Achieving First-Order Statistical Improvements in Data-Driven Optimization

Henry Lam[*] KHL2114@COLUMBIA.EDU
Tianyu Wang[*] TW2837@COLUMBIA.EDU
Department of Industrial Engineering and Operations Research, Columbia University

Abstract

Recent proliferation of data-optimization integration has led to a range of methods that aim to improve the statistical performance of data-driven optimization decisions. However, while many of these methods are motivated intuitively from a robustness or regularization perspective, their resulting statistical benefits are often less clear and, even if available, are argued in a case-by-case fashion. We provide a systematic dissection of data-driven optimization formulations using the view of "directionally perturbed" empirical optimization, which demonstrably covers most of the existing formulations. On the negative side, we argue that under mild smoothness conditions, any such formulations can result in at best second-order improvements. On the positive side, we show that in the presence of auxiliary information such as the availability of additional unsupervised data, we can construct a principled methodology by building connections to the concept of Monte Carlo control variate, to achieve general first-order improvements in terms of solution regret.

1. Introduction

We consider data-driven stochastic optimization problem of the form:

$$\min_{\theta\in\Theta}\left\{Z(\theta) := \mathbb{E}_{\xi\sim\mathbb{P}^*}[\ell(\theta;\xi)]\right\}, \tag{1}$$

where $\ell(\theta;\cdot)$ is a known cost function, $\Theta := \{\theta \in \mathbb{R}^{D_\theta} | F_j(\theta) \leq 0, j \in J\}$ is the set of feasible decisions, and ξ is a random perturbation distributed according to the distribution $\mathbb{P}^*$. However, the decision maker only has access to n iid samples $\mathcal{D}_n := \{\xi_i\}_{i=1}^n$. The goal is to use the data $\mathcal{D}_n$ to identify a decision with the lowest expected cost under the *true* distribution $\mathbb{P}^*$. This problem setup is widely adopted in practice from empirical risk minimization in machine learning [22], to general stochastic optimization problems [32] with applications such as supply chain management [4] and portfolio optimization [8].

Among all data-driven optimization methods, the most straightforward approach is to replace the unknown $\mathbb{P}^*$ with the empirical measure $\hat{\mathbb{P}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i}$ in (1) and obtain an empirical solution $\hat{\theta}$, i.e., solve $\hat{\theta} \in \mathrm{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(\theta;\xi_i)$. It is widely acknowledged that the empirical solution performs poorly under limited samples [18]. To improve performance, various approaches have been developed to obtain alternative solution θ_{RM}, which we refer to broadly as *robust methods*. These include, in the optimization literature, regularized and distributionally robust optimization (DRO) [26] to tackle data uncertainty and, in the statistics/machine learning literature, over-identified generalized method of moments (GMM) [21], shrinkage [17] and transfer learning [6] when we

[*] Authors ordered alphabetically.

have related data or information. While empirical studies often observe that these methods could outperform the empirical solution in terms of excess risk, i.e., $Z(\theta_{RM}) < Z(\hat{\theta})$, a fundamental gap remains: *When should these methods outperform and by how much, and what determines the order of improvement over the empirical solution?*

Our Contributions. In this paper, we address these questions by systematically characterizing the statistical performance of solutions generally perturbed from the empirical solution, which we argue to encompasses essentially all of the robust methods known in the literature and beyond. We analyze their impact on the excess risks:

$$R(\theta) = Z(\theta) - Z(\theta^*), \text{ where } \theta^* \in \operatorname*{argmin}_{\theta\in\Theta} Z(\theta). \tag{2}$$

Throughout the main body of the paper, our comparison centers on the expected excess risk, contrasting $\mathbb{E}_{\mathcal{D}_n}[R(\theta_{RM})]$ with $\mathbb{E}_{\mathcal{D}_n}[R(\hat{\theta})]$. Our key findings are: Robust methods yield first-order improvements only when the auxiliary information they leverage (e.g., moment conditions) is correct under $\mathbb{P}^*$; otherwise, improvements are at best higher-order and insignificant compared with the empirical solution. Building on this analysis, we propose a principled framework for utilizing the auxiliary information and constructing perturbations that achieve the largest possible first-order improvement, establishing a novel connection between robust methods and Monte Carlo control variates. To the best of our knowledge, we are the first to provide definitive performance comparisons between various robust methods and empirical solutions for general cost functions under different data environments. Related literature is reviewed in Appendix A, with additional results and proofs in Appendices B–E, and experimental results in Appendix F.

2. Preliminaries

In this section, we formalize the notion of improvement relative to the empirical solution.

Definition 1 (Orders of Improvements) *For empirical solution $\hat{\theta}$, we say the excess risk rate of $R(\hat{\theta})$ (in (2)) decays polynomially in n with exponent γ and constant C if $\mathbb{E}_{\mathcal{D}_n}[R(\hat{\theta})] = C/n^{\gamma} + o(n^{-\gamma})$.*

For a robust-method solution θ_{RM} with corresponding exponent γ_{RM} and constant C_{RM}, we classify its performance improvement as: (i) First-order improvement: $\gamma_{RM} > \gamma$ or $\gamma_{RM} = \gamma$ and $C_{RM} < C$; (ii) Second-order improvement: all other cases where $\mathbb{E}_{\mathcal{D}_n}[R(\theta_{RM})] < \mathbb{E}_{\mathcal{D}_n}[R(\theta)]$.

Under this definition, if only second-order improvement is achievable, then the relative gain satisfies $\frac{\mathbb{E}_{\mathcal{D}_n}[R(\hat{\theta})-R(\theta_{RM})]}{\mathbb{E}_{\mathcal{D}_n}[R(\hat{\theta})]} \to 0$ as $n \to \infty$; otherwise this ratio converges to a strictly positive limit.

Assumption 1 (Optimality Condition) *θ^* satisfies the KKT conditions for $\min_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}^*}\ell(\theta;\xi)$. The empirical solution $\hat{\theta}$ satisfies $\mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\hat{\theta};\xi)] = \min_{\theta\in\Theta} \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta;\xi)] + o_p\left(n^{-1}\right)$.*

Assumption 2 (Local Quadratic Growth) *$Z(\theta)$ and each constraint $F_j(\theta)$ in Θ are twice continuously differentiable in a neighborhood of θ^*. The Lagrangian $L(\theta,\alpha) = Z(\theta) + \sum_{j\in J} \alpha_j F_j(\theta)$ satisfies a strong second–order sufficient condition at (θ^*,α^*), where α^* denotes the associated KKT multiplier: There exists $\mu > 0$ such that $v^\top \nabla^2_{\theta\theta} L(\theta^*,\alpha^*)\, v \geq \mu\|v\|_2^2$ for all feasible directions v in the tangent space at θ^*.*

Assumptions 1- 2 are standard for smooth objectives: The first provides local quadratic curvature for the augmented Lagrangian; the second ensures regularity of the population solution and that $\hat{\theta}$ is attained up to negligible error. Under these, the empirical solution admits the asymptotic linearization $\hat{\theta} - \theta^* = \frac{1}{n}\sum_{i=1}^{n} IF(\xi_i) + o_p\left(n^{-1/2}\right)$ with a mean-zero *influence function* $IF(\xi)$ [11, 20].

3. Perturbed Solutions: Connections with Robust Methods and Statistical Analysis

We consider the following perturbed solution to understand whether robust methods achieve the first-order performance improvement:

Definition 2 ("Directionally Perturbed" Empirical Solution) *The perturbation of the empirical solution is defined as:*

$$\hat{\theta}_{\lambda,M} = \Pi_\Theta\left(\hat{\theta} + \frac{\lambda}{n}\sum_{i=1}^{n} M(\hat{\theta};\xi_i)\right). \tag{3}$$

where Π_Θ is the projection operator to Θ. Here, $\lambda \in \mathbb{R}$ is an adjustment scale and $M(\theta;\xi)$ is a perturbation function that is differentiable in θ.

The scaling λ in (3) is chosen for analysis. More general adjustments beyond the scale λ are treated in Section 4. Many robust methods admit approximations of the form (3).

Theorem 3 (Unification of Robust Methods) *For any $\lambda > 0$, the solutions $\hat{\theta}_\lambda$ of robust methods admit the representation $\hat{\theta}_{\lambda,M} = \Pi_\Theta\left(\hat{\theta} + \frac{1}{n}\sum_{i=1}^{n}\tilde{H}(\lambda;\xi_i)\widetilde{M}(\theta^*;\xi_i) + o_p\left(\lambda \vee \frac{1}{\sqrt{n}}\right)\right)$. In many robust methods, $\tilde{H}(\lambda;\xi)$ is constant in ξ.*

The representation in Theorem 3 covers multiple streams of robust methods: (i) Optimization: explicit regularization methods, Wasserstein DRO methods, (generalized) f-divergence (such as χ^2-divergence, Conditional Value-at-Risk (CVaR)); (ii) Econometrics: Generalized method of moment; (iii) Statistical methods: shrinkage and transfer learning methods. Related small-λ expansions appear in prior work [15, 28] for a subset of methods we consider, while Theorem 3 is stated for general λ, including constant order and not necessarily shrinking to zero.

Example 1 (Auxiliary Information for DRO) *Consider DRO problems with $\Theta = \mathbb{R}^{D_\theta}$, i.e., $\hat{\theta}_\lambda \in \operatorname{argmin}_\theta \max_{\mathbb{P}:d(\mathbb{P},\hat{\mathbb{P}}_n)\le\lambda} \mathbb{E}_{\mathbb{P}}[\ell(\theta;\xi)]$. Then the perturbation function is: (i) When d is χ^2-divergence [10], $\widetilde{M}(\theta;\xi) \propto \nabla_\theta \ell(\theta;\xi)\ell(\theta;\xi)$; (ii) When d is a CVaR distance [30], $\widetilde{M}(\theta;\xi) \propto \nabla_\theta \ell(\theta;\xi)\mathbf{1}_{\ell(\theta;\xi)>\eta^*(\lambda)}$ when $\eta^*(\lambda)$ is the $(1-\lambda)$-quantile of $\ell(\theta;\xi)$ under $\xi \sim \mathbb{P}^*$.*

We present the following result with respect to the statistical improvement of robust methods.

Theorem 4 (Orders of Performance Improvement) *Suppose Assumptions 1 and 2 hold. Denote $\Gamma := Cov_{\mathbb{P}^*}\left[IF(\xi), M(\theta^*;\xi)\right], \Sigma_0 := \mathbb{E}_{\mathbb{P}^*}[IF(\xi)IF(\xi)^\top]$ and $\pi(\theta) := \mathbb{E}_{\mathbb{P}^*}[M(\theta;\xi)]$. Then (i) If $\pi(\theta^*) = 0$ and the* ***non-orthogonality condition*** *holds: Tr $\left[\left(\nabla_\theta \pi(\theta^*)\Sigma_0 + \Gamma\right)\nabla^2_{\theta\theta} Z(\theta^*)\right] \neq 0$, then there exists a constant λ such that $\mathbb{E}_{\mathcal{D}_n}[R(\hat{\theta}) - R(\hat{\theta}_\lambda)] = \Theta(1/n) > 0$; (ii) Otherwise, $\max_\lambda \mathbb{E}_{\mathcal{D}_n}[R(\hat{\theta}) - R(\hat{\theta}_\lambda)] = o(1/n)$.*

Since $\mathbb{E}_{\mathcal{D}_n}[R(\hat{\theta})] = \Theta(1/n)$ under Assumption 2 [28], Theorem 4 implies that a constant λ yields a first-order improvement in expected risk under the case (i), suggesting the need for nonlocal adjustments to obtain substantial gains. Otherwise, at best a second-order improvement is achievable.

We also extend these comparisons to weighted empirical optimization solutions and comparisons of distributional aspects beyond the mean in Appendix D, where first-order rates of the empirical solution $n^{-\gamma}$ may be slower than n^{-1} yet the conclusions above still apply.

Connecting the perturbation form in Definition 2 with Theorem 3, many existing robust methods cannot markedly outperform the empirical solution unless $\mathbb{E}_{\mathbb{P}^*}[\tilde{M}(\theta^*;\xi)] = 0$ and the non-orthogonality condition is met, together with an appropriate choice of λ. When M is θ-independent, that is $\nabla\pi(\theta) = 0$, the non-orthogonality condition reduces to $\mathbb{E}_{\mathbb{P}^*}[\tilde{M}(\theta;\xi)^\top IF(\xi)] \neq 0$ since $\nabla^2_{\theta\theta}Z(\theta^*)$ is positive semi-definite. That is, whenever the auxiliary moment carries signal different from $IF(\xi)$, one expects a nonnegligible improvement.

Corollary 5 (Linear Regression with Laplace Noise) *If $\ell(\theta;\xi) = (\theta^\top X - Y)^2$ with $\xi = (X,Y)^\top$ and $Y = \theta^\top X + \epsilon$ with symmetric Laplace noise ϵ, a CVaR regularizer with a negative scale $\lambda^* \in (-1,0)$ yields a first-order improvement over the empirical solution $\hat{\theta}$. The negative CVaR regularizer is computable via $\hat{\theta}_\lambda = 2\hat{\theta} - \hat{\theta}_{-\lambda}$ when $\lambda < 0$.*

4. Achieving First-order Improvement

Given the importance of the correct auxiliary information from robust methods, we now show how to use some auxiliary information, such as partial moment knowledge or invariant representations, to construct an optimal first-order perturbation of the empirical solution.

When is first-order improvement possible? In general, if there is no M such that $\mathbb{E}[M(\theta^*;\xi)] = 0$, then any robust method induced by such moments yields at most a second-order improvement, cf. Theorem 4. In contrast, if there exists M with $\mathbb{E}[M(\theta^*;\xi)] = 0$, we can choose an adjustment matrix H (when $H = \lambda I$, it reduces to (3)) so that the perturbed estimator attains the best first-order improvement over $\hat{\theta}$:

$$\hat{\theta}_{H,M} = \Pi_\Theta\left(\hat{\theta} + \frac{H}{n}\sum_{i=1}^{n} M(\hat{\theta};\xi_i)\right). \tag{4}$$

In many stochastic optimization problems, beyond $\mathcal{D}_n$ one may have distributional moment knowledge with respect to ξ, i.e., $M(\theta;\xi) \equiv M(\xi)$, as in moment–based ambiguity sets in DRO [9]. In general machine learning applications, there may be multiple sources of auxiliary information [3] with one special example being semi-supervised learning [33].

Example 2 (Moment Invariance from Related Domains) *Given large unlabeled or weakly labeled samples from K related domains, $\{(\theta_i, \xi_{i,j})\}_{i\in[K];j\in[N]}$ with $N \gg n$, there exists some M so that an estimating equation $\mathbb{E}_{\mathbb{P}_i}[M(\theta_i;\xi)] = 0$ holds in domain i.*

Our main procedure integrates moment information or related-domain side information to improve the empirical solution for downstream decision making, as described below.

First-Order-Improving Perturbation. Let $\mathcal{M}_\phi$ be a candidate class for $M(\cdot)$ indexed by ϕ. Motivated by a second–order expansion of the performance gap:

$$Z(\hat{\theta}_{H,M}) - Z(\theta^*) \approx \tfrac{1}{2}\left(\hat{\theta}_{H,M} - \theta^*\right)^\top \nabla^2_{\theta\theta}Z(\theta^*)\left(\hat{\theta}_{H,M} - \theta^*\right),$$

we directly minimize the dominant quadratic term to obtain the optimal $\hat{\theta}_{H,M}$ as in (4):

$$(\hat{H}, \widehat{M}) \in \operatorname*{argmin}_{H,\, M \in \mathcal{M}_\phi(\delta)} \left[\sum_{i=1}^{n} \left\| \widehat{IF}(\xi_i) + H\, M\left(\hat{\theta}; \xi_i\right) \right\|^2_{\hat{I}(\hat{\theta})} \right] \text{ s.t. } \hat{\theta} + \frac{H}{n} \sum_{i=1}^{n} M\left(\hat{\theta}; \xi_i\right) \in \Theta, \tag{5}$$

where $\hat{I}(\hat{\theta}) = \nabla_{\theta\theta} \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\hat{\theta}; \xi)]$ that approximates $\nabla^2_{\theta\theta} Z(\theta^*)$ and $\widehat{IF}(\xi)$ estimates the influence function $IF(\xi)$. For example, when $\Theta = \mathbb{R}^{D_\theta}$ [25], $IF(\xi) = -[\nabla^2_{\theta\theta} Z(\theta^*)]^{-1} \nabla_\theta \ell(\theta^*; \xi)$ and $\widehat{IF}(\xi) = -[\hat{I}(\hat{\theta})]^{-1} \nabla_\theta \ell(\hat{\theta}; \xi)$. $\mathcal{M}_\phi(\delta)$ is set to contain at least one M with $\mathbb{E}_{\mathbb{P}^*}[M(\theta^*; \xi)] = 0$ with probability at least $1 - \delta$. For example, $\mathcal{M}(\delta) = \Big\{ M \in \mathcal{M}_\phi : \max_{i \in [K]} \left\| \sum_{j=1}^{N} M\left(\theta_i; \xi_{i,j}\right)/N \right\|_2^2 \leq \epsilon_K\left(\mathcal{M}_\phi, \delta\right) \Big\}$ with $\epsilon_K\left(\mathcal{M}_\phi, \delta\right) = \text{Comp}\left(\mathcal{M}_\phi\right) \log(K/\delta)/N$, where $\text{Comp}(\mathcal{M}_\phi)$ captures the complexity of the class and the bound absorbs distributional constants for M [5, 37].

Theorem 6 (Performance Guarantee for the Perturbed Solution) *Suppose Assumptions 1, 2 and 5 hold. and that* $\sup_{H \in \mathcal{H}} \|H\| < \infty$. *Let* (H^*, M^*) *be an oracle solution to:*

$$(H^*, M^*) \in \arg \min_{H, M \in \mathcal{M}_\phi} \mathbb{E}_{\mathbb{P}^*}\left[\|IF(\xi) + HM(\theta^*; \xi)\|^2_{\nabla^2_{\theta\theta} Z(\theta^*)} \right],$$

Let $(\hat{H}, \widehat{M})$ *be any empirical solution to* (5) *with* $\widehat{M} \in \mathcal{M}(\delta)$ *for* $\delta = \Theta(1/n)$ *in Example 2. Then* $R(\hat{\theta}_{\hat{H}, \widehat{M}}) = R(\hat{\theta}_{H^*, M^*}) + o_p(1/n)$.

The performance of the data-driven perturbation $\hat{\theta}_{\hat{H}, \widehat{M}}$ closely matches the oracle performance, indicating that it achieves the maximum possible first-order improvement. In practice, optimizing (5) is typically done via alternating minimization over H and M.

Control variates and semi-parametric efficiency. Finally, we connect our principal first-order improvements with control variates and semi-parametric efficiency. First, we recenter the influence-function representation with its mean by viewing the target as the mean of the influence function, $\theta^* = \mathbb{E}[IF(\xi)]$, so the empirical estimator can be regarded a Monte Carlo average $\hat{\theta} \approx \frac{1}{n} \sum_{i=1}^{n} IF(\xi_i)$, with each ξ_i producing one realization $IF(\xi_i)$. The perturbed estimator that achieves first-order improvement is a calibrated control-variates correction,

$$\hat{\theta}_{H,M} \approx \frac{1}{n} \sum_{i=1}^{n} (IF(\xi_i) + H\, M(\theta^*; \xi_i)) =: \tilde{\theta}_{H,M^*}, \tag{6}$$

where the auxiliary moment $M(\theta^*; \xi)$ has known (typically zero) expectation and the calibration matrix H is chosen to exploit its covariance with $IF(\xi)$ to reduce variance—exactly the classical control-variates principle in Monte Carlo simulation [29, 31]. Thus, with limited samples, the calibrated estimator lowers variance (and hence risk) relative to the empirical mean while preserving unbiasedness to first order, yielding the first-order gains established above.

Besides, our result does not contradict with the local asymptotic minimax normality that the lower bound of empirical solutions (or M-estimators) cannot be improved (i.e, Chapter 15 in [38]). Instead, we are restricted to a space where some moment equation holds. If the moment function $M(\theta; \cdot)$ is θ-independent, we can show that the corresponding estimator $\tilde{\theta}_{H,M^*}$ defined in (6) attains semi-parametric efficiency in the projected space $\{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[M(\theta^*; \xi)] = 0\}$ [23] with details in Theorem 22 of Appendix E.

5. Conclusion

In this paper, we characterize when robust methods beat the empirical estimator and by how much: only provided with auxiliary moment and non-orthogonality conditions, constant-size regularization delivers first-order gains, and provide a practical recipe achieving oracle-level performance. Future work includes extending to smoother losses, developing nonasymptotic empirical-versus-robust comparisons, and designing practical moment equations for related-data settings.

References

[1] Michael Albert, Max Biggs, Ningyuan Chen, and Guan Wang. Post-estimation adjustments in data-driven decision-making with applications in pricing. *arXiv preprint arXiv:2507.20501*, 2025.

[2] Edward Anderson and Andy Philpott. Improving sample average approximation using distributional robustness. *INFORMS Journal on Optimization*, 4(1):90–124, 2022.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Gah-Yi Ban and Cynthia Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2019.

[5] Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 2005.

[6] Hamsa Bastani. Predicting with proxies: Transfer learning in high dimension. *Management Science*, 67(5):2964–2984, 2021.

[7] Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

[8] Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-variance portfolio selection with wasserstein distances. *Management Science*, 68(9):6382–6410, 2022.

[9] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

[10] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019.

[11] John C. Duchi and Feng Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1), 2021. doi: 10.1214/19-AOS1831.

[12] Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

[13] Qi Feng and J George Shanthikumar. The framework of parametric and nonparametric operational data analytics. *Production and Operations Management*, 32(9):2685–2703, 2023.

[14] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 2022.

[15] Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. Calibration of distributionally robust empirical optimization models. *Operations Research*, 69(5):1630–1650, 2021.

[16] Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. A data-driven approach to beating saa out of sample. *Operations Research*, 2023.

[17] Vishal Gupta and Nathan Kallus. Data Pooling in Stochastic Optimization. *Management Science*, 68(3):1595–1615, 2022. doi: 10.1287/mnsc.2020.3933.

[18] Vishal Gupta and Paat Rusmevichientong. Small-Data, Large-Scale Linear Optimization with Uncertain Objectives. *Management Science*, 67(1):220–241, 2021. doi: 10.1287/mnsc.2019.3554.

[19] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[20] Frank R. Hampel. The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962.

[21] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054, 1982.

[22] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[23] Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022.

[24] Nan Jiang and Weijun Xie. Distributionally favorable optimization: A framework for data-driven decision-making with endogenous outliers. *SIAM Journal on Optimization*, 34(1): 419–458, 2024.

[25] Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

[26] Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *arXiv preprint arXiv:2411.02549*, 2024.

[27] Henry Lam. Sensitivity to serial dependency of input processes: A robust approach. *Management Science*, 64(3):1311–1327, 2018.

[28] Henry Lam. On the impossibility of statistically improving empirical optimization: A second-order stochastic dominance perspective. *arXiv preprint arXiv:2105.13419*, 2021.

[29] Barry L Nelson. Control variate remedies. *Operations Research*, 38(6):974–992, 1990.

[30] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

[31] Reuven Y Rubinstein and Ruth Marcus. Efficiency of multivariate control variates in monte carlo simulation. *Operations Research*, 33(3):661–677, 1985.

[32] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[33] Shanshan Song, Yuanyuan Lin, and Yong Zhou. A general m-estimation theory in semi-supervised framework. *Journal of the American Statistical Association*, 119(546):1065–1075, 2024.

[34] Tobias Sutter, Bart PG Van Parys, and Daniel Kuhn. A general framework for optimal data-driven optimization. *arXiv preprint arXiv:2010.06606*, 2020.

[35] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[36] Bart PG Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 2020.

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

[38] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Appendix A. Related Literature

Recent work in diagnosing statistical limits of data-driven optimization have been examined the performance of existing data-driven solutions, or comparing robust methods with the empirical solution. Sutter et al. [34], Van Parys et al. [36] showed that some variants of DRO achieve the optimal Pareto frontier between the out-of-sample performance and disappointment in the cost prediction. However, if we are only interested in the prescriptive side in bounding the out-of-sample performance, Lam [28] proves an impossibility result: if one cares only about prescriptive performance guarantees, the empirical solution stochastically dominates a broad class of data-driven methods asymptotically. Gotoh et al. [15, 16] analyze the expected cost improvement of f-divergence DRO and optimistic variants, quantifying their trade-offs with the empirical solution. Other works design estimators with provable improvements. Feng and Shanthikumar [13] propose Operational Data Analytics (ODA), which assumes the underlying parametric distribution class and integrates into the downstream objective [13], while Albert et al. [1] develop a shrinkage estimator that achieves second-order improvement in predict-then-optimize tasks.

Appendix B. Formal Details of Robust Methods under Theorem 3

In the following, we present the detailed moment constructions for each robust method in Appendix B.1 and proof sketches establishing their unification in Appendix B.2.

B.1. Formulations and Equivalence of Robust Methods

B.1.1. REGULARIZATION / DISTRIBUTIONALLY ROBUST APPROACHES

Proposition 7 (Explicit Regularizer) *The explicit regularized solution $\hat{\theta}_\lambda$ computed by solving:*

$$\hat{\theta}_\lambda \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta;\xi)] + \lambda \mathbb{E}_{\hat{\mathbb{P}}_n}[M(\theta;\xi)] \text{ for some } \lambda > 0$$

satisfies the expansion in Theorem 3 for some $\tilde{H}$ and $\tilde{M}$ when $\lambda = o(1)$ or $\theta_\lambda^* = \theta^*$.

This formulation covers data-independent regularizers of the form $M(\theta;\xi) = M(\theta)$. Beyond classical regularization terms studied in statistics, the close connections between explicit regularization and distributionally robust optimization (DRO) [14, 27] motivate us to also consider DRO formulations based on Wasserstein distance and f-divergences.

Definition 8 (Wasserstein Distance) *The p-Wasserstein distance ($p \in \mathbb{N}$) between two distributions $\mu, \nu \in \mathcal{P}(\mathcal{Y})$ with respect to the l_1-norm (i.e., $\|\cdot\|_1$) is defined as:*

$$\mathcal{W}_p(\mathbb{P},\mathbb{Q}) := \inf_{\xi \in \Xi(\mathbb{P},\mathbb{Q})} \left(\mathbb{E}_{(Y_1,Y_2)\sim\xi}\left[\|Y_1 - Y_2\|_1^p\right]\right)^{\frac{1}{p}}, \tag{7}$$

where ξ is a joint distribution of $(Y_1 \in \mathcal{Y}, Y_2 \in \mathcal{Y})$ from $\Xi(\mu,\nu)$ and $\Xi(\mathbb{P},\mathbb{Q})$ denote the set of all joint distributions with marginal distributions $\mathbb{P}$ and $\mathbb{Q}$.

Example 3 (p-Wasserstein-DRO Regularizer) *Given p-Wasserstein distance $W_p(\mathbb{P},\mathbb{Q})$ defined in Definition 8, the p-Wasserstein DRO solution computed by solving:*

$$\hat{\theta}_\lambda \in \operatorname*{argmin}_{\theta \in \Theta} \sup_{\mathbb{Q}: W_p(\mathbb{Q},\hat{\mathbb{P}}_n) \le \lambda} \mathbb{E}_{\mathbb{Q}}[\ell(\theta;\xi)].$$

For $p \geq 1$, applying Theorem 8.7 in [26], we have:

$$\sup_{\mathbb{Q}:W_p(Q,\hat{\mathbb{P}}_n)\leq\lambda} \mathbb{E}_{\mathbb{Q}}[\ell(\theta;\xi)] = \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta;\xi)] + \lambda\mathbb{E}_{\hat{\mathbb{P}}_n}[\|\nabla\ell(\theta;\xi)\|_\infty^q]^{\frac{1}{q}} + o(\lambda).$$

If we ignore the higher-order term $o(\lambda)$, the p-Wasserstein-DRO regularizer becomes one special case of the standard implicit regularization problem.

We also allow other ambiguity sets such as f-divergence DROs.

Definition 9 (f-divergence) *Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions and $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$. For a convex function $f : [0,\infty) \to (-\infty,\infty]$ such that $f(x)$ is finite $\forall x > 0, f(1) = 0$, the f-divergence of $\mathbb{P}$ from $\mathbb{Q}$ is defined as:*

$$d_f(\mathbb{P},\mathbb{Q}) = \int f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)d\mathbb{Q} = \mathbb{E}_{\mathbb{Q}}\left[f\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)\right].$$

We obtain the χ^2-divergence by setting $f(x) = \frac{1}{2}(x-1)^2$.

Proposition 10 (χ^2-divergence-DRO Regularizer) *Given the χ^2-divergence in Definition 9, the χ^2-divergence DRO solution $\hat{\theta}_\lambda$ computed by solving:*

$$\min_{\theta\in\Theta} \max_{\mathbb{Q}:\chi^2(\mathbb{Q},\hat{\mathbb{P}}_n)\leq\lambda} \mathbb{E}_{\mathbb{Q}}[\ell(\theta;\xi)]$$

satisfies the expansion in Theorem 3 for some $\tilde{H}$ and $\tilde{M}$ when $\lambda = o(1)$ or $\theta_\lambda^* = \theta^*$.

Above, as long as λ is not large enough [10, 27], χ^2-divergence DRO solution allows the exact equivalence with the variance regularization:

$$\max_{\chi^2(\mathbb{Q},\hat{\mathbb{P}}_n)\leq\lambda} \mathbb{E}_{\mathbb{Q}}[\ell(\theta;\xi)] = \mathbb{E}_{\hat{\mathbb{P}}_n}[\ell(\theta;\xi)] + \sqrt{\lambda \mathrm{Var}_{\hat{\mathbb{P}}_n}[\ell(\theta;\xi)]}. \tag{8}$$

We also allow some $f(\cdot)$ that is generalized beyond the standard form in Theorem 9.

Proposition 11 (CVaR-DRO Regularizer) *Consider the Conditional Value-at-Risk (CVaR) objective with the parameter $\lambda \in [0,1)$*

$$CVaR_\lambda(\theta;\mathbb{P}) = \min_{\eta\in\mathbb{R}}\left\{\frac{1}{1-\lambda}\mathbb{E}_{\mathbb{P}}[(\ell(\theta;\xi)-\eta)_+] + \eta\right\}.$$

the CVaR solution $\hat{\theta}_\lambda$ computed by solving $\hat{\theta}_\lambda \in \mathrm{argmin}_{\theta\in\Theta} CVaR_\lambda(\theta;\hat{\mathbb{P}}_n)$ satisfies the expansion in Theorem 3 for some $\tilde{H}$ and $\tilde{M}$ when $\lambda = o(1)$ or $\theta_\lambda^* = \theta^*$.

The two augmented asymptotic normalities follow tools from Theorem 5.31 in [35].

Beyond distributionally robust formulations, Jiang and Xie [24] propose a *distributionally favorable* framework, which replaces the supremum over $\mathcal{P}$ by an infimum, while keeping the ambiguity set fixed. This can also be incorporated into Theorem 3. For example, in the χ^2-divergence case,

$$\min_{\theta\in\Theta} \min_{\chi^2(\mathbb{Q},\hat{\mathbb{P}}_n)\leq\lambda} \mathbb{E}_{\mathbb{Q}}[\ell(\theta;\xi)],$$

which simply changes the right-hand side of (8) from $+\sqrt{\lambda\,\mathrm{Var}_{\hat{\mathbb{P}}_n}[\ell(\theta;\xi)]}$ to $-\sqrt{\lambda\,\mathrm{Var}_{\hat{\mathbb{P}}_n}[\ell(\theta;\xi)]}$.

Comparison with existing work on local perturbation. Our expansion goes beyond the local analyses in Anderson and Philpott [2], Gotoh et al. [16], where the regularization term or ambiguity set shrinks to zero as $n \to \infty$. For instance, in unconstrained optimization, the general f-divergence penalization studied by Gotoh et al. [15] takes the form

$$\hat{\theta}_\lambda = \hat{\theta} + \frac{\lambda}{f''(1)} \cdot \mathbb{E}_{\mathbb{P}} I(\hat{\theta})^{-1} \mathrm{Cov}_{\hat{\mathbb{P}}_n}\big(h(\hat{\theta};\xi), \nabla h(\hat{\theta};\xi)\big) + o(\lambda),$$

which is a special case of our problem instance in Theorem 3.

B.1.2. ECONOMETRICS APPROACH

Method of moments is a classical approach in econometrics for parameter estimation based on moment equations [21]. We show that the over-identified generalized method of moments (GMM) can be reformulated within the framework of Theorem 3.

Proposition 12 (GMM) *Let $C \in \mathbb{R}^{2D_\theta \times 2D_\theta}$ be a fixed weighting matrix and define*

$$\hat{g}(\theta) = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \nabla_\theta \ell(\theta;\xi_i) \\ \frac{1}{n}\sum_{i=1}^n M(\theta;\xi_i) \end{bmatrix} \in \mathbb{R}^{2D_\theta}.$$

The GMM estimator $\hat{\theta}_{GMM}$ is obtained by solving

$$\min_\theta \ \hat{g}(\theta)^\top C \hat{g}(\theta).$$

Then $\hat{\theta}_{GMM}$ admits the expansion in Theorem 3 for suitable choices of $\tilde{H}$ and $\tilde{M}$.

Expressing the solution in the unified perturbation form of (3) and Theorem 3 offers two main advantages over the standard GMM formulation. First, it facilitates analysis of constrained or nonsmooth problems, where $\nabla \ell$ may arise from noncontinuous objectives. Second, the unified view is more flexible: it naturally handles moment conditions that do not depend on θ, provides robustness to misspecification, and enables fast adaptation in streaming-data settings without repeated optimization given the empirical solution $\hat{\theta}$.

B.1.3. STATISTICAL APPROACHES

Theorem 3 also encompasses several classical statistical estimators used in data integration.

Proposition 13 (Shrinkage Estimator) *Given the empirical solution $\hat{\theta}$, a shrinkage estimator takes the form*

$$\hat{\theta}_\lambda = \Big(I + \tfrac{H(\lambda)}{\|\hat{\theta}\|_2^2}\Big)\hat{\theta},$$

where $H(\lambda)$ is a fixed function of λ.

This estimator is directly motivated by the James–Stein shrinkage rule and has also been employed in the general stochastic optimization settings [17].

Proposition 14 (Transfer Learning Estimator) *Let θ^* denote a parameter estimated from a source distribution. The transfer-learning estimator*

$$\hat{\theta}_\lambda \in \underset{\theta\in\Theta}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n}\ell(\theta;\xi_i)+\lambda\|\theta-\theta^*\|_2^2$$

admits the expansion of Theorem 3 for suitable $\tilde{H}$ and $\tilde{M}$.

This formulation represents a simple instance of the broader transfer-learning literature [6] and can also be interpreted as an implicit regularization, corresponding to Proposition 7 with $M(\theta;\xi)=\|\theta-\theta^*\|_2^2$.

B.2. Proof of Theorem 3

Across all robust methods, we focus on the case where Θ is unconstrained. However, for general constrained problems, a similar proof technique applies for the expansion via constructing the Lagrangian multiplier as well. We first denote $I(\theta)=\nabla_{\theta\theta}\mathbb{E}_{\mathbb{P}^*}[\ell(\theta^*;\xi)]$ as the Hessian under the optimal θ^*.

B.2.1. Explicit or Variance Regularization.

We consider two regimes: (i) $\lambda=o(1)$; and (ii) $\theta_\lambda^*=\theta^*$ for explicit regularization in Proposition 7 and variance regularization in Proposition 10.

For the explicit regularizer of Proposition 7, we have:

$$\widetilde{M}(\theta^*;\xi):=\nabla_\theta M(\theta^*;\xi)\in\mathbb{R}^{D_\theta},\qquad \tilde{H}(\lambda):=-\lambda I(\theta^*)^{-1}\in\mathbb{R}^{D_\theta\times D_\theta}.$$

If $M(\theta;\xi)\equiv M(\theta)$ is data independent, then $\widetilde{M}(\theta^*;\xi)\equiv\nabla_\theta M(\theta^*)$ is constant, so the sum gives the standard penalty shift $-\lambda I(\theta^*)^{-1}\nabla_\theta M(\theta^*)$.

Write $\hat{Z}\left(\theta\right):=\mathbb{E}_{\hat{\mathbb{P}}_n}\ell\left(\theta;\xi\right)$ and consider

$$\hat{\theta}_\lambda\in\arg\min_{\theta\in\Theta}\left\{\hat{Z}\left(\theta\right)+\lambda\,\mathbb{E}_{\hat{\mathbb{P}}_n}M\left(\theta;\xi\right)\right\},\qquad \hat{\theta}\in\arg\min_{\theta\in\Theta}\hat{Z}\left(\theta\right).$$

The first-order condition at $\hat{\theta}_\lambda$ is

$$\nabla_\theta\hat{Z}\left(\hat{\theta}_\lambda\right)+\lambda,\mathbb{E}_{\hat{\mathbb{P}}_n}\nabla_\theta M\left(\hat{\theta}_\lambda;\xi\right)=0.$$

Let $\Delta:=\hat{\theta}_\lambda-\hat{\theta}$. Taylor expand both gradients at $\hat{\theta}$:

$$\nabla_\theta\hat{Z}\left(\hat{\theta}_\lambda\right)=\underbrace{\nabla_\theta\hat{Z}\left(\hat{\theta}\right)}_{=0}+\nabla_\theta^2\hat{Z}\left(\hat{\theta}\right)\Delta+R_1,$$

$$\mathbb{E}_{\hat{\mathbb{P}}_n}\nabla_\theta M\left(\hat{\theta}_\lambda;\xi\right)=\mathbb{E}_{\hat{\mathbb{P}}_n}\nabla_\theta M\left(\hat{\theta};\xi\right)+\mathbb{E}_{\hat{\mathbb{P}}_n}\nabla_\theta^2 M\left(\bar{\theta};\xi\right)\Delta+R_2,$$

where $\bar{\theta}$ lies on the segment between $\hat{\theta}$ and $\hat{\theta}_\lambda$, and R_1,R_2 collect higher-order terms. Plugging into the FOC and regrouping,

$$\left[\nabla_\theta^2\hat{Z}\left(\hat{\theta}\right)+\lambda\,\mathbb{E}_{\hat{\mathbb{P}}_n}\nabla_\theta^2 M\left(\bar{\theta};\xi\right)\right]\Delta=-\,\lambda\,\mathbb{E}_{\hat{\mathbb{P}}_n}\nabla_\theta M\left(\hat{\theta};\xi\right)+\tilde{R},$$

with $\tilde{R} := -(R_1 + \lambda R_2)$. Solving for Δ and keeping only the leading term,

$$\Delta = -\left(\nabla_\theta^2 \hat{Z}\left(\hat{\theta}\right)\right)^{-1} \lambda\, \mathbb{E}_{\hat{\mathbb{P}}_n} \nabla_\theta M\left(\hat{\theta};\xi\right) + o_p\left(\lambda \vee n^{-1/2}\right).$$

Now replace the sample quantities by their population counterparts at θ^*, we have:

$$\nabla_\theta^2 \hat{Z}\left(\hat{\theta}\right) = I\left(\theta^*\right) + o_p(1), \quad \mathbb{E}_{\hat{\mathbb{P}}_n} \nabla_\theta M\left(\hat{\theta};\xi\right) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta M\left(\theta^*;\xi_i\right) + o_p(1).$$

Therefore,

$$\hat{\theta}_\lambda - \hat{\theta} = \frac{1}{n}\sum_{i=1}^n \underbrace{\Big(-\lambda\, I\left(\theta^*\right)^{-1}\Big)}_{=:\, \tilde{H}(\lambda)} \underbrace{\nabla_\theta M\left(\theta^*;\xi_i\right)}_{=:\, \widetilde{M}(\theta^*;\xi_i)} + o_p\left(\lambda \vee n^{-1/2}\right).$$

For the variance regularizer of Proposition 10, under (8) holds, following the same result given FOC above, we show:

$$M(\theta^*;\xi) = \big(\ell(\theta^*;\xi) - \mathbb{E}_{\mathbb{P}^*}[\ell(\theta^*;\xi)]\big)\nabla_\theta \ell(\theta^*;\xi) \in \mathbb{R}^{D_\theta},$$
$$\tilde{H}(\lambda) = -I(\theta^*)^{-1}\frac{\sqrt{\lambda}}{\sqrt{\mathrm{Var}_{\mathbb{P}^*}[\ell(\theta^*;\xi)]}} \in \mathbb{R}^{D_\theta \times D_\theta}.$$

B.2.2. CVaR

Let the unregularized empirical minimizer $\hat{\theta}$ satisfy the usual linearization

$$\hat{\theta} - \theta^* \;=\; -\frac{1}{n}\sum_{i=1}^n I(\theta^*)^{-1}\, \nabla_\theta \ell(\theta^*;\xi_i) \;+\; o_p(n^{-1/2}).$$

Consider the CVaR program at level $\lambda \in (0,1)$:

$$\min_{\theta,\eta} \;\; \eta + \frac{1}{1-\lambda}\,\mathbb{E}\big[(\ell(\theta;\xi) - \eta)^+\big].$$

Let $(\theta^*_\lambda, \eta^*_\lambda)$ be the population solution, where $\eta^*_\lambda = \mathrm{VaR}_\lambda(\ell(\theta^*_\lambda;\xi))$. Define the estimating map

$$\psi(\theta,\eta;\xi) := \begin{bmatrix} \frac{1}{1-\lambda}\,\mathbf{1}_{\{\ell(\theta;\xi)>\eta\}}\;\nabla_\theta \ell(\theta;\xi) \\ 1 - \frac{1}{1-\lambda}\,\mathbf{1}_{\{\ell(\theta;\xi)>\eta\}} \end{bmatrix}, \qquad \mathbb{E}[\psi(\theta^*_\lambda, \eta^*_\lambda;\xi)] = 0.$$

Let $J_\lambda := \nabla_{(\theta,\eta)}\mathbb{E}[\psi(\theta^*_\lambda,\eta^*_\lambda;\xi)]$ and write the influence function for (θ,η) as

$$IF_\lambda^{(\theta,\eta)}(\xi) \;=\; -J_\lambda^{-1}\,\psi(\theta^*_\lambda,\eta^*_\lambda;\xi).$$

Extracting the θ-block gives

$$IF_\lambda^{(\theta)}(\xi) \;=\; -A(\lambda)^{-1}\,\frac{1}{1-\lambda}\,\mathbf{1}_{\{\ell(\theta^*_\lambda;\xi)>\eta^*_\lambda\}}\;\nabla_\theta \ell(\theta^*_\lambda;\xi),$$

where

$$A(\lambda) \;:=\; \nabla_\theta\,\mathbb{E}\Big[\frac{1}{1-\lambda}\,\mathbf{1}_{\{\ell(\theta;\xi)>\eta^*_\lambda\}}\;\nabla_\theta \ell(\theta;\xi)\Big]\Big|_{\theta=\theta^*_\lambda}.$$

By standard M-estimation algebra,

$$\hat{\theta}_\lambda - \theta^*_\lambda = -\frac{1}{n}\sum_{i=1}^{n} IF^{(\theta)}_\lambda(\xi_i) \; + \; o_p(n^{-1/2}).$$

Under $\theta^*_\lambda = \theta^*$, this yields

$$\hat{\theta}_\lambda - \theta^* \; = \; \frac{1}{n}\sum_{i=1}^{n} \underbrace{\Big[- A(\lambda)^{-1}\frac{1}{1-\lambda}\mathbf{1}_{\{\ell(\theta^*;\xi_i)>\eta^*_\lambda\}} + I(\theta^*)^{-1}\Big]}_{=:\,\tilde{H}(\lambda)} \underbrace{\Big[\nabla_\theta \ell(\theta^*;\xi_i)\Big]}_{=:\,\widetilde{M}(\theta^*;\xi_i)} + o_p(n^{-1/2}).$$

B.2.3. GMM

First we denote:

$$g_1(\theta;\xi) := \nabla_\theta \ell(\theta;\xi) \in \mathbb{R}^{D_\theta}, \qquad g_2(\theta;\xi) := M(\theta;\xi) \in \mathbb{R}^{D_\theta}, \qquad g(\theta;\xi) := \begin{bmatrix} g_1(\theta;\xi) \\ g_2(\theta;\xi) \end{bmatrix} \in \mathbb{R}^{2D_\theta}.$$

Let

$$\hat{g}(\theta) := \frac{1}{n}\sum_{i=1}^{n} g(\theta;\xi_i), \qquad \hat{g}_j(\theta) := \frac{1}{n}\sum_{i=1}^{n} g_j(\theta;\xi_i) \;\; (j=1,2)\,,$$

and denote C by its diagonal component:

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \qquad C_{jk} \in \mathbb{R}^{D_\theta \times D_\theta}.$$

The empirical solution $\hat{\theta}$ satisfies $\hat{g}_1(\hat{\theta}) = 0$, and write the Jacobians at $\hat{\theta}$ as

$$G_1 := \nabla_\theta \hat{g}_1(\hat{\theta}) \in \mathbb{R}^{D_\theta \times D_\theta}, \qquad G_2 := \nabla_\theta \hat{g}_2(\hat{\theta}) \in \mathbb{R}^{D_\theta \times D_\theta}, \qquad \hat{D} := \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \in \mathbb{R}^{2D_\theta \times D_\theta}.$$

We consider the first-order expansion around $\hat{\theta}$. That is, $\nabla_\theta \hat{Q}\left(\hat{\theta}_{GMM}\right) = 0$, i.e.,

$$0 = 2\hat{D}\left(\tilde{\theta}\right)^\top C\hat{g}\left(\hat{\theta}_{GMM}\right),$$

for some $\tilde{\theta}$ between $\hat{\theta}$ and $\hat{\theta}_{GMM}$. Linearizing $\hat{g}$ at $\hat{\theta}$ and using $\hat{g}_1(\hat{\theta}) = 0$,

$$0 \approx 2\Big\{\hat{D}^\top C\big[\hat{g}(\hat{\theta}) + \hat{D}(\hat{\theta}_{GMM} - \hat{\theta})\big]\Big\} = 2\Big\{ \underbrace{\hat{D}^\top C\hat{g}(\hat{\theta})}_{\text{depends only on } \hat{g}_2(\hat{\theta})} + \underbrace{\hat{D}^\top C\hat{D}}_{\text{curvature at } \hat{\theta}} (\hat{\theta}_{GMM} - \hat{\theta})\Big\}.$$

Compute the two blocks explicitly:

$$\hat{D}^\top C\hat{g}(\hat{\theta}) = G_1^\top\left(C_{11}\hat{g}_1(\hat{\theta}) + C_{12}\hat{g}_2(\hat{\theta})\right) + G_2^\top\left(C_{21}\hat{g}_1(\hat{\theta}) + C_{22}\hat{g}_2(\hat{\theta})\right) = \underbrace{\left(G_1^\top C_{12} + G_2^\top C_{22}\right)}_{=:B_n}\hat{g}_2(\hat{\theta}),$$

$$\hat{D}^\top C\hat{D} = G_1^\top C_{11}G_1 + G_1^\top C_{12}G_2 + G_2^\top C_{21}G_1 + G_2^\top C_{22}G_2 =: H_n.$$

Assuming H_n is nonsingular, we obtain

$$\hat{\theta}_{GMM} - \hat{\theta} = -H_n^{-1} B_n \hat{g}_2(\hat{\theta}) + o_p\left(n^{-1/2}\right).$$

We use the linear representation in sample averages. Since $\hat{g}_2(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^n M\left(\hat{\theta}; \xi_i\right)$, we can write

$$\hat{\theta}_{GMM} - \hat{\theta} = \frac{1}{n}\sum_{i=1}^n \underbrace{\left(-H_n^{-1} B_n\right)}_{=: H_n^\star} M\left(\hat{\theta}; \xi_i\right) \;+\; o_p\left(n^{-1/2}\right).$$

Finally, under LLN, we have: $G_1 \xrightarrow{p} A := \mathbb{E}\left[\nabla_\theta g_1\left[\theta^*; \xi\right]\right]$, $G_2 \xrightarrow{p} B := \mathbb{E}\left(\nabla_\theta g_2\left(\theta^*; \xi\right)\right)$, and $H_n^\star \xrightarrow{p} H^\star$ with

$$H^\star = -\left(A^\top C_{11} A + A^\top C_{12} B + B^\top C_{21} A + B^\top C_{22} B\right)^{-1}\left(A^\top C_{12} + B^\top C_{22}\right).$$

Hence the asymptotic linear form is

$$\hat{\theta}_{GMM} - \hat{\theta} = \frac{1}{n}\sum_{i=1}^n H^\star M\left(\hat{\theta}; \xi_i\right) + o_p\left(n^{-1/2}\right)$$

The last equality follows from the fact that $\mathbb{E}_{\mathbb{P}^*}[M(\theta; \xi)] = 0$. Equivalently, we obtain the perturbed formulation by defining $\widetilde{H}(\lambda; \xi) = H^*$ and $\widetilde{M}(\theta^*; \xi) = M(\theta; \xi)$.

B.2.4. SHRINKAGE

Compared with

$$\hat{\theta} = \theta^* + \frac{1}{n}\sum_{i=1}^n IF(\xi_i) + o_p\left(\frac{1}{\sqrt{n}}\right),$$

we have:

$$\|\hat{\theta}\|_2^2 = \|\theta^*\|_2^2 + 2\theta^{*\top}\left(\frac{1}{n}\sum_{i=1}^n IF(\xi_i)\right) + o_p\left(\frac{1}{\sqrt{n}}\right),$$
$$\frac{1}{\|\hat{\theta}\|_2^2} = \frac{1}{\|\theta^*\|_2^2} - \frac{2\theta^{*\top}\left(\frac{1}{n}\sum_{i=1}^n IF(\xi_i)\right)}{\|\theta^*\|_2^4} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Therefore, the shrinkage estimator

$$\hat{\theta}_\lambda = \left(I + \frac{H(\lambda)}{\|\hat{\theta}\|_2^2}\right)\hat{\theta}$$

admits the expansion

$$\begin{aligned}\hat{\theta}_\lambda = \hat{\theta} &+ \frac{H(\lambda)}{\|\theta^*\|_2^2}\theta^* \\ &+ \left(\frac{H(\lambda)}{\|\theta^*\|_2^2} - \frac{2H(\lambda)\theta^*\theta^{*\top}}{\|\theta^*\|_2^4}\right)\left(\frac{1}{n}\sum_{i=1}^n IF(\xi_i)\right) + o_p\left(\lambda \vee \frac{1}{\sqrt{n}}\right).\end{aligned}$$

Equivalently, we obtain the perturbed formulation by defining

$$\widetilde{M}(\theta^*;\xi_i) := \theta^* + IF(\xi_i) - \frac{2}{\|\theta^*\|_2^2}\theta^*\theta^{*\top} IF(\xi_i), \qquad \tilde{H}(\lambda;\xi) := \frac{H(\lambda)}{\|\theta^*\|_2^2}.$$

Appendix C. Proof of Theorem 4

Before going to the proof of Theorem 4, we first describe the following two lemmas:

Lemma 15 (Smoothness) $\sup_{\|\theta-\theta^*\|\le r_n} \left\| \bar{M}_n(\theta) - \bar{M}_n(\theta^*) - \nabla_\theta \bar{M}_n(\theta^*)\,(\theta-\theta^*) \right\|_2 = o_p\left(n^{-1/2}\right)$ *for some* $r_n \downarrow 0$ *with* $\mathbb{P}\left(\|\hat{\theta}-\theta^*\| \le r_n\right) \to 1$.

Proof of Lemma 15. Apply Taylor's theorem with integral remainder to each summand $M(\theta;\xi_i)$ between θ^* and θ:

$$M(\theta;\xi_i) - M(\theta^*;\xi_i) - \nabla_\theta M(\theta^*;\xi_i)(\theta-\theta^*) = \int_0^1 (1-t)\,(\theta-\theta^*)^\top \nabla^2_{\theta\theta} M\big(\theta^* + t(\theta-\theta^*);\xi_i\big)\,(\theta-\theta^*)\,dt.$$

Taking norms, sup over $\|\theta-\theta^*\| \le r_n$, averaging over i, and using the envelope bound yields the displayed inequality with $\bar{K}_n$. By the LLN, $\bar{K}_n \xrightarrow{p} \mathbb{E}[K(\xi)] < \infty$, hence the remainder is $O_p(r_n^2)$ uniformly on the ball. Choosing $\alpha \in (1/4, 1/2)$ ensures $r_n \downarrow 0$, $r_n^2 = o(n^{-1/2})$, and $n^{1/2} r_n \to \infty$, so a $\sqrt{n}$–consistent $\hat{\theta}$ lies in the ball with probability tending to one. □

Lemma 16 (Asymptotic Normality of Perturbed Solution) *If* $\sqrt{n}(\hat{\theta}-\theta^*) \Rightarrow \Sigma(\theta^*;0)$. *Consider* $\hat{\theta}_\lambda = \hat{\theta} + \lambda \bar{M}_n(\hat{\theta})$. *Then:*

$$\sqrt{n}(\hat{\theta}_\lambda - \theta^*_\lambda) \Rightarrow \mathcal{N}\big(0, \Sigma(\theta^*;\lambda)\big),$$

where

$$\Sigma(\theta^*;\lambda) = (I_d + \lambda\nabla_\theta\pi(\theta^*))\Sigma(\theta^*;0)(I_d + \lambda\nabla_\theta\pi(\theta^*))^\top + \lambda^2\Omega_M + 2\lambda\mathrm{Sym}\big((I_d + \lambda\nabla_\theta\pi(\theta^*))\Gamma\big),$$

with $\mathrm{Sym}(A) = \frac{1}{2}(A + A^\top)$, $\theta^*_\lambda = \theta^* + \lambda\mathbb{E}_{\mathbb{P}^*}[M(\theta^*;\xi)]$, $\Omega_M := \mathit{Var}_{\mathbb{P}^*}[M(\theta^*;\xi)]$ *and* $\Gamma := \mathit{Cov}_{\mathbb{P}^*}[IF(\xi), M(\theta^*;\xi)]$,

For the value of $\Sigma(\theta^*;0)$, when Θ is unconstrained, following the standard condition in the asymptotic of stochastic optimization (i.e., Chapter 5 of [35]), we have:

$$\Sigma(\theta^*;0) = (I(\theta^*))^{-1} J(\theta^*) (I(\theta^*))^{-1},$$

where $J(\theta^*) = \mathbb{E}_{\mathbb{P}^*}[\nabla_\theta \ell(\theta^*;\xi)\nabla_\theta \ell(\theta^*;\xi)^\top]$.

For the general constrained $\Theta = \{\theta | F_j(\theta) \le 0, j \in J\}$, from [11], we have:

$$\Sigma(\theta^*;0) = P_F (I(\theta^*))^{-1} P_F J(\theta^*) P_F (I(\theta^*))^{-1} P_F,$$

where $P_F = I - C^\top (CC^\top)^\dagger C$, $C = (\nabla_\theta F_j(\theta))_{j\in J^*}$, and $J^* = \{j \in J : F_j(\theta^*) = 0\}$.

Proof of Lemma 16. Define

$$\hat{\theta}_\lambda := \hat{\theta} + \lambda\bar{M}_n(\hat{\theta}), \qquad \theta^*_\lambda := \theta^* + \lambda\pi(\theta^*).$$

Let $G_n(\theta) := \theta + \lambda \bar{M}_n(\theta)$ and $G(\theta) := \theta + \lambda\pi(\theta)$. Then

$$\sqrt{n}\left(\hat{\theta}_\lambda - \theta^*_\lambda\right) = \sqrt{n}\left(G_n(\hat{\theta}) - G(\theta^*)\right) = \underbrace{\sqrt{n}\left(G_n(\hat{\theta}) - G_n(\theta^*)\right)}_{(A)} + \underbrace{\sqrt{n}\left(G_n(\theta^*) - G(\theta^*)\right)}_{(B)}.$$

For the part (A), following the mean-value expansion and Lemma 15,

$$G_n(\hat{\theta}) - G_n(\theta^*) = \left(I_d + \lambda \nabla_\theta \bar{M}_n(\theta^*)\right)\left(\hat{\theta} - \theta^*\right) + r_n, \quad \text{with } \|r_n\|_2 = o_p\left(n^{-1/2}\right).$$

Multiplying by $\sqrt{n}$, we have:

$$\sqrt{n}\left(G_n(\hat{\theta}) - G_n(\theta^*)\right) = \left(I_d + \lambda \nabla_\theta \bar{M}_n(\theta^*)\right)\sqrt{n}\left(\hat{\theta} - \theta^*\right) + o_p(1).$$

For the part (B), by definition,

$$\sqrt{n}\left(G_n(\theta^*) - G(\theta^*)\right) = \lambda\sqrt{n}\left(\bar{M}_n(\theta^*) - \pi(\theta^*)\right).$$

Combining (A)–(B) yields

$$\sqrt{n}\left(\hat{\theta}_\lambda - \theta^*_\lambda\right) = \left(I_d + \lambda \nabla_\theta \bar{M}_n(\theta^*)\right)\sqrt{n}\left(\hat{\theta} - \theta^*\right) + \lambda\sqrt{n}\left(\bar{M}_n(\theta^*) - \pi(\theta^*)\right) + o_p(1).$$

By Slutsky's theorem and Lemma 15, $\nabla_\theta \bar{M}_n(\theta^*) = \frac{1}{n}\sum_{i=1}^n \nabla_\theta M(\theta^*;\xi_i) \xrightarrow{p} \nabla_\theta \pi(\theta^*) \in \mathbb{R}^{d\times d}$.

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \begin{bmatrix} IF(\xi_i) \\ M(\theta^*;\xi_i) - \pi(\theta^*) \end{bmatrix} \Rightarrow \mathcal{N}\left(0, \begin{bmatrix} \Sigma(\theta^*;0) & \Gamma \\ \Gamma^\top & \Omega_M \end{bmatrix}\right).$$

Applying the continuous mapping theorem gives the Gaussian limit

$$\sqrt{n}(\hat{\theta}_\lambda - \theta^*_\lambda) \Rightarrow \mathcal{N}\big(0, \Sigma(\theta^*;\lambda)\big),$$

where $\Sigma(\theta^*;\lambda) = (I_d + \lambda\nabla_\theta\pi(\theta^*))\Sigma(\theta^*;0)(I_d+\lambda\nabla_\theta\pi(\theta^*))^\top + \lambda^2\Omega_M + 2\lambda\mathrm{Sym}\big((I_d+\lambda\nabla_\theta\pi(\theta^*))\Gamma\big)$.
□

Then we move to the proof of Theorem 4.

Proof of Theorem 4. Denote $\pi(\theta) = \mathbb{E}_{\mathbb{P}^*}[M(\theta;\xi)]$ and $Z(\theta) = \mathbb{E}_{\mathbb{P}^*}[\ell(\theta;\xi)]$. Besides,

$$I(\theta) = \nabla^2_{\theta\theta} Z(\theta), J(\theta) = \mathbb{E}_{\mathbb{P}^*}[\nabla_\theta \ell(\theta;\xi)\nabla_\theta\ell(\theta;\xi)^\top].$$

We first want to obtain the exact upper and lower bound relationship between $\mathbb{E}_{\mathcal{D}_n}[Z(\hat{\theta}_\lambda)]$ and $\mathbb{E}_{\mathcal{D}_n}[Z(\hat{\theta})]$.

For general $\hat{\theta}_\lambda = \hat{\theta} + \frac{\lambda}{n}\sum_{i\in[n]} M(\hat{\theta};\xi_i) = \hat{\theta} + \lambda\bar{M}_n(\hat{\theta})$, from Theorem 16, we have:

$$\sqrt{n}(\hat{\theta}_\lambda - \theta^*_\lambda) \Rightarrow \mathcal{N}\big(0, \Sigma(\theta^*;\lambda)\big),$$

where $\Sigma(\theta^*;\lambda) = (I_d + \lambda\nabla_\theta\pi(\theta^*))\Sigma(\theta^*;0)(I_d+\lambda\nabla_\theta\pi(\theta^*))^\top + \lambda^2\Omega_M + 2\lambda\mathrm{Sym}\big((I_d+\lambda\nabla_\theta\pi(\theta^*))\Gamma\big)$ with $\mathrm{Sym}(A) = \frac{1}{2}(A + A^\top)$. Letting $\lambda = 0$ in Theorem 16, we have:

$$\sqrt{n}(\hat{\theta} - \theta^*) \Rightarrow N(0, \Sigma(\theta^*;0)). \tag{9}$$

Taken the asymptotic normality of $\hat{\theta}_\lambda$ and $\hat{\theta}$ over D_n, we have:

$$\mathbb{E}[Z(\hat{\theta}_\lambda)] = Z(\theta^*_\lambda) + \frac{\text{Tr}[\Sigma(\theta^*;\lambda)I(\theta^*_\lambda)]}{2n} + o\left(\frac{1}{n}\right), \tag{10}$$

$$\mathbb{E}[Z(\hat{\theta})] = Z(\theta^*) + \frac{\text{Tr}[\Sigma(\theta^*;0)I(\theta^*)]}{2n} + o\left(\frac{1}{n}\right). \tag{11}$$

Then we compare $Z(\theta^*_\lambda)$ and $Z(\theta^*)$.

When the decision space is unconstrained such that $\Theta = \mathbb{R}^{D_\theta}$, since $Z(\cdot)$ is strongly convex from Assumption 2 (with $\theta_1 = \theta^*_\lambda, \theta_2 = \theta^*$), we have:

$$C_1\lambda^2\|\pi(\theta^*)\|_2^2 \le Z(\theta^*_\lambda) - Z(\theta^*) - \overset{\top}{\underbrace{\nabla_\theta Z(\theta^*)}_{=0}}(\theta^*_\lambda - \theta^*) \le C_2\lambda^2\|\pi(\theta^*)\|_2^2. \tag{12}$$

Denote $\rho(\lambda) = \text{Tr}[\Sigma(\theta^*;\lambda)I(\theta^*_\lambda) - \Sigma(\theta^*;0)I(\theta^*)]$. Since $|\lambda| < \infty$, then $\rho(\lambda)$ is finite. Combining all the equalities and inequalities above, we have:

$$\mathbb{E}[Z(\hat{\theta}_\lambda)] \ge \mathbb{E}[Z(\hat{\theta})] + \frac{\rho(\lambda)}{2n} + C_1\lambda^2\|\pi(\theta^*)\|_2^2 + o\left(\frac{1}{n}\right). \tag{13}$$

$$\mathbb{E}[Z(\hat{\theta}_\lambda)] \le \mathbb{E}[Z(\hat{\theta})] + \frac{\rho(\lambda)}{2n} + C_2\lambda^2\|\pi(\theta^*)\|_2^2 + o\left(\frac{1}{n}\right). \tag{14}$$

For problems under a general constrained decision space, Equations (10) and (11) still hold. And we need to show that the Equation (13) hold. Taken $\theta_1 = \theta^*_\lambda, \theta_2 = \theta^*$ and Lagrangian multipliers $\{\alpha^*_j\}_{j\in B}$ from Assumption 2, for the left side, we have:

$$\begin{aligned} C_1\lambda^2\|\pi(\theta^*)\|_2^2 &\le Z(\theta^*_\lambda) + \sum_{j\in J}\alpha^*_j F_j(\theta^*_\lambda) - Z(\theta^*) - \sum_{j\in J}\alpha_j F_j(\theta^*) - (\nabla_\theta Z(\theta^*) + \sum_{j\in J}\alpha_j \nabla_\theta F_j(\theta^*))^\top(\theta^*_\lambda - \theta^*) \\ &= Z(\theta^*_\lambda) + \sum_{j\in J}\alpha^*_j F_j(\theta^*_\lambda) - Z(\theta^*) \\ &\le Z(\theta^*_\lambda) - Z(\theta^*), \end{aligned}$$

where the first equality is due to the KKT condition that $\nabla_\theta Z(\theta^*) + \sum_{j\in J}\alpha^*_j \nabla_\theta F_j(\theta^*) = 0$ and $\sum_{j\in J}\alpha^*_j F_j(\theta^*) = 0$ from Assumption 1; and the second equality follows from $F_j(\theta) \le 0$ for each $j \in J$ at θ^*_λ with the non-negative Lagrangian multiplier $\alpha^*_j \ge 0$ for each $j \in J$.

Then giving (13) (and (14)), we consider the following two cases:

1. $\pi(\theta^*) \neq 0$. This is equivalent to saying $\theta^*_\lambda \neq \theta^*$. We consider the case of $\lambda = o(1)$ and $\lambda = \Theta(1)$:

   First, if $\lambda = \Theta(1)$, then from (13), we have: $\mathbb{E}[Z(\hat{\theta}_\lambda)] \ge \mathbb{E}[Z(\hat{\theta})] + C_1\lambda^2\|\pi(\theta^*)\|_2^2 + \Theta(1/n) > \mathbb{E}[Z(\hat{\theta})]$. In this case, $\hat{\theta}_\lambda$ does not provide any expected improvement to $\hat{\theta}$,

   Then, if $\lambda = o(1)$, then we expand $\rho(\lambda)$. Recall the formula of $\rho(\lambda)$, we immediately observe that $\rho(0) = 0$. Therefore, we have:

   $$\rho(\lambda) = \rho'(0)\lambda + o(\lambda)$$

   $$Z(\theta^*_\lambda) = Z(\theta^*) + \frac{\lambda^2}{2}\pi(\theta^*)^\top\left(I(\theta^*) + \sum_{j\in B^*}\alpha_j\nabla^2_{\theta\theta}F_j(\theta^*)\right)\pi(\theta^*) + o(\lambda^2).$$

where the second equality directly follows from the second-order Taylor expansion. More specifically, $\rho'(0)$ is calculated as:

$$\begin{aligned}\rho'(0) &= \text{Tr}[\Sigma(\theta^*;0)\cdot I'(\theta^*)] + \text{Tr}[\Sigma'(\theta^*;0)\cdot I(\theta^*)]\\ &= \pi(\theta^*)^\top \nabla_\theta \text{Tr}[\Sigma(\theta^*;0)I(\theta^*)]|_{\theta=\theta^*} + 2\text{Tr}(\Sigma(\theta^*;0)\nabla_\theta \pi(\theta^*)I(\theta^*)).\end{aligned}$$

This way, we can change the inequalities of (13) and (14) as the following exact expansion:

$$\mathbb{E}[Z(\hat{\theta}_\lambda)] = \mathbb{E}[Z(\hat{\theta})] + \frac{\rho'(0)\lambda}{n} + \frac{\lambda^2}{2}\pi(\theta^*)^\top I(\theta^*)\pi(\theta^*) + o(\lambda^2) + o\left(\frac{1}{n}\right).$$

The relative improvement function is:

$$G(\lambda) = \frac{\rho'(0)\lambda}{n} + \lambda^2\left(\frac{1}{2}\pi(\theta^*)^\top\left(I(\theta^*) + \sum_{j\in B^*}\alpha_j \nabla^2_{\theta\theta}F_j(\theta^*)\right)\pi(\theta^*) + o(1)\right).$$

This is is a quadratic function with respect to λ. And the optimal

$$\lambda^* = -\frac{\rho'(0)}{2n\pi^\top\left(I(\theta^*) + \sum_{j\in B^*}\alpha_j\nabla^2_{\theta\theta}F_j(\theta^*)\right)\pi} = O\left(\frac{1}{n}\right)$$

and the corresponding $G(\lambda^*) = O\left(\frac{1}{n^2}\right)$. Any other values $\lambda \in (-\infty,\infty)$ leads to a $G(\lambda) < G(\lambda^*)$.

2. $\pi(\theta^*) = 0$. This implies that $\theta^*_\lambda = \theta^*$. Comparing (13) and (14), we have:

$$\begin{aligned}\rho(\lambda) &= \text{Tr}[(\Sigma(\theta^*;\lambda) - \Sigma(\theta^*;0))I(\theta^*)]\\ &= a\lambda + (b + o(1))\lambda^2,\end{aligned}$$

where:

$$\begin{aligned}a &= \text{Tr}\left[(\nabla_\theta\pi(\theta^*)\Sigma_0 + 2\text{Sym}(\Gamma))\, I(\theta^*)\right]\\ b &= \text{Tr}\left[\left(\nabla_\theta\pi(\theta^*)\Sigma_0\nabla_\theta\pi(\theta^*)^\top + 2\text{Sym}\left[\nabla_\theta\pi(\theta^*)\Gamma\right] + \Omega_M\right]\right) I(\theta^*)\right].\end{aligned}$$

Therefore

$$\mathbb{E}_{\mathcal{D}_n}[R(\hat{\theta}_\lambda) - R(\hat{\theta})] = \frac{a}{2n}\lambda + \frac{b}{2n}\lambda^2 + o(n^{-1}).$$

Note $\Omega_M \succeq 0$ contributes positively. Then the quadratic in λ is strictly convex, and the optimal regularization is the *constant* choice

$$\lambda^* = -\frac{a}{2b}, \qquad \mathbb{E}_{\mathcal{D}_n}[R(\hat{\theta}_{\lambda^*}) - R(\hat{\theta})] = -\frac{a^2}{8bn} + o\left(n^{-1}\right).$$

This implies that choosing $\lambda = -\frac{a}{2b}$ leads to a first-order improvement of order $o(n^{-1})$. As long as $a \neq 0$, the perturbed solution can lead to some first-order improvement. □

We also highlight that compared with the empirical optimization, including a data-driven regularization is necessary. That is, the function M needs to depend on the data first. Suppose we fix the regularization direction towards the empirical solution, $\hat{\theta}_\lambda = \hat{\theta} + \lambda\pi, \forall\lambda \geq 0$, which cannot lead to first-order improvement.

Appendix D. Generalization of First-order Improvements in Section 3

Our insights of first-order improvements can also be generalized to data-driven optimization problems with side information or so-called contextual stochastic optimization, optimization under risk functions. In the following, for simplicity, we only consider the unconstrained case $\Theta = \mathbb{R}^{D_\theta}$ and the case that the moment function $M(\theta; \xi)$ is independent of the data-driven parameter $\hat{\theta}$. However, our results still apply when $M(\theta; \xi)$ is a function of θ and ξ and general constrained problems as our main results in Section 3.

D.1. Weighted Empirical Optimizations

In contextual stochastic optimization problems, the distribution of ξ is a function of a covariate $u \in \mathbb{R}^{D_u}$ [7, 12]. The ground-truth distribution $\mathbb{P}^*_{\xi|u}$ is unknown; instead, the decision maker only has data $\mathcal{D}_n = \{(u_i, \xi_i)\}_{i=1}^n$ consisting of iid samples from the joint distribution $\mathbb{P}^*_{(u,\xi)} = \mathbb{P}^*_u \times \mathbb{P}^*_{\xi|u}$. The decision maker observes the covariate $U = u$ before making the decision.

In particular, we consider the class of weighted empirical optimization procedures in [7].

Definition 17 (Data-Driven Weighted Empirical Optimization Procedures) *The empirical decision rule*

$$\hat{\theta}(u) \in \operatorname*{argmin}_{\theta \in \Theta} \sum_{i \in [n]} w_{n,i}(u) \ell(\theta; \xi_i), \tag{15}$$

where $\{w_{n,i}(u)\}_{i \in [n]}$ are weights determined by $\mathcal{D}_n$ and u.

The key observation is to notice that for regularized k-Nearest Neighbor, Nadaraya-Watson kernel estimator and some regular random forests satisfy the following condition:

Assumption 3 (Influence Function Decomposition) *For any u and write*

$$Z_u(\theta) := \mathbb{E}_{\xi \sim \mathbb{P}^*_{\xi|u}}[\ell(\theta; \xi)], \qquad \theta^*(u) \in \arg\min_{\theta} Z_u(\theta),$$

and assume Z_u is twice continuously differentiable with positive definite Hessian

$$I_u := \nabla^2_{\theta\theta} Z_u(\theta^*(u)).$$

Let the estimator $\hat{\theta}(u)$ satisfy the IF decomposition

$$\hat{\theta}(u) - \theta^*(u) = \frac{1}{n} \sum_{i=1}^n IF_{n,u}(u_i, \xi_i) + b_{n,u} + r_{n,u},$$

where at each point u, $\mathbb{E}_{(\tilde{u},\xi) \sim \mathbb{P}^*}[IF_{n,u}(\tilde{u}, \xi)] = 0$, $Var_{\mathbb{P}^*}[IF_{n,u}(\tilde{u}, \xi)]/n^{1-2\gamma} \to \Sigma_u$, the decision bias term $b_{n,u} = O(n^{-\gamma})$, and $r_{n,u} = o_p(n^{-\gamma})$ with $\gamma \leq \frac{1}{2}$.

Proposition 18 (First-order improvement for weighted empirical solution) *Suppose Assumptions 2, 1 (when we replace $Z(\theta)$ with $Z_u(\theta)$) and Assumption 3 hold. Consider the perturbed estimator*

$$\hat{\theta}_\lambda(u) := \hat{\theta}(u) + \lambda_{n,u} \bar{M}_{n,u}.$$

where

$$\bar{M}_{n,u} := \frac{1}{n} \sum_{i=1}^n M_u(u_i, \xi_i), \quad \mu_M(u) := \mathbb{E}_{\mathbb{P}^*}[M_u(U, \xi)], \quad \Omega_M(u) := Var_{\mathbb{P}^*}(M_u(U, \xi)).$$

Define the performance gap of a estimator $\theta(u)$ at u by $R(\theta(u)) := Z_u(\theta(u)) - Z_u(\theta^*(u))$.

(i) If $\mu_M(u) = 0$ and Tr$(I_u \Gamma_{n,u}) \neq 0$, where

$$\Gamma_{n,u} := \mathit{Cov}(IF_{n,u}(U,\xi), M_u(U,\xi)) = O(n^{1-2\gamma}).$$

for each n, there exists $\lambda^*_{n,u} = -\dfrac{\mathit{Tr}(I_u\Gamma_{n,u})}{\mathit{Tr}(I_u\Omega_M(u))}$ such that

$$\mathbb{E}[R(\hat{\theta}_{\lambda^*_{n,u}}(u))] - \mathbb{E}[R(\hat{\theta}(u))] = -\frac{\mathit{Tr}(I_u\Gamma_{n,u})^2}{2\mathit{Tr}(I_u\Omega_M(u))} \cdot \frac{1}{n} + o\Big(n^{-2\gamma}\Big).$$

In particular, a first-order improvement of $\Theta(n^{-2\gamma})$ is achieved.

(ii) If $\mu_M(u) = 0$ and Tr$(I_u\Gamma_{n,u}) = 0$, then:

$$\mathbb{E}[R(\hat{\theta}_\lambda(u))] - \mathbb{E}[R(\hat{\theta}(u))] = \frac{1}{2}\mathit{Tr}(I_u\Omega_M(u)) \cdot \frac{\lambda^2_{n,u}}{n} + o\Big(\frac{1}{n}\Big) \;\geq\; 0,$$

so the best constant choice is $\lambda_{n,u} = 0$.

(iii) If $\mu_M(u) \neq 0$, we can achieve the second-order improvement of the order $o(n^{-2\gamma})$.

Proof of Theorem 18. A second-order Taylor expansion of Z_u at $\theta^*(u)$ gives, for any random $\tilde{\theta}$ close to $\theta^*(u)$,

$$R(\tilde{\theta}) = Z_u(\tilde{\theta}) - Z_u(\theta^*(u)) = \frac{1}{2}(\tilde{\theta} - \theta^*(u))^\top I_u(\tilde{\theta} - \theta^*(u)) + o_p\big(\|\tilde{\theta} - \theta^*(u)\|^2\big).$$

Taking expectations and applying the bias-variance decomposition,

$$\mathbb{E}[R(\tilde{\theta})] = \frac{1}{2}\mathrm{Tr}(I_u\mathrm{Var}(\tilde{\theta})) + \frac{1}{2}(\mathbb{E}[\tilde{\theta}] - \theta^*(u))^\top I_u(\mathbb{E}[\tilde{\theta}] - \theta^*(u)) + o(\mathbb{E}\|\tilde{\theta} - \theta^*(u)\|^2).$$

Apply this to $\tilde{\theta} = \hat{\theta}(u)$ and to $\tilde{\theta} = \hat{\theta}_\lambda(u)$.

Step 1: Moments of $\hat{\theta}(u)$. From the IF decomposition with $\gamma \leq \frac{1}{2}$,

$$\mathrm{Var}(\hat{\theta}(u)) = \frac{1}{n^{2\gamma}}\Sigma_u + o\Big(n^{-2\gamma}\Big), \qquad \mathbb{E}[\hat{\theta}(u)] - \theta^*(u) = b_{n,u} + o(n^{-\gamma}),$$

so

$$\mathbb{E}[R(\hat{\theta}(u))] = \frac{1}{2n}\mathrm{Tr}(I_u\Sigma_u) + \frac{1}{2}b^\top_{n,u}I_u b_{n,u} + o(n^{-2\gamma}).$$

Step 2: Moments of $\hat{\theta}_\lambda(u) = \hat{\theta}(u) + \lambda_{n,u}\bar{M}_{n,u}$. Write

$$\Gamma_{n,u} := \mathrm{Cov}(IF_{n,u}(U,\xi), M_u(U,\xi)), \quad \Omega_M(u) := \mathrm{Var}(M_u(U,\xi)).$$

Since $\hat{\theta}(u) = \theta^*(u) + \frac{1}{n}\sum IF_{n,u} + b_{n,u} + r_{n,u}$ and $\bar{M}_{n,u} = \frac{1}{n}\sum M_u$, we have

$$\begin{aligned}\mathrm{Var}(\hat{\theta}_\lambda(u)) &= \mathrm{Var}(\hat{\theta}(u)) + 2\lambda_{n,u}\mathrm{Cov}(\hat{\theta}(u), \bar{M}_{n,u}) + \lambda^2_{n,u}\mathrm{Var}(\bar{M}_{n,u}) \\ &= \frac{1}{n}\Big(n^{1-2\gamma}\Sigma_u + 2\lambda_{n,u}\Gamma_{n,u} + \lambda^2_{n,u}\Omega_M(u)\Big) + o\Big(n^{-2\gamma}\Big).\end{aligned}$$

Moreover,

$$\mathbb{E}[\hat{\theta}_\lambda(u)] - \theta^*(u) = (\mathbb{E}[\hat{\theta}(u)] - \theta^*(u)) + \lambda_{n,u}\mathbb{E}[\bar{M}_{n,u}] = b_{n,u} + \lambda_{n,u}\mu_M(u) + o(n^{-\gamma}).$$

Step 3: Risk difference. Subtract the expansions:

$$\begin{aligned}\Delta_n(\lambda_{n,u}; u) &:= \mathbb{E}[R(\hat{\theta}_\lambda(u))] - \mathbb{E}[R(\hat{\theta}(u))] \\ &= \frac{1}{2n}\mathrm{Tr}(I_u(2\lambda_{n,u}\Gamma_{n,u} + \lambda_{n,u}^2\Omega_M(u))) + \frac{1}{2}(\lambda_{n,u}\mu_M(u))^\top I_u(\lambda_{n,u}\mu_M(u)) \\ &\quad + \lambda_{n,u}\mu_M(u)^\top I_u b_{n,u} + o(n^{-2\gamma}).\end{aligned}$$

Case (i): $\mu_M(u) = 0$ and $b_u := \mathrm{Tr}(I_u\Gamma_{n,u}) \neq 0$. Set $\lambda_{n,u} \equiv \lambda$ constant. Then

$$\Delta_n(\lambda; u) = \frac{1}{2n}(2\lambda\mathrm{Tr}(I_u\Gamma_{n,u}) + \lambda^2\mathrm{Tr}(I_u\Omega_M(u))) + o\Big(n^{-2\gamma}\Big).$$

This quadratic in λ is minimized at $\lambda^*(u) = -\dfrac{\mathrm{Tr}(I_u\Gamma_{n,u})}{\mathrm{Tr}(I_u\Omega_M(u))}$, yielding

$$\Delta_n(\lambda^*; u) = -\frac{\mathrm{Tr}(I_u\Gamma_{n,u})^2}{2\mathrm{Tr}(I_u\Omega_M(u))} \cdot \frac{1}{n} + o\Big(n^{-2\gamma}\Big),$$

which proves the first-order improvement.

Case (ii): $\mu_M(u) = 0$ and $\mathrm{Tr}(I_u\Gamma_{n,u}) = 0$. Then $\Delta_n(\lambda; u) = \frac{1}{2n}\lambda^2\mathrm{Tr}(I_u\Omega_M(u)) + o(n^{-2\gamma}) \geq 0$ for any constant λ, so the best constant choice is $\lambda = 0$.

Case (iii): $\mu_M(u) \neq 0$. The term $\frac{1}{2}\lambda_{n,u}^2\mu_M(u)^\top I_u\mu_M(u)$ is nonnegative and of order 1 when $\lambda_{n,u}$ is constant, which prevents $\Theta(n^{-2\gamma})$ improvement. Choosing some $\lambda_{n,u} = \Theta(n^{-2\gamma}$ suppresses this bias but then any improvement is at most $\Theta(n^{-2\gamma})$.

These three cases establish the claim. □

D.2. Risk Functions

In this section, we extend to the problem instance that compares the distributional aspect information of $Z(\hat{\theta}_\lambda)$ and $Z(\hat{\theta})$ beyond the expectation, i.e., comparing $\mathbb{E}[g(Z(\hat{\theta}_\lambda))]$ and $\mathbb{E}[g(Z(\hat{\theta}))]$ for general $g(\cdot)$.

Assumption 4 (Condition on $g(\cdot)$) *The function $g : \mathbb{R} \to \mathbb{R}$ is monotonically nondecreasing and twice continuously differentiable.*

This includes the expected performance gap comparison with $g(x) = x$.

Proposition 19 (First-order improvement for general risk functions) *Suppose Assumptions 2, 1 and 4 hold. If $\mathbb{E}_{\mathbb{P}^*}[M(\theta^*; \xi)] = 0$ and $\mathbb{E}_{\mathbb{P}^*}[M(\theta^*; \xi)^\top IF(\xi)] \neq 0$, there exists some $\lambda^* = \Theta(1)$ such that we have: $\mathbb{E}_{\mathcal{D}_n}[g(Z(\hat{\theta}_{\lambda^*}))] - \mathbb{E}_{\mathcal{D}_n}[g(Z(\hat{\theta}))] = \Theta(1/n) < 0$. Otherwise the improvement is of higher order, at most $o(1/n)$.*

Proof of Theorem 19. We decompose

$$\mathbb{E}_{\mathcal{D}_n}[g(Z(\hat{\theta}_\lambda))] = \underbrace{\mathbb{E}_{\mathcal{D}_n}[g(Z(\hat{\theta}_\lambda))] - \mathbb{E}_{\mathcal{D}_n}[g(Z(\theta^*_\lambda))]}_{(A)} + \underbrace{\mathbb{E}_{\mathcal{D}_n}[g(Z(\theta^*_\lambda))] - \mathbb{E}_{\mathcal{D}_n}[g(Z(\theta^*))]}_{(B)}.$$

Step 1. Term (A). Apply a second-order Taylor expansion:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}[g(Z(\hat{\theta}_\lambda))] = \mathbb{E}_{\mathcal{D}_n}[g(Z(\theta^*_\lambda))] + g'(Z(\theta^*_\lambda))\mathbb{E}_{\mathcal{D}_n}[Z(\hat{\theta}_\lambda) - Z(\theta^*_\lambda)] \\ + \frac{1}{2}g''(Z(\theta^*_\lambda))\mathbb{E}_{\mathcal{D}_n}[(Z(\hat{\theta}_\lambda) - Z(\theta^*_\lambda))^2] + o\Big(\mathbb{E}_{\mathcal{D}_n}[(Z(\hat{\theta}_\lambda) - Z(\theta^*_\lambda))^2]\Big).\end{aligned}$$

Recall the proof in Theorem 4, we have:

$$\mathbb{E}_{\mathcal{D}_n}[Z(\hat{\theta}_\lambda) - Z(\theta^*_\lambda)] = \frac{1}{2n}\mathrm{Tr}(\Sigma(\theta^*;\lambda)I(\theta^*_\lambda)) + o\Big(\frac{1}{n}\Big),$$

Besides, taken the first-order Taylor expansion of $Z(\hat{\theta}_\lambda) - Z(\theta^*_\lambda)$, we have:

$$Z(\hat{\theta}_\lambda) - Z(\theta^*_\lambda) = \nabla_\theta Z(\theta^*_\lambda)^\top(\hat{\theta}_\lambda - \theta^*_\lambda) + o_p(n^{-1/2}).$$

Combining it with Theorem 16, this gives rise to:

$$\mathbb{E}_{\mathcal{D}_n}[(Z(\hat{\theta}_\lambda) - Z(\theta^*_\lambda))^2] = \frac{1}{n}\nabla_\theta Z(\theta^*_\lambda)^\top\Sigma(\theta^*;\lambda)\nabla_\theta Z(\theta^*_\lambda) + O\Big(\frac{1}{n^2}\Big).$$

Step 2. Term (B). Expand around θ^*. First, we consider $\lambda = o(1)$, which gives rise to:

$$Z(\theta^*_\lambda) = Z(\theta^*) + \frac{1}{2}\lambda^2\pi^\top I(\theta^*)\pi + o(\lambda^2).$$

Therefore

$$\mathbb{E}_{\mathcal{D}_n}[g(Z(\theta^*_\lambda))] - \mathbb{E}_{\mathcal{D}_n}[g(Z(\theta^*))] = g'(Z(\theta^*))\frac{1}{2}\lambda^2\pi^\top I(\theta^*)\pi + o(\lambda^2).$$

For general $\lambda = \Theta(1)$, we utilize the fact that $g(\cdot)$ is nondecreasing from Assumption 4. In order for possible performance improvement, we require that $\theta^*_\lambda = \theta^*$ and $\pi(\theta^*) = 0$.

Step 3. Combine. When $\lambda = o(1)$, subtracting the expansion for $\hat{\theta}$ (with $\lambda = 0$) from that for $\hat{\theta}_\lambda$. If we further denote $\rho(\lambda) = \mathrm{Tr}[\Sigma(\theta^*;\lambda)I(\theta^*_\lambda) - \Sigma(\theta^*;0)I(\theta^*)] = \rho'(0)\lambda + o(\lambda)$, then we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}[g(Z(\hat{\theta}_\lambda))] - \mathbb{E}_{\mathcal{D}_n}[g(Z(\hat{\theta}))] &= \frac{g'(Z(\theta^*))}{2n}(\rho'(0)\lambda + o(\lambda)) \\ &\quad + \frac{g''(Z(\theta^*))}{2n}\Big(\nabla_\theta Z(\theta^*_\lambda)^\top\Sigma(\theta^*;\lambda)\nabla_\theta Z(\theta^*_\lambda)\Big) \\ &\quad + \frac{g'(Z(\theta^*))}{2}\lambda^2\pi^\top I(\theta^*)\pi + o(\lambda^2) + O\Big(\frac{\lambda^2}{n}\Big). \\ &= C\frac{\lambda}{n} + D\lambda^2 + o\left(\frac{1}{n}\right),\end{aligned} \tag{16}$$

where

$$C = \frac{1}{2}\Big(g'(Z(\theta^*))\rho'(0) + g''(Z(\theta^*))\Sigma(\theta^*;0)I(\theta^*)\pi(\theta^*)\Big), \qquad D = \frac{g'(Z(\theta^*))}{2}\pi^\top I(\theta^*)\pi,$$

Above inside the expansion of $\nabla_\theta Z(\theta^*_\lambda)^\top \Sigma(\theta^*;\lambda)\nabla_\theta Z(\theta^*_\lambda)$, we simplify it following from the formula of $\Sigma(\theta^*;\lambda)$ in Theorem 16 with $\nabla_\theta \pi(\theta^*) = 0$ and the fact that:

$$\nabla_\theta Z(\theta^*_\lambda) = \nabla_\theta Z(\theta^*) + \lambda I(\theta^*)\pi(\theta^*) + o(\lambda) = \lambda I(\theta^*)\pi(\theta^*) + o(\lambda).$$

In this case, the optimal improvement is taken when $\lambda = -\frac{C}{2nD} = O(1/n)$ and the improvement is of the second-order.

When $\lambda = \Theta(1)$, the second-order term associated with $g''(\cdot)$ becomes zero and the performance gap in (16) becomes the first-order difference:

$$\mathbb{E}_{\mathcal{D}_n}[g(Z(\hat{\theta}_\lambda))] - \mathbb{E}_{\mathcal{D}_n}[g(Z(\hat{\theta}))] = \frac{g'(Z(\theta^*))}{2n}\rho(\lambda),$$

and the minimizer of $\rho(\lambda)$ with respect to λ is taken as the same as in Theorem 4. □

Appendix E. Proofs in Section 4

First, we have the following guarantee for the estimated uncertainty region $M(\delta)$ for consistence:

Assumption 5 (Source-domain Uniform Concentration) *Assume there exist constants $B, \sigma > 0$ such that, for every $i \in [K]$ and every $M \in \mathcal{F}_\phi$, each coordinate of $M(\theta;\xi;\pi_i)$ is sub-Gaussian with proxy σ and $\|M(\theta;\xi;\pi_i)\|_2 \le B$ almost surely. Let Comp$(\mathcal{F}_\phi)$ denote a capacity measure for $\mathcal{F}_\phi$ (e.g., squared Gaussian/Rademacher complexity or a metric-entropy integral). Then there exists a universal $C > 0$ such that, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\max_{i\in[K]} \sup_{M\in\mathcal{F}_\phi} \left\| \frac{1}{N}\sum_{j=1}^N \Big(M(\theta;\xi_{i,j};\pi_i) - \mathbb{E}_{P_i}[M(\theta;\xi;\pi_i)] \Big) \right\|_2 \le C\left(\sqrt{\frac{Comp(\mathcal{F}_\phi) + \log(K/\delta)}{N}} + \frac{\log(K/\delta)}{N} \right).$$

Lemma 20 (High-probability inclusion of invariant M) *Fix $\delta \in (0,1)$. Suppose Assumption 5 holds and some $M \in \mathcal{F}_\phi$ satisfies the moment equation $\mathbb{E}_{\mathbb{P}_i}\big[M(\theta;\xi;\pi_i)\big] = 0$ for all $i \in [K]$ and the target domain. Define*

$$\epsilon_K(\mathcal{F}_\phi,\delta) := C^2 \frac{Comp(\mathcal{F}_\phi) + \log(K/\delta)}{N}, \qquad \mathcal{M}(\delta) := \Big\{ M \in \mathcal{F}_\phi : \max_{i\in[K]} \Big\| \frac{1}{N}\sum_{j=1}^N M(\theta;\xi_{i,j};\pi_i) \Big\|_2^2 \le \epsilon_K(\mathcal{F}_\phi,\delta) \Big\}.$$

Then $M \in \mathcal{M}(\delta)$ with probability at least $1-\delta$.

Proof of Lemma 20. First, we know $\mathbb{E}_{\mathbb{P}_i} M(\theta;\xi;\pi_i) = 0$ for $\mathbb{P}$ can be any distribution $\mathbb{P}_i$ from the source domain and the target distribution $\mathbb{P}^*$. Apply Assumption 5 and take a union bound over $i \in [K]$ to get, with probability at least $1-\delta$,

$$\max_{i\in[K]} \left\| \frac{1}{N}\sum_{j=1}^N M(\theta;\xi_{i,j};\pi_i) \right\|_2 \;\le\; C\left[\sqrt{\frac{\mathrm{Comp}(\mathcal{F}_\phi) + \log(K/\delta)}{N}} + \frac{\log(K/\delta)}{N} \right].$$

Squaring both sides and using $(a+b)^2 \le 2a^2 + 2b^2$ yields

$$\max_{i\in[K]} \left\| \frac{1}{N}\sum_{j=1}^N M(\theta;\xi_{i,j};\pi_i) \right\|_2^2 \;\le\; C^2 \frac{\mathrm{Comp}(\mathcal{F}_\phi) + \log(K/\delta)}{N} = \epsilon_K(\mathcal{F}_\phi,\delta),$$

which is exactly the defining inequality for $M \in \mathcal{M}(\delta)$. □

Lemma 21 (Target-side plug-in consistency) *Recall $IF(\xi)$ denote the influence function of $\hat{\theta}$ and $I(\theta^*)$ the population curvature matrix entering the decision risk expansion. We have:*

$$\|\widehat{IF} - IF\|_{L^2(P^*)} = o_p(1), \|\hat{I}(\hat{\theta}) - I(\theta^*)\|_{\mathrm{op}} = o_p(1)$$

Proof of Theorem 6. First, we consider the first-order risk expansion. Following the same second-order Taylor expansion of $Z(\theta)$ around θ^* and the influence-function representation as in Theorem 4, we have that for any bounded H and any M,

$$R(\hat{\theta}_{H,M}) \;=\; R(\hat{\theta}) \;+\; \frac{1}{2n}\mathbb{E}_{\mathbb{P}^*}\left[\|IF(\xi) + HM(\xi;\pi_{\mathrm{tgt}})\|^2_{I(\theta^*)}\right] \;+\; o_p\Big(\frac{1}{n}\Big). \tag{17}$$

Intuitively, the perturbation $n^{-1}\sum_i M(\cdot)$ shifts the first-order term of $\hat{\theta}$ from $IF(\xi)$ to $IF(\xi) + HM(\xi)$; the curvature $I(\theta^*)$ weights the quadratic risk.

Then we bound the uniform convergence of the empirical objective. Let

$$\mathcal{F}(H,M) := \mathbb{E}_{\mathbb{P}^*}\left[\|IF(\xi){+}HM(\xi;\pi_{\mathrm{tgt}})\|^2_{I(\theta^*)}\right], \qquad \widehat{\mathcal{F}}_n(H,M) := \frac{1}{n}\sum_{i=1}^{n}\|\widehat{IF}(\xi_i){+}HM(\hat{\theta};\xi_i;\pi_{\mathrm{tgt}})\|^2_{\hat{I}(\hat{\theta})}.$$

By Lemma 21 and standard arguments (triangle inequality, Lipschitzness of $v \mapsto \|v\|_A^2$ in both v and A on bounded sets), we get

$$\sup_{H\in\mathcal{H},\; M\in\mathcal{M}(\delta)} \left|\widehat{\mathcal{F}}_n(H,M) - \mathcal{F}(H,M)\right| \;=\; o_p(1).$$

Because $N \gg n$ and by Lemma 20 with $\delta = \Theta(1/n)$, the additional error from using $\widehat{M} \in \mathcal{M}(\delta)$ (estimated from the K source domains) is also $o_p(1)$ uniformly over M.

Then we transfer the optimality gap from the empirical objective to the population one. Let $(\hat{H}, \widehat{M})$ minimize (5) (equivalently, minimize $\widehat{\mathcal{F}}_n$ up to an irrelevant scale) over $\mathcal{H} \times \mathcal{M}(\delta)$. Then, for any population oracle (H^*, M^*),

$$\begin{aligned}
&\mathcal{F}(\hat{H},\widehat{M}) - \mathcal{F}(H^*,M^*)\\
&\le \left[\mathcal{F}(\hat{H},\widehat{M}) - \widehat{\mathcal{F}}_n(\hat{H},\widehat{M})\right] + \left[\widehat{\mathcal{F}}_n(\hat{H},\widehat{M}) - \widehat{\mathcal{F}}_n(H^*,M^*)\right] + \left[\widehat{\mathcal{F}}_n(H^*,M^*) - \mathcal{F}(H^*,M^*)\right]\\
&\le 2\sup_{\mathcal{H}\times\mathcal{M}(\delta)} |\widehat{\mathcal{F}}_n(\hat{H},\widehat{M}) - \mathcal{F}(H,M)| = o_p(1).
\end{aligned}$$

Finally, for the true performance gap, we plug the expansion (17) for both $(\hat{H},\widehat{M})$ and (H^*,M^*):

$$\begin{aligned}
R(\hat{\theta}_{\hat{H},\widehat{M}}) - R(\hat{\theta}_{H^*,M^*}) &= \frac{1}{n}\Big[\mathcal{F}(\hat{H},\widehat{M}) - \mathcal{F}(H^*,M^*)\Big] + o_p\Big(\frac{1}{n}\Big)\\
&= \frac{1}{n}o_p(1) + o_p\Big(\frac{1}{n}\Big) \;=\; o_p\Big(\frac{1}{n}\Big),
\end{aligned}$$

which proves the claim. □

Theorem 22 (Semiparametric efficiency of the optimal augmentation) *Suppose Assumptions 2 and 1 hold and $\Theta = \mathbb{R}^{D_\theta}$, the estimator with*

$$H^* = -\Gamma\Omega^{-1}$$

is regular, asymptotically linear with influence function

$$\psi_{H^*}(\xi) = IF(\xi) - \Gamma\Omega^{-1}M(\theta^*;\xi),$$

and thus

$$\sqrt{n}(\tilde{\theta}_{H^*} - \theta^*) \;\Rightarrow\; \mathcal{N}\big(0,\; \Sigma_0 - \Gamma\Omega^{-1}\Gamma^\top\big).$$

Moreover, ψ_{H^*} equals the efficient influence function (EIF) of the semiparametric model that augments the baseline with the restriction $\mathbb{E}[M(\theta^*;\xi)] = 0$. Consequently, $\tilde{\theta}_{H^*}$ attains the semiparametric efficiency bound

$$\Sigma_{\text{eff}} = \Sigma_0 - \Gamma\Omega^{-1}\Gamma^\top.$$

Proof of Theorem 22. First, it is easy to see the influence function $\tilde{\theta}_H$ is

$$\psi_H(\xi) = IF(\xi) + HM(\theta^*;\xi).$$

Let $\Sigma(H) := \text{Var}(\psi_H) = \Sigma_0 + H\Omega H^\top + H\Gamma^\top + \Gamma H^\top$. This is a convex quadratic in H. Differentiating $\text{tr}\Sigma(H)$ with respect to H and setting to zero yields the normal equations

$$2H\Omega + 2\Gamma = 0 \quad \Rightarrow \quad H^* = -\Gamma\Omega^{-1},$$

using $\Omega \succ 0$. Substituting gives the minimized covariance

$$\Sigma(H^*) = \Sigma_0 - \Gamma\Omega^{-1}\Gamma^\top.$$

In the semiparametric model that incorporates the valid restriction $\mathbb{E}[M(\theta^*;\xi)] = 0$, the nuisance tangent space contains the span of $M(\theta^*;\xi)$. The efficient influence function is the $L^2(\mathbb{P}^*)$ projection of any regular influence function onto the orthogonal complement of this space. The orthogonal projection of IF off the span of M is

$$IF(\cdot) - \Pi_{\text{span}(M)}IF(\cdot) = IF - \Gamma\Omega^{-1}M(\theta^*;\xi) = \psi_{H^*}.$$

Therefore ψ_{H^*} is the EIF. By semiparametric efficiency theory, any regular estimator has asymptotic covariance at least Var(EIF), and equality holds if and only if its influence function equals the EIF. Since $\tilde{\theta}_{H^*}$ has IF ψ_{H^*}, it attains the bound $\Sigma_0 - \Gamma\Omega^{-1}\Gamma^\top$. □

Remark 23 (On constraints and singular Ω) *If Θ is constrained but θ^* lies in its interior, the projection Π_Θ is asymptotically inactive and the proof is unchanged. If Ω is singular, the same argument goes through by restricting H to the column space of Ω and replacing Ω^{-1} with the Moore–Penrose pseudoinverse $\Omega^\dagger$; the bound becomes $\Sigma_0 - \Gamma\Omega^\dagger\Gamma^\top$ on that subspace.*

Appendix F. Experiments

We validate how perturbed solutions in the standard linear regression and weighted empirical optimization example dempnstrate the first-order performance improvement over the empirical solution.

Table 1: Loss comparison in terms of whether adding the moment condition can achieve the first-order improvement (under a well-specified linear model).

Loss Types / Noise	Normal	Laplace	Exponential	t-distribution
LAD	Yes	No	Yes	Yes
OLS	No	Yes	No	Yes

F.1. Linear Regression

For $\xi = (X, Y)$ with the true data generating process $Y = (\theta^*)^\top X + \epsilon$, consider the OLS loss $\ell(\theta; \xi) = (\theta^\top X - Y)^2$, where X denotes the feature and Y denotes the label., The noise ϵ may follow different noise specifications (normal, Laplace, recentered exponential distribution).

We apply the moment equation $M(\theta; \xi) = \ell(\theta; \xi) \cdot \nabla_\theta \ell(\theta; \xi)$ for different losses ℓ to understand whether the moment equation provides the first-order improvement over the empirical solution.

Above in Table 1, for the recentered exponential noise, the condition $\mathbb{E}[M(\theta; \xi)] = 0$ no longer holds under OLS loss. For OLS-Normal / LAD-Laplace, the condition $\mathbb{E}[M(\theta^*; \xi)] = 0$ holds but we cannot compute $M(\theta; \xi)$ by empirical calculation since we can only observe $\hat{\theta}$ and the non-orthogonal condition does not hold.

We consider the OLS loss and take the corresponding empirical estimator to be $\hat{\theta}_{\mathrm{OLS}}$. Beyond the perturbed estimator induced by $M(\theta; \xi) = X(\theta^\top X - Y)^3$, we evaluate the following procedures to assess whether the theory's predicted improvements materialize in practice:

1. **χ^2-DRO estimator.** We search $\lambda \in [0, 0.2]$ on a grid with step size 0.002 and solve χ^2- DRO problem. For negative values $\lambda \in (-0.2, 0)$, we use the symmetry identity $\hat{\theta}_\lambda = 2\hat{\theta}_{\mathrm{OLS}} - \hat{\theta}_{-\lambda}$ to obtain the estimate.

2. **CVaR-DRO estimator.** We search $\lambda \in (0, 1)$ with step size 0.05 and solve the CVaR-DRO problem. For $\lambda \in (-1, 0)$, we again use $\hat{\theta}_\lambda = 2\hat{\theta}_{\mathrm{OLS}} - \hat{\theta}_{-\lambda}$.

3. **Fusion estimator.** To provide an intuitive baseline for effect size, we consider a convex combination of LAD and OLS with step size of λ being 0.01:

$$\theta_\lambda = (1 - \lambda)\, \hat{\theta}_{\mathrm{LAD}} + \lambda \hat{\theta}_{\mathrm{OLS}}, \qquad \lambda \in [0, 1].$$

Table 2: Performance Gap of OLS Loss where ϵ follows Laplace distribution

	Original	Simple-M	χ^2	CVaR	Fusion
50	0.7151	0.4739	0.5177 (-0.066)	0.4840 (-0.38)	0.3979 (0.0)
100	0.2892	0.2211	0.2298 (-0.032)	0.1920 (-0.32)	0.1681 (0.12)
200	0.1525	0.1378	0.1388 (-0.018)	0.1110 (-0.38)	0.0953 (0.11)
400	0.0848	0.0661	0.0676 (-0.034)	0.0572 (-0.3)	0.0434 (0.01)

Across Tables 2–4, the brackets in the χ^2, CVaR, and Fusion estimators report the oracle-optimal λ selected from the grid search under expected performance. The empirical findings are consistent with the conceptual summary in Table 1, and both sets of results highlight that exploiting structural knowledge of the noise distribution can substantially improve efficiency.

Table 3: Performance Gap of OLS Loss where ϵ follows t-distribution

	Original	Simple-M	χ^2	CVaR	Fusion
50	0.6925	0.5207	0.5014 (-0.052)	0.6046 (-0.4)	0.5345 (0.33)
100	0.3530	0.2433	0.2447 (-0.032)	0.2561 (-0.26)	0.2494 (0.27)
200	0.1815	0.1314	0.1333 (-0.02)	0.1414 (-0.26)	0.1353 (0.33)
400	0.0851	0.0698	0.0732 (-0.006)	0.0711 (-0.32)	0.0684 (0.35)

Table 4: Performance Gap of OLS Loss where ϵ follows normal distribution

	Original	Simple-M	χ^2	CVaR	Fusion
50	0.3726	0.4339	0.3720 (-0.002)	0.3576 (0.640)	0.3725 (1.0)
100	0.1262	0.1590	0.1262 (0.000)	0.1241 (0.840)	0.126 (1.0)
400	0.0492	0.0627	0.0490 (0.002)	0.0491 (0.900)	0.0492 (1.0)

F.2. Weighted (Contextual) Newsvendor

We focus on the following feature-based (univariate) newsvendor problem with

$$\ell(\theta;\xi) = c\theta - p\min\{\theta,\xi\}.$$

Above we set $c = 4, p = 10$. We generate a 10-dimensional covariate ($D_u = 10$), which is uniformly sampled in $[0,1]^{D_u}$. And $\xi|u \overset{d}{=} 5 + Bu + \sin\|u\|_2 + \epsilon$ where $\epsilon \sim N(0,1)$.

Recall the data-driven weighted empirical optimization solution in Definition 17. For each covariate u, we obtain the empirical solution by:

$$\hat{\theta}(u) \in \operatorname*{argmin}_{\theta\in\Theta} \sum_{i\in[n]} w_{n,i}(u)\ell(\theta;\xi).$$

We focus on the performance improvement of two weighted empirical optimization procedure [19]:

1. k-NN estimator, $w_{n,i}(u) = \mathbf{1}_{\{u_i \text{ is a kNN of } u\}}$ with $k_n = \lceil 2\sqrt{n}\,\rceil$;
2. Nadaraya-Watson kernel estimator, $w_{n,i}(u) = \frac{K((u-u_i)/h_n))}{\sum_{j\in[n]} K((u-u_j)/h_n)}$ with $K(u) = \mathbf{1}_{\{\|u\|_2\le 1\}}$ and $h_n = 1.5n^{-\frac{1}{D_u+2}}$.

In this case, we set the moment equation as the knowledge of the conditional mean $\mathbb{E}[\xi|u]$ for each u. And the conditional mean "noisy oracle" is extracted from either of the following:

M1: A noisy oracle estimator that outputs $\mathbb{E}[\xi|u] + \eta$, where $\eta \sim N(0, 25/n)$;

M2: Suppose we have $\{(u_i,\xi_i)\}_{i\in[m]}$ where the true conditional distribution shares the same $\mathbb{E}[\xi|u]$ but has other different quantile informations. $m = 2n$. We fit the dataset with a randomforest regressor.

For each new covariate u, we proceed given the additional conditional mean oracle as follows:

1. Obtain the corresponding estimated influence function $\{\widehat{IF}_{n,u}(U,\xi)\}_{i\in[n]}$, where $\widehat{IF}_{n,u}(U_i,\xi_i) = -\frac{c-p}{p\hat{f}(\hat{\theta}(u))}\mathbf{1}_{\{u \text{ is kNN},\xi>\hat{\theta}(u)\}} - \frac{c}{p\hat{f}(\hat{\theta}(u))}\mathbf{1}_{\{u \text{ is kNN},\xi\leq\hat{\theta}(u)\}}$, where $\hat{f}(\cdot)$ is the estimated density of $\mathbb{P}_{\xi|u}$;

2. Compute $M(\theta;\xi_i) = \xi_i - \hat{\mathbb{E}}[\xi|z]$, $\forall i \in [n]$ and solve the following optimization problem:

$$\min_H \left\{ \sum_{i\in[n]} \|\widehat{IF}_{n,u}((u_i,\xi_i)) + H(\xi_i - \hat{\mathbb{E}}[\xi|z])\|_2^2, \text{ s.t., } \hat{\theta}(z) + \frac{H}{n}\sum_{i\in[n]} w_{n,i}(u)(\xi_i - \hat{\mathbb{E}}[\xi|z]) \in \Theta \right\}$$

3. Obtain the perturbed solution $\hat{\theta}(U) + \frac{H}{n}\sum_{i\in[n]} w_{n,i}(u)(\xi_i - \hat{\mathbb{E}}[\xi|z])$.

We evaluate the performance of each method $\hat{\theta}(u)$—the baseline estimator and the two perturbed versions (M1, M2)—using the metric $\mathcal{G}(\hat{\theta}(u))$, with results reported in Table 5. Each entry represents an average over $N = 500$ problem instances: the linear component B is fixed across all cases, while the solution $\hat{\theta}(u)$ is computed for each model under a distinct u sampled uniformly from $[0,1]^{D_u}$.

$$\mathcal{G}(\hat{\theta}(u)) = \frac{1}{N}\sum_{i\in[N]}\left(\mathbb{E}_{\mathbb{P}_{\xi|u_i}}[\ell(\hat{\theta}(u_i);\xi)] - \min_{\theta\in\Theta}\mathbb{E}_{\mathbb{P}_{\xi|u_i}}[\ell(\theta;\xi)]\right).$$

We find that incorporating perturbations via the corresponding moment equations (M1 or M2) leads to statistically significant performance gains: both the kNN and kernel estimators achieve lower expected cost compared to the original baseline.

Table 5: Performance Comparison of Different Policies (averaged over 500 problem instances)

	kNN			Kernel		
n	Original	M1	M2	Original	M1	M2
100	4.48	2.56	4.09	4.45	2.78	4.33
200	4.32	2.30	3.94	4.33	2.70	4.13
400	3.93	2.07	3.84	4.30	2.68	3.90