Fast Variability Approximation: Speeding up Divergence-Based Distributionally Robust Optimization via Directed Perturbation

Henry Lam, Mohamed Lakhnichi
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027
{khl2114, ml5000}@columbia.edu

Abstract

Distributionally Robust Optimization (DRO) has become a popular paradigm for decision-making under uncertainty, especially when the uncertainty raises from the underlying distributions in stochastic problems. While DRO has been known to enjoy a range of robustness and statistical advantages, it also pays the cost of additional computational overheads. Moreover, this cost can amplify extensively in the phase of hyperparameter tuning. We show that, in the case of ϕ-divergence uncertainty set, simply perturbing an empirical optimizer (i.e., solution from sample average approximation) in a statistically guided fashion achieves almost the same generalization effect as DRO. Importantly, this perturbation avoids the expensive overheads of DRO as long as the problem is smooth enough, which allows suitable gradient extraction via direct computation or resampling-based methods.

1 Introduction

Distributionally Robust Optimization (DRO) [4, 10, 7, 19] has become a popular approach for decision-making under uncertainty, especially when the underlying distribution in a stochastic problem is not fully informed. More concretely, suppose we are faced with the stochastic optimization problem $\min_{x\in\mathcal{X}} \mathbb{E}_Q[\ell(x,\zeta)]$ with a decision x, loss function $\ell(x,\zeta)$ and uncertain parameter ζ, but Q that controls the expectation $\mathbb{E}$ is not fully known. Then, we can consider DRO which is typically formulated as

$$\min_{x\in\mathcal{X}}\{Z^{\text{DRO}}(\rho,x) := \sup_{Q\in\mathcal{U}_\rho} \mathbb{E}_Q[\ell(x,\zeta)]\}, \qquad \mathcal{U}_\rho = \{Q : D_\phi(Q\|P) \le \rho\}\,.$$

Here, the set $\mathcal{U}_\rho$ is known as the uncertainty or ambiguity set, and represents the set of distributions that are intuitively "likely" to be the ground-truth distribution. That is, DRO is a min-max framework that advocates the use of decision that optimizes loss under the worst-case scenario, where the worst case here is over the distribution space within the uncertainty set.

Key questions arising from the DRO framework above are roughly of two types: 1) *robustness and statistical properties:* How to construct $\mathcal{U}_\rho$, and what advantages it would give to the obtained solution; 2) *computation:* How to solve the DRO problem. For the first line of study, the classical motivation of DRO, which is also evident as intuited by the worst-case nature of the setup, is that DRO solution can protect against unexpected distribution shift within the uncertainty set. That is, DRO solution enjoy a performance bound that it cannot perform worse than the optimal value obtained from the DRO formulation [1, 4, 21, 2]. This is an argument on robustness, and many of the recent works have drifted to understand the statistical properties of DRO. In particular, when data are available (and

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

that there is no distribution shift per se), using a neighborhood ball as the geometry of the uncertainty set can be viewed as a regularization on the variability of the loss function [5, 6, 3, 14, 16]. This implies, as an advantage of DRO, that its solutions have better statistical generalization than empirical optimizer (EO) that simply uses sample average to replace the unknown Q, in situations where the loss function has a small variance [5].

On the computational side, compared to EO which involves only a "min", the "min-max" operation in DRO adds to additional overheads. To this end, there have been significant algorithmic advances. In this paper, we focus on ϕ-divergence uncertainty set, which is a common type of uncertainty set and has the statistical advantages described above. In this setting, methods including stochastic mirror descent, re-weighted gradient descent, and dual reformulations, have been developed to enhance computation of the solution. Table 1 summarizes several known algorithms in convex settings.

Table 1: Common algorithms for solving ϕ-divergence DRO

Algorithm	Uncertainty Set	Hyperparameters	Principle	Nature
Two-player mirror descent [17]	ϕ-divergence	ρ, primal/dual learning rates	Primal-dual iterative updates	Solver
Accelerated SGD [15]	CVaR, χ^2	ρ, α, learning rate	Lagrange dual via SGD	Solver
RSCDRO [18]	KL divergence	ρ, momentum, learning rate	Recursive variance reduction	Solver
Re-weighted GD [11]	KL divergence	γ, τ, learning rate	Dynamic sample re-weighting	Solver
This work	ϕ-divergence	ρ	Lagrange dual + perturbation expansion	**Approximation**

Despite the elegance of these approaches, it is also the case that, expectedly, DRO is more challenging to solve than EO. The challenge is not only about the algorithmic complexity, but also the increased set of algorithmic hyperparameters including learning rates, clipping thresholds, re-weighting exponents, and batch sizes, which need to be fine-tuned via cross-validation. In cases where the radius of the uncertainty set itself needs to be fine-tuned, the altogether-involved cross-validation will impose a significant computational load than EO that is free of the radius parameter, especially in large-scale settings. At a high level, this computational overhead appears logical, as it can be viewed as the price to pay for the robustness/statistical gain. However, and we will explain further below, this known computation-robustness trade-off is not necessarily optimized. In fact, our main goal in this paper is to tackle the question: *Is there an alternative approach that achieves similar robustness/statistical effects as DRO, while avoiding its computational cost?*

2 Fast Variability Approximation: Basic Ideas

We propose *fast variability approximation (FVA)* as a simple alternative approach to ϕ-divergence DRO that achieves a "best of both worlds": It attains similar robustness/statistical benefits as DRO, while having a computation cost that is substantially less, arguably much closer to EO than the existing DRO algorithms. On a high level, FVA conducts a *perturbation* on the EO solution. That is, it first solves EO (using any existing EO algorithm), then perturbs it in a statistically guided manner that improves EO and towards the performance of DRO. More precisely, suppose we can solve the "un-robustified" problem $x^* = \arg\min_{x\in\mathcal{X}}\{Z(x) := \mathbb{E}_P[\ell(x,\zeta)]\}$. Then, we perturb x^* to obtain the FVA solution

$$x^{\mathrm{FVA}}(\rho) = x^* - \sqrt{2\phi^{*\prime\prime}(0)\rho}\,\delta(x^*), \tag{1}$$

where

$$\delta(x^*) = \left(\nabla_x^2 \mathbb{E}_P[\ell(x^*,\zeta)]\right)^{-1} \frac{\mathrm{Cov}_P(\ell(x^*,\zeta), \nabla_x \ell(x^*,\zeta))}{\sqrt{\mathrm{Var}_P(\ell(x^*,\zeta))}}.$$

We will describe a basic guarantee of $x^{\mathrm{FVA}}(\rho)$. First, we state some regularity conditions on the loss function $\ell(x,\zeta)$:

Assumption 1 (Regularity Conditions on the Loss Function)**.**

1. *$\ell(.,\zeta)$ is convex and differentiable almost everywhere in $x \in \mathcal{X}$*
2. *$\ell(.,\zeta)$ is $\mathcal{L}^2 - Lipschitz$ and $\nabla_x \ell(.,\zeta)$ is $\mathcal{L}^1 - Lipschitz$.*

3. *For any bounded function $g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ that is continuously differentiable with bounded partial derivatives, the function $\mathbb{E}_P[g(x, \ell(x,\zeta))\nabla_x \ell(x,\zeta)]$ is continuously differentiable in $x \in \mathcal{X}$.*

4. *For every $x \in \mathcal{X}$: $Var_P(\ell(x,\zeta)) > 0$.*

5. *$\nabla_x^2 \mathbb{E}[\ell(x^*,\zeta)]$ is positive definite.*

6. *$\mathcal{X}$ is compact and convex.*

Theorem 1 (Worst-Case Performance Guarantee of FVA)**.** *Under Assumption 1, there exists $\bar{\rho} > 0$ and a constant $M > 0$ such that: 1) As ρ shrinks to 0, it holds that $Z^{\mathrm{DRO}}(\rho, x^{\mathrm{FVA}}(\rho)) = Z^{\mathrm{DRO}}(\rho, x^{\mathrm{DRO}}(\rho)) + O(\rho)$. 2) For all $\rho \le \bar{\rho}$ and any distribution $\tilde{Q} \ll P$ satisfying $D_\phi(\tilde{Q}\|P) \le \rho$, it holds that $\mathbb{E}_{\tilde{Q}}[\ell(x^{\mathrm{FVA}}(\rho),\zeta)] \le Z^{\mathrm{DRO}}(\rho, x^{\mathrm{DRO}}(\rho)) + \rho M$.*

Theorem 1 establishes that FVA has a similar robustness guarantee as DRO. Specifically, the first part stipulates that $x^{\mathrm{FVA}}(\rho)$ has a similar worst-case objective performance as $x^{\mathrm{DRO}}(\rho)$, in the sense of up to an $O(\rho)$ error as $\rho \to 0$. The second part further makes clear that the performance of $x^{\mathrm{FVA}}(\rho)$, regardless of the distribution $\tilde{Q}$ within the uncertainty set, is bounded by the DRO optimal value up to a similar $o(\rho)$ error. The latter is essentially the standard robustness guarantee of DRO. That is, FVA attains a similar protection mechanism as DRO against distribution shift.

The rationale of FVA hinges on the Taylor-type approximation of divergence-based DRO, namely that $Z^{\mathrm{DRO}}(\rho, x)$ is approximately equal to the *variance-regularized* objective [12, 13, 5, 8, 9]

$$Z^{Reg} = Z(x) + \sqrt{2\phi^{*\prime\prime}(0)\rho Var_P(\ell(x,\zeta))} \tag{2}$$

Solution (1) then can be viewed as running a Newton-like step starting from x^* to solve for the variance-regularized objective. That is, this step locally minimizes a quadratic approximation of (2). As a result, $x^{\mathrm{FVA}}(\rho)$ can be shown to be close to $x^{\mathrm{DRO}}(\rho)$ as $\rho \to 0$.

3 Data-Driven Fast Variability Approximation

We now turn our attention to the data-driven scenario, where there is no anticipated distribution shift, but the true distribution P is unknown and only sample data are available. In this case, we first construct the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{\zeta_i}$, where δ_{ζ_i} denotes the Dirac measure at ζ_i. The empirical DRO problem is

$$\hat{x}^{\mathrm{DRO}}(\rho) = \arg\min_{x\in\mathcal{X}} \left\{ Z_n^{\mathrm{DRO}}(\rho, x) := \sup_{Q: D_\phi(Q\|P_n)\le\rho} \mathbb{E}_Q[\ell(x,\zeta)] \right\}$$

In a similar fashion as above, we construct the data-driven FVA solution by first solving the empirical optimization $\hat{x}^{\mathrm{EO}} = \arg\min_{x\in\mathcal{X}} \mathbb{E}_{P_n}[\ell(x,\zeta)]$, and then perturbing to

$$\hat{x}^{\mathrm{FVA}}(\rho) := \hat{x}^{\mathrm{EO}} - \sqrt{2\phi^{*\prime\prime}(0)\rho}(\nabla_x^2 \mathbb{E}_{P_n}[\ell(\hat{x}^{\mathrm{EO}},\zeta)])^{-1} \frac{Cov_{P_n}(\ell(\hat{x}^{\mathrm{EO}},\zeta), \nabla_x \ell(\hat{x}^{\mathrm{EO}},\zeta))}{\sqrt{Var_{P_n}(\ell(\hat{x}^{\mathrm{EO}},\zeta))}} \tag{3}$$

To state the statistical guarantee brought by $\hat{x}^{\mathrm{FVA}}(\rho)$, we consider the function class:

$$\mathcal{F} := \{\, g(|\ell(x,\cdot)|, \|\nabla_x \ell(x,\cdot)\|_2,\ \|\nabla_x^2 \ell(x,\cdot)\|_2, \|\nabla_x^3 \ell(x,\cdot)\|_2) : x \in \mathcal{X}, g : \mathbb{R}^4 \to \mathbb{R} \text{ is Lipschitz}\}$$

and the empirical Rademacher complexity [20] of the class $\mathcal{F}$: $\mathcal{R}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(\zeta_i)\right]$ where $\epsilon_i \in \{-1, 1\}$ are i.i.d. independent of the ζ_i's that follows the Rademacher distribution ($P(\epsilon = 1) = P(\epsilon = -1) = 1/2$). With these, and with the regularity conditions on the loss function similar to Assumption 1 (depicted in Assumption 2 in Appendix A), we have the following high-probability bounds:

Theorem 2. *If Assumption 2 holds, there exist constants $\rho_m, \mu_1, \tilde{\mu}_1$ that depend on ℓ and its derivatives, and universal constants $\tilde{M}, C_0, \tilde{C}_0$ so that the following holds with probability at least $1-\delta$:*

$$\begin{cases} ||\hat{x}^{\mathrm{DRO}}(\rho) - \hat{x}^{\mathrm{EO}}||_2 \le \sqrt{\rho}\frac{\tilde{M}}{m_\phi \sigma^*} \cdot \left[\mu_1 + C_0\left(2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(4/\delta)}{n}}\right)\right] \\ ||\hat{x}^{\mathrm{DRO}}(\rho) - \hat{x}^{\mathrm{FVA}}(\rho)||_2 \le \rho\frac{\tilde{M}}{2m_\phi \sigma^*} \left[\tilde{\mu} + \tilde{C}_0\left(2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(4/\delta)}{n}}\right)\right] \end{cases} \tag{4}$$

for all $\rho \le \rho_m$ and $\rho_m = O\left(\frac{m_\phi^2 (\sigma^*)^2}{M_\phi^2}\right)$.

The first conclusion of Theorem 2 establishes that the deviation between $\hat{x}^{\mathrm{EO}}$ and $\hat{x}^{\mathrm{DRO}}(\rho)$ scales as $O(\sqrt{\rho})$, up to a finite-sample factor of $O(\mathcal{R}_n(\mathcal{F}) + \sqrt{\log(1/\delta)/n})$. On the other hand, the second conclusion shows that the deviation between FVA and DRO is much smaller, within $O(\rho)$, up to the same order of factor. This translates into the following generalization bound:

Corollary 1. *Under Assumption 2, with probability at least* $1-\delta$*,*

$$Z(\hat{x}^{\mathrm{FVA}}(\rho_n)) \leq Z_n^{\mathrm{DRO}}(\rho_n, \hat{x}^{\mathrm{FVA}}(\rho_n)) + C_0\frac{\rho}{n} + C_1\frac{\tilde{M}}{2m_\phi\sigma^*}\left[\tilde{\mu} + \tilde{C}_0\left(2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(4/\delta)}{n}}\right)\right]\frac{\rho}{n}, \tag{5}$$

where $\rho_n = \rho/n$*.*

To understand the significance of Corollary 1, it is known that the exact ϕ-divergence DRO solution satisfies $Z(\hat{x}^{\mathrm{DRO}}(\rho_n)) \leq Z^{\mathrm{DRO}}(\rho_n, \hat{x}^{\mathrm{DRO}}(\rho_n)) + C_0\frac{\rho}{n}$ where $\hat{x}^{\mathrm{DRO}}(\rho_n)$ is the empirical DRO solution with radius $\rho_n = \rho/n$. This bound can be converted into a regret bound, namely $Z(\hat{x}^{\mathrm{DRO}}(\rho_n)) - Z(x^*)$, where x^* is the population optimal solution, that places DRO in favor of empirical optimization when $Var(\ell(x^*, \zeta))$ is small [5]. In particular, in the latter situation, the dominating term in the regret bound will become proportional to $Var(\ell(x^*, \zeta))$. Corollary 1 implies that the same benefit holds for FVA, as the additional term in (5) compared to the known bound is of higher order.

4 Computational Comparisons

Now that we illustrated how FVA enjoys similar statistical guarantees as DRO in the small-radius regime, we turn to the computational comparison to argue how FVA has the "best of both worlds". The computation of FVA requires two main steps. First is solving the empirical optimization, and second is to compute the perturbation direction. Compared to DRO, the first step avoids the intricate saddle-point or reweighting procedures inherent in existing DRO solvers, and requires less hyperparameters such as momentum terms and clipping thresholds. Moreover, if we want to fine-tune the radius parameter in DRO using cross-validation, it requires re-solving the DRO (number of folds) $\times$ (discretization grid size to select radius) times. In contrast, FVA only requires solving the empirical optimization and computing the perturbation direction (number of folds) times, without computationally scaling with the radius grid size, because FVA solution has a simple linear form in the radius.

Lastly, in computing the perturbation direction, the main overhead is the inversion of the Hessian of the expected loss. This inversion can be computationally demanding in high dimensions. Moreover, for some problems (such as the newsvendor problem), while the Hessian can exist, approximating it using Monte Carlo or data directly is challenging due to the prohibition in the interchange of derivative and expectation. In such situations, a practical and theoretically justifiable alternative is to use resampling, based on the known approximation of the empirical solution:

$$\hat{x}^{\mathrm{EO}} = x^* - \left(\nabla_x^2 \mathbb{E}_{P_n}[\ell(x^*, \zeta)]\right)^{-1}\left(\mathbb{E}_{P_n}[\nabla\ell(x^*, \zeta)] - \mathbb{E}_P[\nabla\ell(x^*, \zeta)]\right) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

This expansion suggests a bootstrap-based procedure: draw empirical distributions $(P^{(b)})_{1\leq b\leq B}$ by resampling from the available data, solve the empirical optimization problem under each $P^{(b)}$ to obtain solution $\hat{x}^b$, and regress each coordinate $(\hat{x}_i^b)_{1\leq b\leq B}$ against the corresponding empirical losses $(\mathbb{E}_{P^{(b)}}[\ell(\hat{x}^{\mathrm{EO}}, \zeta)])_{1\leq b\leq B}$. The resulting slope $\hat{\beta}_i$ approximates the perturbation direction, yielding the FVA solution after aggregating across coordinates. This approach obtains the FVA solution by solving multiple EO problems, without requiring explicit Hessian inversion.

In the appendix, we provide comprehensive technical details for the paper's results. Section A states the regularity conditions for the data-driven setting. Section B elucidates the derivation of finite-sample guarantees in Theorem 2.

References

[1] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2): 341–357, 2013.

[2] D. Bertsimas, V. Gupta, and N. Kallus. Robust sample average approximation. *Mathematical Programming*, 171(1):217–282, 2018.

[3] J. Blanchet, Y. Kang, and K. Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

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

[5] J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

[6] R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 72(3):1177–1191, 2024.

[7] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4-part-1):902–917, 2010.

[8] J.-y. Gotoh, M. J. Kim, and A. E. Lim. Robust empirical optimization is almost the same as mean–variance optimization. *Operations research letters*, 46(4):448–452, 2018.

[9] J.-y. Gotoh, M. J. Kim, and A. E. Lim. Calibration of distributionally robust empirical optimization models. *Operations Research*, 69(5):1630–1650, 2021.

[10] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

[11] R. Kumar, A. S. Suggala, P. Ravikumar, and M. Fazel. Stochastic re-weighted gradient descent via distributionally robust optimization. *Transactions on Machine Learning Research*, 2024.

[12] H. Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

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

[14] H. Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

[15] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[16] P. Mohajerin Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[17] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in Neural Information Processing Systems*, 29:2208–2216, 2016.

[18] Q. Qi, J. Lyu, Z. Zhang, and Y. Wang. Stochastic constrained dro with a complexity independent of sample size. *arXiv preprint arXiv:2210.05740*, 2022.

[19] H. Rahimian and S. Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022.

[20] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[21] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

A Regularity Assumptions for the Data-Driven Setting

Assumption 2 (Conditions for Finite-sample Bounds)**.**

1. *$\ell(.,\zeta)$ is convex and a.s. three times continuously differentiable in $x \in \mathcal{X}$.*
2. *$\ell(.,\zeta)$, $\nabla_x \ell(.,\zeta)$, $\nabla_x^2 \ell(.,\zeta)$, $\nabla_x \ell(.,\zeta)\nabla_x \ell(.,\zeta)^T$ and $\nabla_x^3 \ell(.,\zeta)$ are $\mathcal{L}^1 - Lipschitz$.*
3. *For every $x \in \mathcal{X}:\ Var_P(\ell(x,\zeta)) > 0$.*
4. *There exists $\sigma^* > 0$ such that $\mathbb{E}_P[\nabla_x^2 \ell(x^*,\zeta)], \mathbb{E}_{P_n}[\nabla_x^2 \ell(\hat{x}^{\mathrm{EO}},\zeta)] \geq \sigma^*$.*
5. *ϕ^* is three times continuously differentiable and we denote*

$$\begin{cases} m_\phi := \min\{\min_{(x,\lambda)} |\phi^{*'}(g(x,\lambda))|,\ g : \mathcal{X} \times \mathbb{R} \to \mathbb{R} \textit{ bounded}\} > 0 \\ M_\phi := \max\{\max_{(x,\lambda)}\{|\phi^{*''}(g(x,\lambda))|, |\phi^{*'''}(g(x,\lambda))|\},\ g : \mathcal{X} \times \mathbb{R} \to \mathbb{R} \textit{ bounded}\} < \infty \end{cases}$$

B Some Explanations on the Derivation of Finite-Sample Results

The proof of Theorem 2 combines several techniques from empirical process theory, perturbation analysis, and nonlinear functional analysis. A key step is to characterize the empirical DRO optimizer as the solution to an implicit system of equations involving the dual variables, which requires establishing differentiability of the optimal dual multipliers with respect to the robustness parameter. This is accomplished via an application of the implicit function theorem, but a nontrivial technical challenge arises in ensuring that the relevant Jacobian remains invertible uniformly in a neighborhood of the empirical risk minimizer.

We address this by carefully bounding the perturbation of the empirical Hessian and exploiting curvature conditions that ensure stability of the solution mapping. In particular, the admissible robustness radius $\rho_m$is governed by the interplay between the curvature parameter and the regularity constants associated with ϕ^*. When the Hessian is ill-conditioned, i.e., σ^* approaches zero, the resulting ρ_m becomes vanishingly small, thereby restricting the applicability of the finite-sample guarantees in Theorem 2 to very conservative robustness levels. Conversely, when the loss landscape exhibits well-behaved curvature, the bound allows for a substantially larger robustness radius, thereby broadening the regime in which the perturbation approximation remains accurate. Moreover, the constant prefactor in the PAC bounds is given by $\frac{\tilde{M}}{m_\phi \sigma^*}$, which is inversely proportional to ρ_m. Hence, a larger admissible robustness radius directly translates into tighter finite-sample guarantees. This reveals a precise trade-off: strong curvature not only enlarges the robustness regime in which the perturbation expansion remains valid but simultaneously sharpens the associated statistical error bounds, while flat loss landscapes yield vacuous or excessively loose guarantees.