A Distributionally Robust Estimator that Dominates the Empirical Average

Nikolas Koumpis
Yale University
nikolaos.koumpis@yale.edu

Dionysis Kalogerias
Yale University
dionysis.kalogerias@yale.edu

Abstract

We leverage the duality between risk-averse and distributionally robust optimization (DRO) to devise a distributionally robust estimator that strictly outperforms the empirical average for all probability measures having triple squared variance greater than the central fourth-order moment.

1 Introduction

The aim of this paper is to deploy the basic insights of DRO in order to devise an empirical estimator that outperforms the (commonly used) sample average w.r.t. the mean squared error; the latter being a standard figure of merit in statistics and machine learning [Girshick and Savage, 1951, James and Stein, 1961, Devroye et al., 2003, Wang and Bovik, 2009, Verhaegen and Verdult, 2007, Kuhn et al., 2019]. To set up the stage for our investigation, let $\mathrm{B}(\mathbb{R}^d)$ be the set of Borel probability measures in $\mathbb{R}^d$ and $D = \{X_1, \ldots, X_n\}$ a set of independent identically distributed (i.i.d.) random vectors realizing $X \sim \mu \in \mathrm{B}(\mathbb{R}^d)$. The corresponding to D empirical measure is $\mu_n = n^{-1}\sum_{k=1}^{n} \delta_{X_i}$, where δ_x is the Delta measure of unit mass located at $x \in \mathbb{R}^d$.

By defining the real-valued map $f(\xi, \widehat{\mathrm{X}}) \equiv \mathbb{E}_\xi\|\widehat{\mathrm{X}} - X\|_2^2$, $\xi \in \mathrm{B}(\mathbb{R}^d)$, we obtain the following characterization of the empirical average

$$\alpha(D; \mu_n) = \frac{1}{n}\sum_{i}^{n} X_i = \underset{\widehat{\mathrm{X}} \in \mathbb{R}^d}{\operatorname{argmin}} f(\mu_n, \widehat{\mathrm{X}}), \tag{1}$$

as the solution to the surrogate of the "true" problem

$$\min_{\widehat{\mathrm{X}} \in \mathbb{R}^d} f(\mu, \widehat{\mathrm{X}}), \tag{2}$$

when μ is unknown. Plausibly then, we are interested in the performance of (1) on μ:

$$\pi_\alpha(\mu, \mu_n) \equiv f(\mu, \alpha(D; \mu_n)) - f^\star, \tag{3}$$

where[1] $f^\star = f(\mu, \mathbb{E}X)$ is the optimal value of (2). Although (1) dominates among unbiased estimators for normally distributed low dimensional data, in the case of skewed or heavy tailed measures, it exhibits a rather sub-optimal performance and therefore, despite its simplicity, it leaves room for improvement. Such an improvement could be potentially achieved by an appropriate re-weighting $\nu_n \in \mathrm{B}(\mathbb{R}^d)$, $\nu_n \ll \mu_n$ on D, such that

$$\pi_\alpha(\mu, \nu_n) \leq \pi_\alpha(\mu, \mu_n). \tag{4}$$

The most prominent approach for doing so is via distributionally robust optimization. As a paradigm, DRO [Scarf et al., 1957] has gained interest due to its connections to regularization, generalization, robustness and statistics [Staib and Jegelka, 2019], [Blanchet et al., 2021], [Bertsimas and Sim, 2004], [Blanchet et al., 2021], [Blanchet and Shapiro, 2023], [Nguyen et al., 2023], [Blanchet et al., 2024].

Let $\mathrm{F}(\beta, \rho)$ be the field of DRO problem instances P of the form

$$\min_{\widehat{\mathrm{X}} \in \mathbb{R}^d} \left\{ \sup_{\xi \ll \mu} f(\xi, \widehat{\mathrm{X}}) \ : \ \xi \in \beta_\rho(\mu) \right\}, \tag{5}$$

where $\beta_\rho(\mu) \subseteq \mathrm{B}(\mathbb{R}^d)$ is the so called uncertainty set of radius $\rho \geq 0$ centered at μ. The two main approaches for representing the uncertainty set, recently unified by [Blanchet et al., 2023], are the divergence approach, and the Wasserstein approach. In the former one, distributional shifts are measured in terms of likelihood ratios [Bertsimas and Sim, 2004, Bayraksan and Love, 2015, Namkoong and Duchi, 2016, Duchi and Namkoong, 2018, Van Parys et al., 2021], while in the latter, the uncertainty set is represented by a Wasserstein ball [Esfahani and Kuhn, 2015, Gao and Kleywegt, 2023, Lee and Raginsky, 2018, Kuhn et al., 2019].

Let $(\widehat{\mathrm{X}}^\star, \xi^\star)$ be a solution of $\mathrm{P} \in \mathrm{F}(\beta, \rho)$. Based on the discussion so far, we are interested in an *ideal* sub-field of instances $\mathrm{F}^\star(\beta, \rho) \subset \mathrm{F}(\beta, \rho)$ formed by the following design specifications: First, if $\mathrm{P} \in \mathrm{F}^\star(\beta, \rho)$, then there exist a saddle point $(\widehat{\mathrm{X}}^\star(\xi^\star), \xi^\star(\widehat{\mathrm{X}}^\star))$. Then, the first design specification implies that necessarily, it is going to be

$$\widehat{\mathrm{X}}^\star = \widehat{\mathrm{X}}^\star(\xi^*) = \mathbb{E}_{\xi^\star(\widehat{\mathrm{X}}^\star)} X. \tag{6}$$

In some cases, robustness against distributional shifts is as safeguarding against risky events (naturally coordinated by the distribution's tail). That is, solving a DRO problem is equivalent to solving a risk-averse problem. This brings us to the second specification which requires $\widehat{\mathrm{X}}^\star$ to be the closed form solution of a risk-averse problem (over μ). Combining the aforesaid specifications would allow us to ***implement*** the empirical counterpart $\alpha(D; \xi^\star_{\rho,n})$ of $\widehat{\mathrm{X}}^\star$ from data D drawn from μ. That being said, in this paper we study the following motivating question:

Is there an instance $\mathrm{P} \in \mathrm{F}^\star(\beta, \rho)$ *such that* $\pi_\alpha(\mu, \xi^\star_{\rho,n}(\beta)) < \pi_\alpha(\mu, \mu_n)$*?*

The aforesaid dual relation between risk-averse and DRO suggests a place to start our design. In particular, our methodology is as follows: We are going to state a risk-averse

[1] We assume that the first order moment is finite

problem (which in principle is easier to design) that is solvable in closed form. Then, we are going to show that under certain conditions, its solution also solves a DRO problem. Then, we are going to verify our initial assertion by studying the performance of the corresponding empirical solution. We start by considering the following risk-constrained extension of (2):

$$\begin{aligned} \min_{\widehat{\mathrm{X}}\in\mathbb{R}^d} \quad & f(\mu,\widehat{\mathrm{X}}) \\ \text{s.t.} \quad & \mathbb{E}_\mu\Big(\|X-\widehat{\mathrm{X}}\|_2^2 \;-\; \mathbb{E}_\mu\|X-\widehat{\mathrm{X}}\|_2^2\Big)^2 \le \varepsilon. \end{aligned} \tag{7}$$

The risk-averse problem (7) is a quadratically constrained quadratic program that in principle can be solved in closed form.

Risk-averse optimization refers to the optimization of risk measures. Unlike expectations, these are (often) convex functionals on appropariate spaces of random variables and deliver optimizers that take into account the risky events of the underlying distribution. Unlike expectations, a convex risk measure can be represented through the Legendre transform as the supremum of its affine minorants. This representation essentially characterizes the risk measure as a supremum over a certain class of affine functionals, specifically over a subset of the dual space of signed measures [Shapiro et al., 2021]. For the particular class of coherent risk measures, this set contains probability measures and thus, safeguarding for risky events is as being robust against distributional shifts. In other words, solving a risk-constrained problem on the population at hand is as solving a min-max problem where expectations are taken for an appropriate re-weighting of that population.

That said, the challenge is that minimization of a risk measure does not always correspond to a DRO formulation. As an example, problem (7) is closely related to the mean-variance risk-measure [Shapiro et al., 2021, p. 275], which does not lead to a DRO formulation in general. Mean-variance optimization was initially utilized by [Markowitz, 1952], in portfolio selection. Later on, [Abeille et al., 2016] deploys mean-variance for dynamic portfolio allocation problems, and [Kalogerias et al., 2020] stressed the importance of safeguarding against risky events in mean squared error Bayesian estimation. DRO with χ^2-divergence is almost equivalent as controlling by the variance [Gotoh et al., 2018, Lam, 2016, Duchi and Namkoong, 2019] were the worst case measure is computed exactly in $O(n\log(n))$ [Staib and Jegelka, 2019]. Although minimization of the mean-variance risk-measure does not correspond to an element of $\mathrm{F}(\beta,\rho)$, the mean deviation risk measure of order $p=2$ does (see Theorem 1).

At this point, we state informally the first Theorem (see Theorem 1) of this paper which establishes that the *candidate* problem (7) corresponds to an instance $\mathrm{P}\in\mathrm{F}^\star(\beta,\rho)$:

Informal Theorem 1. *Let* $\Lambda\equiv\Big\{\lambda\ge 0 \;:\; \lambda\le 1/\gamma^*\Big\}$, *be a subset of admissible Lagrange multipliers associated with the constraint of (7), with* γ^* *being a computable constant. Then, for the corresponding levels of risk, problem (7) is the same as the* χ^2-*DRO problem*

$$\min_{\widehat{\mathrm{X}}\in\mathbb{R}^d}\left\{\sup_{\nu\ll\mu}\ \mathbb{E}_\nu\|\widehat{\mathrm{X}}-X\|_2^2 \;:\; \chi^2(\nu||\mu)\le\lambda\right\},\ \lambda\in\Lambda. \tag{8}$$

Theorem 1 states that for the corresponding to Λ levels of risk ε of (7), the closed form solution of (7) solves all $\mathrm{P}\in\mathrm{F}(\chi^2,1/\gamma^*)$.

Subsequently, we verify our initial assertion by answering the main question of this work in the one dimensional case. For all probability measures with finite second order moment, let N denote the subset of those with the property that the triple squared variance exceeds the central fourth order moment, $3m_2^2 > m_4$. We have the following Theorem (see Theorem 2):

Informal Theorem 2. *Let $\widehat{\mathrm{X}}^{\star}_{n,\bar{\lambda}}$ be the empirical solution of (7), and $\bar{\mathrm{X}}_n$ be the empirical average. Fix an $\epsilon > 0$. Then, for probability measures $\mu \in N$, and for any $n \geq 3$,*

$$\mathbb{E}|\widehat{\mathrm{X}}^{\star}_{n,\bar{\lambda}} - X|^2 < \mathbb{E}|\bar{\mathrm{X}}_n - X|^2, \tag{9}$$

for all $\bar{\lambda} < \frac{1}{n}\left[\frac{8(3m_2^2 - m_4)}{\mathbb{E}\mathrm{T}_n^2 + \epsilon}\right]$.

1.1 Summary of contributions

Problem (7) can be recast to a quadratically constrained quadratic program (Lemma 1), and the resulting risk-constrained estimator (Proposition 1) corresponds to a stereographic projection of the mmse sphere in the set of square integrable functions $\mathcal{L}_2$ onto $\mathbb{R}^d$. We find that the components of the resulting bias between the aforesaid stereographic projection and the traditionally deployed orthogonal projection depend on the individual central third order moments and on cross third order terms between the components (Corollary 1). In the one dimensional case, the risk-averse estimator performs a shift of the expectation by a fraction of the Fisher's moment coefficient of skewness. Under certain conditions, the risk-constrained estimator is written as an expectation w.r.t. a measure resulting after a re-weighting (Radon-Nikodym derivative) on the initial measure (Lemma 2). By considering the differential structure of the space of probability measures with finite second order moment under the 2−Wasserstein metric, we give meaning to the directionality of the aforesaid Radon-Nikodym derivative as the negative gradient of the χ^2-divergence (see sub-section 2.1). We show that the aforesaid risk-constrained estimator solves the χ^2-distributionally robust mean squared error problem for all χ^2-radii that do not exceed a computable constant (Theorem 1). Lastly, for the one dimensional case we prove strict domination for the estimator w.r.t. the quadratic loss against the empirical average for a rather large set of probability measures (Theorem 2).

All proofs of the results presented hereafter can be found in the supplementary material.

2 Risk-constrained minimum mean squared error estimators

We begin analyzing the risk-constrained mmse problem (7) by first referring the reader to [Kalogerias et al., 2019, Lemma 1], where the well-definiteness of the statement (7) is ensured under the following regularity condition

Assumption 1. *It is true that $\int \|x\|_2^3 \, \mathrm{d}\mu(x) < +\infty$.*

By noticing that the constraint in (7) can be written as a perfect square, we establish a quadratic reformulation that on the one hand, allows us to interpret the solution to (7) as a stereographic projection, and on the other, sets the stage for a derivative-free solution:

Lemma 1 (The risk-constrained estimator as stereographic projection)**.** *Under Assumption 1, problem (7) is well-defined and equivalent to the convex quadratic program*

$$\begin{aligned} \min_{\widehat{\mathrm{X}}\in\mathbb{R}^d} \quad & \mathbb{E}_\mu\|X-\widehat{\mathrm{X}}\|_2^2 \\ \text{s.t.} \quad & \|\widehat{\mathrm{X}}-\tfrac{1}{4}\Sigma^{-1}\zeta^-\|_{4\Sigma}^2 \leq \bar{\varepsilon} \end{aligned} \tag{10}$$

where $\zeta^- = 2\Big(\mathbb{E}[\|X\|_2^2X]-\mathbb{E}\|X\|_2^2\mathbb{E}X\Big)$, $\bar{\varepsilon}=\varepsilon+(\zeta^-)^\top\frac{\Sigma^{-1}}{4}\zeta^- - \alpha$, $\alpha=\mathbb{E}(||X||^2-\mathbb{E}||X||^2)^2$ *and* $\Sigma\succcurlyeq 0$ *is the covariance matrix of* X [2]*.*

Lemma 1 shows the equivalence of (7) to the convex quadratic program (10) allowing to interpret the former in geometric terms based on the standard inner product in $\mathbb{R}^d$. Within this setting, the solution to (7) is a stereographic projection of the mmse−sphere onto $\mathbb{R}^d$. Further, we can do slightly more by applying Pythagora's theorem to the objective in (10) to obtain

$$\begin{aligned} \min_{\widehat{\mathrm{X}}\in\mathbb{R}^d} \quad & \|\widehat{\mathrm{X}}-\mathbb{E}X\|_2^2 \\ \text{s.t.} \quad & \|\widehat{\mathrm{X}}-\tfrac{1}{4}\Sigma^{-1}\zeta^-\|_{4\Sigma}^2 \leq \bar{\varepsilon}, \end{aligned}, \tag{11}$$

where by ζ^- we mean the opposite vector to ζ. According to Lemma 1, the solution of (11) (and therefore of (7)) is the point lying on the ellipse of radius $\sqrt{\bar{\varepsilon}}$ and of center $\frac{1}{4}\Sigma^{-1}\zeta$, with the least distance from the orthogonal projection $\mathbb{E}X$. Being guided by the geometric picture, the problem is feasible for any $\bar{\varepsilon}\geq 0$. In particular, for all $\bar{\varepsilon}\geq\bar{\varepsilon}_c=\|\mathbb{E}X-\frac{1}{4}\Sigma^{-1}\zeta^-\|_{4\Sigma}^2$, $\widehat{\mathrm{X}}^*=\mathbb{E}X$, while for $\bar{\varepsilon}=0$, a solution exists only when $\mathbb{E}X=\frac{1}{4}\Sigma^{-1}\zeta^-$. Moving forward, for all $\bar{\varepsilon}\in(0,\bar{\varepsilon}_c)$ the solution is the unique touching point of the circle and the ellipse and it is obtained by equating the "equilibrium forces". However, based on Lemma 1, we may proceed derivative-free (see Appendix A).

Proposition 1 (The risk-constrained mmse estimator)**.** *Under the Assumption 1, the optimal solution to (7) is*

$$\widehat{\mathrm{X}}^\star_{\lambda^\star}=\Big(I+4\lambda^*\Sigma\Big)^{-1}\Bigg(\mathbb{E}X+2\lambda^*\Big(\mathbb{E}[\|X\|_2^2X]-\mathbb{E}\|X\|_2^2\mathbb{E}X\Big)\Bigg), \tag{12}$$

where $\lambda^*\geq 0$ *is the solution to the corresponding dual problem.*

Because by construction, (7) aims to reduce the estimation error variability, it is expected to do so by shifting the optimal estimates towards the areas suffering high loss, that is, towards the tail of the underlying distribution. The next result characterizes the bias of (12): The risk constrained estimator is a shifted version of the conditional expectation by a fraction of the Fisher's moment coefficient of skewness plus the cross third-order statistics of X.

[2]By $\|\cdot\|_Q$ we mean the quadratic form $x^\top Qx$ w.r.t. the positive semi-definite matrix $Q\in\mathbb{R}^{n\times n}$.

Corollary 1 (Risk-constrained mmse estimator and Fisher's skewness)**.** *The estimator (12) operates according to*

$$\widehat{\mathrm{X}}^{\star}_{i,u,\lambda} = \mathbb{E}X_{i,u} + \frac{\lambda\sigma_i\sqrt{\sigma_i}}{1+2\lambda\sigma_i}\mathbb{E}\left(\frac{X_{i,u}-\mathbb{E}X_{i,u}}{\sqrt{\sigma_i}}\right)^3 + \mathrm{T}_{i,\lambda}, \tag{13}$$

where $X_{i,u}$, and $\widehat{\mathrm{X}}^{\star}_{i,u,\lambda}$ denotes the ith component of $X_u \equiv U^\top X$, and $U^\top \widehat{\mathrm{X}}^{*}_{\lambda}$, respectively, U the orthonormal matrix with columns the eigen-vectors of Σ_X, and $\mathrm{T}_{i,\lambda} = \frac{\lambda}{1+2\lambda\sigma_i}\sum_{k\neq i}\mathbb{E}X^2_{k,u}\mathbb{E}X_{i,u} - \mathbb{E}[X^2_{k,u}X_{i,u}]$.

In an effort to minimize the squared error on average while maintaining its variance small, (12) attributes the estimation error made by the conditional expectation to the skewness of the conditional distribution along the eigen-directions of the conditional expected error. In particular, in the one-dimensional case, cross third-order statistics vanish and therefore,

$$\widehat{\mathrm{X}}^{\star}_{\lambda} = \mathbb{E}X + \frac{\lambda\sigma\sqrt{\sigma}}{1+2\lambda\sigma}\mathbb{E}\left(\frac{X-\mathbb{E}X}{\sqrt{\sigma}}\right)^3, \ \lambda \in [0,+\infty). \tag{14}$$

Corollary 1 reveals the "internal mechanism" with which (12) biases towards the tail of the underline distribution. Per error eigen-direction, (12) compensates for the high losses relative to the expectation, by a fraction of the Fisher's skewness coefficient and a term T_i, that captures the cross third-order statistics of the state components. Although (13) involves the cross third-order statistics of the projected state, these statistics cannot emerge artificially from a coordinate change such as $U^\top X$, but are intrinsic to the state itself. Besides, a meaningfull measure of risk should always be coordinate change invariant.

Motivated by the geometric picture, and the Lagrangean of (7), we note that there exists a unique random element such that (12) is viewed as its corresponding average. Fortunately, such an element can be expressed as a Radon-Nikodym derivative w.r.t. μ.

Lemma 2 (Risk constrained estimator as a weighted average)**.** *The primal optimal solution of (7) can be written as*

$$\widehat{\mathrm{X}}^{\star}_{\lambda} = \mathbb{E}_{\xi(\widehat{\mathrm{X}}^{\star}_{\lambda},\lambda)}X, \tag{15}$$

where

$$\frac{\mathrm{d}\xi}{\mathrm{d}\mu}(X) \equiv 1 + 2\lambda\Big(\|X-\widehat{\mathrm{X}}^{\star}_{\lambda}\|_2^2 - \mathbb{E}\|X-\widehat{\mathrm{X}}^{\star}_{\lambda}\|_2^2\Big). \tag{16}$$

In particular, for $\lambda \leq 1/\gamma^*$, $\gamma^* = 2\Big(\mathbb{E}\|\widehat{\mathrm{X}}^{\star}_{\infty} - \widehat{\mathrm{X}}^{\star}_{0}\|_2^2 + \mathrm{mmse}(\mu)\Big)$, it follows that $\frac{\mathrm{d}\xi}{\mathrm{d}\mu} \geq 0$.

Combining Lemma 2 and (14), we obtain [3]

$$\mathbb{E}_{\xi_\lambda}X = \mathbb{E}_{\mu}X + \frac{\lambda\sigma\sqrt{\sigma}}{1+2\lambda\sigma}\mathbb{E}_{\mu}\left(\frac{X-\mathbb{E}_{\mu}X}{\sqrt{\sigma}}\right)^3, \ \lambda \leq 1/\gamma^* \tag{17}$$

[3] Where we explicitly refer to the measure w.r.t. which expectations are taken.

In words, averaging on the re-distribution ξ is the same as biasing the expectation w.r.t. the initial measure μ towards the tail of μ based on a fraction of the Fisher's moment coefficient of skewness. This fact can be considered as a strong indication of another manifestation of duality between risk-averse and distributionally robust optimization. In addition, for the one dimensional case, assuming that μ is positively skewed, we obtain the bound

$$\mathbb{E}_{\xi(\gamma^*)}X - \mathbb{E}_\mu X \leq \frac{\sigma\sqrt{\sigma}}{2\left(\frac{\sigma}{2}\mathbb{E}\left|\frac{X-\mathbb{E}X}{\sqrt{\sigma}}\right|^3 + \mathrm{mmse}(\mu)\right) + 2\sigma}\mathbb{E}\left(\frac{X-\mathbb{E}X}{\sqrt{\sigma}}\right)^3. \tag{18}$$

That is, the expectations of neighboring populations that can be tracked with re-weighting are determined by a fraction of the skewness of the population at hand. In the case of empirical measures, (18) implies that lack of mass in the area of the tail is seen as the presence of risky events.

2.1 The Fisher's coefficient of skewness and the $2-$Wasserstein gradient of the χ^2-divergence

According to the previous discussion, we observe that the shift of the expectation in (13), and (17) attributes directionality to the mass re-distribution of Lemma 2. In this subsection we make this observation more formal through the differentiable structure of the $2-$Wasserstein space $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ of probability measures absolutely continuous w.r.t. Lebesgue measure with finite second moment, equipped with the $2-$Wasserstein metric W_2. In particular, the dynamic formulation of the Wasserstein distance [Benamou and Brenier, 2000] entails the definition of a (pseudo) Riemannian structure [Chewi, 2023]. At every point $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$, the tangent space $T_\mu\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) := \left\{\dot{\mu}:\mathbb{R}^d \to \mathbb{R}\middle| \int \dot{\mu} = 0\right\}$ contains the re-allocation directions and it is parameterized by convex functions $\omega(x)$ on $\mathbb{R}^d$, through the elliptic operator $\omega \mapsto -\nabla\cdot(\mu\nabla\omega)$ [Otto, 2001]. The metric $g : \mathrm{T}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \times \mathrm{T}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to C^\infty(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d))$ [4] has value at the point $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ [Otto, 2001]:

$$g(\dot{\mu}_1, \dot{\mu}_2)(\mu) := \int_{\mathbb{R}^d}\langle\nabla\omega_1(x), \nabla\omega_2(x)\rangle\mu(x)dx = \mathbb{E}_\mu\langle\nabla\omega_1(x), \nabla\omega_2(x)\rangle.$$

Moving forward, let $f \in C^\infty(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d))$. Then, the gradient w.r.t. the defined metric, $\mathrm{grad}f : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathrm{T}\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is the unique vector field satisfying

$$\mathrm{d}f(X)(\mu) = g(\mathrm{grad}f, X)(\mu),$$

where $\mathrm{d}f : \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \to \mathrm{T}^*\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is the differential of f. To derive the gradient $(\mathrm{grad}f)(\mu)$, consider the curves $\mu_t : [0,1] \to \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ progressing through μ and realize the tangent vector $\dot{\mu}_t = \nabla\cdot(\mu_t\nabla\omega(x))$ for some $\omega(x)$. It is

$$\begin{aligned}\mathrm{d}f(\dot{\mu})(\mu) &= \lim_{t\to 0}\frac{f(\mu + t\dot{\mu}_t) - f(\mu)}{t} = \int\frac{\delta f}{\delta\mu}(x)\dot{\mu}_t dx = -\int\frac{\delta f}{\delta\mu}(x)\nabla\cdot(\mu\nabla\omega(x))dx\\ &= \int\langle\nabla\frac{\delta f}{\delta\mu}(x), \nabla\omega(x)\rangle\mu dx.\end{aligned}$$

[4] By $C^\infty(\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d))$ we denote the class of smooth real-valued maps defined on $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$).

We identify $(\mathrm{grad} f)(\mu) \equiv \nabla \frac{\delta f}{\delta \mu}(x)$, or equivalently $(\mathrm{grad} f)(\mu) = -\nabla \cdot (\mu \nabla \frac{\delta f}{\delta \mu}(x))$, where $\frac{\delta f}{\delta \mu}$ denotes the first variation of f w.r.t. μ. For the particular case of the χ^2-divergence from ξ to μ, we obtain $\mathrm{grad}\chi^2(\cdot\|\mu)(\xi) = 2\nabla \frac{\mathrm{d}\xi}{\mathrm{d}\mu}$, which when evaluated to (16) in the one dimensional case (for simplicity), yields $\mathrm{grad}\chi^2(\cdot\|\mu)(\xi_\lambda) = 4\lambda(X - \widehat{\mathrm{X}}^*_\lambda)$. By taking expectations

$$-\mathbb{E}_\mu[\mathrm{grad}\chi^2(\cdot\|\mu)(\xi_\lambda)] = \frac{\lambda^2 \sigma \sqrt{\sigma}}{1+\lambda\sigma} \mathbb{E}\left(\frac{X - \mathbb{E}X}{\sqrt{\sigma}}\right)^3, \lambda \leq 1/2\gamma^*. \tag{19}$$

Subsequently, by combining (14) and (19) we obtain

$$\mathbb{E}_{\xi_\lambda} X - \mathbb{E}_\mu X = -\frac{\mathbb{E}_\mu \mathrm{grad}\chi^2(\cdot\|\mu)(\xi_\lambda)}{\lambda}, \ \lambda \in (0, 1/2\gamma^*).$$

The change of measure is performed in the direction of the negative 2$-$Wasserstein gradient of the χ^2-divergence. For a more comprehensive review in the theory of optimal transport and applications we refer the reader to [Ambrosio et al., 2005, Villani et al., 2009, Wibisono, 2018, Figalli and Glaudo, 2021, Chewi, 2023].

3 Distributionally robust mean square error estimation

In this section we show that (15) (and therefore the corresponding solution of (7)) is the solution to a χ^2- DRO problem for all χ^2-radii less than a computable constant. We start, by taking into account the equivalence between (7) and the mean-deviation risk measure of order $p = 2$ (see e.g. [Shapiro et al., 2021]). To not overload the notation, let us declare $\|X - \widehat{\mathrm{X}}\|_2^2 = Z_{\widehat{\mathrm{X}}}$. Then, (7) reads

$$\begin{aligned} \min_{\widehat{\mathrm{X}} \in \mathbb{R}^d} \quad & \mathbb{E}_\mu \, Z_{\widehat{\mathrm{X}}} \\ \text{s.t.} \quad & \mathrm{var}_\mu Z_{\widehat{\mathrm{X}}} \leq \varepsilon, \end{aligned} \tag{20}$$

or equivalently

$$\begin{aligned} \min_{\widehat{\mathrm{X}} \in \mathbb{R}^d} \quad & \mathbb{E}_\mu \, Z_{\widehat{X}} \\ \text{s.t.} \quad & \sqrt{\mathrm{var}_\mu Z_{\widehat{\mathrm{X}}}} \leq \sqrt{\varepsilon}. \end{aligned} \tag{21}$$

It is a standard exercise to verify that the constraint in (21) is convex w.r.t. $\widehat{\mathrm{X}}$ which leads us to the following result.

Lemma 3 (Relation between the Lagrange multipliers of (20), and (21))**.** *Let λ^*_{mv} and λ^*_{md} be the Lagrange multipliers associated with the constraint of (20), and (21) respectively. Then, for all $\varepsilon \geq \varepsilon_{\min}$ it holds*

$$\lambda^*_{md}(\varepsilon) = 2\lambda^*_{mv}(\varepsilon)\sqrt{\varepsilon}. \tag{22}$$

Moving forward, the problem is equivalent to

$$-\lambda^*_{md}(\varepsilon)\sqrt{\varepsilon} + \min_{\widehat{X}\in\mathbb{R}^d} \mathbb{E}_\mu Z_X + \lambda^*_{md}(\varepsilon)\sqrt{\operatorname{var}_\mu Z_{\widehat{\mathrm{X}}}},$$

or to

$$-\lambda^*_{md}(\varepsilon)\sqrt{\varepsilon} + \min_{\widehat{X}\in\mathbb{R}^d} \rho(Z_{\widehat{\mathrm{X}}}),$$

where

$$\rho(Z_{\widehat{\mathrm{X}}}) \equiv \mathbb{E}_\mu Z_{\widehat{\mathrm{X}}} + \lambda^*_{md}(\varepsilon)\sqrt{\operatorname{var}_\mu Z_{\widehat{X}}}, \tag{23}$$

is the mean-deviation risk measure of order $p = 2$ w.r.t. μ (see e.g. [Shapiro et al., 2021]), with sensitivity constant, the optimal Lagrange multiplier $\lambda^*(\varepsilon)$.

3.1 Dual representation of the mean-deviation risk measure of order $p = 2$

Equation (23) expresses the mean-deviation risk measure in its primal form, where the sensitivity constant is identified with the Lagrange multiplier associated with the constraint in (7). Aiming to study robustness in distributional shifts, the place to start is with the dual representation of (23). Following [Shapiro et al., 2021], the variance in (23), can be expressed via the norm in $\mathcal{L}_2$

$$\operatorname{var}_\mu Z_{\widehat{\mathrm{X}}} = (\mathbb{E}_\mu |Z_{\widehat{\mathrm{X}}} - \mathbb{E}_\mu Z_{\widehat{\mathrm{X}}}|^2)^{1/2} = \|Z_{\widehat{\mathrm{X}}} - \mathbb{E}_\mu Z_{\widehat{\mathrm{X}}}\|_{\mathcal{L}_2},$$

the latter being identified with its dual norm $\|Z\|_{\mathcal{L}_2^*} = \sup_{\|\xi\|_{\mathcal{L}_2}\leq 1}\langle \xi, Z\rangle$, where $||\xi||_{\mathcal{L}_2} \equiv ||\xi||_2 = \sqrt{\mathbb{E}_\mu \xi^2}$, and $\langle \xi, Z\rangle = \mathbb{E}_\mu \xi Z$. Thus, we may write

$$\operatorname{var}_\mu Z_{\widehat{\mathrm{X}}} = \sup_{||\xi||_2\leq 1} \langle \xi - \mathbb{E}_\mu \xi, Z_{\widehat{\mathrm{X}}}\rangle,$$

and subsequently (23) as:

$$\begin{aligned}
\rho(Z_{\widehat{\mathrm{X}}}) &= \langle 1, Z_{\widehat{\mathrm{X}}}\rangle + \lambda^*_{md}(\varepsilon) \sup_{||\xi||_2\leq 1} \langle \xi - \mathbb{E}_\mu \xi, Z_{\widehat{\mathrm{X}}}\rangle \\
&= \sup_{||\xi||_2\leq 1} \langle Z_{\widehat{\mathrm{X}}}, 1\rangle + \lambda^*_{md}(\varepsilon)\langle \xi - \mathbb{E}_\mu \xi, Z_{\widehat{\mathrm{X}}}\rangle \\
&= \sup_{||\xi||_2\leq 1} \langle 1 + \lambda^*_{md}(\varepsilon)(\xi - \mathbb{E}_\mu \xi), Z_{\widehat{\mathrm{X}}}\rangle \\
&= \sup_{\xi'\in\Xi} \langle \xi', Z_{\widehat{\mathrm{X}}}\rangle, \;\; \Xi \equiv \left\{\xi' \in \mathcal{L}_2^* : \xi' = 1 + \lambda^*_{md}(\varepsilon)(\xi - \mathbb{E}_\mu \xi), ||\xi||_2 \leq 1\right\}.
\end{aligned} \tag{24}$$

From (24), $\mathbb{E}_\mu \xi' = 1$, while $\mathbb{E}_\mu \xi^2 \leq 1$ implies $\mathbb{E}_\mu(\xi' - 1)^2 \leq \lambda^{2*}_{md}(\varepsilon)$. Therefore, $\Xi \subseteq \mathrm{U}(\lambda^*_{md})$, where

$$\mathrm{U}(\lambda^*_{md}) \equiv \left\{\xi \in \mathcal{L}_2^* \;:\; \mathbb{E}_\mu \xi = 1, \; \mathbb{E}_\mu(\xi - 1)^2 \leq \lambda^{2*}_{md}(\varepsilon)\right\}. \tag{25}$$

In addition, let $\xi' = \xi - 1 \Leftrightarrow \xi = 1 + \xi' + \mathbb{E}\xi'$. Then $\lambda_{md}^{2*}(\varepsilon) \geq \mathbb{E}(\xi - 1)^2 = \mathbb{E}\xi'^2$, which implies that $\Xi = \mathrm{U}(\lambda_{md}^*)$. Thus, we may write

$$\rho(Z_{\widehat{\mathrm{X}}}) = \sup_{\xi \in \mathrm{U}(\lambda_{md}^*)} \langle \xi, Z_{\widehat{\mathrm{X}}} \rangle. \tag{26}$$

Based on the previous discussion, (7) is equivalent to

$$\min_{\widehat{\mathrm{X}} \in \mathbb{R}^d} \sup_{\xi \in \mathrm{U}(\lambda_{md}^*)} \langle \xi, Z_{\widehat{\mathrm{X}}} \rangle. \tag{27}$$

The equality constraint in (25) can be absorbed and the constraint set may be re-expressed as

$$\mathrm{U}(\lambda_{md}^*) \equiv \left\{ \nu \in \mathcal{L}_2^* \ : \ \mathbb{E}_\mu (\mathrm{d}\nu / \mathrm{d}\mu - 1)^2 \leq \lambda_{md}^{2*}(\varepsilon) \right\},$$

and subsequently (27) as

$$\min_{\widehat{\mathrm{X}} \in \mathbb{R}^d} \left\{ \sup_{\nu \in \mathcal{L}_2^*} \mathbb{E}_\nu Z_{\widehat{\mathrm{X}}} \ : \ \chi^2(\nu || \mu) \leq \lambda_{md}^{2*}(\varepsilon) \right\}, \tag{28}$$

where $\chi^2(\nu||\mu)$ is the chi-square divergence between the probability measure μ, and the signed measure $\nu \in \mathcal{L}_2^*$. The inner optimization problem is of a linear objective subject to a quadratic constraint over the infinite dimensional (dual) vector space $\mathcal{L}_2^*$, and can be easily handled by variational calculus. In particular, by introducing an additional Lagrange multiplier, for dualizing the quadratic constraint, it is easy to show that the supremum is attained at

$$\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu} \right)^* = \lambda_{md}^*(\varepsilon) \frac{Z_{\widehat{\mathrm{X}}} - \mathbb{E}_\mu Z_{\widehat{\mathrm{X}}}}{\sqrt{\mathrm{var}_\mu Z_{\widehat{\mathrm{X}}}}} + 1. \tag{29}$$

Please note that the objective of the minimization problem in (28) is not convex since the expectation is taken w.r.t. signed measures. At this point, the reader might find it instructive comparing (29) to (16) through (22), and ask the question: Is there a subset of admissible Lagrange multipliers for which (28) (and therefore (7)) can be formulated as a distributionally (over probability measures) robust optimization problem? We answer this question with the following result.

Theorem 1 (Distributionally robust mean square error esimator)**.** *Let* $\Lambda \equiv \left\{ \lambda \geq 0 \ : \ \lambda \leq 1/\gamma^* \right\}$*, be a subset of admissible Lagrange multipliers associated with the constraint of (7), with* γ^* *as in Lemma 2. Then, for the corresponding levels of risk, problem (1) is the same as the* χ^2-*DRO problem*

$$\min_{\widehat{\mathrm{X}} \in \mathbb{R}^d} \left\{ \sup_{0 \leq \nu \in \mathcal{L}_2^*} \mathbb{E}_{\sim\nu} \|X - \widehat{\mathrm{X}}\|_2^2 \ : \ \chi^2(\nu||\mu) \leq \lambda \right\}, \ \lambda \in \Lambda. \tag{30}$$

Based on Theorem 1, for the corresponding levels of risk to Λ, the solution to the distributionally robust problem (30) is the optimal solution to (7), $\widehat{\mathrm{X}}_\lambda^\star$. The solution to the χ^2-squared mmse DRO problem, biases -up to a change of co-ordinates- towards the areas of high loss, with a fraction of the Fisher's moment coefficient of skewness plus some additional cross third-order statistics.

4 Mean squared error of the empirical distributionally robust mse estimator

Now that the risk-constrained estimator has all the desired features, we want to study its performance w.r.t. the mean squared error. We do so in two steps. First, we consider a simplified version of the empirical distributionally robust estimator and we show that it outperforms the empirical average over a rather large class of models. Second, we argue that with high probability this estimator is the empirical version of the distributionally robust estimator (14). The next results refer to the one dimensional case ($d = 1$), with the extension to other dimensions left open for future exploration.

By considering (14) w.r.t. μ_{n}, we define the following algorithm

$$\widehat{\mathrm{X}}^{\star}_{n,\bar{\lambda}}(D) \equiv \frac{1}{n}\sum_i X_i + \bar{\lambda}\frac{1}{n}\sum_i \left(X_i - n^{-1}\sum_i X_i\right)^3, \tag{31}$$

where $\bar{\lambda}$ is a non-negative and deterministic parameter that replaces the first factor in the second term of (14). For ease of notation, let us first denote $\bar{\mathrm{X}}_n(D) = n^{-1}\sum_i X_i$, $\mathrm{T}_n(D) = n^{-1}\sum_i (X_i - \bar{\mathrm{X}}_n)^3$. We call (31) the simplified empirical distributionally robust estimator, which after the above notation reads $\widehat{\mathrm{X}}^{\star}_{n,\bar{\lambda}} = \bar{\mathrm{X}}_n + \bar{\lambda}\mathrm{T}_n$, and achieves mean squared error

$$\begin{aligned} \mathbb{E}(\widehat{\mathrm{X}}^{\star}_{n,\bar{\lambda}} - X)^2 &= \mathbb{E}(\bar{\mathrm{X}}_n - X)^2 + \bar{\lambda}^2\mathbb{E}\mathrm{T}_n^2 - 2\bar{\lambda}\mathbb{E}\Big[(X - \bar{\mathrm{X}}_n)\mathrm{T}_n\Big] \\ &\le \mathbb{E}(\bar{\mathrm{X}}_n - X)^2 + \bar{\lambda}^2(\mathbb{E}\mathrm{T}_n^2 + \epsilon) - 2\bar{\lambda}\mathbb{E}\Big[(X - \bar{\mathrm{X}}_n)\mathrm{T}_n\Big], \end{aligned} \tag{32}$$

for any $\epsilon > 0$. Strict domination follows for all those positive $\bar{\lambda}$ such that

$$\bar{\lambda} < 2\mathbb{E}\Big[(X - \bar{\mathrm{X}}_n)\mathrm{T}_n\Big]/(\mathbb{E}\mathrm{T}_n^2 + \epsilon), \tag{33}$$

only under reasonable conditions that render the right-hand side of (33) strictly positive. Let Q be the set of all probability measures with finite second order moment. On top of that, let $N = \{\mu \in Q \mid\ 3m_2^2 > m_4\}$ be those measures in Q, for which the triple squared variance m_2^2 exceeds the central fourth order moment m_4. Then, strict domination of (31) is demonstrated by the following

Theorem 2 (Strict domination of the simplified empirical DRMSEE)**.** *Let $\widehat{\mathrm{X}}^{\star}_{n,\bar{\lambda}}$ be as in (31), and $\bar{\mathrm{X}}_n$ be the empirical average. Fix an $\epsilon > 0$. Then, for probability measures $\mu \in N$, and for any $n \ge 3$,*

$$\mathbb{E}|\widehat{\mathrm{X}}^{\star}_{n,\bar{\lambda}} - X|^2 < \mathbb{E}|\bar{\mathrm{X}}_n - X|^2, \tag{34}$$

for all $\bar{\lambda} < \frac{1}{n}\left[\frac{8(3m_2^2 - m_4)}{\mathbb{E}\mathrm{T}_n^2 + \epsilon}\right]$.

Gaussian measures are among those satisfying $3m_2^2 = m_4$. In this case $\bar{\lambda} = 0$, the condition (32) is satisfied with equality for any $n \ge 1$, and the mse performance of (31) reduces to that of the empirical average. On the contrary, when the data are generated from models $\mu \in N$, (31) outperforms the average as soon as the parameter $\bar{\lambda}$ is chosen inside the prescribed interval. Domination remains feasible for arbitrarily large data sets as soon as λ is sent to zero with a rate of at least $O(1/n)$. Interestingly, although (31) reduces to the empirical average for datasets with zero (empirical) skewness, the defining property of the set N does not involve the model's skewness but the projection $\langle X - \bar{\mathrm{X}}_n, \mathrm{T}_n \rangle = \mathbb{E}\Big[(X - \bar{\mathrm{X}}_n)\mathrm{T}_n\Big]$ which turns out to involve only the second and fourth central moments via the difference $3m_2^2 - m_4$.

Moving forward, for the case where the random variable is supported over a compact subset of $\mathbb{R}^d$, we can bound the polynomial terms of γ^* in (2), and also the variance in the denominator of the first factor of (14). This assumption is not absolutely necessary and it can be avoided by showing that with high probability (31) is a DRO estimator. Since Theorem 2 allows $n \geq 3$, this can be done by controlling the empirical variance in the factor $\lambda/(1 + 2\lambda\sigma)$ of (14) as well as, the polynomial terms in the expectation (w.r.t. μ_n) related to the threshold γ^*. Then, for a fixed model, as $\lambda \xrightarrow{n\to\infty} 0$ one has to trade performance for a distributional robustness, the latter being favored by the logarithmic dependency in the number of samples.

4.1 Numerical example

We put the estimator into work, by considering a Gaussian mixture model $\mu_1 = 3, \mu_2 = 0.1, \sigma_1 = 0.5, \sigma_2 = 1.7$, and mixing proportions $\pi_1 = 0.3, \pi_2 = 1 - \pi_1$. The condition where the tripled square of the variance is greater than the fourth central moment is satisfied with $3m_2^2 - m_4 = 11.4622$. Specifically, the variance of the mixture is calculated to be $m_2 = 3.826$, and the fourth central moment is approximately $m_4 = 32.46$. We draw $n = 10$ i.i.d. samples. Then $\widehat{\mathrm{X}}^\star_{10,\bar{\lambda}} = 1.963 - \bar{\lambda}3.61$, and the mse reads $\mathrm{mse}(\bar{\lambda}) = \mathbb{E}_\mu(1.963 - \bar{\lambda}3.61 - X)^2$.

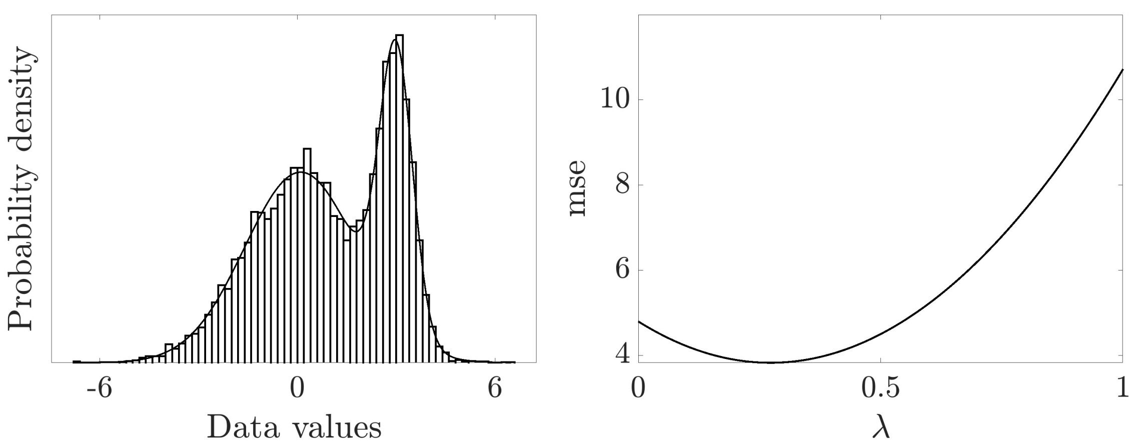

Figure 1: Gaussian mixture model (left), mean squared error (right)

5 Conclusion and future work

In this work we deployed the main insights of DRO to come up with a distributionally robust empirical representation of expectations that dominates the empirical average w.r.t. the mean squared error for a particular class of data generating distributions. Our methodology relied on two design specifications. We confined our design space into DRO problems where the solution is a saddle point, and also we took advantage of the duality between risk-averse and DRO formulations to ask for a closed form solution. Both specifications combined allow for an *implementable* representation of expectation based on data drawn from the unknown measure. In future work, we would also like to study the multidimensional case of Theorem (2).

References

Marc Abeille, Alessandro Lazaric, Xavier Brokmann, et al. Lqg for portfolio optimization. *arXiv preprint arXiv:1611.00997*, 2016.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *The operations research revolution*, pages 1–19. INFORMS, 2015.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1): 35–53, 2004.

Jose Blanchet and Alexander Shapiro. Statistical limit theorems in distributionally robust optimization. In *2023 Winter Simulation Conference (WSC)*, pages 31–45. IEEE, 2023.

Jose Blanchet, Karthyek Murthy, and Viet Anh Nguyen. Statistical analysis of wasserstein distributionally robust estimators. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*, pages 227–254. INFORMS, 2021.

Jose Blanchet, Daniel Kuhn, Jiajin Li, and Bahar Taskesen. Unifying distributionally robust optimization via optimal transport theory. *arXiv preprint arXiv:2308.05414*, 2023.

Jose Blanchet, Jiajin Li, Sirui Lin, and Xuhui Zhang. Distributionally robust optimization and robust statistics. *arXiv preprint arXiv:2401.14655*, 2024.

Sinho Chewi. *An optimization perspective on log-concave sampling and beyond*. PhD thesis, Massachusetts Institute of Technology, 2023.

Luc Devroye, Dominik Schäfer, László Györfi, and Harro Walk. The estimation problem of minimum mean squared error. *Statistics & Decisions*, 21(1):15–28, 2003.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

Alessio Figalli and Federico Glaudo. *An invitation to optimal transport, Wasserstein distances, and gradient flows*. 2021.

Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

MA Girshick and LJ Savage. Bayes and minimax estimates for quadratic loss functions. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 53–74. University of California Press, 1951.

Jun-ya Gotoh, Michael Jong Kim, and Andrew EB Lim. Robust empirical optimization is almost the same as mean–variance optimization. *Operations research letters*, 46(4): 448–452, 2018.

William James and Charles Stein. Estimation with quadratic loss. In *Breakthroughs in statistics: Foundations and basic theory*, pages 443–460. Springer, 1961.

Dionysios S Kalogerias, Luiz FO Chamon, George J Pappas, and Alejandro Ribeiro. Risk-aware mmse estimation. *arXiv preprint arXiv:1912.02933*, 2019.

Dionysios S Kalogerias, Luiz FO Chamon, George J Pappas, and Alejandro Ribeiro. Better safe than sorry: Risk-aware nonlinear bayesian estimation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5480–5484. IEEE, 2020.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh 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.

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

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 31, 2018.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. ISSN 00221082, 15406261. URL `http://www.jstor.org/stable/2975974`.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

Viet Anh Nguyen, Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Bridging bayesian and minimax mean square error estimation via wasserstein distributionally robust optimization. *Mathematics of Operations Research*, 48(1):1–37, 2023.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

Herbert E Scarf, KJ Arrow, and S Karlin. *A min-max solution of an inventory problem.* Rand Corporation Santa Monica, 1957.

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

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Bart PG Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402, 2021.

Michel Verhaegen and Vincent Verdult. *Filtering and system identification: a least squares approach.* Cambridge university press, 2007.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

A Appendix A (Proofs of Section 2)

Proof of Lemma 1. The square of the constraint of (7) is written as[5]:

$$\begin{aligned}\Big(\|X-\widehat{\mathrm{X}}\|_2^2-\mathbb{E}||X-\widehat{\mathrm{X}}||_2^2\Big)^2 &= (\|X\|_2^2+\mathbb{E}\|X\|_2^2)^2\\ &+4\widehat{\mathrm{X}}^\top(X-\mathbb{E}X)(X-\mathbb{E}X)^\top\widehat{\mathrm{X}}\\ &+4(||X||_2^2-\mathbb{E}\{||X||_2^2\})(\mathbb{E}X-X)^\top\widehat{\mathrm{X}}\\ &=\begin{bmatrix}1 & \widehat{\mathrm{X}}^\top\end{bmatrix}\begin{bmatrix}\alpha(X) & \zeta(X)^\top\\ \zeta(X) & \gamma(X)\end{bmatrix}\begin{bmatrix}1\\ \widehat{\mathrm{X}}\end{bmatrix},\end{aligned} \tag{35}$$

where

$$\begin{aligned}\alpha(X)&=(||X||_2^2-\mathbb{E}\|X\|_2^2)^2\\ \zeta(X)&=2(\|X\|_2^2-\mathbb{E}\{\|X\|_2^2\})(\mathbb{E}X-X)\\ \gamma(X)&=4(X-\mathbb{E}X)(X-\mathbb{E}X)^\top.\end{aligned}$$

Thus, by taking expectations

$$\mathbb{E}\Big(||X-\widehat{\mathrm{X}}||_2^2-\mathbb{E}_\mu||X-\widehat{\mathrm{X}}||_2^2\Big)^2=\begin{bmatrix}1 & \widehat{\mathrm{X}}^\top\end{bmatrix}\begin{bmatrix}\alpha & \zeta^\top\\ \zeta & \Gamma\end{bmatrix}\begin{bmatrix}1\\ \widehat{\mathrm{X}}\end{bmatrix}, \tag{36}$$

where $\alpha=\mathbb{E}\alpha(X)$, $\zeta=\mathbb{E}\zeta(X)$, and $\Gamma=4\Sigma$. Equation (36) rests on the permutation-invariance property of the trace. Without loss of generality, let us assume that the covariance matrix Σ is strictly positive, and consider its Schur complement w.r.t. Γ thus obtaining the following standard factorization (see e.g. [Verhaegen and Verdult, 2007, Lemma 2.3]):

$$\mathbb{E}\Big(\|X-\widehat{\mathrm{X}}\|_2^2-\mathbb{E}\|X-\widehat{\mathrm{X}}\|_2^2\Big)^2=\begin{bmatrix}1\\ \Gamma^{-1}\zeta-\widehat{\mathrm{X}}\end{bmatrix}^\top\begin{bmatrix}\alpha-\zeta^\top\Gamma^{-1}\zeta & 0\\ 0 & \Gamma\end{bmatrix}\begin{bmatrix}1\\ \Gamma^{-1}\zeta-\widehat{\mathrm{X}}\end{bmatrix}. \tag{37}$$

□

Proof of Proposition 1. From Lemma 1, problem (7) can equivalently be written as:

$$\begin{array}{cl}\min\limits_{\widehat{\mathrm{X}}\in\mathbb{R}^d} & \|\widehat{\mathrm{X}}-\mathbb{E}X\|_2^2\\ \text{s.t.} & \|\widehat{\mathrm{X}}-\frac{1}{4}\Sigma^{-1}\zeta^-\|_{4\Sigma}^2\le\bar{\varepsilon}\end{array},\ \bar{\varepsilon}\ge 0\ =\ \min_{\widehat{\mathrm{X}}\in\mathbb{R}^d}\sup_{\lambda\ge 0}\mathrm{L}(\widehat{\mathrm{X}},\lambda),$$

where $\mathrm{L}:\mathbb{R}^d\times\mathbb{R}^+\to\mathbb{R}$ is the Lagrangean of (11) given by

$$\mathrm{L}(\widehat{\mathrm{X}},\lambda)=\|\widehat{\mathrm{X}}-\mathbb{E}X\|_2^2+\lambda\ \|\widehat{\mathrm{X}}-\frac{1}{4}\Sigma^{-1}\zeta^-\|_{4\Sigma}^2-\lambda\bar{\varepsilon}, \tag{38}$$

and $\lambda\in\mathbb{R}^+$ is the multiplier associated with the constraint of (11). Slater's condition is directly verified from (11) and therefore, due to convexity, the dual-optimal value is attained, and (11) has zero duality gap:

$$\min_{\widehat{\mathrm{X}}\in\mathbb{R}^d}\sup_{\lambda\ge 0}\mathrm{L}(\widehat{\mathrm{X}},\lambda)=\sup_{\lambda\ge 0}\min_{\widehat{\mathrm{X}}\in\mathbb{R}^d}\mathrm{L}(\widehat{\mathrm{X}},\lambda).$$

[5] All expectations are taken w.r.t. μ.

The Lagrangian reads

$$\mathrm{L}(\widehat{\mathrm{X}}, \lambda) = \begin{bmatrix} \widehat{\mathrm{X}}^\top & 1 \end{bmatrix} \begin{bmatrix} I + 4\lambda\Sigma & -\Big(\mathbb{E}X + \lambda\zeta^-\Big) \\ -\Big(\mathbb{E}X + \lambda\zeta^-\Big)^\top & \|\mathbb{E}X\|_2^2 + \lambda\|\frac{1}{4}\Sigma^{-1}\zeta^-\|_2^2 \end{bmatrix} \begin{bmatrix} \widehat{\mathrm{X}} \\ 1 \end{bmatrix} - \lambda\bar{\varepsilon},$$

and by performing a similar decomposition as in (37) we obtain

$$\begin{aligned} \mathrm{L}(\widehat{\mathrm{X}}, \lambda) = & \|\widehat{\mathrm{X}} - (I + 4\lambda\Sigma)^{-1}(\mathbb{E}X + \lambda\zeta^-)\|_{I+4\lambda\Sigma}^2 \\ & - \Big(\mathbb{E}X + \lambda\zeta^-\Big)^\top \Big(I + 4\lambda\Sigma\Big)^{-1} \Big(\mathbb{E}X + \lambda\zeta^-\Big) \\ & + \|\mathbb{E}X\|_2^2 + \lambda\|\frac{1}{4}\Sigma^{-1}\zeta^-\|_\Sigma^2 - \lambda\bar{\varepsilon}. \end{aligned} \tag{39}$$

The last two terms in (39) do not depend on $\widehat{\mathrm{X}}$, and thus

$$\widehat{\mathrm{X}}_\lambda^\star = (I + 4\lambda\Sigma)^{-1}(\mathbb{E}X + \lambda\zeta^-), \tag{40}$$

is primal-optimal. □

Proof of Corollary 1. Let us denote $\Pi(\lambda) = (I + 4\lambda\Sigma)^{-1}$. Differentiation of (40) w.r.t. $\lambda \in (0, +\infty)$, gives

$$\mathrm{d}_\lambda \widehat{\mathrm{X}}_\lambda^\star = -4\Pi(\lambda)\Sigma\Pi(\lambda)(\mathbb{E}\{X\} + \lambda\zeta^-) + \Pi(\lambda)\zeta^-, \tag{41}$$

where the commutator $[\Pi, \Sigma] \equiv \Pi\Sigma - \Sigma_{X|X}\Pi = 0$, and therefore we may write

$$\begin{aligned} \mathrm{d}_\lambda \widehat{\mathrm{X}}_\lambda^\star &= \Pi(\lambda)^2\Big(- 4\Sigma(\mathbb{E}X + \lambda\zeta^-) + \Pi(\lambda)^{-1}\zeta^-\Big) \\ &= -4\Pi(\lambda)^2\Sigma\Big(\mathbb{E}X - (4\Sigma)^{-1}\zeta^-\Big). \end{aligned} \tag{42}$$

It is worth noticing that the dependency on λ has transferred in Π, and that $\mathbb{E}X = \widehat{\mathrm{X}}_0^\star$, and $(4\Sigma)^{-1}\zeta^- = \widehat{\mathrm{X}}_\infty^\star$. Thus, we declare $\widehat{\Delta X}^\star_{0\infty} = (4\Sigma)^{-1}\zeta^- - \mathbb{E}X$. By integrating (42) w.r.t. λ we obtain

$$\widehat{\mathrm{X}}_\lambda^\star = \widehat{\mathrm{X}}_0^\star + U\mathrm{H}(\lambda)U^\top \widehat{\Delta X}^\star_{0\infty}. \tag{43}$$

where $\mathrm{H}(\lambda) = \mathrm{diag}(\frac{4\lambda\sigma}{1+4\lambda\sigma})$, and $\Sigma = U\Lambda U^\top$ refers to the spectral decomposition of Σ. That is, the optimal estimates are shifted versions of the conditional expectation by a transformed version of the vector $\widehat{\Delta X}^\star_{0\infty}$. Recall that this vector encodes the difference between the center of the circle and the center of the ellipsoid. From (43) we obtain

$$U^\top \widehat{\mathrm{X}}_\lambda^\star = U^\top \widehat{\mathrm{X}}_0^\star + \mathrm{H}(\lambda)\Big(U^\top \widehat{\mathrm{X}}_0^\star - U^\top \widehat{\mathrm{X}}_\infty^\star\Big) \tag{44}$$

Further,

$$U^\top \widehat{\mathrm{X}}_0^\star = U^\top \mathbb{E}X = \mathbb{E}X_u, \tag{45}$$

where $X_u \equiv U^\top X$. In addition, since

$$U^\top \mathbb{E}[\|X\|_2^2 X] = \mathbb{E}[\|UU^\top X\|_2^2 U^\top X],$$

and $U^\top$ is unitary,

$$U^\top \widehat{\mathrm{X}}_\infty^\star = \frac{1}{2}\Lambda^{-1}\Big(\mathbb{E}[\|X_u\|_2^2 X_u] - \mathbb{E}\|X_u\|_2^2 \mathbb{E}X_u\Big) \tag{46}$$

By $\widehat{\mathrm{X}}_{i,u,\lambda}^\star$, and $X_{i,u}$ we mean the ith component of $\widehat{\mathrm{X}}_{u,\lambda}^\star \equiv U^\top \widehat{\mathrm{X}}_\lambda^\star$, and X_u, respectively. Also, note that since the singular values of the covariance matrix are coordinate-change invariant, $\Lambda = \Lambda_{X_u}$, or $\mathrm{var}[X_i] = \mathrm{var}[X_{i,u}] \equiv \sigma_i$. Thus,

$$\begin{aligned}
\widehat{\mathrm{X}}_{i,u,\lambda}^\star &= \mathbb{E}X_{i,u} + \frac{2\lambda\sigma_i}{1+2\lambda\sigma_i}\Bigg(\mathbb{E}X_{i,u} - \frac{1}{2\sigma_i}\Big(\sum_k \mathbb{E}[X_{k,u}^2 X_{i,u}] - \sum_k \mathbb{E}X_{k,u}^2 \mathbb{E}X_{i,u}\Big)\Bigg)\\
&= \mathbb{E}X_{i,u}\\
&+ \frac{2\lambda\sigma_i}{1+2\lambda\sigma_i}\Bigg(\mathbb{E}X_{i,u} - \frac{1}{2\sigma_i}\Big(\mathbb{E}X_{i,u}^3 - \mathbb{E}X_{i,u}^2 \mathbb{E}X_{i,u}\Big)\Bigg)\\
&+ \frac{2\lambda\sigma_i}{1+2\lambda\sigma_i}\Bigg(-\frac{1}{2\sigma_i}\Big(\sum_{k\neq i} \mathbb{E}[X_{k,u}^2 X_{i,u}] - \sum_{k\neq i} \mathbb{E}X_{k,u}^2 \mathbb{E}X_{i,u}\Big)\Bigg)
\end{aligned} \tag{47}$$

Completing the third power in the second term of (47), and subsequently setting the last term equal to $\mathrm{T}_{i,\lambda}$ concludes the proof. $\square$

Proof of Lemma 2. The optimallity condition for the corresponding to (7) Lagrangean yields:

$$\mathrm{d}_{\widehat{\mathrm{X}}}\mathrm{L}(\widehat{\mathrm{X}}^\star, \lambda) = 0,$$

or

$$2\mathbb{E}(\widehat{\mathrm{X}}_\lambda^\star - X) + \lambda\mathbb{E}\Big\{2\Big(\|X - \widehat{\mathrm{X}}_\lambda^\star\|_2^2 - \mathbb{E}\|X - \widehat{\mathrm{X}}_\lambda^\star\|_2^2\Big)\Big(2(X - \widehat{\mathrm{X}}_\lambda^\star) - 2\mathbb{E}(X - \widehat{\mathrm{X}}_\lambda^\star)\Big)\Big\} = 0,$$

or

$$2\mathbb{E}(\widehat{\mathrm{X}}_\lambda^\star - X) - \lambda\mathbb{E}\Big\{2\Big(\|X - \widehat{\mathrm{X}}_\lambda^\star\|_2^2 - \mathbb{E}\|X - \widehat{\mathrm{X}}_\lambda^\star\|_2^2\Big)\Big(2X - 2\mathbb{E}X\Big)\Big\} = 0,$$

or

$$\mathbb{E}(\widehat{\mathrm{X}}_\lambda^\star - X) - 2\lambda\mathbb{E}\Big\{\Big(\|X - \widehat{\mathrm{X}}_\lambda^\star\|_2^2 - \mathbb{E}\|X - \widehat{\mathrm{X}}_\lambda^\star\|_2^2\Big)X\Big\} = 0,$$

or

$$\widehat{\mathrm{X}}_\lambda^\star = \mathbb{E}\Big\{\Big[1 + 2\lambda\Big(\|X - \widehat{\mathrm{X}}_\lambda^\star\|_2^2 - \mathbb{E}\|X - \widehat{\mathrm{X}}_\lambda^\star\|_2^2\Big)\Big]X\Big\}, \tag{48}$$

The Radon-Nikodym derivative in (48) takes positive as well as negative values. However, $\frac{\mathrm{d}\xi}{\mathrm{d}\mu} \geq 1 - 2\lambda\mathbb{E}||X - \widehat{\mathrm{X}}^\star_\lambda||_2^2$, and therefore, $\lambda \leq 1/2\mathbb{E}||X - \widehat{\mathrm{X}}^\star_\lambda||_2^2$, $\lambda \in [0, +\infty)$ renders the factor in (48) a positive density. This is guaranteed when

$$\lambda \leq \inf_{\lambda\in[0,+\infty)} \frac{1}{2\mathbb{E}\|X - \widehat{\mathrm{X}}^\star_\lambda\|_2^2} = \frac{1}{2\mathbb{E}\|X - \widehat{\mathrm{X}}^\star_\infty\|_2^2} = \frac{1}{2\Big(\mathbb{E}\|\widehat{\Delta X}^\star_{0\infty}\|_2^2 + \mathrm{mmse}(\mu)\Big)} \equiv 1/\gamma^\star,$$

where the first equality follows because the objective increases with $\lambda \geq 0$. $\square$

B Appendix B (Proofs of Sections 3, 4)

Proof of Theorem 1. Convexity of (20), as well as of (21) and therefore Slater's condition imply that the dual optimal is attained. The KKT conditions for (20), and (21) are $\mathrm{d}_{\widehat{\mathrm{X}}}\mathbb{E}Z_{\widehat{\mathrm{X}}^\star} = -\lambda^\star_{mv}(\varepsilon)\nabla\mathrm{var}Z_{\widehat{\mathrm{X}}^\star}$, and $\mathrm{d}_{\widehat{\mathrm{X}}}\mathbb{E}Z_{\widehat{\mathrm{X}}^\star} = -\lambda^\star_{md}(\varepsilon)\nabla\mathrm{var}Z_{\widehat{\mathrm{X}}^\star}(2\sqrt{\mathrm{var}Z_{\widehat{\mathrm{X}}^\star}})^{-1}$, respectively. Complementary slackness condition for (21) yields $\lambda^\star_{md}(\varepsilon) = 2\lambda^\star_{mv}(\varepsilon)\sqrt{\varepsilon}$. $\square$

Proof of Theorem 1. It is true that

$$\min_{\widehat{\mathrm{X}}\in\mathbb{R}^d} \sup_{\nu\in\mathrm{U}(\lambda^\star_{md})} \mathbb{E}_\nu||X - \widehat{\mathrm{X}}||_2^2 = \mathbb{E}_{\nu^\star(\lambda^\star_{md})}\|X - \widehat{\mathrm{X}}_{\lambda^\star_{mv}}\|_2^2 \tag{49}$$

$$= \mathbb{E}_{\nu^\star(\lambda^\star_{md})}\|X - \mathbb{E}_{\frac{d\xi}{d\mu}(\lambda^\star_{mv})}X\|_2^2 \tag{50}$$

$$= \mathbb{E}_{\nu^\star(\lambda^\star_{md})}\|X - \mathbb{E}_{\nu^\star(\lambda^\star_{md})}X\|_2^2 \tag{51}$$

Lemma 2 justifies the second equality, while (51) follows from (50) by matching (29), with (15) for $\lambda = \lambda^\star_{mv}$, and subsequently applying (22). In addition, for every $\lambda^\star_{md} \leq 2\sqrt{\varepsilon}/\gamma^\star$

$$\min_{\widehat{\mathrm{X}}\in\mathbb{R}^d} \sup_{\nu\in\mathrm{U}(\lambda^\star_{md})} \mathbb{E}_\nu\|X - \widehat{\mathrm{X}}\|_2^2 \geq \min_{\widehat{\mathrm{X}}\in\mathbb{R}^d} \sup_{0\leq\nu\in\mathrm{U}(\lambda^\star_{md})} \mathbb{E}_\nu\|X - \widehat{\mathrm{X}}\|_2^2 \tag{52}$$

$$= \sup_{0\leq\nu\in\mathrm{U}(\lambda^\star_{md})} \min_{\widehat{\mathrm{X}}} \mathbb{E}_\nu\|X - \widehat{\mathrm{X}}\|_2^2 \tag{53}$$

$$= \sup_{0\leq\nu\in\mathrm{U}(\lambda^\star_{md})} \mathbb{E}_\nu\|X - \mathbb{E}_\nu X\|_2^2$$

$$\geq \mathbb{E}_{\nu^\star(\lambda^\star_{md})}\|X - \mathbb{E}_{\nu^\star(\lambda^\star_{md})}X\|_2^2. \tag{54}$$

Equality (53) follows from (52) because the conditions of Sion's minimax theorem are satisfied for the concave-convex minimax problem (in the corresponding strong and weak* topologies). Lastly, from Lemma 2 the expectation in (54) is w.r.t. positive measure and all the above inequalities are in fact equalities, and the result follows from (52). $\square$

Proof of Theorem 2. It is $\bar{\mathrm{X}}_n = n^{-1}\sum_i X_i$, $\mathrm{T}_n = n^{-1}\sum_i(X_i - \bar{\mathrm{X}}_n)^3$ and $\widehat{\mathrm{X}}^\star_{n,\bar{\lambda}} = \bar{\mathrm{X}}_n + \bar{\lambda}\mathrm{T}_n$. The condition for strict domination reads

$$\bar{\lambda} < \frac{2\mathbb{E}\Big[(X - \bar{\mathrm{X}}_n)\mathrm{T}_n\Big]}{\mathbb{E}\mathrm{T}_n^2 + \epsilon} = \frac{2n\mathbb{E}\Big[(X - n^{-1}\sum_i X_i)\sum_i(X_i - n^{-1}\sum_i X_i)^3\Big]}{\mathbb{E}\Big[\Big(\sum_i(X_i - n^{-1}\sum X_i)^3\Big)^2\Big] + \epsilon}. \tag{55}$$

For the numerator of (55) we have:

$$\mathbb{E}\Big[(X-\bar{\mathrm{X}}_n)\sum_i (X_i-\bar{\mathrm{X}}_n)^3\Big] = \mathbb{E}\Big[X\sum_i (X_i-\bar{\mathrm{X}}_n)^3\Big] - \mathbb{E}\Big[\bar{\mathrm{X}}_n\sum_i (X_i-\bar{\mathrm{X}}_n)^3\Big] \tag{56}$$

We compute each term of (56). For the first one we have

$$\begin{aligned}\mathbb{E}\Big[X\sum_i (X_i-\bar{\mathrm{X}}_n)^3\Big] &= \mathbb{E}\Big[X\sum_i (X_i^3-3X_i^2\bar{\mathrm{X}}_n+3X_i\bar{X}_n^2-\bar{\mathrm{X}}_n^3)\Big]\\ &= \mathbb{E}X\sum_i^n \Big(\mathbb{E}X_i^3-3\mathbb{E}[X_i^2\bar{\mathrm{X}}_n]+3\mathbb{E}[X_i\bar{X}_n^2]-\mathbb{E}\bar{\mathrm{X}}_n^3\Big).\end{aligned} \tag{57}$$

We compute each term inside the parenthesis in (57): For the first we have $\mathbb{E}X_i^3=\mathbb{E}X^3$, while for the second:

$$\begin{aligned}-3\mathbb{E}[X_i^2\bar{\mathrm{X}}_n] &= -3n^{-1}\mathbb{E}[X_i^2(X_i+\sum_{j=1,j\neq i}^n X_j)]\\ &= -3n^{-1}\mathbb{E}X^3-3n^{-1}(n-1)\mathbb{E}X^2\mathbb{E}X.\end{aligned} \tag{58}$$

For the third term inside the parenthesis of (57) we use

$$\mathbb{E}\Bigg(\sum_{j=1,j\neq i}^n X_j\Bigg)^2 = (n-1)\mathbb{E}X^2+(\mathbb{E}X)^2(n-1)(n-2) \tag{59}$$

to obtain

$$3\mathbb{E}[X_i\bar{\mathrm{X}}_n^2] = 3n^{-2}\mathbb{E}X^3+9n^{-2}(n-1)\mathbb{E}X\mathbb{E}X^2+3n^{-2}(n-1)(n-2)(\mathbb{E}X)^3. \tag{60}$$

Lastly, the fourth term reads:

$$\mathbb{E}\bar{\mathrm{X}}_n^3 = n^{-2}\mathbb{E}X^3+3n^{-2}(n-1)\mathbb{E}X^2\mathbb{E}X+n^{-2}(\mathbb{E}X)^3\Big(n^2-3n+2\Big). \tag{61}$$

Thus, by (58), (60), (61), the parenthesis in (57) reads:

$$\begin{aligned}\Big(\mathbb{E}X_i^3-3\mathbb{E}[X_i^2\bar{\mathrm{X}}_n]+3\mathbb{E}[X_i\bar{\mathrm{X}}_n^2]-\mathbb{E}\bar{\mathrm{X}}_n^3\Big) &= \mathbb{E}X^3\Big[2n^{-2}-3n^{-1}+1\Big]\\ &\quad+\mathbb{E}X^2\mathbb{E}X\Big[3n^{-2}(n-1)\Big]\\ &\quad+(\mathbb{E}X)^3 2n^{-2}\Big(n^2-3n+2\Big).\end{aligned} \tag{62}$$

The summation in (57) increases order by n:

$$\begin{aligned}\sum_{i=1}^n\Big(\mathbb{E}X_i^3-3\mathbb{E}[X_i^2\bar{\mathrm{X}}_n]+3\mathbb{E}[X_i\bar{\mathrm{X}}_n^2]-\mathbb{E}\bar{\mathrm{X}}_n^3\Big) &= \mathbb{E}X^3\Big[2n^{-1}+n-3\Big]\\ &\quad+\mathbb{E}X^2\mathbb{E}X\Big[3n^{-1}(n-1)\Big]\\ &\quad+(\mathbb{E}X)^3 2n^{-1}\Big(n^2-3n+2\Big).\end{aligned} \tag{63}$$

Therefore, for the first term in (56) we have:

$$
\begin{aligned}
\mathbb{E}\Big[X\sum_i (X_i-\bar{\mathrm{X}}_n)^3\Big] &= \mathbb{E}X\mathbb{E}X^3\left(2n^{-1}+n-3\right) \\
&\qquad + \mathbb{E}X^2(\mathbb{E}X)^2 3n^{-1}(n-1) + (\mathbb{E}X)^4 2n^{-1}\left(n^2-3n+2\right) \\
&= \mathbb{E}X\Big(\mathbb{E}X^3 - 3\mathbb{E}X^2\mathbb{E}X + 2(\mathbb{E}X)^3\Big)n^{-1}(n-1)(n-2) \\
&= \mathbb{E}X\mathbb{E}(X-\mathbb{E}X)^3 n^{-1}(n-1)(n-2) \qquad (64)
\end{aligned}
$$

Moving forward, for the second term of (56) we have

$$
\begin{aligned}
&\mathbb{E}\Big[\bar{\mathrm{X}}_n\sum_i (X_i-\bar{\mathrm{X}}_n)^3\Big] = \mathbb{E}\Big[\bar{\mathrm{X}}_n\sum_i\left(X_i^3 - 3X_i^2\bar{\mathrm{X}}_n + 3X_i\bar{\mathrm{X}}_n^2 - \bar{\mathrm{X}}_n^3\right)\Big] \\
&= \sum_i\left(\mathbb{E}[X_i^3\bar{\mathrm{X}}_n] - 3\mathbb{E}[X_i^2\bar{\mathrm{X}}_n^2] + 3\mathbb{E}[X_i\bar{\mathrm{X}}_n^3] - \mathbb{E}\bar{\mathrm{X}}_n^4\right) \qquad (65)
\end{aligned}
$$

We compute the terms inside the parenthesis in (65). For the first one we have:

$$
\begin{aligned}
\mathbb{E}[X_i^3\bar{\mathrm{X}}_n] &= n^{-1}\mathbb{E}[X_i^3\left(X_i + \sum_{j=1,j\neq i}^n X_j\right)] \\
&= n^{-1}\mathbb{E}X^4 + n^{-1}(n-1)\mathbb{E}X^3\mathbb{E}X. \qquad (66)
\end{aligned}
$$

For the second:

$$
\begin{aligned}
&-3\mathbb{E}[X_i^2\bar{\mathrm{X}}_n^2] = -3n^{-2}\mathbb{E}[X_i^2\left(X_i + \sum_{j=1,j\neq i} X_j\right)^2] \\
&\overset{(59)}{=} -3n^{-2}\mathbb{E}X^4 - 6n^{-2}(n-1)\mathbb{E}X^3\mathbb{E}X - 3n^{-2}(n-1)(\mathbb{E}X^2)^2 - 3n^{-2}(n-1)(n-2)\mathbb{E}X^2(\mathbb{E}X)^2 \qquad (67)
\end{aligned}
$$

The third term in (65) reads:

$$
\begin{aligned}
&3\mathbb{E}[X_i\bar{\mathrm{X}}_n^3] = 3n^{-3}\mathbb{E}[X_i(\sum_j X_j)^3] \\
&= 3n^{-3}\mathbb{E}[X_i(X_i + \sum_{j=1,j\neq i} X_j)^3] \\
&= 3n^{-3}\mathbb{E}X^4 + 12n^{-3}(n-1)\mathbb{E}X^3\mathbb{E}X + 9n^{-3}(n-1)(\mathbb{E}X^2)^2 \\
&\quad + 18n^{-3}(n-1)(n-2)\mathbb{E}X^2(\mathbb{E}X)^2 + 3n^{-3}(\mathbb{E}X)^4(n-1)(n-2)(n-3), \qquad (68)
\end{aligned}
$$

where we expanded the third power and subsequently used that

$$
\mathbb{E}\left(\sum_{j=1,j\neq i}^n X_j\right)^3 = (n-1)\mathbb{E}X^3 + 3(n-1)(n-2)\mathbb{E}X^2\mathbb{E}X + (\mathbb{E}X)^3(n-1)(n-2)(n-3). \qquad (69)
$$

The last term in the parenthesis of (65) is

$$-\mathbb{E}\bar{\mathrm{X}}_n^4 = -n^{-3}\mathbb{E}X^4 - 4n^{-3}(n-1)\mathbb{E}X^3\mathbb{E}X - 3n^{-3}(n-1)(\mathbb{E}X^2)^2$$
$$- 6n^{-3}(n-1)(n-2)\mathbb{E}X^2(\mathbb{E}X)^2 - n^{-3}(n-1)(n-2)(n-3)(\mathbb{E}X)^4, \quad (70)$$

where we used

$$\mathbb{E}\bar{\mathrm{X}}_n^4 = n^{-4}\sum_{k_1,\dots,k_n\geq 0}\frac{4!}{k_1!\dots k_n!}\mathbb{E}[X^{k_1}\cdot\dots\cdot X^{k_n}], \quad (71)$$

where $k_1,\dots,k_n \geq 0$ are all combinations that sum up to n. By gathering all the terms we have that the parenthesis in (65) reads:

$$\begin{aligned}\Big(\mathbb{E}[X_i^3\bar{\mathrm{X}}_n] - 3\mathbb{E}[X_i^2\bar{\mathrm{X}}_n^2] + 3\mathbb{E}[X_i\bar{\mathrm{X}}_n^3] - \mathbb{E}\bar{\mathrm{X}}_n^4\Big) &= n^{-3}\mathbb{E}X^4(n-1)(n-2)\\ &+ n^{-3}\mathbb{E}X^3\mathbb{E}X(n-1)(n-2)(n-4)\\ &- 3(\mathbb{E}X^2)^2n^{-3}(n-1)(n-2)\\ &- 3\mathbb{E}X^2(\mathbb{E}X)^2n^{-3}(n-1)(n-2)(n-4)\\ &+ 2(\mathbb{E}X)^4n^{-3}(n-1)(n-2)(n-3) \quad (72)\end{aligned}$$

Lastly, the outer summation in (65) increases the order by n.

$$\begin{aligned}\mathbb{E}\left[\bar{\mathrm{X}}_n\sum_i\Big(X_i - \bar{\mathrm{X}}_n\Big)^3\right] &= n^{-2}\mathbb{E}X^4(n-1)(n-2)\\ &+ n^{-2}\mathbb{E}X^3\mathbb{E}X(n-1)(n-2)(n-4)\\ &- 3(\mathbb{E}X^2)^2n^{-2}(n-1)(n-2)\\ &- 3\mathbb{E}X^2(\mathbb{E}X)^2n^{-2}(n-1)(n-2)(n-4)\\ &+ 2(\mathbb{E}X)^4n^{-2}(n-1)(n-2)(n-3)\\ &= \mathbb{E}X\mathbb{E}(X-\mathbb{E}X)^3n^{-1}(n-1)(n-2)\\ &+ n^{-2}(n-1)(n-2)\Big(\mathbb{E}(X-\mathbb{E}X)^4 - 3[\mathbb{E}(X-\mathbb{E}X)^2]^2\Big) \quad (73)\end{aligned}$$

By plugging (64), and (73) into (56) we obtain

$$\begin{aligned}\mathbb{E}\Big[(X-\bar{\mathrm{X}}_n)\sum_i(X_i-\bar{\mathrm{X}}_n)^3\Big] &= \mathbb{E}X\mathbb{E}(X-\mathbb{E}X)^3n^{-1}(n-1)(n-2)\\ &- \mathbb{E}X\mathbb{E}(X-\mathbb{E}X)^3n^{-1}(n-1)(n-2)\\ &- n^{-2}(n-1)(n-2)\Big(\mathbb{E}(X-\mathbb{E}X)^4 - 3[\mathbb{E}(X-\mathbb{E}X)^2]^2\Big) \quad (74)\end{aligned}$$

or

$$\mathbb{E}\Big[(X-\bar{\mathrm{X}}_n)\sum_i(X_i-\bar{\mathrm{X}}_n)^3\Big] = n^{-2}(n-1)(n-2)\Big(3m_2^2 - m_4\Big), \quad (75)$$

where m_2, and m_4 denote the second and fourth central moments of the probability measure μ, respectively. Strictly positive values for $\bar{\lambda}$ are feasible for all those models with $3m_2^2 > m_4$. Lastly, the condition (55) reads:

$$\begin{aligned}\bar{\lambda} &\leq \frac{2\mathbb{E}\Big[(X - \bar{\mathrm{X}}_n)\mathrm{T}_n\Big]}{\mathbb{E}\mathrm{T}_n^2 + \epsilon} = \frac{2(3m_2^2 - m_4)}{\mathbb{E}\mathrm{T}_n^2 + \epsilon}\frac{n^2 - 3n + 2}{n^3} \\ &\leq \frac{1}{n}\left[\frac{8(3m_2^2 - m_4)}{\mathbb{E}\mathrm{T}_n^2 + \epsilon}\right]. \qquad (76)\end{aligned}$$

□