N and Paul Grigas (2022). "Smart "predict, then optimize"". In: *Management Science* 68.1, pp. 9–26.

Elmachtoub, Adam N. et al. (2023). *Estimate-Then-Optimize versus Integrated-Estimation-Optimization versus Sample Average Approximation: A Stochastic Dominance Perspective*. arXiv: `2304.06833 [stat.ML]`.

Elvik, Rune et al. (2009). "The handbook of road safety measures, Bingley". In: *UK: Emerald Group Publishing Limited* 4.5, p. 12.

Feng, Qi and J George Shanthikumar (2023). "The framework of parametric and non-parametric operational data analytics (ODA)". In: *Available at SSRN 4400555*.

Gallicchio, Claudio and Simone Scardapane (2020). "Deep randomized neural networks". In: *Recent Trends in Learning From Data: Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019)*. Springer, pp. 43–68.

Grigas, Paul, Meng Qi, et al. (2021). "Integrated conditional estimation-optimization". In: *arXiv preprint arXiv:2110.12351*.

Guo, Wenshuo, Michael Jordan, and Angela Zhou (2022). "Off-policy evaluation with policy-dependent optimization response". In: *Advances in Neural Information Processing Systems* 35, pp. 37081–37094.

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

– (2022b). "Simplifying the Analysis of the Stein-Correction in the Small-Data, Large-Scale Optimization Regime". In.

Gupta, Vishal and Nathan Kallus (2022). "Data pooling in stochastic optimization". In: *Management Science* 68.3, pp. 1595–1615.

– (2023). "Contextual Data Pooling for Panel Data". Working paper.

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

Hu, Yichun, Nathan Kallus, and Xiaojie Mao (2022). "Fast rates for contextual linear optimization". In: *Management Science* 68.6, pp. 4236–4245.

Ignatiadis, Nikolaos and Stefan Wager (2019). "Covariate-Powered empirical Bayes estimation". In: *Advances in Neural Information Processing Systems* 32.

Ito, Shinji, Akihiro Yabe, and Ryohei Fujimaki (2018). "Unbiased objective estimation in predictive optimization". In: *International Conference on Machine Learning*. PMLR, pp. 2176–2185.

Jacot, Arthur, Franck Gabriel, and Clément Hongler (2018). "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in Neural Information Processing Systems* 31.

Kavran, Andrew J and Aaron Clauset (2021). "Denoising large-scale biological data using network filters". In: *BMC Bioinformatics* 22, pp. 1–21.

Kim, Ki-Su et al. (2020). "Prediction of ocean weather based on denoising autoencoder and convolutional LSTM". In: *Journal of Marine Science and Engineering* 8.10, p. 805.

Kontorovich, Aryeh (2014). "Concentration in unbounded metric spaces and algorithmic stability". In: *International Conference on Machine Learning*. PMLR, pp. 28–36.

Kuntzman, Gersh (2022). *stuckatdot: It takes Years (and years) to get a speed hump in this city*. URL: https://nyc.streetsblog.org/2022/11/11/stuckatdot-it-takes-years-and-years-to-get-a-speed-hump-in-this-city/.

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

Liyanage, Liwan H and J George Shanthikumar (2005). "A practical inventory control policy using operational statistics". In: *Operations Research Letters* 33.4, pp. 341–348.

New York, City of (2023). *Motor Vehicle Collisions (Crashes) Dataset*. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.

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

Rabiner, Lawrence R and Bernard Gold (1975). "Theory and application of digital signal processing". In: *Englewood Cliffs: Prentice-Hall*.

Rahimi, Ali and Benjamin Recht (2008). "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning". In: *Advances in Neural Information Processing Systems* 21.

Seeger, Matthias (2004). "Gaussian processes for machine learning". In: *International journal of neural systems* 14.02, pp. 69–106.

Shalev-Shwartz, Shai et al. (2010). "Learnability, stability and uniform convergence". In: *The Journal of Machine Learning Research* 11, pp. 2635–2670.

Tian, Chunwei et al. (2020). "Deep learning on image denoising: An overview". In: *Neural Networks* 131, pp. 251–275.

Vershynin, Roman (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Vol. 47. Cambridge University Press.

Wahba, Grace (1990). *Spline models for observational data*. SIAM.

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

Wilder, Bryan, Bistra Dilkina, and Milind Tambe (2019). “Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 1658–1665.

Wilson, Kevin W et al. (2008). “Speech denoising using nonnegative matrix factorization with priors”. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 4029–4032.

Yu, Bin (2013). “Stability”. In: *Bernoulli*, pp. 1484–1500.

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

Online Appendix: Decision-Aware Denoising

Appendix A: Background Results

Recall, a mean-zero random variable X is subgaussian with variance proxy κ^2 if $\mathbb{E}\left[\exp(\lambda X)\right] \le \exp\left(\frac{\lambda^2\kappa^2}{2}\right)$. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a function. We say that f is 1-Lipschitz with respect to the $\boldsymbol{c}$-weighted ℓ_1 norm if

$$|f(x_1,\ldots,x_n) - f(y_1,\ldots,y_n)| \le \sum_{i=1}^{n} c_i \left|x_i - y_i\right|,$$

for some $\boldsymbol{c} \in \mathbb{R}^n_+$. Kontorovich (2014) proved that such functions applied to subGaussian random variables concentrate.

Theorem A.1 (Kontrovich 2014) *Let $\boldsymbol{X} = (X_1,\ldots,X_n)$ be a vector of independent random variables such that $X_i - \mathbb{E}\left[X_i\right]$ is subGaussian with variance proxy κ_i^2. Suppose $f(\cdot)$ is 1-Lipschitz with respecct to the $\boldsymbol{c}$-weighted ℓ_1 norm. Then, there exists a universal constant C such that for any $0 < \epsilon < \frac{1}{2}$, with probability at least $1-\epsilon$,*

$$|f(\boldsymbol{X}) - \mathbb{E}\left[f(\boldsymbol{X})\right]| \;\le\; C\sqrt{\sum_{i=1}^{n} c_i^2 \kappa_i^2} \cdot \sqrt{\log(2/\epsilon)}.$$

Remark 1. The original result in Kontorovich, 2014 is stated with respect the sub-Gaussian diameters of the X_i. Note however that if $X_i - \mathbb{E}\left[X_i\right]$ is sub-Gaussian with variance proxy κ_i^2, then X_i has sub-Gaussian diameter at most a constant times κ_i.

Appendix B: Section 3 Proofs

B.1. Auxilliary Lemmas

Many of our results leverage that the function $V(\boldsymbol{z},\rho) = \min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{z}^\top \boldsymbol{x} + \rho \sum_{j=1}^{n} \phi(x_j)$ is Lipschitz.

Lemma B.1 (Lipschitz Bounds on V) *Let $V(\boldsymbol{z},\rho) = \min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{z}^\top \boldsymbol{x} + \rho \sum_{j=1}^{n} \phi(x_j)$ and $\mathcal{X} \subseteq [0,1]^n$. Then, the following holds,*

i) $|V(\boldsymbol{z},\rho) - V(\boldsymbol{y},\rho)| \;\le\; \|\boldsymbol{z}-\boldsymbol{y}\|_1$.

ii) $|V(\boldsymbol{z},\rho) - V(\boldsymbol{z},\rho')| \;\le\; n\phi_{\max}\left|\rho-\rho'\right|$, *where* $\phi_{\max} = \sup_{x\in[0,1]}|\phi(x)|$

Proof of Lemma B.1:

Part i) We first upper-bound the difference. We see,

$$\begin{aligned} V(\boldsymbol{z},\rho) - V(\boldsymbol{y},\rho) \;&=\; \underbrace{\boldsymbol{z}^\top \hat{\boldsymbol{x}}(\boldsymbol{z},\rho) + \rho\sum_{j=1}^{n}\phi(\hat{x}_j(\boldsymbol{z},\rho)) - \boldsymbol{z}^\top \hat{\boldsymbol{x}}(\boldsymbol{y},\rho) - \rho\sum_{j=1}^{n}\phi(\hat{x}_j(\boldsymbol{y},\rho))}_{\le 0,\text{ by optimality of } \hat{\boldsymbol{x}}(\boldsymbol{z},\rho)} + (\boldsymbol{z}-\boldsymbol{y})^\top \hat{\boldsymbol{x}}(\boldsymbol{y},\rho) \\ &\le \|\boldsymbol{z}-\boldsymbol{y}\|_1 \|\hat{\boldsymbol{x}}(\boldsymbol{y},\rho)\|_\infty \;\le\; \|\boldsymbol{z}-\boldsymbol{y}\|_1 \end{aligned}$$

where the first inequality holds by Holder's inequality and the second inequality holds because $\|\hat{\boldsymbol{x}}(\boldsymbol{y},\rho)\|_\infty \le 1$. By symmetry, this completes the proof of Part i).

Part ii) We apply a similar argument. Letting $\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{j=1}^n \phi(x_j)$,

$$V(\boldsymbol{z},\rho) - V(\boldsymbol{z},\rho') \;=\; \underbrace{\boldsymbol{z}^\top \hat{\boldsymbol{x}}(\boldsymbol{z},\rho) + \rho\boldsymbol{\phi}(\hat{\boldsymbol{x}}(\boldsymbol{z},\rho)) - \boldsymbol{z}^\top \hat{\boldsymbol{x}}(\boldsymbol{z},\rho') - \rho\boldsymbol{\phi}(\hat{\boldsymbol{x}}(\boldsymbol{z},\rho'))}_{\le 0,\text{ by optimality of } \hat{\boldsymbol{x}}(\boldsymbol{z},\rho)} + (\rho-\rho')\boldsymbol{\phi}(\hat{\boldsymbol{x}}(\boldsymbol{z},\rho'))$$
$$\le\; |\rho-\rho'| \sum_{j=1}^n \sup_{x\in[0,1]} |\phi(x)| \;=\; n\phi_{\max}\,|\rho-\rho'|\,.$$

By symmetry, we can also bound $V(\boldsymbol{z},\rho') - V(\boldsymbol{z},\rho)$, completing the proof. □

Since $V(\cdot)$ is 1-Lipschitz relative to $\|\cdot\|_1$, we can bound the expected error introduced by proxying $\boldsymbol{\xi}$ by δ_h in our proof:

Lemma B.2 (Distribution Replication Bound) *The following holds,*

$$|\mathbb{E}\left[V(\boldsymbol{T}(\boldsymbol{Z}) + h\boldsymbol{\xi}) - V(\boldsymbol{T}(\boldsymbol{Z}) + \boldsymbol{\delta}_h)\right]| \;\le\; \mathbb{W}_1\left(\boldsymbol{T}(\boldsymbol{Z}) + h\boldsymbol{\xi},\, \boldsymbol{T}(\boldsymbol{Z}) + \boldsymbol{\delta}_h\right).$$

Proof of Lemma B.2: The result is immediate given Lemma B.1 and integral-probability metric representation of the Wasserstein distance (Wainwright, 2019, pg. 76).

□

B.2. Proof for Theorem 3.2

We fix $\boldsymbol{\theta}$ througout and drop it from the notation. Recall $\boldsymbol{Z} = \boldsymbol{\xi} + \boldsymbol{\mu}$. Let $\bar{\boldsymbol{\xi}}$ be an i.i.d. copy of $\boldsymbol{\xi}$. Let $\boldsymbol{\xi}^k$ denote the vector $\boldsymbol{\xi}$ with element k replaced with $\bar{\xi}_k$. By the Efron-Stein Inequality,

$$\mathbb{V}\mathrm{ar}(D(\boldsymbol{Z})) \le \frac{1}{2}\sum_{k=1}^n \mathbb{E}\left[\left(D(\boldsymbol{\xi}+\boldsymbol{\mu}) - D(\boldsymbol{\xi}^k+\boldsymbol{\mu})\right)^2\right]$$

Focusing on the term inside the square, we see

$$\begin{aligned}\left|D(\boldsymbol{\xi}+\boldsymbol{\mu}) - D(\boldsymbol{\xi}^k+\boldsymbol{\mu})\right| \le\;& \frac{1}{h}\left|\mathbb{E}\left[V(\boldsymbol{T}(\boldsymbol{\xi}+\boldsymbol{\mu}) + \boldsymbol{\delta}_h) - V(\boldsymbol{T}(\boldsymbol{\xi}^k+\boldsymbol{\mu}) + \boldsymbol{\delta}_h)\,\middle|\,\boldsymbol{\xi},\boldsymbol{\xi}^k\right]\right| \\ &+ \frac{1}{h}\left|V(\boldsymbol{T}(\boldsymbol{\xi}+\boldsymbol{\mu})) - V(\boldsymbol{T}(\boldsymbol{\xi}^k+\boldsymbol{\mu}))\right| \\ \le\;& \frac{2}{h}\|\boldsymbol{T}(\boldsymbol{\xi}+\boldsymbol{\mu}) - \boldsymbol{T}(\boldsymbol{\xi}^k+\boldsymbol{\mu})\|_1 \\ \le\;& \frac{2}{h}C_{\boldsymbol{T}}\|\boldsymbol{\xi} - \boldsymbol{\xi}^k\|_1 \\ \le\;& \frac{2}{h}C_{\boldsymbol{T}}\left|\xi_k - \bar{\xi}_k\right|.\end{aligned}$$

The second inequality holds by Lemma B.1 and the third holds by Assumption 3.1.

Plugging in the upper-bound,

$$\mathbb{E}\left[\left(D(\boldsymbol{\xi}+\boldsymbol{\mu}) - D(\boldsymbol{\xi}^k+\boldsymbol{\mu})\right)^2\right] \;\le\; \mathbb{E}\left[\frac{4}{h^2}\left|\xi_k - \bar{\xi}_k\right|^2 C_{\boldsymbol{T}}^2\right] \;\le\; \frac{8C_{\boldsymbol{T}}^2}{\nu_k h^2}$$

Letting $\nu_{\min} = \min_k \nu_k$, we see,

$$\mathbb{V}\text{ar}(D(\boldsymbol{Z})) \leq \frac{4nC_{\boldsymbol{T}}^2}{\nu_{\min} h^2}$$

This completes the proof.

B.3. Proof for Theorem 3.4

We leverage Theorem A.1 on the function $D(\boldsymbol{Z})$. Let $\boldsymbol{Z}^k$ be identical to $\boldsymbol{Z}$ except possibly in the k^{th} component. Following the same steps of the proof of Theorem 3.2, we have

$$\left|D(\boldsymbol{Z}) - D(\boldsymbol{Z}^k)\right| \leq \frac{2}{h}\left|Z_k - \bar{Z}_k\right| C_{\boldsymbol{T}},$$

We take $c_i = \frac{2}{h} C_{\boldsymbol{T}}$ and $\kappa(Z_i) \leq \kappa$.

Applying Theorem A.1 yields the result.

B.4. The One Shot and Multi-Shot VGC under Decoupling

In Section 3.3, we showed that in special cases where both corrections apply, the multi-shot VGC can be seen as a special case of the one-shot VGC. We next show that when the optimization problem fully decouples, so that $\hat{x}_j(\boldsymbol{t}, 0)$ only depends on t_j but not t_k for $k \neq j$, for all j, then the one-shot VGC is equivalent to the multi-shot VGC.

Proposition B.3 (One-Shot vs. Multi-Shot VGC for Decoupled Problems) *Suppose that $T(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) = \boldsymbol{Z} + \ell(\bar{\boldsymbol{\theta}})$ for some $\ell(\cdot)$, $\rho = 0$, and $Z_j \sim \mathcal{N}(\mu_j, 1)$ for each j (not necessarily independent). Let $\boldsymbol{\delta}$ be any random variable such that $\delta_j \sim \mathcal{N}(0, h^2 + 2h)$. Finally, assume that $\hat{x}_j(\boldsymbol{t}, 0)$ only depends on t_j for each j. Then, the one-shot VGC with stepsize h and perturbation $\boldsymbol{\delta}$ is equivalent ot the multi-shot VGC with step size h.*

Proof. We make two observations necessary for the proof: First, since $\hat{x}_j(\boldsymbol{t}, 0)$ only depends on t_j, $\hat{x}_j(\boldsymbol{t}, 0) = \hat{x}(t_j \boldsymbol{e}_j, 0)$ for each j. Second, $V(\boldsymbol{t} + \delta_j \boldsymbol{e}_j) - V(\boldsymbol{t}) = (t_j + \delta_j)\hat{x}((t_j + \delta_j)\boldsymbol{e}_j) - t_j \hat{x}(t_j \boldsymbol{e}_j)$ since all the terms $k \neq j$ in the two sums are identical and drop out.

Then, starting with the one-shot VGC, and suppressing $\boldsymbol{\theta}$ in the notation,

$$\begin{aligned}
D(\boldsymbol{Z}) &\equiv \frac{1}{h}\mathbb{E}\left[V(\boldsymbol{Z} + \boldsymbol{\ell} + \boldsymbol{\delta}) - V(\boldsymbol{Z} + \boldsymbol{\ell}) \mid \boldsymbol{Z}\right] \\
&= \frac{1}{h}\sum_{j=1}^{n} \mathbb{E}\left[(Z_j + \ell_j + \delta_j)\hat{x}_j(\boldsymbol{Z} + \boldsymbol{\ell} + \boldsymbol{\delta}) - (Z_j + \ell_j)\hat{x}_j(\boldsymbol{Z} + \boldsymbol{\ell}) \mid \boldsymbol{Z}\right] \\
&= \frac{1}{h}\sum_{j=1}^{n} \mathbb{E}\left[(Z_j + \ell_j + \delta_j)\hat{x}_j(\boldsymbol{Z} + \boldsymbol{\ell} + \delta_j \boldsymbol{e}_j) - (Z_j + \ell_j)\hat{x}_j(\boldsymbol{Z} + \boldsymbol{\ell}) \mid \boldsymbol{Z}\right] \\
&= \frac{1}{h}\sum_{j=1}^{n} \mathbb{E}\left[V(\boldsymbol{Z} + \boldsymbol{\ell} + \delta_j \boldsymbol{e}_j) - V(\boldsymbol{Z} + \boldsymbol{\ell}) \mid \boldsymbol{Z}\right],
\end{aligned}$$

where the second-to-last and last equalities use our two observations from the beginning of the proof. Recognizing the definition of the multi-shot VGC completes the proof. □

Appendix C: Section 4 Proofs

C.1. Proof for Theorem 4.3

Recall from Eq. (12), the estimation error can be decomposed into the following two components:

$$\sup_{\boldsymbol{\theta}\in\Theta}\underbrace{\left|\boldsymbol{Z}^\top\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})-D(\boldsymbol{Z},\boldsymbol{\theta})-\boldsymbol{\mu}^\top\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}))\right|}_{\text{Estimation Error}}$$

$$\leq\sup_{\boldsymbol{\theta}\in\Theta}\underbrace{\left|\boldsymbol{\xi}^\top\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})-\frac{1}{h}\left(V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho)\right)\right|}_{\text{Finite Difference Error}}$$

$$+\sup_{\boldsymbol{\theta}\in\Theta}\underbrace{\left|\frac{1}{h}\left(V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho)\right)-D(\boldsymbol{Z},\boldsymbol{\theta})\right|}_{\text{Replication Error}}.$$

We bound the finite difference error by solution stability in the main text (c.f. Eq. (13)). Hence, we focus on on bounding the replication error.

To bound replication error, first note that by definition,

$$\left|\frac{V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho)}{h}-D(\boldsymbol{Z},\boldsymbol{\theta},h)\right|=\left|\frac{V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)|\boldsymbol{Z}]}{h}\right|.$$

Applying the triangle inequality to the right side for a fixed $\boldsymbol{\theta}$,

$$\begin{aligned}\text{Replication Error}_{\boldsymbol{\theta}}\leq\ &\underbrace{\left|\frac{V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)]}{h}\right|}_{\text{(i)}}\\&+\underbrace{\left|\frac{\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)|\boldsymbol{Z}]-\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)]}{h}\right|}_{\text{(ii)}}\\&+\underbrace{\left|\frac{\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)]}{h}\right|}_{\text{(iii)}}\end{aligned}$$

Bounding Component (i). We first fix $\boldsymbol{\theta}=(\bar{\boldsymbol{\theta}},\rho)$ and prove a point-wise bound. We will apply Theorem A.1. To that end, define $\boldsymbol{\xi}^k=\left(\xi_1,\ldots,\bar{\xi}_k,\ldots,\xi_n\right)^\top$ and $\boldsymbol{Z}^k=\left(Z_1,\ldots,\bar{Z}_k,\ldots,Z_n\right)^\top$ where $\bar{\xi}_k$ and $\bar{Z}_k$ are i.i.d. copies of ξ_k and Z_k, respectively. Then,

$$\begin{aligned}\left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-V(\boldsymbol{T}(\boldsymbol{Z}^k,\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi}^k,\rho)\right|&\leq\|\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi}-\boldsymbol{T}(\boldsymbol{Z}^k,\bar{\boldsymbol{\theta}})-h\boldsymbol{\xi}^k\|_1\\&\leq(C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}})+h)\|\boldsymbol{\xi}-\boldsymbol{\xi}^k\|_1\\&\leq(C_{\boldsymbol{T}}+h)\left|\xi_k-\bar{\xi}_k\right|,\end{aligned}$$

where $C_{\boldsymbol{T}}=\max_{\bar{\boldsymbol{\theta}}}C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}})$. The first inequality follows from Lemma B.1 and the second inequality follows from Assumption 3.1. Since ξ_k are sub-Gaussian, we can then apply Theorem A.1 with $c_k=C_{\boldsymbol{T}}+h$ to show the following holds with probability at least $1-\epsilon$,

$$\left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)-\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi},\rho)]\right|\leq C\cdot(C_{\boldsymbol{T}}+h)\,\kappa\sqrt{n\log\left(\frac{2}{\epsilon}\right)}\tag{20}$$

for a fixed $\boldsymbol{\theta}$ and universal constant C.

Now consider the covering $\Theta_0(\varepsilon)$. Letting $\boldsymbol{\theta}_0$ be the element in $\Theta_0(n^{-1/2})$ closest to $\boldsymbol{\theta}$, we have that for any $\boldsymbol{t}_h \in \mathbb{R}^n$,

$$\begin{aligned}
&\left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{t}_h,\rho)-\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{t}_h,\rho)]\right| \\
&\qquad \le \sup_{\boldsymbol{\theta}\in\Theta_0(n^{-1/2})} \left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{t}_h,\rho)-\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{t}_h,\rho)]\right| && \text{(21a)} \\
&\qquad\quad + \left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{t}_h,\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)+\boldsymbol{t}_h,\rho_0)\right| && \text{(21b)} \\
&\qquad\quad + \left|\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{t}_h,\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)+\boldsymbol{t}_h,\rho_0)]\right|. && \text{(21c)}
\end{aligned}$$

Equation (20) and the union bound show that for $\boldsymbol{t}_h = h\boldsymbol{\xi}$ that with probability at least $1-\epsilon$

$$\text{Eq. (21a)} \le C(C_{\boldsymbol{T}}+h)\cdot\kappa\sqrt{n\log\left(\left|\Theta_0\left(n^{-1/2}\right)\right|\right)\log\left(\frac{2}{\epsilon}\right)},$$

where $|\Theta_0(n^{-1/2})|$ is the cardinality of $\Theta_0(n^{-1/2})$.

To bound Eq. (21b), recall that $\boldsymbol{\theta}=(\bar{\boldsymbol{\theta}},\rho)$ and $\boldsymbol{\theta}_0=(\bar{\boldsymbol{\theta}}_0,\rho_0)$, so with probability at least $1-\epsilon$,

$$\begin{aligned}
\left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{t}_h,\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)+\boldsymbol{t}_h,\rho_0)\right| &\le \left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{t}_h,\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)+\boldsymbol{t}_h,\rho)\right| \\
&\quad + \left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)+\boldsymbol{t}_h,\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)+\boldsymbol{t}_h,\rho_0)\right| \\
&\overset{(a)}{\le} \|\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})-\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)\|_1 + n\phi_{\max}|\rho-\rho_0| \\
&\overset{(b)}{\le} n^{-1/2}C\left(\|\boldsymbol{Z}\|_1+1\right)+n^{-1/2}\cdot n\phi_{\max} \\
&\overset{(c)}{\le} \frac{C}{\sqrt{n}}\left(\kappa\sqrt{n\log\left(\frac{1}{\epsilon}\right)}+n\left(\nu_{\min}^{-1/2}+\|\boldsymbol{\mu}\|_\infty+\phi_{\max}\right)+1\right) \\
&\overset{(d)}{\le} C_1\kappa\sqrt{n\log\left(\frac{1}{\epsilon}\right)}
\end{aligned}$$

where $\phi_{\max}=\max_{x\in[0,1]}\phi(x)$ and $C_1 = 4C\left(\nu_{\min}^{1/2}+\|\boldsymbol{\mu}\|_\infty+\phi_{\max}+1\right)$. We see (a) holds by the Lipschitz results of V in Lemma B.1, (b) holds by applying Assumption 4.2 with $\epsilon = n^{-1/2}$ and universal constant $C\ge 1$, (c) holds since $C \ge 1$ and

$$\|\boldsymbol{Z}\|_1 \le \|\boldsymbol{\xi}\|_1+\|\boldsymbol{\mu}\|_1 \le (\|\boldsymbol{\xi}\|_1-\mathbb{E}\|\boldsymbol{\xi}\|_1)+\mathbb{E}\|\boldsymbol{\xi}\|_1+n\|\boldsymbol{\mu}\|_\infty \le \kappa\sqrt{n\log\left(\frac{1}{\epsilon}\right)}+n\nu_{\min}^{1/2}+n\|\boldsymbol{\mu}\|_\infty$$

where the last inequality applies Theorem A.1 to $\|\boldsymbol{\xi}\|_1-\mathbb{E}\|\boldsymbol{\xi}\|_1$ with $c_j = 1$ for all j and that $\mathbb{E}|\xi_j| \le \mathbb{E}\left|\sqrt{\xi_j^2}\right| \le \sqrt{\mathbb{E}\xi_j^2} \le \nu_{\min}^{-1/2}$. For the former, we see $c_j = 1$ since $\|\boldsymbol{\xi}-\boldsymbol{\xi}^k\|_1 = \left|\xi_k-\bar{\xi}_k\right|$. Inequality (d) holds for $\epsilon < 1/e$ and noting $a+b\le 2ab$ for $a\ge 1, b\ge 1$.

A similar argument can be applied to bound Eq. (21c),

$$\begin{aligned}
\mathbb{E}\left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{t}_h,\rho)-V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)+\boldsymbol{t}_h,\rho_0)\right| &\le \mathbb{E}\left[n^{-1/2}C_\Theta\left(\|\boldsymbol{Z}\|_1+1\right)+n^{-1/2}\cdot n\phi_{\max}\right] \\
&\le C_\Theta n^{-1/2}\left(n\nu_{\min}^{1/2}+n\|\boldsymbol{\mu}\|_\infty+n\phi_{\max}+1\right) \\
&\le C_1\kappa\sqrt{n}
\end{aligned}$$

Combining Eq. (21a), Eq. (21b), and Eq. (21c), collecting terms and applying a supremum over θ, we have for some universal constant C,

$$\text{(i)} \;\leq\; C\frac{C_1(C_{\boldsymbol{T}}+h)\kappa}{h}\sqrt{n\cdot\log\left(\left|\Theta_0\left(n^{-1/2}\right)\right|\right)\cdot\log\left(\frac{2}{\epsilon}\right)}$$

Bounding Component (ii). We can make an argument similar to (i). By Jensen's inequality,

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta}\left|\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)|\boldsymbol{Z}]-\mathbb{E}[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)]\right|\\
&\quad=\sup_{\boldsymbol{\theta}\in\Theta}\left|\mathbb{E}\left[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)-\mathbb{E}\left[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)\,\middle|\,\boldsymbol{\delta}_h\right]\,\middle|\,\boldsymbol{Z}\right]\right|\\
&\quad\leq\mathbb{E}\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)-\mathbb{E}\left[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)\,\middle|\,\boldsymbol{\delta}_h\right]\right|\,\middle|\,\boldsymbol{Z}\right]
\end{aligned}$$

We then bound the term inside the expectation. For a fixed $\boldsymbol{\delta}_h$, we apply the bounds to Eq. (21) for $\boldsymbol{t}_h=\boldsymbol{\delta}_h$. We only need to adjust the bound to Eq. (21a) since it depends on the choice $\boldsymbol{t}_h$. Note that,

$$\begin{aligned}
\left|\mathbb{E}\left[V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)\,\middle|\,\boldsymbol{Z}\right]-\mathbb{E}\left[V(\boldsymbol{T}(\boldsymbol{Z}^k,\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h,\rho)\,\middle|\,\boldsymbol{Z}^k\right]\right| &\leq \left\|\mathbb{E}\left[\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h-\boldsymbol{T}(\boldsymbol{Z}^k,\bar{\boldsymbol{\theta}})-\boldsymbol{\delta}_h\,\middle|\,\boldsymbol{Z},\bar{Z}_k\right]\right\|_1\\
&\leq C_{\boldsymbol{T}}(\bar{\boldsymbol{\theta}})\|\boldsymbol{\xi}-\boldsymbol{\xi}^k\|_1\\
&\leq C_{\boldsymbol{T}}\left|\xi_k-\bar{\xi}_k\right|,
\end{aligned}$$

Thus, for $\boldsymbol{t}_h=\boldsymbol{\delta}_h$ and applying Theorem A.1 with $c_k=C_{\boldsymbol{T}}$, we have

$$\text{Eq. (21a)}\leq CC_{\boldsymbol{T}}\cdot\kappa\sqrt{n\log\left(\left|\Theta_0\left(n^{-1/2}\right)\right|\right)\log\left(\frac{2}{\epsilon}\right)}.$$

Since the bounds for Eq. (21b) and Eq. (21c) do not depend on $\boldsymbol{t}_h$, we have

$$\text{(ii)} \;\leq\; C\frac{C_1 C_{\boldsymbol{T}}\kappa}{h}\sqrt{n\cdot\log\left(\left|\Theta_0\left(n^{-1/2}\right)\right|\right)\cdot\log\left(\frac{2}{\epsilon}\right)}$$

Bounding Component (iii). Component (iii) is bounded by Lemma B.2.

Combining (i), (ii), and (iii), we prove the result.

C.2. Proof for Proposition 4.4.

First observe that since $h>0$ and $\boldsymbol{L}(\bar{\boldsymbol{\theta}})\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\boldsymbol{L}(\bar{\boldsymbol{\theta}})^\top$ is positive semidefinite, the matrix $h(\boldsymbol{L}(\bar{\boldsymbol{\theta}})\boldsymbol{\Sigma}+\boldsymbol{\Sigma}\boldsymbol{L}(\bar{\boldsymbol{\theta}})^\top)+h^2\boldsymbol{\Sigma})$ is also positive semidefinite and $\boldsymbol{\delta}_h$ is well-defined.

Since the sum of two multivariate Gaussians is also a multivariate Gaussian, we see both $\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi}$ and $\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h$ are distributed as $\mathcal{N}(\boldsymbol{\mu},(\boldsymbol{L}+h\boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{L}+h\boldsymbol{I})^\top)$. Thus, their 1-Wasserstein distance is 0.

Appendix D: Continuous Policy Proofs

D.1. Proof for Lemma 5.1

We first establish the following well-known result:

Lemma D.1 (Regularized Plug-in Policies are ρ-Lipschitz) *When $\mathcal{X}$ is convex and $\rho > 0$, $\|\hat{\boldsymbol{x}}(\boldsymbol{t},\rho) - \hat{\boldsymbol{x}}(\boldsymbol{t}',\rho)\| \le \frac{1}{\rho}\|\boldsymbol{t}-\boldsymbol{t}'\|_*$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.*

Proof. Let $h_\rho(\boldsymbol{x},\boldsymbol{t}) \equiv \sum_{j=1}^n t_j x_j + \rho\phi(x_j)$. Since $h_\rho(\boldsymbol{x},\boldsymbol{t})$ is strongly convex over $\mathcal{X}$ and by the optimality of $\hat{\boldsymbol{x}}(\boldsymbol{t},\rho)$ and $\hat{\boldsymbol{x}}(\boldsymbol{y},\rho)$, we see

$$h_\rho(\hat{\boldsymbol{x}}(\boldsymbol{t},\rho),\boldsymbol{y}) - h_\rho(\hat{\boldsymbol{x}}(\boldsymbol{y},\rho),\boldsymbol{y}) \ge \frac{\rho}{2}\left\|\hat{\boldsymbol{x}}(\boldsymbol{t},\rho) - \hat{\boldsymbol{x}}(\boldsymbol{y},\rho)\right\|^2$$

$$h_\rho(\hat{\boldsymbol{x}}(\boldsymbol{y},\rho),\boldsymbol{t}) - h_\rho(\hat{\boldsymbol{x}}(\boldsymbol{t},\rho),\boldsymbol{t}) \ge \frac{\rho}{2}\left\|\hat{\boldsymbol{x}}(\boldsymbol{t},\rho) - \hat{\boldsymbol{x}}(\boldsymbol{y},\rho)\right\|^2$$

Adding the two inequalities, we see

$$\begin{aligned}
\rho\left\|\hat{\boldsymbol{x}}(\boldsymbol{t},\rho) - \hat{\boldsymbol{x}}(\boldsymbol{y},\rho)\right\|^2 &\le h_\rho(\hat{\boldsymbol{x}}(\boldsymbol{t},\rho),\boldsymbol{y}) - h_\rho(\hat{\boldsymbol{x}}(\boldsymbol{y},\rho),\boldsymbol{y}) + h_\rho(\boldsymbol{x}(\boldsymbol{y},\rho),\boldsymbol{t}) - h_\rho(\boldsymbol{x}(\boldsymbol{z},\rho),\boldsymbol{t}) \\
&= \sum_{j=1}^n y_j(\hat{x}_j(\boldsymbol{t},\rho) - \hat{x}_j(\boldsymbol{y},\rho)) + t_j(\hat{x}_j(\boldsymbol{y},\rho) - \hat{x}_j(\boldsymbol{t},\rho)) \\
&= \sum_{j=1}^n (y_j - t_j)(\hat{x}_j(\boldsymbol{t},\rho) - \hat{x}_j(\boldsymbol{y},\rho)) \\
&\le \|\boldsymbol{y}-\boldsymbol{t}\|_*\left\|\hat{\boldsymbol{x}}(\boldsymbol{t},\rho) - \hat{\boldsymbol{x}}(\boldsymbol{y},\rho)\right\|,
\end{aligned}$$

where the last line holds by Holder's inequality. Rearranging, we get our intended result.

We can now simply prove Lemma 5.1.

Proof of Lemma 5.1. The result follows directly from Cauchy-Schwarz since

$$\mathsf{SS}(\boldsymbol{\xi},h,\boldsymbol{\theta}) \;\le\; \|\boldsymbol{\xi}\|_*\|\hat{\boldsymbol{x}}(T(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho) - \hat{\boldsymbol{x}}(T(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) + h\boldsymbol{\xi},\rho)\| \;\le\; \frac{h}{\rho}\|\boldsymbol{\xi}\|_*^2.$$

D.2. Proof for Theorem 5.2

The proof is immediate from combining Theorem 4.3 and Lemma 5.1 and collecting constants.

D.2.1. Proof Corollary 5.3.

Proof. First consider the case where $\phi(\cdot)$ is 1-strongly convex with respect to $\|\cdot\|_1$. To form a more explicit bound on the solution stability, we analyze $\|\boldsymbol{\xi}\|_*^2 = \|\boldsymbol{\xi}\|_\infty^2$. Since $\boldsymbol{\xi}$ is mean-zero, $\|\boldsymbol{\xi}\|_\infty = \|\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]\|_\infty$,. Furthermore, since $\xi_j - \mathbb{E}[\xi_j]$ is sub-Gaussian with variance proxy κ^2, we have that with probability at least $1-\epsilon$, there exists an absolute constant C_1 such that

$$|\xi_j - \mathbb{E}[\xi_j]| \le C_1\kappa\sqrt{\log(n/\epsilon)} \le C_1\kappa\sqrt{\log n \log(1/\epsilon)},$$

simultaneously for all j. Hence, with probability at least $1-\epsilon$,

$$\|\boldsymbol{\xi}\|_\infty^2 \ \le C_1^2 \kappa^2 \log n \log(1/\epsilon).$$

Next, we substitute this bound into Theorem 5.2 and optimize the choice of h. Recall, by assumption the Wasserstein contribution is zero. Hence, the terms depending on h are (neglecting constants)

$$\frac{h}{\rho_{\min}} \log n + \frac{C_{\boldsymbol{T}}}{h}\sqrt{n \log n}.$$

Choosing $h = \rho_{\min}^{1/2} n^{1/4}$ optimizes this quantity (up to logarithmic factors). Substituting back into the bound and collecting constants proves the first part of the corollary.

We next consider the case that $\phi(\cdot)$ is 1-strongly convex with respect to $\|\cdot\|_2$. To bound solution stability more explicitly, we bound $\|\boldsymbol{\xi}\|_*^2 = \|\boldsymbol{\xi}\|_2^2 = \sum_{j=1}^n \xi_j^2$. Since ξ_j is sub-Gaussian with parameter κ, Vershynin, 2018, Lemma 2.7.6 shows that there exists an absolute constant C_1 such that that $\|\xi_j^2\|_{\Psi_1} \le C_1 \kappa^2$. (Recall, for any random variable Y, $\|Y\|_{\Psi_1} \equiv \inf\{t > 0 : \mathbb{E}[\exp(|Y|/t)] \le 2\}$, c.f., Vershynin (2018, pg. 31).) Hence, by the triangle inequality, $\|\sum_{j=1}^n \xi_j^2\|_{\Psi_1} \le C_1 n \kappa^2$. Finally, by Markov's Inequality, with probability at least $1-\epsilon$, $\sum_{j=1}^n \xi_j^2 \ \le \ C_2 n \kappa^2 \log(1/\epsilon)$, for some absolute constant C_2.

Next we plug this upper bound into Theorem 5.2 and optimize the choie of h. Again, the Wasserstein contribution is zero. Neglecting constants, the remaining terms are

$$\frac{h}{\rho_{\min}} n + \frac{C_{\boldsymbol{T}}}{h}\sqrt{n \log n}.$$

Choosing $h = \rho_{\min}^{1/2} n^{-1/4}$ optimizes the boudn up to logarithmic factors. Substituting in and collecting terms completes the proof. □

D.3. Proof of Theorem 5.6.

Proof of Lemma 5.5. By Cauchy-Schwarz,

$$\begin{aligned}
|\mathsf{SS}(\boldsymbol{\xi}, h, \boldsymbol{\theta})| \ &\le \ \|\boldsymbol{\xi}\|_* \|\hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}), 0) - \hat{\boldsymbol{x}}(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + h\boldsymbol{\xi}, 0)\|_* \\
&\le \ \frac{2h\|\boldsymbol{\xi}\|_*^2\|}{\alpha\left(\|\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\|_* + \|\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + h\boldsymbol{\xi}\|_*\right)} \\
&\le \ \frac{2h\|\boldsymbol{\xi}\|_*^2\|}{\alpha\|\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\|_*}
\end{aligned}$$

This completes the proof. □

Proof of Theorem 5.6. The proof follows from substituting Lemma 5.5 into Theorem 4.3 and collecting terms. □

Appendix E: Proofs and Additional Materials for Block Decoupled Problems

E.1. What are typical values of $T_{\max}$ and $\mathcal{X}_{\max}$?

In this subsection, we revisit some examples from Section 2.3 to provide explicit bounds on the constants $T_{\max}$ and $\mathcal{X}_{\max}$.

Example E.1 (Clustering Revisited) Recall our clustering based policy class from Example 2.7. Let $\boldsymbol{x}^{CL}(\boldsymbol{Z},\bar{\theta})$ be the resulting policy when clustering into $\bar{\theta}$ clusters. We form a policy class by allowing the number of clusters to vary between $\bar{\theta}_{\min}$ and $\bar{\theta}_{\max}$.

Then, $T_{\max}$ corresponds to the size of the largest cluster. Hence, if $\bar{\theta}_{\min} = O(n)$, i.e., we only want to average over "local" neighborhoods, then $T_{\max} = O(1)$. Furthermore, the number of possible policies $\mathcal{X}_{\max}$ is bounded by the number of possible clusterings $\bar{\theta}_{\max} - \bar{\theta}_{\min}$ which is in turn at most n. Hence, in the typical case where $S_{\max} = 1$ and we constrain $\bar{\boldsymbol{\theta}}_{\min} = O(n)$, Corollary 6.7 bounds the error by a term of size $\tilde{O}_p(n^{3/4})$.

In the previous example, Θ is finite, since we can have at most n clusters. We next provide an example where Θ is not finite, but the induced set of plug-in policies is finite.

Example E.2 (Kernel Smoothers Revisited) Consider the kernel regression policy from Example 2.4 with the box-kernel with bandwidth $\bar{\theta}$, i.e., $K_{\bar{\theta}}(\boldsymbol{W}_j,\boldsymbol{W}_k) = \mathbb{I}\left\{\|\boldsymbol{W}_j - \boldsymbol{W}_k\| \le \bar{\theta}\right\}$. Let $\boldsymbol{x}^{KR}(\boldsymbol{Z},\bar{\theta})$ be the corresponding plug-in policy. Consider the policy class $\mathcal{X}^{KR}(\boldsymbol{Z}) = \{\boldsymbol{x}^{KR}(\boldsymbol{Z},\bar{\theta}) : 0 \le \bar{\theta} \le \bar{\theta}_{\max}\}$.

Then,

$$T_{\max} = \max_j \sum_{i=1}^n \max_{\bar{\theta}\in[0,\bar{\theta}_{\max}]} \mathbb{I}\left\{K_{\bar{\theta}}\left(\boldsymbol{W}_j,\boldsymbol{W}_i\right) \neq 0\right\} = \max_j \sum_{i=1}^n \mathbb{I}\left\{\|\boldsymbol{W}_j - \boldsymbol{W}_i\| \le \bar{\theta}_{\max}\right\},$$

which decreases with $\bar{\theta}_{\max}$. In other words, for small enough $\bar{\theta}_{\max}$, i.e., "local", neighborhoods, $T_{\max}$ can be made $O(1)$.

That said, $[0,\bar{\theta}]$ is an infinite set. However, the set of kernel weights induced by $\theta \in [0,\bar{\theta}]$ is finite. Namely, partition the real number line at the $O(n^2)$ points $\|\boldsymbol{W}_i - \boldsymbol{W}_j\|$. Then, for any θ_1,θ_2 in the same interval, $K_{\theta_1}(\boldsymbol{W}_i,\boldsymbol{W}_j) = K_{\theta_2}(\boldsymbol{W}_i,\boldsymbol{W}_j)$. Letting $\mathcal{T}_\Theta^{KR}(\boldsymbol{Z}) = \{\boldsymbol{T}^{KR}(\boldsymbol{Z},\theta) : \theta\in\Theta\}$, we see $|\mathcal{T}_\Theta^{KR}(\boldsymbol{Z})| \le n(n+1)/2+1$ by counting the number of intervals. We can then bound $\mathcal{X}_{\max}$ as follows,

$$|\mathcal{X}_\Theta^{KR}(\boldsymbol{Z})| = \left|\left\{\hat{\boldsymbol{x}}(\boldsymbol{t}) : \boldsymbol{t}\in\mathcal{T}_\Theta^{KR}(\boldsymbol{Z})\right\}\right| \le |\mathcal{T}_\Theta^{KR}(\boldsymbol{Z})| \le \frac{n(n+1)}{2}+1.$$

In settings where neither Θ or $\mathcal{T}_\Theta(\boldsymbol{Z})$ are finite, Gupta, Huang, and Rusmevichientong (2022a) describe how to bound $\mathcal{X}_{\max}$ under a general assumption on the plug-in class. Roughly, they show that $\|\mathcal{X}_\Theta(\boldsymbol{Z})\| = O(n^{\dim(\Theta)})$ for separable affine plug-in classes, or in words, they show $\mathcal{X}_{\max}$ is polynomial in n if the dimension of $\boldsymbol{\theta}\in\Theta$ is fixed. Similar techniques can be applied to bound the affine plug-ins considered in our paper. We omit the details.

E.2. Proof for Bounding Solution Stability, Lemma 6.5

We bound the solution stability by decomposing it into the following components,

$$\begin{aligned}
&\sup_{\boldsymbol{\theta}\in\Theta} \left|\boldsymbol{\xi}^\top\Big(\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\big) - \hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi}\big)\Big)\right| \\
&\quad\le \sup_{\bar{\boldsymbol{\theta}}\in\Theta} \left|\mathbb{E}\left[\boldsymbol{\xi}^\top\Big(\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\big) - \hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi}\big)\Big)\right]\right| \\
&\qquad + \sup_{\boldsymbol{\theta}\in\Theta} \left|\boldsymbol{\xi}^\top\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\big) - \mathbb{E}\left[\boldsymbol{\xi}^\top\boldsymbol{x}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\big)\right]\right| \\
&\qquad + \sup_{\boldsymbol{\theta}\in\Theta} \left|\boldsymbol{\xi}^\top\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi}\big) - \mathbb{E}\left[\boldsymbol{\xi}^\top\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi}\big)\right]\right| \\
&\quad\le \sup_{\boldsymbol{\theta}\in\Theta} \left|\mathbb{E}\left[\boldsymbol{\xi}^\top\Big(\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\big) - \hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi}\big)\Big)\right]\right| \qquad (22a) \\
&\qquad + 2\sup_{\substack{\boldsymbol{\theta}\in\Theta \\ h'\in\{0,h\}}} \left|\boldsymbol{\xi}^\top\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h'\boldsymbol{\xi}\big) - \mathbb{E}\left[\boldsymbol{\xi}^\top\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h'\boldsymbol{\xi}\big)\right]\right| \qquad (22b)
\end{aligned}$$

We bound Eq. (22a) in Lemma E.3 and bound Eq. (22b) by invoking Lemma E.5 with $\kappa = 1/\sqrt{\nu_{\min}}$ and using the crude bound on the chromatic number given after the theorem. Combining the two bounds gives us our bound on the solution stability.

E.3. Bound on Eq. (22a)

Lemma E.3 (Expected Solution Stability) *Suppose Assumptions 6.2 to 6.4 hold. Then, for any $0 \le h \le S_{\max}^{-1}$, there exists a constant C (depending on $\nu_{\min}, \nu_{\max}, \sigma_{\min}$) such that,*

$$\sup_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}} \left|\mathbb{E}\left[\boldsymbol{\xi}^\top\Big(\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\big) - \hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h\boldsymbol{\xi}\big)\Big)\right]\right| \le CS_{\max} nh\sqrt{\log\left(\frac{1}{h}\right)}$$

Proof for Lemma E.3. Fix $\bar{\boldsymbol{\theta}}$ and suppress it in the notation. Rewrite the argument of the supremum as a sum over blocks S_k:

$$\begin{aligned}
&\left|\sum_{k=1}^K \mathbb{E}\left[\left(\boldsymbol{\xi}^k\right)^\top\left(\hat{\boldsymbol{x}}^k\big(\boldsymbol{T}(\boldsymbol{Z})\big) - \hat{\boldsymbol{x}}^k\big(\boldsymbol{T}(\boldsymbol{Z})+h\boldsymbol{\xi}\big)\right)\right]\right| \\
&\qquad = \left|\sum_{k=1}^K \mathbb{E}\left[\left(\boldsymbol{\xi}^k\right)^\top\left(\hat{\boldsymbol{x}}^k\big(\boldsymbol{T}^k(\boldsymbol{Z})\big) - \hat{\boldsymbol{x}}^k\big(\boldsymbol{T}^k(\boldsymbol{Z})+h\boldsymbol{\xi}^k\big)\right)\right]\right| \\
&\qquad = \left|\sum_{k=1}^K \mathbb{E}\left[\mathbb{E}\left[\left(\boldsymbol{\xi}^k\right)^\top\left(\hat{\boldsymbol{x}}^k\big(\boldsymbol{T}^k(\boldsymbol{Z})\big) - \hat{\boldsymbol{x}}^k\big(\boldsymbol{T}^k(\boldsymbol{Z})+h\boldsymbol{\xi}^k\big)\right)\middle|\,\overline{\boldsymbol{Z}^k}\right]\right]\right|
\end{aligned}$$

where $\overline{\boldsymbol{Z}^k} \in \mathbb{R}^{n-|S_k|}$ correspond to the components of $\boldsymbol{Z}$ *not* in S_k. The first equality holds since the k^{th} block only depends on the objective cost components in the k^{th} block and the second by law of iterated expectations.

We now focus on bounding the conditional expectation

$$\mathbb{E}\left[\left(\boldsymbol{\xi}^k\right)^\top\left(\hat{\boldsymbol{x}}^k\big(\boldsymbol{T}^k(\boldsymbol{Z})\big) - \hat{\boldsymbol{x}}^k\big(\boldsymbol{T}^k(\boldsymbol{Z})+h\boldsymbol{\xi}^k\big)\right)\middle|\,\overline{\boldsymbol{Z}^k}\right] \qquad (23)$$

for a fixed k. We first note

$$\boldsymbol{T}^k(\boldsymbol{Z}) = \boldsymbol{e}^k \circ (\boldsymbol{L}\boldsymbol{Z} + \boldsymbol{l}) = \boldsymbol{L}^k \boldsymbol{Z}^k + \overline{\boldsymbol{L}^k}\,\overline{\boldsymbol{Z}^k} + \boldsymbol{l}^k = \boldsymbol{L}^k \boldsymbol{\xi}^k + \boldsymbol{L}^k \boldsymbol{\mu}^k + \overline{\boldsymbol{L}^k}\,\overline{\boldsymbol{Z}^k} + \boldsymbol{l}^k$$

where $\boldsymbol{e}^k$ is a vector where component j is 1 if $j \in S_k$ and 0 otherwise. The matrix $\boldsymbol{L}^k \in \mathbb{R}^{|S_k|\times|S_k|}$ is the square submatrix induced by the rows and columns in S_k. (Recall, this matrix is invertible by Assumption 6.3.) The matrix $\overline{\boldsymbol{L}^k} \in \mathbb{R}^{|S_k|\times(n-|S_k|)}$ is the matrix with rows corresponding S_k and columns corresponding to S_k^c. We then fix a k and $\overline{\boldsymbol{Z}^k}$, and temporarily suppress k in the notation. We define $\boldsymbol{y}_0 = \boldsymbol{L}\boldsymbol{\xi}$, $\boldsymbol{y}_h = \boldsymbol{L}\boldsymbol{\xi} + h\boldsymbol{\xi}$ and $g(\boldsymbol{y}) = \hat{\boldsymbol{x}}(\boldsymbol{y} + \boldsymbol{L}\boldsymbol{\mu} + \overline{\boldsymbol{L}}\,\overline{\boldsymbol{Z}} + \boldsymbol{l}) \in \mathbb{R}^{|S_k|}$. Then, the conditional expectation Eq. (23) is

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_0\right)^\top g(\boldsymbol{y}_0)\right] - \mathbb{E}\left[\left((\boldsymbol{L}+h\boldsymbol{I})^{-1}\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\right]\right| \\
&\quad\le \underbrace{\left|\mathbb{E}\left[\left(\left(\boldsymbol{L}^{-1} - (\boldsymbol{L}+h\boldsymbol{I})^{-1}\right)\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\right]\right|}_{\text{Term (i)}} + \underbrace{\left|\mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_0\right)^\top g(\boldsymbol{y}_0)\right] - \mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\right]\right|}_{\text{Term (ii)}}
\end{aligned}$$

Bounding Term (i)

Since $\|g(\boldsymbol{y})\|_\infty \le 1$, we have that

$$\begin{aligned}
\left|\left[\left(\left(\boldsymbol{L}^{-1} - (\boldsymbol{L}+h\boldsymbol{I})^{-1}\right)\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\right]\right| &\le \left\|\left(\boldsymbol{L}^{-1} - (\boldsymbol{L}+h\boldsymbol{I})^{-1}\right)\boldsymbol{y}_h\right\|_1 = h\|\boldsymbol{L}^{-1}\boldsymbol{\xi}\|_1 \\
&\le h\sqrt{|S_k|}\|\boldsymbol{L}^{-1}\boldsymbol{\xi}\|_2 \le \frac{h\sqrt{|S_k|}}{\sigma_{\min}}\|\boldsymbol{\xi}\|_2
\end{aligned}$$

where last inequality holds by definition of the operator norm on the matrix $\boldsymbol{L}^{-1}$.

Hence, we have

$$\left|\mathbb{E}\left[\left(\left(\boldsymbol{L}^{-1} - (\boldsymbol{L}+h\boldsymbol{I})^{-1}\right)\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\right]\right| \le \frac{h\sqrt{|S_k|}}{\sigma_{\min}}\mathbb{E}\left[\|\boldsymbol{\xi}\|_2\right] \le \frac{h|S_k|}{\sigma_{\min}\sqrt{\nu_{\min}}}$$

Bounding Term (ii)

To bound the second term, we would like to appeal to a total variation argument, however the terms in the expectation are not bounded. We thus consider truncating the terms at U (a constant we will determine later).

$$\begin{aligned}
&\left|\mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_0\right)^\top g(\boldsymbol{y}_0)\right] - \mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\right]\right| \\
&\quad\le \left|\mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\mathbb{I}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1 > U\right\}\right]\right| && \text{(24a)} \\
&\qquad+ \left|\mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_0\right)^\top g(\boldsymbol{y}_0)\mathbb{I}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_0\|_1 > U\right\}\right]\right| && \text{(24b)} \\
&\qquad+ \left|\mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_0\right)^\top g(\boldsymbol{y}_0)\mathbb{I}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_0\|_1 \le U\right\}\right] - \mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\mathbb{I}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1 \le U\right\}\right]\right| && \text{(24c)}
\end{aligned}$$

We first bound Eq. (24a). By Holder's inequality, we have

$$\left|\mathbb{E}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\mathbb{I}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1 > U\right\}\right]\right| \le \mathbb{E}\left[\left(\left(\boldsymbol{L}^{-1}\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\right)^2\right]^{1/2}\mathbb{P}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1 > U\right\}^{1/2} \quad (25)$$

after observing $\mathbb{E}\left[\mathbb{I}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1 > U\right\}^2\right]^{1/2} = \mathbb{P}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1 > U\right\}^{1/2}$.

To bound Eq. (25), we first bound the term inside the expectation. Applying Holder's inequality again shows

$$\left(\left(\boldsymbol{L}^{-1}\boldsymbol{y}_h\right)^\top g(\boldsymbol{y}_h)\right)^2 \;\le\; \|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1^2 \|g(\boldsymbol{y}_h)\|_\infty^2 \;\le\; \|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1^2,$$

where the second holds since $\|g(\boldsymbol{y}_h)\|_\infty \le 1$. For $0 \le h \le 1$ and $\sigma_{\min} \le 1$, the last term can be further bounded as follows,

$$\begin{aligned}
\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1^2 &= \|\boldsymbol{L}^{-1}(\boldsymbol{L} + h\boldsymbol{I})\boldsymbol{\xi}\|_1^2 \\
&\le 2\|\boldsymbol{\xi}\|_1^2 + 2h^2\|\boldsymbol{L}^{-1}\boldsymbol{\xi}\|_1^2 && \text{(Triangle Inequality and } (a+b)^2 \le 2a^2 + 2b^2) \\
&\le 2|S_k|\|\boldsymbol{\xi}\|_2^2 + 2|S_k|h^2\|\boldsymbol{L}^{-1}\boldsymbol{\xi}\|_2^2 \\
&\le 2|S_k|\|\boldsymbol{\xi}\|_2^2 + \frac{2|S_k|h^2}{\sigma_{\min}^2}\|\boldsymbol{\xi}\|_2^2 \\
&\le \frac{4|S_k|}{\sigma_{\min}^2}\|\boldsymbol{\xi}\|_2^2, && (\text{since } \sigma_{\min} \le 1, h \le 1).
\end{aligned}$$

Applying this bound to the the probability term as well, we have

$$\mathbb{P}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1 > U\right\}^{1/2} \le \mathbb{P}\left\{\frac{2\sqrt{|S_k|}}{\sigma_{\min}}\|\boldsymbol{\xi}\|_2 > U\right\}^{1/2} = \mathbb{P}\left\{\|\boldsymbol{\xi}\|_2 > \frac{\sigma_{\min}U}{2\sqrt{|S_k|}}\right\}^{1/2}.$$

We can then bound the probability as follows,

$$\mathbb{P}\left\{\|\boldsymbol{\xi}\|_2 > \frac{\sigma_{\min}U}{2\sqrt{|S_k|}}\right\} = \mathbb{P}\left\{\|\boldsymbol{\xi}\|_2^2 > \frac{\sigma_{\min}^2U^2}{4|S_k|}\right\} \le 2\exp\left(\frac{-\sigma_{\min}^2\nu_{\min}U^2}{C_1|S_k|^2}\right)$$

for some absolute constant C_1. The inequality holds by first noting $\|\boldsymbol{\xi}\|_2 - \mathbb{E}[\|\boldsymbol{\xi}\|_2]$ is sub-Gaussian with variance proxy $O(\sqrt{2|S_k|/\nu_{\min}})$ by Gupta and Rusmevichientong (2021, Lemma A.1). Then, via Vershynin (2018, Lemma 2.7.6), this implies $\|\boldsymbol{\xi}\|_2^2$ is sub-exponential satisfying

$$\mathbb{P}\left\{\|\boldsymbol{\xi}\|_2^2 > t\right\} \;\le\; 2\exp\left(\frac{-t\nu_{\min}}{C_1|S_k|}\right)$$

for some absolute constant C_1.

Plugging in our bounds to Eq. (25), we have

$$\text{Eq. (24a)} \le \mathbb{E}\left[\frac{4|S_k|}{\sigma_{\min}^2}\|\boldsymbol{\xi}\|_2^2\right]^{1/2}\mathbb{P}\left\{\|\boldsymbol{\xi}\|_2 > \frac{\sigma_{\min}U}{2\sqrt{|S_k|}}\right\}^{1/2} \le \frac{2\sqrt{2}|S_k|}{\sigma_{\min}\sqrt{\nu_{\min}}}\exp\left(\frac{-\sigma_{\min}^2\nu_{\min}U^2}{C_2|S_k|^2}\right)^{1/2},$$

for some absolute constant C_2.

To bound Eq. (24b), we can repeat the same argument but take $h = 0$ throughout. Consequently,

$$\text{Eq. (24b)} \;\le\; \frac{2|S_k|}{\sigma_{\min}\sqrt{\nu_{\min}}}\exp\left(\frac{-\sigma_{\min}^2\nu_{\min}U^2}{C_2|S_k|^2}\right)^{1/2}.$$

Combining the two bounds, we have for $\sigma_{\min} \leq 1$ and $0 \leq h \leq 1$,

$$\text{Eq. (24a)} + \text{Eq. (24b)} \leq \frac{5|S_k|}{\sigma_{\min}\sqrt{\nu_{\min}}} \exp\left(\frac{-\sigma_{\min}^2 \nu_{\min} U^2}{C_2 |S_k|^2}\right)^{1/2}.$$

Finally, we bound Eq. (24c) using a Total Variation argument. We first note the function $\boldsymbol{y} \mapsto \left(\boldsymbol{L}^{-1}\boldsymbol{y}\right)^\top g(\boldsymbol{y})$ is bounded by U for all $\boldsymbol{y}$ as

$$\left|\left(\boldsymbol{L}^{-1}\boldsymbol{y}\right)^\top g(\boldsymbol{y}) \mathbb{I}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1 \leq U\right\}\right| \;\leq\; \mathbb{I}\left\{\|\boldsymbol{L}^{-1}\boldsymbol{y}_h\|_1 \leq U\right\} \|\boldsymbol{L}^{-1}\boldsymbol{y}\|_1 \|g(\boldsymbol{y})\|_\infty \;\leq\; U.$$

Thus, we have

$$\text{Eq. (24c)} \;\leq\; U \cdot \mathsf{TV}(\boldsymbol{y}_0, \boldsymbol{y}_h),$$

where $\mathsf{TV}(\boldsymbol{y}_0, \boldsymbol{y}_h)$ is the total variation distance between $\boldsymbol{y}_0$ and $\boldsymbol{y}_h$. By Lemma E.4 (c.f. Remark 2, this term is at most $U \cdot C_3|S_k|h$ for the constant C_3 described in the lemma.

Putting it all together in Eq. (24), we have

$$\text{Term (ii)} \;\leq\; UC_4|S_k|h + C_4|S_k| \exp\left(\frac{-U^2}{C_4|S_k|^2}\right)^{1/2}.$$

for some constant C_4 (depending on $\nu_{\min}$, $\nu_{\max}$, $\sigma_{\min}$). Choosing $U = C_4\,|S_k|\,\sqrt{\log h^{-2}}$ and collecting constants shows

$$\text{Term (ii)} \;\leq\; C_5\sqrt{\log(1/h)}\,|S_k|^2\,h + C_5\,|S_k|\,h \;\leq\; C_5\,|S_k|^2\,h\sqrt{\log(1/h)},,$$

for some C_5 (depending on $\nu_{\min}$, $\nu_{\max}$, $\sigma_{\min}$).

Combining Term (i) and Term (ii) and summing across k, we have

$$\sum_{k=1}^{K} C_6 h|S_k| + C_6|S_k|^2 h\sqrt{\log\left(\frac{1}{h}\right)} \;\leq\; 2C_6 S_{\max} n h\sqrt{\log\left(\frac{1}{h}\right)},$$

for a constant C_6. This completes the proof.

□

The next lemma bounds the total variation distance between two random variables of interest. When we invoke the lemma, the role of $\boldsymbol{L}$ will be played by $\boldsymbol{L}^k$. We phrase the lemma with out this k notation for simplicity.

Lemma E.4 (Bound on TV distance) *Assume $\boldsymbol{L}^{-1}$ exists for $\boldsymbol{L} \in \mathbb{R}^{d\times d}$. Let*

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \operatorname{diag}\left((1/\nu_j)_{j=1}^d\right), \quad \boldsymbol{y}_0 = \boldsymbol{L}\boldsymbol{\xi}, \quad \boldsymbol{y}_h = (\boldsymbol{L} + h\boldsymbol{I})\boldsymbol{\xi}.$$

Suppose the matrix $\boldsymbol{L}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{L}^\top$ is positive semi-definite. Then, for $0 \leq h \leq 1$, there exists a constant C (depending on $\sigma_{\min}$, $\nu_{\min}$, $\nu_{\max}$) such that

$$\mathsf{TV}(\boldsymbol{y}_0, \boldsymbol{y}_h) \;\leq\; Cdh,$$

where $\nu_{\min} = \min_i \nu_i$, $\nu_{\max} = \max_i \nu_i$, and $\sigma_{\min} = \min\{1, \sigma_d(\boldsymbol{L})\}$ bounds the smallest singular value of $\boldsymbol{L}$.

Proof for Lemma E.4. Let

$$\boldsymbol{\Sigma}_0 = \boldsymbol{L}\boldsymbol{\Sigma}\boldsymbol{L}^\top, \quad \boldsymbol{\Sigma}_h = (\boldsymbol{L}+h\boldsymbol{I})\boldsymbol{\Sigma}(\boldsymbol{L}+h\boldsymbol{I})^\top,$$

and λ_i be the eigvenvalues of $\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_h - \boldsymbol{I}$. Since $\boldsymbol{y}_0$ and $\boldsymbol{y}_h$ are multivariate Gaussians, Devroye, Mehrabian, and Reddad (2018, Theorem 1.1) proves that

$$\mathsf{TV}(\boldsymbol{y}_0, \boldsymbol{y}_h) \le \sqrt{\sum_{i=1}^d \lambda_i^2}.$$

To bound the right side, we will first argue that $\lambda_i \ge 0$ for all i. Write $\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_h - \boldsymbol{I} = \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma}_h - \boldsymbol{\Sigma}_0)$. Notice $\boldsymbol{\Sigma}_h - \boldsymbol{\Sigma}_0 = h(\boldsymbol{L}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{L}^\top) + h^2\boldsymbol{\Sigma}$ is the sum of two positive semidefinite matrices, and hence positive semidefinite. It is also symmetric. On the other hand, $\boldsymbol{\Sigma}_0^{-1}$ is symmetric and positive semidefinite. Finally, the product of two positive semidefinite, symmetric matrices is positive semidefinite. This proves that $\lambda_i \ge 0$ for each i.

We next bound the summation inside the square root by first observing,

$$\begin{aligned} 0 \ \le\ \sum_{i=1}^d \lambda_i^2 \ &\le\ \left(\sum_{i=1}^d \lambda_i\right)^2 && \text{(since } \lambda_i \ge 0\text{)} \\ &=\ \operatorname{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma}_h - \boldsymbol{\Sigma}_0)\right)^2. \end{aligned}$$

Since $\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma}_h - \boldsymbol{\Sigma}_0) = \boldsymbol{\Sigma}_0^{-1}(h\boldsymbol{L}\boldsymbol{\Sigma} + h\boldsymbol{\Sigma}\boldsymbol{L}^\top + h^2\boldsymbol{\Sigma})$, we have

$$\begin{aligned} \operatorname{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma}_h - \boldsymbol{\Sigma}_0)\right) &= \operatorname{tr}\left(h\boldsymbol{\Sigma}_0^{-1}\boldsymbol{L}\boldsymbol{\Sigma}\right) + \operatorname{tr}\left(h\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}\boldsymbol{L}^\top\right) + \operatorname{tr}\left(h^2\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}\right) \\ &= h\cdot\operatorname{tr}\left((\boldsymbol{L}^\top)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{\Sigma}\right) + h\cdot\operatorname{tr}\left((\boldsymbol{L}^\top)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}^{-1}\boldsymbol{\Sigma}\boldsymbol{L}^\top\right) \\ &\qquad + h^2\cdot\operatorname{tr}\left((\boldsymbol{L}^\top)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}^{-1}\boldsymbol{\Sigma}\right) \\ &= h\cdot\operatorname{tr}\left((\boldsymbol{L}^\top)^{-1}\right) + h\cdot\operatorname{tr}\left(\boldsymbol{L}^{-1}\right) + h^2\cdot\operatorname{tr}\left((\boldsymbol{L}^\top)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}^{-1}\boldsymbol{\Sigma}\right) \end{aligned}$$

where the first equality holds since the trace is a linear mapping, the second equality holds by plugging in the definition of $\boldsymbol{\Sigma}_0$, the third equality holds by applying the cyclic property of the trace.

Next, note that multiplying by $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$ scales the diagonals by at most $1/\nu_{\min}$ or $\nu_{\max}$, respectively. Hence,

$$\operatorname{tr}\left((\boldsymbol{L}^\top)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}^{-1}\boldsymbol{\Sigma}\right) \ \le\ \frac{\operatorname{tr}\left((\boldsymbol{L}^\top)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{L}^{-1}\right)}{\nu_{\min}} \ =\ \frac{\operatorname{tr}\left(\boldsymbol{L}^{-1}(\boldsymbol{L}^\top)^{-1}\boldsymbol{\Sigma}^{-1}\right)}{\nu_{\min}} \ \le\ \frac{\nu_{\max}}{\nu_{\min}}\operatorname{tr}\left((\boldsymbol{L}^\top\boldsymbol{L})^{-1}\right).$$

Thus,

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma}_h - \boldsymbol{\Sigma}_0)\right) \ \le\ 2h\cdot\left|\operatorname{tr}\left(\boldsymbol{L}^{-1}\right)\right| + \frac{h^2\nu_{\max}}{\nu_{\min}}\cdot\operatorname{tr}\left((\boldsymbol{L}^\top\boldsymbol{L})^{-1}\right).$$

We can further bound

$$\operatorname{tr}\left((\boldsymbol{L}^\top \boldsymbol{L})^{-1}\right) \;=\; \sum_{i=1}^{d} \frac{1}{\sigma_i^2} \;\leq\; \frac{d}{\sigma_{\min}^2}$$

since the trace is the sum of the squared singular values of $\boldsymbol{L}^{-1}$ which are inverse of the singular values of $\boldsymbol{L}$.

Lastly, we bound $|\operatorname{tr}(\boldsymbol{L}^{-1})|$. Letting $\lambda_j^{\boldsymbol{L}}$ be the eigenvalues of $\boldsymbol{L}$, we see

$$\left|\operatorname{tr}(\boldsymbol{L}^{-1})\right| \;=\; \left|\sum_{j=1}^{d} \frac{1}{\lambda_j^{\boldsymbol{L}}}\right| \;\leq\; \frac{d}{\sigma_{\min}}$$

where $\lambda_{\min}^{\boldsymbol{L}} = \min_j \lambda |\lambda_j^{\boldsymbol{L}}|$. The first equality holds by definition of trace and noting the eigenvalues of $\boldsymbol{L}^{-1}$ are the inverses the eigenvalues of $\boldsymbol{L}$. The last inequality holds since $\sigma_{\min} \leq \left|\lambda_j^{\boldsymbol{L}}\right|$ for all j.

In summary, we have shown

$$\mathsf{TV}\left(\boldsymbol{y}_0, \boldsymbol{y}_h\right) \leq \frac{2hd}{\sigma_{\min}} + \frac{h^2 d \nu_{\max}}{\nu_{\min}\sigma_{\min}^2} \leq dh\left(\frac{2}{\sigma_{\min}} + \frac{2\nu_{\max}}{\nu_{\min}\sigma_{\min}^2}\right)$$

This completes the proof. □

Remark 2 (Applying Lemma E.4). We observe that under Assumptions 6.2 to 6.4, to the random variables induced by $\mathcal{S}^k$, i.e., that

$$\mathsf{TV}(\boldsymbol{L}^k \boldsymbol{\xi}^k, (\boldsymbol{L}^k + h\boldsymbol{I})\boldsymbol{\xi}^k) \;\leq C\,|S_k|\,h.$$

To do so, we simply verify the assumptions of the lemma. The matrix $\boldsymbol{L}^k$ is invertible by Assumption 6.3. By Assumption 6.4, the matrix $\boldsymbol{L\Sigma} + \boldsymbol{\Sigma L}^\top \in \mathbb{R}^{n\times n}$ is positive semidefinite. We claim that this implies that $\boldsymbol{L}^k\boldsymbol{\Sigma}^k + \boldsymbol{\Sigma}^k(\boldsymbol{L}^k)^\top \in \mathbb{R}^{|S_k|\times|S_k|}$ is also positive semidefinite, $\boldsymbol{\Sigma}^k$ is the diagonal covariance matrix of $\boldsymbol{\xi}^k$. To see this, observe that the ij^{th} component of $\boldsymbol{L\Sigma}$ can be written as $L_{ij}\sigma_j$ since $\boldsymbol{\Sigma}$ is a diagonal matrix. Consequently, $\boldsymbol{L}^k\boldsymbol{\Sigma}^k$ is a principle submatrix of $\boldsymbol{L\Sigma}$. Similarly, we see $\boldsymbol{\Sigma}^k(\boldsymbol{L}^k)^\top$ is a principle submatrix of $\Sigma\boldsymbol{L}^\top$. Thus, $\boldsymbol{L}^k\boldsymbol{\Sigma}^k + \boldsymbol{\Sigma}^k(\boldsymbol{L}^k)^\top$ is a principle submatrix of $\boldsymbol{L\Sigma} + \boldsymbol{\Sigma L}^\top$. Finally, the principle sub-matrices of a positive semidefinite matrix are also positive semidefinite. Thus, under Assumptions 6.2 to 6.4, we can apply the lemma to the subcomponents pertaining to S_k.

E.4. Bound on Eq. (22b)

To bound Eq. (22b) for block decoupled problems (i.e., Problem (14) with $m = 0$ and $|\mathcal{X}_0| = 0$), the primary challenge is that for $k \neq \bar{k}$, even though $\boldsymbol{\xi}^k$ and $\boldsymbol{\xi}^{\bar{k}}$ are independent, it may still be that $\hat{\boldsymbol{x}}^k(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}))$ and $\hat{\boldsymbol{x}}^{\bar{k}}(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}))$ are correlated, since $\boldsymbol{L}(\bar{\boldsymbol{\theta}})$ may smooth across subproblems. Hence, we cannot treat $\sum_{k=1}^{K} {\boldsymbol{\xi}^k}^\top \hat{\boldsymbol{x}}^k(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) - \mathbb{E}\left[{\boldsymbol{\xi}^k}^\top \hat{\boldsymbol{x}}^k(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\right]$ as a sum of independent random variables.

To circumvent this issue, we will identify groups of subproblems such that subproblems in the same group are independent. We construct these groups by looking at the coloring of a particular graph. We will then sum over subproblems of a particular color (which will be independent by construction), and use a union bound over the colors.

To that end, we introduce the block graph $G(V,E)$. The vertex $V=\{0,\dots,K\}$ corresponds to our disjoint partition $\{S_0,\dots,S_K\}$. An edge $e(k,\bar{k})$ exists between vertices k and $\bar{k}$ if there exists $i\in S_k$, $j\in S_{\bar{k}}$, $l\in\{1,\dots,n\}$, and $\bar{\boldsymbol{\theta}}\in\bar{\Theta}$ such that $L_{il}(\bar{\boldsymbol{\theta}})\neq 0$ and $L_{jl}(\bar{\boldsymbol{\theta}})\neq 0$. In words, this means that $T_i(\boldsymbol{Z},\bar{\boldsymbol{\theta}})$ and $T_j(\boldsymbol{Z},\bar{\boldsymbol{\theta}})$ both depend on ξ_l, despite i and j belonging to different subproblems.

Let $\chi(G)$ be the chromatic number of $G(V,E)$ and fix a vertex coloring for the graph. We claim that if $k\neq\bar{k}$ have the same color, then ${\boldsymbol{\xi}^k}^\top\hat{\boldsymbol{x}}^k(\boldsymbol{T}(\boldsymbol{Z},\boldsymbol{\theta}))$ and ${\boldsymbol{\xi}^k}^\top\hat{\boldsymbol{x}}^k(\boldsymbol{T}(\boldsymbol{Z},\boldsymbol{\theta}))$ must be independent. Indeed, suppose this was not the case. Then, there must exist an $i\in\mathcal{S}_k$ and a $j\in\mathcal{S}_{\bar{k}}$ such that $T_i(\boldsymbol{Z},\bar{\boldsymbol{\theta}})$ and $T_j(\boldsymbol{Z},\bar{\boldsymbol{\theta}})$ both depend on some component ξ_l, which suggests $e(k,\bar{k})$ is present, and hence k and $\bar{k}$ cannot be the same color. This is a contradiction.

Let $R_1,\dots,R_{\chi(G)}$ be the partition of V corresponding to the different colors. Using this partition, we prove the following result under the weaker Assumption 3.3 rather than the stronger Assumption 6.4 assumed in Theorem 6.6.

Lemma E.5 (Convergence of Solution Stability to Expectation) *Suppose Assumption 3.3 holds. Then, there exists a constant C such that with probability at least $1-\epsilon$,*

$$\begin{aligned}\sup_{\substack{\bar{\boldsymbol{\theta}}\in\bar{\Theta}\\ h'\in\{0,h\}}}&\left|\boldsymbol{\xi}^\top\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h'\boldsymbol{\xi}\big)-\mathbb{E}\left[\boldsymbol{\xi}^\top\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h'\boldsymbol{\xi}\big)\right]\right|\\ &\leq\; C\kappa S_{\max}\sqrt{\chi(G)n\log(\mathcal{X}_{\max})}\log\left(\chi(G)\right)\log(1/\epsilon).\end{aligned}$$

Proof for Lemma E.5. We first decompose Eq. (22b) into the sets determined by the vertex coloring as follows,

$$\begin{aligned}&\sup_{\substack{\bar{\boldsymbol{\theta}}\in\bar{\Theta}\\ h'\in\{0,h\}}}\left|\boldsymbol{\xi}^\top\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h'\boldsymbol{\xi}\big)-\mathbb{E}\left[\boldsymbol{\xi}^\top\hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h'\boldsymbol{\xi}\big)\right]\right|\\ &\qquad\leq\sum_{l=1}^{\chi(G)}\sup_{\substack{\bar{\boldsymbol{\theta}}\in\bar{\Theta}\\ h'\in\{0,h\}}}\sum_{k\in R_l}(\boldsymbol{\xi}^k)^\top\hat{\boldsymbol{x}}^k\left(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h'\boldsymbol{\xi}\right)-\mathbb{E}\left[(\boldsymbol{\xi}^k)^\top\hat{\boldsymbol{x}}^k\left(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h'\boldsymbol{\xi}\right)\right].\end{aligned}$$

We bound each term of the outer summation using Theorem A.1 of Gupta, Huang, and Rusmevichientong (2022a) since $f_k(\boldsymbol{Z},\bar{\boldsymbol{\theta}},h')=(\boldsymbol{\xi}^k)^\top\hat{\boldsymbol{x}}^k\left(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+h'\boldsymbol{\xi}\right)$ for $k\in R_l$ are independent. Fixing an l, we take

$$\boldsymbol{F}(\boldsymbol{Z})=\left(\|\boldsymbol{\xi}^k\|_1\right)_{k\in R_l}\in\mathbb{R}^{|R_l|}$$

We can bound the Orlicz norm as follows,

$$\| \|\boldsymbol{F}(\boldsymbol{Z})\|_2 \|_\Psi \leq \left\| \sqrt{S_{\max}} \|(\boldsymbol{\xi}^k)_{k\in R_l}\|_2 \right\|_\Psi \leq \sqrt{S_{\max}}\kappa\sqrt{2S_{\max}|R_l|} = S_{\max}\kappa\sqrt{2|R_l|}$$

where the first inequality holds by noting $\|\boldsymbol{\xi}^k\|_1 \leq \sqrt{S_{\max}}\|\boldsymbol{\xi}^k\|_2$ and the second inequality holds by Gupta and Rusmevichientong (2021, Lemma A.1).

Note that

$$\left| \left\{ \left((\boldsymbol{\xi}^k)^\top \hat{\boldsymbol{x}}^k \left(\boldsymbol{T}(\boldsymbol{Z}, \boldsymbol{\theta}) + h'\boldsymbol{\xi} \right) \right)_{k\in R_l} : \bar{\boldsymbol{\theta}} \in \bar{\Theta}, h' \in \{0, h\} \right\} \right| \leq 2\mathcal{X}_{\max}.$$

Now apply Gupta and Rusmevichientong (2021, Theorem A.1) to show that there exists an absolute constant C such that probability at least $1-\epsilon$,

$$\begin{aligned}
\sup_{\substack{\bar{\boldsymbol{\theta}}\in\bar{\Theta} \\ h'\in\{0,h\}}} & \left| \sum_{k\in R_l} (\boldsymbol{\xi}^k)^\top \hat{\boldsymbol{x}}^k \left(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + h'\boldsymbol{\xi} \right) - \mathbb{E}\left[(\boldsymbol{\xi}^k)^\top \hat{\boldsymbol{x}}^k \left(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + h'\boldsymbol{\xi} \right) \right] \right| \\
& \leq \; CS_{\max}\kappa\sqrt{|R_l|\log(\mathcal{X}_{\max})}\log\left(\frac{1}{\epsilon}\right).
\end{aligned}$$

To complete the proof, we take a union bound over l allowing us to show that with probability $1-\epsilon$

$$\begin{aligned}
\sup_{\substack{\bar{\boldsymbol{\theta}}\in\bar{\Theta} \\ h'\in\{0,h\}}} & \left| \boldsymbol{\xi}^\top \hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + h'\boldsymbol{\xi} \big) - \mathbb{E}\left[\boldsymbol{\xi}^\top \hat{\boldsymbol{x}}\big(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + h'\boldsymbol{\xi} \big) \right] \right| \\
& \leq \; \sum_{l=1}^{\chi(G)} CS_{\max}\kappa\sqrt{|R_l|\log(\mathcal{X}_{\max})}\log\left(\frac{\chi(G)}{\epsilon}\right). \\
& \leq \; C\chi(G)S_{\max}\kappa\sqrt{\frac{K}{\chi(G)}\log(\mathcal{X}_{\max})}\log\left(\frac{\chi(G)}{\epsilon}\right) \\
& \leq \; C\kappa S_{\max}\sqrt{\chi(G)n\log(\mathcal{X}_{\max})}\log\left(\frac{\chi(G)}{\epsilon}\right)
\end{aligned}$$

where the first inequality is the union bound, and the second inequality uses $\sum_{l=1}^{\chi(G)} |R_l| = K$ and Jensen's inequality. Collecting constants proves the lemma.

□

In the main body, we use a slightly looser bound by further bounding $\chi(G)$. Specifically, by Brook's Theorem, $\chi(G)$ is at most the maximal degree of $G(V,E)$ plus one. Let $T_{\max} = \max_j \sum_{i=1}^n \sup_{\bar{\boldsymbol{\theta}}\in\bar{\Theta}} \mathbb{I}\left\{ \boldsymbol{L}_{ij}(\bar{\boldsymbol{\theta}}) \neq 0 \right\}$. In words, $T_{\max}$ is the maximal number of non-zero elements in a column of $\boldsymbol{L}(\bar{\boldsymbol{\theta}})$.

Then we claim that the maximal degree of $G(V,E)$ is at most $S_{\max}T_{\max}$. To see this, consider S_1. For each $k \in S_1$, $T_k(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$ can depend on at most $T_{\max}$ different components of $\boldsymbol{\xi}$. In a worst-case, these components each belong to different subproblems, creating at most $T_{\max}$ outgoing edges from node 1. This is true for each $k \in S_1$, so S_1 has an outdoing degree of at most $|S_1| T_{\max}$. This is true for each subproblem, yielding the upperbound $S_{\max} T_{\max}$.

For simplicity of exposition, we use this upper bound in the main text, but the chromatic number bound above is often substantively tighter.

E.5. Proof for Block Decoupled Estimation Error, Theorem 6.6

We apply Theorem 4.3 and use Lemma 6.5 to bound the solution stability term. We complete the bound by obtaining $C_{\boldsymbol{T}}$ for the affine plug-in class. We see

$$\begin{aligned}\left\|\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})-\boldsymbol{T}(\boldsymbol{Y},\bar{\boldsymbol{\theta}})\right\|_1 &= \left\|\boldsymbol{L}(\bar{\boldsymbol{\theta}})\,(\boldsymbol{Z}-\boldsymbol{Y})\right\|_1 \\ &\leq \sum_{j=1}^{n}\left\|\boldsymbol{L}_j(\bar{\boldsymbol{\theta}})(Z_j-Y_j)\right\|_1 = \left(\max_j \left\|\boldsymbol{L}_j(\bar{\boldsymbol{\theta}})\right\|_1\right)\|\boldsymbol{Z}-\boldsymbol{Y}\|_1 \leq T_{\max}L_{\max}\|\boldsymbol{Z}-\boldsymbol{Y}\|_1,\end{aligned}$$

so $C_{\boldsymbol{T}} = T_{\max} L_{\max}$.

Appendix F: Weakly-Coupled Policy Proofs

Proof for Theorem 6.11 We can bound the estimation error for weakly-coupled problems as follows,

$$\begin{aligned}&\sup_{\boldsymbol{\theta}\in\Theta}\left|\boldsymbol{\xi}^\top\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})-D^{WC}(\boldsymbol{Z},\boldsymbol{\theta},h)\right| \\ &\quad\leq \sup_{\boldsymbol{\theta}\in\Theta}\left|\boldsymbol{\xi}^\top\left(\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})-\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\bar{\boldsymbol{\theta}})\right)\right| \\ &\qquad+\sup_{\boldsymbol{\theta}\in\Theta}\left|\boldsymbol{\xi}^\top\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\bar{\boldsymbol{\theta}})-\tilde{D}(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\bar{\boldsymbol{\theta}},h)\right| \\ &\quad\leq \underbrace{\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\xi}\|_\infty\left\|\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})-\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\bar{\boldsymbol{\theta}})\right\|_1}_{(a)} \\ &\qquad+\underbrace{\sup_{\boldsymbol{\theta}\in\Theta,\,\boldsymbol{x}^0\in\mathcal{X}^0}\left|\boldsymbol{\xi}^\top\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0),\boldsymbol{x}^0,\bar{\boldsymbol{\theta}})-\tilde{D}(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0),\boldsymbol{x}^0,\bar{\boldsymbol{\theta}},h)\right|}_{(b)}\end{aligned}$$

The first inequality adds and subtracts our approximate policy. The last inequality gives us the key terms we must bound.

To bound (a), we apply Lemma F.1, showing

$$\sup_{\boldsymbol{\theta}\in\Theta}\ \|\boldsymbol{\xi}\|_\infty\left\|\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})-\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda}^0(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\bar{\boldsymbol{\theta}})\right\|_1 \ \leq\ \|\boldsymbol{\xi}\|_\infty\cdot m\cdot S_{\max}$$

Since $\boldsymbol{\xi}$ is a vector of independent sub-Gaussian random variables, with probability at least $1-\epsilon$,

$$\|\boldsymbol{\xi}\|_\infty\ \leq\ C_1\sqrt{\frac{\log n}{\nu_{\min}}\cdot\log\left(\frac{1}{\epsilon}\right)}$$

where C_1 is a universal constant (Wainwright, 2019) . Thus, with probability at least $1-\epsilon$,

$$(a) \;\le\; C_1 \cdot m \cdot S_{\max}\sqrt{\frac{\log n}{\nu_{\min}} \cdot \log\left(\frac{1}{\epsilon}\right)}.$$

To bound (b), we first observe by Lemma F.2 $\|\boldsymbol{\lambda}^0(\boldsymbol{Z},\bar{\boldsymbol{\theta}},\boldsymbol{x}^0)\|_1 \le \frac{2}{\bar{s}}\|\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\|_1$, for every $\bar{\boldsymbol{\theta}},\boldsymbol{x}^0$. Moreover,

$$\|\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\|_1 \le \|\boldsymbol{L}(\bar{\boldsymbol{\theta}})\|_1\|\boldsymbol{Z}\|_1 + \|\boldsymbol{l}(\bar{\boldsymbol{\theta}})\|_1 \le \|\boldsymbol{\mu}\|_1 + \|\boldsymbol{\xi}\|_1 + 1 \le C_\mu(n+1) + n\cdot\|\boldsymbol{\xi}\|_\infty.$$

Plugging in our bound on $\|\boldsymbol{\xi}\|_\infty$, we see with probability at least $1-\epsilon$

$$\|\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\|_1 \le 2C_1C_\mu n\sqrt{\frac{\log n}{\nu_{\min}} \cdot \log\left(\frac{1}{\epsilon}\right)}.$$

Letting $\epsilon = e^{-n}$, we see that for $\lambda_{\max} = 4C_1C_\mu\bar{s}^{-1}\nu_{\min}^{-1/2}n^2$ that $\mathbb{P}\{\|\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0))\|_1 \ge \lambda_{\max}\} < e^{-n}$.

Let $\Lambda = \left\{\boldsymbol{\lambda}\in\mathbb{R}^m_+ : \|\boldsymbol{\lambda}\|_1 \le \lambda_{\max}\right\}$. Then, when $\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0)\in\Lambda$,

$$\begin{aligned}(b) \;&=\; \sup_{\boldsymbol{\theta}\in\Theta,\,\boldsymbol{x}^0\in\mathcal{X}^0}\left|\boldsymbol{\xi}^\top\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda}^0(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0),\boldsymbol{x}^0,\boldsymbol{\theta}) - \tilde{D}(\boldsymbol{Z},\boldsymbol{\lambda}^0(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0),\boldsymbol{x}^0,\boldsymbol{\theta},h)\right| \\ &\le\; \sup_{\boldsymbol{x}^0\in\mathcal{X}^0}\sup_{\substack{\boldsymbol{\theta}\in\Theta\\ \boldsymbol{\lambda}\in\Lambda}}\left|\boldsymbol{\xi}^\top\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}_0,\boldsymbol{\theta}) - \tilde{D}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}_0,\bar{\boldsymbol{\theta}},h)\right|\end{aligned}$$

We complete the bound on (b) by first fixing $\boldsymbol{x}^0$ in the last line and then applying Corollary 6.7 to the block decoupled setting induced by the lifted affine policy class to show

$$\begin{aligned}&\sup_{(\bar{\boldsymbol{\theta}},\boldsymbol{\lambda})\in\tilde{\Theta}}\left|\boldsymbol{\xi}^\top\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}_0,\boldsymbol{\theta}) - \tilde{D}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}_0,\bar{\boldsymbol{\theta}},h)\right| \\ &\quad\le\; C_2S_{\max}^{3/4}n^{3/4}\log n\sqrt{T_{\max}\log\mathcal{X}_{\max}\log\left|\tilde{\Theta}_0\left(n^{-1/2}\right)\right|}\cdot\log\left(\frac{S_{\max}T_{\max}}{\epsilon}\right),\end{aligned}$$

where C_2 is the constant from the corollary and $(\bar{\boldsymbol{\theta}},\boldsymbol{\lambda})\in\tilde{\Theta}\equiv\Theta\times\Lambda$. Note that the new plug-in class $\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\lambda}^\top\boldsymbol{A}^0(\boldsymbol{x}^0)$ indexed by $\tilde{\Theta}$ is still an affine policy class and satisfies Assumption 3.1 with the same Lipschitz constant as the original plug-in class.

We can further bound $\left|\tilde{\Theta}_0(n^{-1/2})\right|$ with respect to the covering $\Theta_0(\epsilon)$. Letting $\Lambda_0(\epsilon)$ be the ϵ-covering of Λ_0, we see

$$\left|\tilde{\Theta}_0(2\epsilon)\right| \;\overset{(i)}{\le}\; |\Theta_0(\epsilon)|\cdot\left|\Lambda_0\left(\frac{\epsilon}{\|\boldsymbol{A}^0\|_\infty n}\right)\right| \;\overset{(ii)}{\le}\; |\Theta_0(\epsilon)|\cdot\left(\frac{3\lambda_{\max}\|\boldsymbol{A}^0\|_\infty n}{\epsilon}\right)^m.$$

The inequality (i) follows because $\Theta_0(\epsilon)\times\Lambda_0\left(\frac{\epsilon}{\|\boldsymbol{A}^0\|_\infty n}\right)$ is a valid covering that satisfies Assumption 4.2 since for each $(\bar{\boldsymbol{\theta}},\boldsymbol{\lambda})\in\tilde{\Theta}$ there exists $(\bar{\boldsymbol{\theta}}_0,\boldsymbol{\lambda}_0)\in\tilde{\Theta}_0(n^{-1/2})$ we have

$$\|\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\lambda}^\top\boldsymbol{A}^0(\boldsymbol{x}^0) - \left(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)+\boldsymbol{\lambda}_0^\top\boldsymbol{A}^0(\boldsymbol{x}^0)\right)\|_1 \le \|\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})-\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}_0)\|_1 + \|\left(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0\right)^\top\boldsymbol{A}^0(\boldsymbol{x}^0)\|_1$$

$$
\begin{aligned}
&\leq C\left(\|\boldsymbol{Z}\|_1+1\right)\epsilon+\sum_{j=1}^{n}\left|\left(\boldsymbol{\lambda}-\boldsymbol{\lambda}_0\right)^\top \boldsymbol{A}_j^0(\boldsymbol{x}^0)\right| \\
&\leq C\left(\|\boldsymbol{Z}\|_1+1\right)\epsilon+n\left\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_0\right\|_1\|\boldsymbol{A}^0\|_\infty \\
&\leq C\left(\|\boldsymbol{Z}\|_1+1\right)\epsilon+\epsilon \\
&\leq 2C\left(\|\boldsymbol{Z}\|_1+1\right)\epsilon.
\end{aligned}
$$

Inequality (ii) applies Pollard (1990, Lemma 4.1).

We can further simplify our bound for $n>2$ by

$$
\begin{aligned}
\log\left|\tilde{\Theta}_0(2n^{-1/2})\right| &\leq \log\left|\Theta_0(n^{-1/2})\right|+\log\left(3\lambda_{\max}\|\boldsymbol{A}^0\|_\infty n^{1/2}\right)^m \\
&\leq \left(\log\left|\Theta_0(n^{-1/2})\right|\right)m\log\left(3\lambda_{\max}\|\boldsymbol{A}^0\|_\infty n^{1/2}\right) \\
&\leq m\left(\log\left|\Theta_0(n^{-1/2})\right|\right)\left(\log\left(3\lambda_{\max}\|\boldsymbol{A}^0\|_\infty\right)+\log n^{1/2}\right) \\
&\leq m\log n\cdot\left(\log\left|\Theta_0(n^{-1/2})\right|\right)\left(\log\left(3\lambda_{\max}\|\boldsymbol{A}^0\|_\infty\right)+1\right).
\end{aligned}
$$

Redefining $2n^{-1/2}\to n^{-1/2}$ shows

$$
\log\left|\tilde{\Theta}_0(n^{-1/2})\right| \;\leq\; C_3 m\log n\cdot\left(\log\left|\Theta_0(n^{-1/2}/2)\right|\right)
$$

for some constant C_3 depending on $\log(\lambda_{\max})$ and $\log(\|\boldsymbol{A}^0\|_\infty)$.

Taking a union bound over $\boldsymbol{x}^0\in\mathcal{X}^0$, and collecting terms, we have

$$
\begin{aligned}
(b) &\leq \sup_{\boldsymbol{x}^0\in\mathcal{X}^0}\sup_{\substack{\boldsymbol{\theta}\in\Theta\\ \boldsymbol{\lambda}\in\Lambda}}\left|\boldsymbol{\xi}^\top\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}_0,\boldsymbol{\theta})-\tilde{D}(\boldsymbol{Z},\boldsymbol{\lambda},\boldsymbol{x}_0,\bar{\boldsymbol{\theta}},h)\right| \\
&\leq\; C_4 S_{\max}^{3/4} n^{3/4}\log^{3/2} n\sqrt{mT_{\max}\log\mathcal{X}_{\max}\log\left|\Theta_0\left(n^{-1/2}/2\right)\right|}\cdot\log\left(\frac{S_{\max}T_{\max}\left|\mathcal{X}^0\right|}{\epsilon}\right)
\end{aligned}
$$

where C_4 depends on $\log\lambda_{\max}$, $\log\|\boldsymbol{A}^0\|_\infty$, $\nu_{\min}$, $\nu_{\max}$, $\sigma_{\min}$, $\|\boldsymbol{\mu}\|_\infty$, $L_{\max}$.

Combining bounds on (a) and (b) completes the proof.

□

Lemma F.1 (Error Approximating Weakly-Coupled by Constraints Policies) *Assume Assumption 6.9 holds. Then,*

$$
\|\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})-\tilde{\boldsymbol{x}}(\boldsymbol{Z},\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta})),\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}),\boldsymbol{\theta})\|_1\leq m\cdot S_{\max},\quad a.s.
$$

Proof for Lemma F.1 To streamline the proof, we fix a $\boldsymbol{\theta}\in\Theta$ and $\boldsymbol{Z}$, dropping them from the notation. We let $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{\lambda}}$ represent the optimal primal solution $\boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta})$ and optimal dual solution $\boldsymbol{\lambda}(\boldsymbol{Z},\boldsymbol{\theta},\boldsymbol{x}^0(\boldsymbol{Z},\boldsymbol{\theta}))$, respectively.

Our goal is to bound

$$\|\bar{\boldsymbol{x}}-\tilde{\boldsymbol{x}}\left(\bar{\boldsymbol{\lambda}},\bar{\boldsymbol{x}}^0\right)\|_1=\sum_{k=0}^{K}\|\bar{\boldsymbol{x}}^k-\tilde{\boldsymbol{x}}^k\left(\bar{\boldsymbol{\lambda}},\bar{\boldsymbol{x}}^0\right)\|_1.$$

Focusing on the k^{th} term, recall $\tilde{\boldsymbol{x}}^k\left(\bar{\boldsymbol{\lambda}},\bar{\boldsymbol{x}}^0\right)$ solves the optimization problem,

$$\min_{\boldsymbol{x}^k\geq\mathbf{0}}\quad\sum_{j\in S_k}(T_j+\bar{\boldsymbol{\lambda}}^\top\boldsymbol{A}_j^0(\bar{\boldsymbol{x}}^0))x_j,\quad\text{s.t.}\quad\boldsymbol{x}^k\in\mathcal{X}^k(\bar{\boldsymbol{x}}^0).\tag{26}$$

We claim that if Problem (26) has a unique solution, then $\bar{\boldsymbol{x}}^k=\tilde{\boldsymbol{x}}^k(\bar{\boldsymbol{\lambda}},\bar{\boldsymbol{x}}^0)$. To prove the claim, we show first show $\bar{\boldsymbol{x}}^k$ is the optimal solution of Problem (26). Let

$$g(\boldsymbol{\lambda})=-\left\langle\boldsymbol{b}(\bar{\boldsymbol{x}}^0),\,\boldsymbol{\lambda}\right\rangle+\sum_{k=1}^{K}\min_{\boldsymbol{x}^k\in\mathcal{X}^k}\left\langle\boldsymbol{T}^k+\boldsymbol{\lambda}^\top\boldsymbol{A}_{S_k}^0(\bar{\boldsymbol{x}}^0),\,\boldsymbol{x}^k\right\rangle$$

be the dual problem objective solved by $\bar{\boldsymbol{\lambda}}$. Using strong duality of Problem (16) with a fixed $\boldsymbol{x}^0$, we can show

$$\begin{aligned}
\sum_{k=1}^{K}\left\langle\boldsymbol{T}^k,\bar{\boldsymbol{x}}^k\right\rangle=&\ g(\bar{\boldsymbol{\lambda}}),\ \text{by strong duality,}\\
=&\ -\left\langle\boldsymbol{b}(\bar{\boldsymbol{x}}^0),\,\bar{\boldsymbol{\lambda}}\right\rangle+\sum_{k=1}^{K}\min_{\boldsymbol{x}^k\in\mathcal{X}^k}\left\langle\boldsymbol{T}^k+\bar{\boldsymbol{\lambda}}^\top\boldsymbol{A}_{S_k}^0(\bar{\boldsymbol{x}}^0),\,\boldsymbol{x}^k\right\rangle\\
\overset{(*)}{\leq}&\ -\left\langle\boldsymbol{b}(\bar{\boldsymbol{x}}^0),\,\bar{\boldsymbol{\lambda}}\right\rangle+\sum_{k=1}^{K}\left\langle\boldsymbol{T}^k+\bar{\boldsymbol{\lambda}}^\top\boldsymbol{A}_{S_k}^0(\bar{\boldsymbol{x}}^0),\,\bar{\boldsymbol{x}}^k\right\rangle,\ \text{since }\bar{\boldsymbol{x}}^k\in\mathcal{X}^k(\bar{\boldsymbol{x}}^0),\\
=&\ \underbrace{\left\langle\boldsymbol{A}^0(\bar{\boldsymbol{x}}^0)\bar{\boldsymbol{x}}-\boldsymbol{b}(\bar{\boldsymbol{x}}^0),\,\bar{\boldsymbol{\lambda}}\right\rangle}_{\leq 0,\text{ since }\bar{\boldsymbol{x}}\in\mathcal{X}^{WC}\text{ and }\bar{\boldsymbol{\lambda}}\geq\mathbf{0}}+\sum_{k=1}^{K}\left\langle\boldsymbol{T}^k,\bar{\boldsymbol{x}}^k\right\rangle\\
\leq&\ \sum_{k=1}^{K}\left\langle\boldsymbol{T}^k,\bar{\boldsymbol{x}}^k\right\rangle.
\end{aligned}$$

Here we have shown that equality holds throughout. Since inequality $(*)$ is an equality, we see that $\bar{\boldsymbol{x}}^k$ minimizes Problem (26) for each k. Since we claim Problem (16) has a unique solution, we have $\bar{\boldsymbol{x}}^k=\tilde{\boldsymbol{x}}^k\left(\bar{\boldsymbol{\lambda}},\bar{\boldsymbol{x}}^0\right)$.

Thus, to bound $\|\bar{\boldsymbol{x}}-\tilde{\boldsymbol{x}}\left(\bar{\boldsymbol{\lambda}},\bar{\boldsymbol{x}}^0\right)\|_1$, we count the number of sub-problems that do not have unique solutions. Let $\mathcal{K}$ be the set of k where $\tilde{\boldsymbol{x}}^k\left(\bar{\boldsymbol{\lambda}},\bar{\boldsymbol{x}}^0\right)$ is not a unique solution. We see $k\in\mathcal{K}$ if:

1. The subproblem cost vector is $\mathbf{0}$, or formally,

$$\begin{pmatrix}T_j\\\boldsymbol{A}_j^0\\0\end{pmatrix}^\top\begin{pmatrix}1\\\boldsymbol{\lambda}\\0\end{pmatrix}=0,\quad\forall j\in S_k$$

2. The subproblem cost vector points in a direction corresponding to a unit vector in $\mathcal{D}^k(\bar{\boldsymbol{x}}^0)$, or formally, there exists $\boldsymbol{d}\in\mathcal{D}^k(\bar{\boldsymbol{x}}^0)$ and $\alpha^k>0$ such that

$$\begin{pmatrix}T_j\\\boldsymbol{A}_j^0\\-d_j\end{pmatrix}^\top\begin{pmatrix}1\\\boldsymbol{\lambda}\\\alpha^k\end{pmatrix}=0,\quad\forall j\in S_k$$

We let $\mathcal{K}_1$ define the set of k that satisfy the first condition.

Combining all the equalities across $k \in \mathcal{K}, j \in S_k$, we see they each represent a linear relation among the vectors of $\mathcal{F}$ in Assumption 6.9. By the assumption, the maximal number of equalities we can have among the vectors is at most the number of variables:

$$\underbrace{m}_{\text{of } \boldsymbol{\lambda}} + \underbrace{|\mathcal{K}| - |\mathcal{K}_1|}_{\text{of } \alpha^k \text{ for } k \in \mathcal{K} \setminus \mathcal{K}_1} + 1.$$

Hence

$$m + |\mathcal{K}| - |\mathcal{K}_1| + 1 \;\geq\; \sum_{k \in \mathcal{K}} |S_k| \;=\; \sum_{k \in \mathcal{K} \setminus \mathcal{K}_1} |S_k| + \sum_{k \in \mathcal{K}_1} |S_k|.$$

Next, we claim that for any $k \in \mathcal{K} \setminus \mathcal{K}_1$, we must have $|S_k| \geq 2$. Indeed, if $|S_k| = 1$, then $\mathcal{X}^k$ is an interval, and the only way to have multiple solutions is if the subproblem cost vector is $\mathbf{0}$, i.e., $k \in \mathcal{K}_1$, a contradiction. As a result, we can further lower bound this last quantity by

$$m + |\mathcal{K}| - |\mathcal{K}_1| + 1 \;\geq\; 2\left(|\mathcal{K}| - |\mathcal{K}_1|\right) + |\mathcal{K}_1|.$$

Rearranging, proves $|\mathcal{K}| \leq m + 1$.

Finally, for $k \in \mathcal{K}$, $\|\bar{\boldsymbol{x}}^k - \tilde{\boldsymbol{x}}^k\left(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{x}}^0\right)\|_1 \leq |S_k|$. Hence, $\|\bar{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\left(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{x}}^0\right)\|_1 \leq m \cdot S_{\max}$. This completes the proof. □

Lemma F.2 (Dual Variables Bounded by Plug-in) *Under Assumption 6.10, we see for all* $\boldsymbol{\theta} \in \{0\} \times \bar{\Theta}$*,* $\boldsymbol{z} \in \mathbb{R}^n$*, and* $\boldsymbol{x}^0 \in \mathcal{X}^0$ *that*

$$\|\boldsymbol{\lambda}(\boldsymbol{z}, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^0)\|_1 \;\leq\; \frac{2}{s}\|\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})\|_1.$$

Proof of Lemma F.2 Fix $\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{x}^0$. To reduce notation, let $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}}\left(\boldsymbol{z}, \boldsymbol{\lambda}\left(\boldsymbol{z}, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^0\right), \boldsymbol{x}^0, \bar{\boldsymbol{\theta}}\right)$ and

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})^\top \boldsymbol{x} + \boldsymbol{\lambda}^\top\left(\boldsymbol{A}^0(\boldsymbol{x}^0)\boldsymbol{x} - \boldsymbol{b}^0(\boldsymbol{x}^0)\right),$$

By optimality,

$$\mathcal{L}\left(\tilde{\boldsymbol{x}}, \boldsymbol{\lambda}\left(\boldsymbol{z}, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^0\right)\right) \;\geq\; \mathcal{L}(\tilde{\boldsymbol{x}}, \mathbf{0}) \;\geq\; -\|\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})\|_1$$

where the last inequality holds since decision variables are bounded between 0 and 1.

Since $\boldsymbol{\lambda}(\boldsymbol{z}, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^0) \geq 0$, we see $\|\boldsymbol{\lambda}(\boldsymbol{z}, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^0)\|_1 = \boldsymbol{e}^\top \boldsymbol{\lambda}(\boldsymbol{z}, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^0)$. Thus,

$$\begin{aligned} \|\boldsymbol{\lambda}(\boldsymbol{z}, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^0)\|_1 \;\leq\; & \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \; \boldsymbol{e}^\top \boldsymbol{\lambda} \\ & \text{s.t.} \;\; \min_{\boldsymbol{x} \in \tilde{\mathcal{X}}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \;\geq\; -\|\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})\|_1, \end{aligned}$$

where $\tilde{\mathcal{X}} = \left\{ \boldsymbol{x} : \boldsymbol{x}^k \in \mathcal{X}^k(\boldsymbol{x}^0), \forall k = 1, \ldots, K \right\}$. We upper bound the optimization problem by relaxing our one constraint with penalty $1/\bar{s} > 0$ to show,

$$
\begin{aligned}
&\|\boldsymbol{\lambda}(\boldsymbol{z}, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^0)\|_1 \\
&\quad \leq \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \; \boldsymbol{e}^\top \boldsymbol{\lambda} + \frac{1}{\bar{s}} \left(\min_{\boldsymbol{x} \in \tilde{\mathcal{X}}} \; \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) + \|\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})\|_1 \right) \\
&\quad = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \; \boldsymbol{e}^\top \boldsymbol{\lambda} + \frac{1}{\bar{s}} \left(\min_{\boldsymbol{x} \in \tilde{\mathcal{X}}} \; \boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})^\top \boldsymbol{x} + \boldsymbol{\lambda}^\top \left(\boldsymbol{A}^0(\boldsymbol{x}^0)\boldsymbol{x} - \boldsymbol{b}^0(\boldsymbol{x}^0) \right) + \|\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})\|_1 \right) \\
&\quad \leq \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \; \boldsymbol{e}^\top \boldsymbol{\lambda} + \frac{1}{\bar{s}} \left(\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})^\top \bar{\boldsymbol{x}} + \boldsymbol{\lambda}^\top \left(\boldsymbol{A}^0(\boldsymbol{x}^0)\bar{\boldsymbol{x}} - \boldsymbol{b}^0(\boldsymbol{x}^0) \right) + \|\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})\|_1 \right), \\
&\quad = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \; \frac{1}{\bar{s}} \left(\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})^\top \bar{\boldsymbol{x}} + \boldsymbol{\lambda}^\top \left(\boldsymbol{A}^0(\boldsymbol{x}^0)\bar{\boldsymbol{x}} + \bar{s}\boldsymbol{e} - \boldsymbol{b}^0(\boldsymbol{x}^0) \right) + \|\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})\|_1 \right)
\end{aligned}
$$

where $\bar{\boldsymbol{x}} \in \mathcal{X}^{WC}$ satisfies our strict feasibility assumption (Assumption 6.10) for $\boldsymbol{x}^0$. The first inequality is our upper-bound from relaxing the one constraint. The first equality expands $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda})$. The second inequality holds since $\bar{\boldsymbol{x}} \in \mathcal{X}^{WC} \subseteq \tilde{\mathcal{X}}$. The second equality collects terms.

Since $\boldsymbol{A}^0(\boldsymbol{x}^0)\bar{\boldsymbol{x}} + \bar{s}\boldsymbol{e} - \boldsymbol{b}^0(\boldsymbol{x}^0) \leq \mathbf{0}$ by the strict feasibility assumption,

$$
\|\boldsymbol{\lambda}(\boldsymbol{z}, \bar{\boldsymbol{\theta}}, \boldsymbol{x}^0)\|_1 \;\leq\; \frac{1}{\bar{s}} \left(\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})^\top \bar{\boldsymbol{x}} + \|\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})\|_1 \right) \;\leq\; \frac{2}{\bar{s}} \|\boldsymbol{T}(\boldsymbol{z}, \bar{\boldsymbol{\theta}})\|_1
$$

Note the bound holds for any choice of $\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{x}^0$. This completes the proof. □

Appendix G: Details for Numerics and Examples

G.1. Implementation Details for Numerics

We describe the implementation of the one-shot VGC, multi-shot VGC, and Stein correction for the speed hump case study. For the lifted policy classes, we let $\lambda(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$ be the optimal dual variable corresponding to the budget constraint for the cost vector $\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}})$.

G.1.1. One-Shot VGC Implementation The primary challenge of implementing the one-shot VGC is choosing a distribution for $\boldsymbol{\delta}_h$ for each policy and evaluating the expectation

$$
\mathbb{E}\left[V(\boldsymbol{T}(\boldsymbol{Z}, \bar{\boldsymbol{\theta}}) + \boldsymbol{\delta}_h, \rho) \,\middle|\, \boldsymbol{Z} \right].
$$

In Section 7.2, for a fixed $\boldsymbol{L}$, we choose $\boldsymbol{\delta}_h \sim \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{\Sigma}}^h\right)$ and evaluate the expectation by Monte Carlo simulation with 25 samples of $\boldsymbol{\delta}_h$. To construct $\hat{\boldsymbol{\Sigma}}^h$, consider the matrix $\boldsymbol{\Sigma}^h = h\left(\boldsymbol{L}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{L}^\top\right) + h^2\boldsymbol{\Sigma}$. If $\boldsymbol{\Sigma}^h$ is positive semi-definite, then $\hat{\boldsymbol{\Sigma}}^h = \boldsymbol{\Sigma}^h$. If $\boldsymbol{\Sigma}^h$ is not positive semi-definite, then we form the eigenvalue decomposition $\boldsymbol{\Sigma}^h = \boldsymbol{Q}^h \boldsymbol{\Lambda}^h \boldsymbol{Q}^{h\top}$ and construct

$$
\hat{\boldsymbol{\Sigma}}^h = \rho \boldsymbol{Q}^h \boldsymbol{\Lambda}_+^h \boldsymbol{Q}^{h\top}, \text{ where } \boldsymbol{\Lambda}_+^h = \operatorname{diag}\left\{ \max(0, \Lambda_{11}^h), \ldots, \max(0, \Lambda_{nn}^h) \right\} \text{ and } \rho = \frac{\operatorname{tr}\left(\boldsymbol{\Lambda}^h\right)}{\operatorname{tr}\left(\boldsymbol{\Lambda}_+^h\right)}.
$$

This method of "correcting" a non-positive definite matrix was also used in Chan and Wood, 1997 when simulating stationary Gaussian random fields.

To estimate the conditional expectation, we simulate $\boldsymbol{\delta}_h^i \sim \mathcal{N}\left(\mathbf{0}, \hat{\mathbf{\Sigma}}^h\right)$ for $i=1,\ldots,S$. The resulting one-shot VGC out-of-sample performance estimate is

$$\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - \frac{\left(\frac{1}{S}\sum_{i=1}^S V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\boldsymbol{\delta}_h^i,\rho)\right) - V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho)}{h}.$$

In our experiments, we choose $S=25$.

G.1.2. Multi-Shot VGC Implementation For the multi-shot VGC, we let $\delta_j^h \sim \mathcal{N}(0,(h^2+2h)/\nu_j)$ and define $\overline{\boldsymbol{\delta}}_h(\bar{\boldsymbol{\theta}})$, $a_j(\bar{\boldsymbol{\theta}}) \sim L_{jj}(\bar{\boldsymbol{\theta}})\delta_j^h \boldsymbol{e}_j$, $L_{jj}(\bar{\boldsymbol{\theta}})$ with probability $1/n$ for every j. We simulate $\overline{\boldsymbol{\delta}}_h^i(\bar{\boldsymbol{\theta}}), a_j^i(\bar{\boldsymbol{\theta}})$ from the distribution. The resulting multi-shot VGC out-of-sample performance estimate is

$$\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - \frac{1}{S}\sum_{i=1}^S \frac{V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}})+\overline{\boldsymbol{\delta}}_h^i(\bar{\boldsymbol{\theta}}),\rho) - V(\boldsymbol{T}(\boldsymbol{Z},\bar{\boldsymbol{\theta}}),\rho)}{a_j^i(\bar{\boldsymbol{\theta}})h/n}.$$

In our experiments, we choose $S=n$.

G.1.3. Stein Correction Implementation We implement the Stein correction on the lifted plug-in policy which was also proposed in Gupta and Rusmevichientong, 2021. Since the lifted policy fully decouples the optimization problem, the decisions correspond to indicator variables. The Stein correction out-of-sample performance estimate is calculated by

$$\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - \sum_{j=1}^n \frac{1}{2h\sqrt{\nu_j}}\left(\mathbb{I}\left\{T_j\left(\boldsymbol{Z}+\frac{h}{\sqrt{\nu_j}},\bar{\boldsymbol{\theta}}\right) - \lambda(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) \geq 0\right\} - \mathbb{I}\left\{T_j\left(\boldsymbol{Z}-\frac{h}{\sqrt{\nu_j}},\bar{\boldsymbol{\theta}}\right) - \lambda(\boldsymbol{Z},\bar{\boldsymbol{\theta}}) \geq 0\right\}\right)$$

G.1.4. Lifted One-Shot VGC To define the one-shot VGC on the lifted plug-in policy, we first let $\tilde{V}(t) = t\cdot\mathbb{I}\{t\geq 0\}$ which represents the optimal plug-in objective value with plug-in t for one element of the fully decoupled optimization problem. Again, we let $\boldsymbol{\delta}_h^i \sim \mathcal{N}\left(\mathbf{0},\hat{\mathbf{\Sigma}}^h\right)$ for $i=1,\ldots,S$ and $S=25$. The one-shot VGC out-of-sample performance estimate is

$$\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - \frac{1}{S}\sum_{i=1}^S \frac{1}{h}\left(\tilde{V}\left(T_j\left(\boldsymbol{Z},\bar{\boldsymbol{\theta}}\right) + \delta_j^{h,i} - \lambda(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\right) - \tilde{V}\left(T_j\left(\boldsymbol{Z},\bar{\boldsymbol{\theta}}\right) - \lambda(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\right)\right)$$

G.1.5. Lifted Multi-Shot VGC Similar to the multi-shot VGC on the original policy, we have $\delta_j^{h,i} \sim \mathcal{N}(0,(h^2+2h)/\nu_j)$ for $i=1,\ldots,S$ and $j=1,\ldots,n$. The out-of-sample estimate is then computed as

$$\boldsymbol{Z}^\top \boldsymbol{x}(\boldsymbol{Z},\boldsymbol{\theta}) - \frac{1}{S}\sum_{i=1}^S \frac{n}{L_{jj}(\bar{\boldsymbol{\theta}})h}\left(\tilde{V}\left(T_j\left(\boldsymbol{Z},\bar{\boldsymbol{\theta}}\right) + L_{jj}(\bar{\boldsymbol{\theta}})\delta_j^{h,i} - \lambda(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\right) - \tilde{V}\left(T_j\left(\boldsymbol{Z},\bar{\boldsymbol{\theta}}\right) - \lambda(\boldsymbol{Z},\bar{\boldsymbol{\theta}})\right)\right)$$

for $S=25$. Note that $T_j\left(\boldsymbol{Z},\bar{\boldsymbol{\theta}}\right) + L_{jj}(\bar{\boldsymbol{\theta}})\delta_j^{h,i} = T_j\left(\boldsymbol{Z}+\delta_j^{h,i},\bar{\boldsymbol{\theta}}\right)$ since we are only looking at the j^{th} element.

G.2. Additional Results

We include a larger version of Fig. 3 that includes the VGC variants with the lifted policy for the reader's convenience.

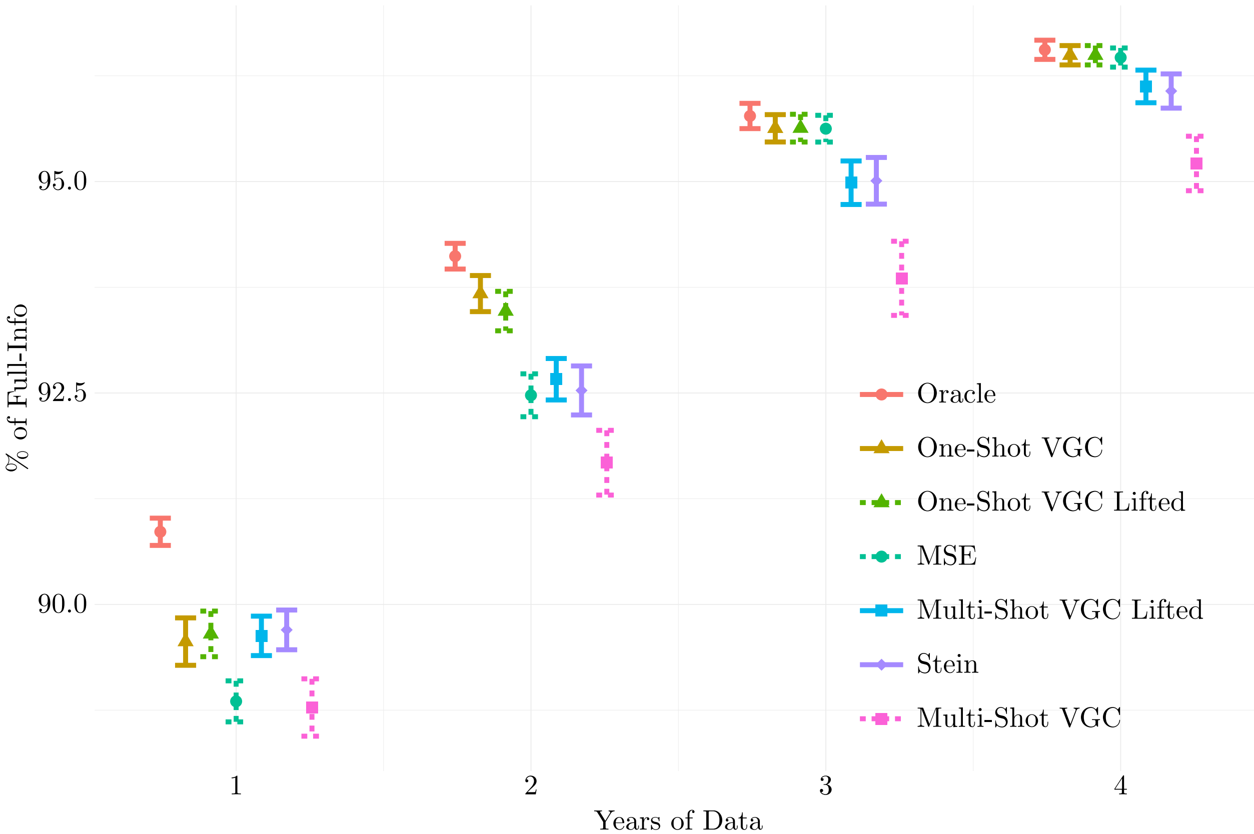

Figure EC.1 **Performance Results**. We compare the estimated expected out-of-sample performance of our method to various benchmarks and various over 100 trials. The error bars are 95% confidence intervals. The experiments vary the amount of data available but keep the number of decisions fixed.