Regret-Optimal Wasserstein-Robust Linear Regression

Sloan Nietert
EPFL
sloan.nietert@epfl.ch

Daniel Kuhn
EPFL
daniel.kuhn@epfl.ch

Abstract

Distributionally robust optimization (DRO) minimizes worst-case risk under distributional uncertainty, but its decisions can be suboptimal in terms of their excess risk, or regret. We study this gap for multivariate least-squares regression, assuming that the observed source law and unknown target law differ by an adversarial 2-Wasserstein perturbation of size ε. We first identify the asymptotic complexity of robust decision-making in the $\varepsilon \to 0$ limit, characterized by a problem-specific condition number κ. Here, a wide spectrum of simple algorithms, including DRO, achieve the instance-optimal rate of $O(\kappa\varepsilon^2)$. For fixed $\varepsilon > 0$, we prove that DRO still achieves the minimax rate if the problem has bounded residual effective rank. On the other hand, we identify significant failure modes where DRO is provably suboptimal by dimension-dependent factors. Fortunately, we prove that two alternative procedures, minimax regret optimization (MRO) and spectral-norm regularization (SNR), attain the minimax rate. We derive an exact convex reformulation of MRO and an accelerated proximal algorithm for SNR, along with high-probability finite-sample guarantees. We support these theoretical results with numerical experiments separating the performance of MRO and SNR from DRO.

1 Introduction

Distributionally robust optimization (DRO) is a flexible tool for hedging against model misspecification in ML, statistics, and operations research (Rahimian and Mehrotra, 2022; Lin et al., 2022; Kuhn et al., 2025). Wasserstein DRO is a well-studied variant that minimizes worst-case risk over all Wasserstein perturbations of the observed data distribution. For many tasks, this objective admits dual formulations, tractable implementations, and regularization interpretations (Kuhn et al., 2019; Gao and Kleywegt, 2022; Gao et al., 2024). Model misspecification comes in many flavors, e.g., sampling error, distribution shift, and data poisoning. In this work, we take an adversarial view: the learner observes a source law and seeks a predictor that performs well under a nearby but otherwise unrestricted target law. As we will see, robust risk at the source (the objective minimized by DRO) is fundamentally distinct from decision quality at the target. Our motivating question is the following:

When does Wasserstein DRO make good decisions in the face of adversarial data perturbations, and when can we efficiently outperform it?

We examine this for the setting of multivariate linear regression. Fix dimensions $p, m \geq 1$ and let $\mathcal{A} := \mathbb{R}^{m\times p}$ be the class of linear predictors. Set $d := p + m$ and write $\mathcal{P}_2(\mathbb{R}^d)$ for the laws of $Z = (X, Y) \in \mathbb{R}^p \times \mathbb{R}^m$ with finite second moment. For $P \in \mathcal{P}_2(\mathbb{R}^d)$, the least-squares risk of $\mathbf{A} \in \mathcal{A}$ is $L(\mathbf{A}, P) := \mathbb{E}_P \|\mathbf{A}X - Y\|^2$, and its *excess risk*, or (ex-ante) *regret*, is

$$R(\mathbf{A}, P) := L(\mathbf{A}, P) - \inf_{\mathbf{A}'\in\mathcal{A}} L(\mathbf{A}', P).$$

This regret is ex-ante, rather than ex-post, because we benchmark against the best decision given knowledge of P and not a realized pair $(X, Y) \sim P$. Both risk and regret depend on P only through

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

its second-moment matrix $\mathbf{\Sigma}_P \coloneqq \mathbb{E}_P[ZZ^\top]$, with blocks $\mathbf{\Sigma}_X^P$, $\mathbf{\Sigma}_Y^P$, and $\mathbf{\Sigma}_{XY}^P = (\mathbf{\Sigma}_{YX}^P)^\top$. Write $\mathcal{P}^* \coloneqq \{P \in \mathcal{P}_2(\mathbb{R}^d) : \mathbf{\Sigma}_X^P \succ \mathbf{0}\}$ for the class of *non-degenerate* laws. For $P \in \mathcal{P}^*$, its unique risk minimizer is $\mathbf{A}_P \coloneqq \mathbf{\Sigma}_{YX}^P(\mathbf{\Sigma}_X^P)^{-1}$ and regret has the closed form

$$R(\mathbf{A}, P) = \left\|(\mathbf{A} - \mathbf{A}_P)(\mathbf{\Sigma}_X^P)^{1/2}\right\|_{\mathrm{F}}^2.$$

We write $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{\mathrm{op}}$ for the Frobenius norm and Euclidean operator norm, respectively.

In our population model, an adversary chooses a source–target pair $P_{\mathrm{s}}, P_{\mathrm{t}} \in \mathcal{P}_2(\mathbb{R}^d)$ satisfying $\mathsf{W}_2(P_{\mathrm{s}}, P_{\mathrm{t}}) \le \varepsilon$ and reveals only the source P_{s}. Given the radius $\varepsilon > 0$, the learner applies an estimator $\hat{\mathbf{A}} : \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{A}$ and incurs target regret $R(\hat{\mathbf{A}}[P_{\mathrm{s}}], P_{\mathrm{t}})$. We measure the *robust regret* of $\hat{\mathbf{A}}$ at a reference law P by allowing P to occupy either the source or target role:

$$R_\varepsilon(\hat{\mathbf{A}}, P) \coloneqq \sup_{Q \in \mathcal{B}_\varepsilon(P)} \max\big\{R(\hat{\mathbf{A}}[P], Q),\ R(\hat{\mathbf{A}}[Q], P)\big\},$$

where $\mathcal{B}_\varepsilon(P) \coloneqq \{Q \in \mathcal{P}_2(\mathbb{R}^d) : \mathsf{W}_2(P, Q) \le \varepsilon\}$. The first term varies the target with fixed source P, while the second varies the source with fixed target P.

Robust estimators. We primarily compare *empirical risk minimization* (ERM), Wasserstein DRO, *minimax regret optimization* (MRO), and *spectral-norm regularization* (SNR):[1]

$$\begin{aligned}
\mathsf{ERM}[P] &\in \arg\min_{\mathbf{A} \in \mathcal{A}} L(\mathbf{A}, P),\\
\mathsf{DRO}_\varepsilon[P] &\in \arg\min_{\mathbf{A} \in \mathcal{A}} \sup_{Q \in \mathcal{B}_\varepsilon(P)} L(\mathbf{A}, Q),\\
\mathsf{MRO}_\varepsilon[P] &\in \arg\min_{\mathbf{A} \in \mathcal{A}} \sup_{Q \in \mathcal{B}_\varepsilon(P)} R(\mathbf{A}, Q),\\
\mathsf{SNR}_\varepsilon[P] &\in \arg\min_{\mathbf{A} \in \mathcal{A}} L(\mathbf{A}, P) + \varepsilon^2 \|\mathbf{A}\|_{\mathrm{op}}^2.
\end{aligned}$$

Here, ERM fits the source law directly, returning $\mathbf{A}_P$ when P is non-degenerate. MRO replaces the risk objective in DRO with regret. Finally, SNR adds a squared operator-norm penalty to ERM.

Local guarantees. To quantify estimation complexity, we first study the local minimax benchmark

$$R_\varepsilon^\star(P) \coloneqq \inf_{\hat{\mathbf{A}}} R_\varepsilon(\hat{\mathbf{A}}, P) = \inf_{\mathbf{A} \in \mathcal{A}} \sup_{Q \in \mathcal{B}_\varepsilon(P)} R(\mathbf{A}, Q). \tag{1}$$

Equality follows by considering P in the source role for the "$\ge$" direction and considering constant estimators for the "$\le$" direction. For non-degenerate P, let $\mathbf{\Sigma}_E^P$ denote the second-moment matrix of the residual $Y - \mathbf{A}_P X$. Our first result shows that the following limit exists and computes it exactly:

$$\kappa_P \coloneqq \lim_{\varepsilon \downarrow 0} \varepsilon^{-2} R_\varepsilon^\star(P) = 1 + \lambda_{\max}\left(\mathbf{A}_P \mathbf{A}_P^\top + \frac{\mathbf{\Sigma}_E^P}{\lambda_{\min}(\mathbf{\Sigma}_X^P)}\right).$$

Thus, the *condition number* κ_P governs the optimal local regret, $\kappa_P \varepsilon^2 + o(\varepsilon^2)$. It reflects the size of the regression coefficients and that of the residual second moment relative to the smallest feature eigenvalue. All four estimators attain this local rate up to absolute multiplicative constants as $\varepsilon \downarrow 0$.

Global guarantees. For fixed $\varepsilon > 0$, we tightly characterize estimator performance as a function of ε and the condition number. For a class $\mathcal{G} \subseteq \mathcal{P}_2(\mathbb{R}^d)$, define the *worst-case* and *minimax* regrets $R_\varepsilon(\hat{\mathbf{A}}, \mathcal{G}) \coloneqq \sup_{P \in \mathcal{G}} R_\varepsilon(\hat{\mathbf{A}}, P)$ and $R_\varepsilon^\star(\mathcal{G}) \coloneqq \inf_{\hat{\mathbf{A}}} R_\varepsilon(\hat{\mathbf{A}}, \mathcal{G})$, so that $R_\varepsilon^\star(P) = R_\varepsilon^\star(\{P\})$. For the well-conditioned family $\mathcal{G}(\kappa) \coloneqq \{P \in \mathcal{P}^* : \kappa_P \le \kappa\}$, our main findings are summarized in Table 1. Namely, the minimax rate is $O(\kappa\varepsilon^2)$ for all $\varepsilon > 0$, and this is attained by MRO and SNR. In contrast, ERM can suffer arbitrarily large regret even for well-conditioned instances. DRO lies between these extremes; when $m = O(1)$ or, more generally, when the residual second moment $\mathbf{\Sigma}_E^P$ has low effective rank, we prove that it achieves the minimax rate. For $\kappa \ge 2$, however, its worst-case robust regret is suboptimal by a multiplicative factor of $\sqrt{\min\{p, m\}} + \sqrt{m/\kappa}$.

For upper bounds, our analysis employs the Bures–W_2 geometry of second-moment matrices to prove a key regret stability lemma. In particular, we show that $R(\mathbf{A}_P, Q) \lesssim \kappa_P \mathsf{W}_2(P, Q)^2$ for every

[1] All four optimization problems admit minimizers; see Section D.1. All of our regret bounds (upper and lower) hold for any minimizer selection rule.

Estimator $\hat{\mathbf{A}}$	ERM	DRO_ε	MRO_ε	SNR_ε
Robust regret	∞	$\left(\sqrt{\min\{p,m\}}\,\kappa+\sqrt{m\kappa}\right)\varepsilon^2$	$\kappa\varepsilon^2$	$\kappa\varepsilon^2$
Minimax optimal?	$\times$	$\times$	✓	✓
Computation	Closed form	Convex reformulation	(New) convex reformulation	Accelerated proximal gradient method (FISTA)

Table 1: Worst-case robust regret $R_\varepsilon(\hat{\mathbf{A}}, \mathcal{G}(\kappa))$ for our running estimators, up to absolute multiplicative constants, for $\varepsilon > 0$ and $\kappa \geq 2$. Computation uses second-moment matrices.

non-degenerate P and arbitrary $Q \in \mathcal{P}_2(\mathbb{R}^d)$. More generally, our stability lemma controls changes in square-root regret through the predictor's operator norm, explaining the role of the SNR penalty. For lower bounds, we provide explicit constructions where these upper bounds are tight.

Computation and finite-sample statistics. We derive a novel, exact convex reformulation of MRO, building upon previous results for DRO (Shafieezadeh-Abadeh et al., 2018), along with an efficient proximal algorithm for SNR, based on FISTA (Beck and Teboulle, 2009). Given a fixed source second-moment matrix and radius $\varepsilon > 0$, our SNR algorithm reaches objective gap $\delta > 0$ in $O(\delta^{-1/2})$ iterations, each using a singular value decomposition. We also extend our results to two finite-sample models: adversarial corruption of independent and identically distributed (i.i.d.) target observations, and i.i.d. source observations with a Wasserstein-perturbed population target. Here, we prove high-probability regret bounds under fourth-moment assumptions, with improved dependence on the confidence level under sub-Gaussian tails. Finally, we provide synthetic numerical experiments illustrating the predicted separation of MRO and SNR from DRO.

1.1 Related Work

Wasserstein DRO (WDRO) for regression and other quadratic models has been studied extensively through the lens of duality, regularization, and robust risk minimization; see, e.g., Nguyen et al. (2022, 2021a); Taşkesen et al. (2023) in addition to the works cited above. Shafieezadeh-Abadeh et al. (2018) explicitly develop WDRO algorithms for our multivariate linear regression setting, applied to a sequential Kalman filtering problem. More broadly, WDRO has gained interest in ML from the perspective of adversarial training (Sinha et al., 2018), and Wasserstein-robust risk has been well explored by several communities (Lee and Raginsky, 2018; Blanchet and Murthy, 2019).

Our focus is different: we evaluate estimators by *excess* risk at the unknown target law, rather than by risk at the observed source or worst-case law. This distinction is essential, since regret-optimal procedures need not coincide with DRO. MRO, also called distributionally robust regret optimization, has garnered significant study. Agarwal and Zhang (2022) compare MRO and DRO with general ambiguity sets and show that heterogeneity in irreducible risk can cause DRO to incur larger regret. In operations research and control, related formulations have been studied for newsvendor models (Perakis and Roels, 2008; Chen and Xie, 2021), more general Wasserstein ambiguity sets (Cho and Yang, 2024; Fiechtner and Blanchet, 2025), and multistage stochastic control (Fiechtner and Blanchet, 2026). However, these works do not establish worst-case robust regret rates for estimator comparison.

Fiechtner and Blanchet (2025) introduce a first-order sensitivity parameter which nullifies for our setting. Theorem 3 in Version 5 of their paper (posted after the conference submission of this work) establishes a second-order sensitivity expansion whose coefficient in fact specializes to our condition number. We provide a self-contained Bures-geometric derivation, an explicit formula for multivariate least squares, and connect this local quantity to global stability bounds. Their all-radius optimality result for ERM concerns losses jointly quadratic in the decision and uncertainty, excluding our setting.

Our model is closely related to adversarially robust statistics under Wasserstein perturbations. Scalar-response linear regression has been studied under 1-Wasserstein (Zhu et al., 2022; Liu and Loh, 2023) and 2-Wasserstein contamination (Chao and Dobriban, 2023). These works establish robust guarantees under different distributional and design assumptions; our global comparisons are instead indexed by the local minimax condition number. Moreover, moving to multivariate responses is crucial for us to separate the worst-case regret of DRO from the minimax rates of MRO and SNR.

1.2 Preliminaries

Notation. Unsubscripted vector norms are Euclidean. Write $\mathbb{S}_+^d$ for the cone of positive semidefinite $d \times d$ matrices and $\mathbb{O}^d := \{\mathbf{U} \in \mathbb{R}^{d\times d} : \mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}_d\}$ for the family of orthogonal matrices. For symmetric matrices $\mathbf{M}, \mathbf{N}$ of the same size, write $\mathbf{M} \preceq \mathbf{N}$ ($\mathbf{M} \prec \mathbf{N}$) when $\mathbf{N} - \mathbf{M}$ is positive semidefinite (positive definite), and define $\succeq, \succ$ analogously. Let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ denote the smallest and largest eigenvalues of a symmetric matrix $\mathbf{M}$. For a real matrix $\mathbf{M}$, let $\mathbf{M}^\dagger$ denote its Moore–Penrose pseudoinverse. If square, write $\mathrm{tr}(\mathbf{M})$ for its trace. Let $\mathbf{0}$ and $\mathbf{I}$ denote the zero and identity matrices, respectively, indexed by dimension when needed. The notation diag constructs a diagonal matrix from scalar arguments and a block-diagonal matrix from square-matrix arguments. For $\mathbf{M} \succeq \mathbf{0}$, $\mathbf{M}^{1/2}$ denotes its unique positive semidefinite square root. Write $a \lesssim b$ if $a \le Cb$ for an absolute constant $C > 0$. Write $a \gtrsim b$ if $b \lesssim a$ and $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$.

Bures–Wasserstein geometry. We employ the *2-Wasserstein distance* defined for $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$\mathsf{W}_2(P, Q) := \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(Z,Z')\sim\pi}\big[\|Z - Z'\|^2\big]^{\frac{1}{2}},$$

where $\Pi(P, Q)$ denotes the set of joint probability laws on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals P and Q. Since our least-squares risk only depends on (uncentered) second moments, it is convenient to switch from the full W_2 geometry on probability measures to its induced geometry on second-moment matrices. In particular, we turn to the *Bures distance*, defined between positive semidefinite $\mathbf{A}, \mathbf{B} \succeq \mathbf{0}$ by

$$\mathsf{B}(\mathbf{A}, \mathbf{B})^2 := \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) - 2\,\mathrm{tr}\big((\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\big).$$

We refer the reader to Bhatia et al. (2019) for a recent survey of this metric and recall several of its important properties below, deferring proofs to Appendix B. We often use an equivalent formulation of the Bures distance as the minimum Frobenius norm between square-root factors of $\mathbf{A}$ and $\mathbf{B}$.

Lemma 1.1 (Procrustes formulation of B)**.** *For all $\mathbf{A}, \mathbf{B} \in \mathbb{S}_+^d$, we have*

$$\mathsf{B}(\mathbf{A}, \mathbf{B}) = \min_{\mathbf{U}\in\mathbb{O}^d} \|\mathbf{A}^{\frac{1}{2}} - \mathbf{B}^{\frac{1}{2}}\mathbf{U}\|_{\mathrm{F}} = \min_{\substack{\mathbf{S},\mathbf{T}\in\mathbb{R}^{d\times d} \\ \mathbf{A}=\mathbf{S}\mathbf{S}^\top, \mathbf{B}=\mathbf{T}\mathbf{T}^\top}} \|\mathbf{S} - \mathbf{T}\|_{\mathrm{F}}.$$

Thus, $\mathsf{B}(\mathbf{A}, \mathbf{B}) \le \varepsilon$ if and only if there exists $\mathbf{\Delta} \in \mathbb{R}^{d\times d}$ with $\|\mathbf{\Delta}\|_{\mathrm{F}} \le \varepsilon$ such that $\mathbf{A} = (\mathbf{B}^{1/2} + \mathbf{\Delta})(\mathbf{B}^{1/2} + \mathbf{\Delta})^\top$. The minimum is attained by $\mathbf{U} = \mathbf{I}_d$ if $\mathbf{A}$ and $\mathbf{B}$ are simultaneously diagonalizable.

More importantly, this metric is precisely that induced by W_2 on the space of second-moment matrices. Namely, setting $\mathcal{B}_\varepsilon(\mathbf{\Sigma}) := \{\mathbf{\Sigma}' \in \mathbb{S}_+^d : \mathsf{B}(\mathbf{\Sigma}', \mathbf{\Sigma}) \le \varepsilon\}$, we have the following.

Lemma 1.2 (W_2–Bures equivalence)**.** *For all $P \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varepsilon \ge 0$, we have*

$$\{\mathbf{\Sigma}_Q : Q \in \mathcal{B}_\varepsilon(P)\} = \mathcal{B}_\varepsilon(\mathbf{\Sigma}_P) = \Big\{(\mathbf{\Sigma}_P^{1/2} + \mathbf{\Delta})(\mathbf{\Sigma}_P^{1/2} + \mathbf{\Delta})^\top : \|\mathbf{\Delta}\|_{\mathrm{F}} \le \varepsilon\Big\}.$$

Throughout, we refer to $\mathbb{S}_+^d$ equipped with B as *Bures space*.

2 Local Asymptotic Complexity and the Condition Number

We begin by exploring the fundamental limits of robust estimation in the asymptotic $\varepsilon \to 0$ regime. Recall the local minimax robust regret $R_\varepsilon^\star(P)$, defined in (1). In general, this quantity depends on the complex interaction between P, R, and the W_2 geometry. However, as $\varepsilon \to 0$, we show that the complexity of robust prediction is dictated by the following limiting quantity, which we compute in closed form under a mild non-degeneracy assumption.

Definition 2.1. Define the *condition number* of $P \in \mathcal{P}^*$ by $\kappa_P := \limsup_{\varepsilon\downarrow 0} \varepsilon^{-2} R_\varepsilon^\star(P)$. For convenience, we set $\kappa_P := \infty$ for $P \notin \mathcal{P}^*$.[2]

Theorem 2.2 (Condition number formula)**.** *If $P \in \mathcal{P}^*$, then $R_\varepsilon^\star(P) = \kappa_P\varepsilon^2 + o(\varepsilon^2)$ as $\varepsilon \downarrow 0$ and*

$$\kappa_P = 1 + \lambda_{\max}\big(\mathbf{A}_P\mathbf{A}_P^\top + \lambda_{\min}(\mathbf{\Sigma}_X^P)^{-1}\mathbf{\Sigma}_E^P\big).$$

[2]This definition and our resulting non-asymptotic analysis can be adapted to the setting where $\mathbf{\Sigma}_X^P$ is singular but the condition number limit is finite (which occurs when the optimal risk is 0). We omit this edge case for ease of presentation.

This implies the useful bounds

$$1 + \max\left\{\|\mathbf{A}_P\|_{\mathrm{op}}^2, \frac{\lambda_{\max}(\mathbf{\Sigma}_E^P)}{\lambda_{\min}(\mathbf{\Sigma}_X^P)}\right\} \le \kappa_P \le 1 + \|\mathbf{A}_P\|_{\mathrm{op}}^2 + \frac{\lambda_{\max}(\mathbf{\Sigma}_E^P)}{\lambda_{\min}(\mathbf{\Sigma}_X^P)}.$$

We derive this formula in Appendix E.1 by separately considering perturbations to the feature coordinates X and the response coordinates Y. The additive 1 term arises due to response perturbations which leave X fixed, while the remaining eigenvalue term results from feature perturbations which leave Y fixed. This latter term is determined by both the size of the optimal coefficient matrix and the scale of the residual variation compared to that of the features.

The proof equates κ_P to a second-order sensitivity of $R(\mathbf{A}_Q, P)$ as Q varies in the neighborhood of P. In Bures space, this amounts to a squared operator norm of a suitable Euclidean derivative, which we compute with some calculus and a scalarization trick that simplifies tensor calculations.

3 Non-Asymptotic Regret Bounds

Although κ is defined by an $\varepsilon \to 0$ limit, we seek regret bounds which hold for all $\varepsilon > 0$. To derive these, we prove a key lemma allowing one to transfer regret bounds between nearby distributions.

Lemma 3.1 (Regret stability)**.** *For all $\mathbf{A} \in \mathcal{A}$, $P \in \mathcal{P}^*$, and $Q \in \mathcal{P}_2(\mathbb{R}^d)$, we have*

$$\left|\sqrt{R(\mathbf{A}, P)} - \sqrt{R(\mathbf{A}, Q)}\right| \le (2\sqrt{\kappa_P} + \|\mathbf{A}\|_{\mathrm{op}})\mathsf{W}_2(P, Q).$$

Moreover, for $\mathbf{A} = \mathbf{A}_P$, we have $R(\mathbf{A}_P, Q) \le 4\kappa_P \mathsf{W}_2(P, Q)^2$.

The proof in Appendix F.1 performs some finite-difference error analysis using a closed form for the regret in terms of Euclidean projections, valid even for degenerate Q. Here, (a variant of) the Davis–Kahan–Wedin sine theorem (Wedin, 1972) serves as a key technical tool, controlling the Frobenius discrepancy between projection matrices onto the row spaces of nearby matrices.

Regret stability implies several important regret bounds for our main estimators.

Lemma 3.2 (Stability implications)**.** *Let $P \in \mathcal{P}^*$ and $\varepsilon > 0$. We then have*

$$R_\varepsilon(\mathsf{MRO}_\varepsilon, P) \lesssim \kappa_P \varepsilon^2, \qquad R_\varepsilon(\mathsf{SNR}_\varepsilon, P) \lesssim \kappa_P \varepsilon^2.$$

In contrast, our ERM and DRO guarantees adapt to the largest condition number in a ball around P:

$$R_\varepsilon(\mathsf{ERM}, P) \lesssim \sup_{Q \in \mathcal{B}_\varepsilon(P)} \kappa_Q \varepsilon^2 \qquad R_\varepsilon(\mathsf{DRO}_\varepsilon, P) \lesssim \sup_{Q \in \mathcal{B}_{2\varepsilon}(P)} \kappa_Q \varepsilon^2$$

If $\varepsilon^2 \le \lambda_{\min}(\mathbf{\Sigma}_X^P)/16$, then these two upper bounds are at most $O(\kappa_P \varepsilon^2)$.

We give the proof for MRO and SNR below, deferring the ERM and DRO arguments to Appendix F.2.

Proof of Lemma 3.2 (MRO and SNR). Let $Q \in \mathcal{B}_\varepsilon(P)$ and fix source–target pair $(P_{\mathrm{s}}, P_{\mathrm{t}})$ equal to either (P, Q) or (Q, P). For MRO, we set $\hat{\mathbf{A}} \coloneqq \mathsf{MRO}_\varepsilon[P_{\mathrm{s}}]$ and bound

$$R(\hat{\mathbf{A}}, P_{\mathrm{t}}) \le \sup_{Q' \in \mathcal{B}_\varepsilon(P_{\mathrm{s}})} R(\hat{\mathbf{A}}, Q') \le \sup_{Q' \in \mathcal{B}_\varepsilon(P_{\mathrm{s}})} R(\mathbf{A}_P, Q') \le 16\kappa_P \varepsilon^2.$$

The first inequality uses $P_{\mathrm{t}} \in \mathcal{B}_\varepsilon(P_{\mathrm{s}})$, the second uses MRO optimality, and the last uses $\mathcal{B}_\varepsilon(P_{\mathrm{s}}) \subseteq \mathcal{B}_{2\varepsilon}(P)$ and Lemma 3.1. Supremizing over Q and source–target pairs gives $R_\varepsilon(\mathsf{MRO}_\varepsilon, P) \lesssim \kappa_P \varepsilon^2$.

For SNR, take instead $\hat{\mathbf{A}} \coloneqq \mathsf{SNR}_\varepsilon[P_{\mathrm{s}}]$. Since risk and regret differ by a constant independent of $\mathbf{A}$,

$$R(\hat{\mathbf{A}}, P_{\mathrm{s}}) + \varepsilon^2 \|\hat{\mathbf{A}}\|_{\mathrm{op}}^2 \le R(\mathbf{A}_P, P_{\mathrm{s}}) + \varepsilon^2 \|\mathbf{A}_P\|_{\mathrm{op}}^2 \le 5\kappa_P \varepsilon^2,$$

where the first inequality uses SNR optimality and the second follows from $P_{\mathrm{s}} \in \mathcal{B}_\varepsilon(P)$, Lemma 3.1, and $\|\mathbf{A}_P\|_{\mathrm{op}}^2 \le \kappa_P$ (by Theorem 2.2). Applying Lemma 3.1 again yields

$$\begin{aligned}\sqrt{R(\hat{\mathbf{A}}, P_{\mathrm{t}})} &\le \sqrt{R(\hat{\mathbf{A}}, P_{\mathrm{s}})} + \big(\|\hat{\mathbf{A}}\|_{\mathrm{op}} + 2\sqrt{\kappa_P}\big)\varepsilon \\ &\le \sqrt{2}\sqrt{R(\hat{\mathbf{A}}, P_{\mathrm{s}}) + \varepsilon^2 \|\hat{\mathbf{A}}\|_{\mathrm{op}}^2} + 2\sqrt{\kappa_P}\varepsilon \le 6\sqrt{\kappa_P}\varepsilon.\end{aligned}$$

The second inequality is Cauchy–Schwarz, and the last uses the objective comparison above. Squaring this SNR bound and supremizing over Q and source–target pairs gives $R_\varepsilon(\mathsf{SNR}_\varepsilon, P) \lesssim \kappa_P \varepsilon^2$. □

By combining the DRO guarantee above with existing DRO regularizer bounds (Gao et al., 2024; Lee and Raginsky, 2018), we obtain an all-radius guarantee adapted to the residual effective rank. Writing $r_{\mathrm{eff}}(\mathbf{M}) \coloneqq \mathrm{tr}(\mathbf{M})/\|\mathbf{M}\|_{\mathrm{op}}$ for $\mathbf{M} \succeq \mathbf{0}$, with $r_{\mathrm{eff}}(\mathbf{0}) \coloneqq 0$, we prove the following in Appendix F.3.

Lemma 3.3 (Rank-adaptive DRO bound)**.** *For $P \in \mathcal{P}^*$ and $\varepsilon > 0$, set $r \coloneqq r_{\mathrm{eff}}(\mathbf{\Sigma}_E^P)$. Then*

$$R_\varepsilon(\mathsf{DRO}_\varepsilon, P) \lesssim \Big(\sqrt{\min\{p, r\}}(\kappa_P - 1) + \sqrt{r(\kappa_P - 1)} + \kappa_P\Big)\varepsilon^2.$$

In particular, DRO achieves regret $O(\kappa_P \varepsilon^2)$ whenever the residual second moment has effective rank $O(1)$. In the worst case, however, the effective rank can be as large as m.

Combining the above, we derive worst-case rates for the family $\mathcal{G}(\kappa) = \{P \in \mathcal{P}^* : \kappa_P \le \kappa\}$.

Theorem 3.4 (Population-limit summary rates)**.** *Fix $\varepsilon > 0$ and $\kappa \ge 1$. Then the minimax rate for the class $\mathcal{G} = \mathcal{G}(\kappa)$ is $R_\varepsilon^\star(\mathcal{G}) \asymp \kappa\varepsilon^2$, and this is achieved by both MRO_ε and SNR_ε. In contrast, we have $R_\varepsilon(\mathsf{ERM}, \mathcal{G}) = \infty$ and*

$$R_\varepsilon(\mathsf{DRO}_\varepsilon, \mathcal{G}) \asymp \Big(\sqrt{\min\{p, m\}}(\kappa - 1) + \sqrt{m(\kappa - 1)} + \kappa\Big)\varepsilon^2.$$

In particular, if $\kappa \ge 2$ and $p \ge m$, then $R_\varepsilon(\mathsf{DRO}_\varepsilon, \mathcal{G}) \asymp \sqrt{m}\kappa\varepsilon^2$.

In summary, MRO and SNR are always minimax optimal, ERM can suffer unbounded regret, and DRO is provably suboptimal by a dimension-dependent factor for $\kappa \ge 2$. The upper bounds in Appendix F.4 apply Lemmas 3.2 and 3.3. The minimax lower bound follows by rescaling the local result from Theorem 2.2; this works because the condition number is invariant to uniform rescaling.

For ERM, we consider a target law P with identically zero responses. A nearby source law with a nearly collapsed feature and a response correlated with that feature forces the ERM solution to diverge, yielding unbounded target regret even though $\kappa_P = 1$. The DRO lower bound requires more care. We employ two constructions with $\lambda_{\min}(\mathbf{\Sigma}_X^P) = \varepsilon^2$ and independent residual noise spread across many response coordinates. In each, the adversary contracts one feature coordinate while leaving the responses unchanged. We then exploit rotational symmetry and a strong duality formula to characterize the DRO minimizers explicitly. The noisy response coordinates determine the minimum value, and we tune the contraction so that attaining this value forces an inflated coefficient on the contracted feature and hence large target regret. This demonstrates how independent residual noise, which contributes an additive constant to target risk and thus cancels from regret, can nevertheless lead to suboptimal DRO decisions.

4 Efficient Computation

We next treat computation, given an observed second-moment matrix $\tilde{\mathbf{\Sigma}} \in \mathbb{S}_+^d$ and radius $\varepsilon > 0$. Since risk and regret only depend on laws via their second moments, we define $L(\mathbf{A}, \mathbf{\Sigma}) \coloneqq L(\mathbf{A}, P)$ for any law P such that $\mathbf{\Sigma}_P = \mathbf{\Sigma}$, and likewise for $R(\mathbf{A}, \mathbf{\Sigma})$. By the Bures–$\mathsf{W}_2$ equivalence (Lemma 1.2), the optimization problems defining our estimators admit equivalent formulations over second-moment matrices, used below. We defer sampling considerations to Section 5.

ERM admits the canonical minimum-Frobenius-norm solution $\tilde{\mathbf{\Sigma}}_{YX}\tilde{\mathbf{\Sigma}}_X^\dagger$, which can be computed via eigendecomposition of $\tilde{\mathbf{\Sigma}}_X$. Alternatively, one may apply any standard convex solver to the objective $\mathbf{A} \mapsto L(\mathbf{A}, \tilde{\mathbf{\Sigma}}) = \mathrm{tr}(\mathbf{A}^\top \mathbf{A}\tilde{\mathbf{\Sigma}}_X) - 2\,\mathrm{tr}(\mathbf{A}\tilde{\mathbf{\Sigma}}_{XY}) + \mathrm{tr}(\tilde{\mathbf{\Sigma}}_Y)$.

Shafieezadeh-Abadeh et al. (2018) developed a convex reformulation for the DRO problem, accompanied by a Frank–Wolfe method with convergence guarantees (see Theorem 3.3 therein). The convex reformulation naïvely requires solving a large semidefinite program, so their Frank–Wolfe method is a noted improvement, albeit with some dependencies on $\lambda_{\min}(\tilde{\mathbf{\Sigma}})^{-1}$ in their convergence rate. We do not seek any further improvements, given the stronger performance of MRO and SNR.

The MRO objective $\mathbf{A} \mapsto \sup_{\mathbf{\Sigma} \in \mathcal{B}_\varepsilon(\tilde{\mathbf{\Sigma}})} R(\mathbf{A}, \mathbf{\Sigma})$ is convex as the pointwise supremum of convex functions in $\mathbf{A}$. However, regret is convex in $\mathbf{\Sigma}$, as the pointwise supremum over $\mathbf{A}'$ of $L(\mathbf{A}, \mathbf{\Sigma}) - L(\mathbf{A}', \mathbf{\Sigma})$ which is linear in $\mathbf{\Sigma}$. This appears problematic, since the inner problem requires maximizing over $\mathbf{\Sigma}$. Nevertheless, we show that it admits a convex reformulation. Here $\mathrm{tr}_p(\mathbf{H})$ denotes the sum of the p largest eigenvalues of a symmetric matrix $\mathbf{H}$, counted with multiplicity.

Theorem 4.1 (Convex reformulation for MRO). *Let $\varepsilon > 0$, $\tilde{\mathbf{\Sigma}} \in \mathbb{S}_+^d$, and $\tilde{\mathbf{S}} \in \mathbb{R}^{d\times d}$ such that $\tilde{\mathbf{\Sigma}} = \tilde{\mathbf{S}}\tilde{\mathbf{S}}^\top$. Decompose*

$$\tilde{\mathbf{S}} = \begin{pmatrix} \tilde{\mathbf{S}}_X \\ \tilde{\mathbf{S}}_Y \end{pmatrix}, \qquad \tilde{\mathbf{S}}_X \in \mathbb{R}^{p\times d}, \qquad \tilde{\mathbf{S}}_Y \in \mathbb{R}^{m\times d}.$$

For $\mathbf{A} \in \mathcal{A}$, consider the MRO objective $F(\mathbf{A}) \coloneqq \sup_{\mathbf{\Sigma}\in\mathcal{B}_\varepsilon(\tilde{\mathbf{\Sigma}})} R(\mathbf{A}, \mathbf{\Sigma})$ and the residual factor $\tilde{\mathbf{E}}_{\mathbf{A}} \coloneqq \tilde{\mathbf{S}}_Y - \mathbf{A}\tilde{\mathbf{S}}_X \in \mathbb{R}^{m\times d}$. Define the extended-real-valued function $D : \mathcal{A} \times \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ by

$$D(\mathbf{A}, \eta) \coloneqq \eta\Big(\varepsilon^2 - \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2 + \mathrm{tr}_p\Big(\tilde{\mathbf{S}}_X^\top\tilde{\mathbf{S}}_X + \tilde{\mathbf{E}}_{\mathbf{A}}^\top\big((\eta-1)\mathbf{I}_m - \mathbf{A}\mathbf{A}^\top\big)^{-1}\tilde{\mathbf{E}}_{\mathbf{A}}\Big)\Big)$$

if $\eta > 1 + \|\mathbf{A}\|_{\mathrm{op}}^2$ and $+\infty$ otherwise. Then D is convex as an extended-value function on $\mathcal{A} \times \mathbb{R}$ and $F(\mathbf{A}) = \inf_{\eta\in\mathbb{R}} D(\mathbf{A}, \eta)$. Consequently, MRO admits the lifted convex reformulation

$$\inf_{\mathbf{A}\in\mathcal{A}} F(\mathbf{A}) = \inf_{\mathbf{A}\in\mathcal{A}} \sup_{\mathbf{\Sigma}\in\mathcal{B}_\varepsilon(\tilde{\mathbf{\Sigma}})} R(\mathbf{A}, \mathbf{\Sigma}) = \inf_{\mathbf{A}\in\mathcal{A},\ \eta\in\mathbb{R}} D(\mathbf{A}, \eta).$$

This proof in Section G.1 extends a known dual formulation for the inner DRO problem

$$\sup_{\mathbf{\Sigma}\in\mathcal{B}_\varepsilon(\tilde{\mathbf{\Sigma}})} L(\mathbf{A}, \mathbf{\Sigma}) = \inf_{\eta > 1+\|\mathbf{A}\|_{\mathrm{op}}^2} \eta\Big(\varepsilon^2 + \mathrm{tr}\Big(\tilde{\mathbf{E}}_{\mathbf{A}}^\top\big((\eta-1)\mathbf{I}_m - \mathbf{A}\mathbf{A}^\top\big)^{-1}\tilde{\mathbf{E}}_{\mathbf{A}}\Big)\Big), \tag{2}$$

which underlies the Frank–Wolfe implementation in Shafieezadeh-Abadeh et al. (2018). In both cases, the variable η is introduced as a Lagrange multiplier for the Bures ball constraint. However, our route to the MRO formula involves an additional maximization problem over rank-p projection matrices. Initially outside the infimum, we employ a non-convex minimax theorem to bring the supremum inside, with the tr_p term arising due to the Ky Fan variational principle.

Theorem 4.1 suggests solving the MRO problem by applying a subgradient method to minimize D over $\mathbf{A}, \eta$. Indeed, subgradients can be computed via eigendecomposition, due to the Ky Fan variational formula for tr_p and Danskin's theorem. Moreover, if $\hat{\mathbf{A}}, \hat{\eta}$ are a feasible, δ-approximate minimizing pair, then $\hat{\mathbf{A}}$ is a δ-approximate MRO solution. We defer a complete convergence analysis to future work, since the minimax rate of MRO is shared by the much simpler SNR, explored next.

In contrast, the SNR objective $G(\mathbf{A}) \coloneqq L(\mathbf{A}, \tilde{\mathbf{\Sigma}}) + \varepsilon^2\|\mathbf{A}\|_{\mathrm{op}}^2$ is convex and enjoys simpler subgradient computations. Indeed, for unit singular vectors $u \in \mathbb{R}^m$, $v \in \mathbb{R}^p$ satisfying $\mathbf{A}v = \|\mathbf{A}\|_{\mathrm{op}}u$, we have[3]

$$2\mathbf{A}\tilde{\mathbf{\Sigma}}_X - 2\tilde{\mathbf{\Sigma}}_{YX} + 2\varepsilon^2\|\mathbf{A}\|_{\mathrm{op}}uv^\top \in \partial G(\mathbf{A}).$$

Some simple bounds on the optimizer norm yield the following for a projected subgradient method.

Proposition 4.2 (SNR subgradient convergence). *Let $\tilde{\mathbf{\Sigma}} \in \mathbb{S}_+^d$ and $\varepsilon > 0$. If $\tilde{\mathbf{\Sigma}}_Y = \mathbf{0}$, then $\mathbf{0}$ minimizes G. Otherwise, define radius $B \coloneqq (d\,\mathrm{tr}\,\tilde{\mathbf{\Sigma}}_Y)^{1/2}/\varepsilon$, Lipschitz bound $M \coloneqq 2(B\|\tilde{\mathbf{\Sigma}}_X\|_{\mathrm{op}} + \|\tilde{\mathbf{\Sigma}}_{YX}\|_{\mathrm{F}} + \varepsilon^2 B)$, and iteration count $T \coloneqq \lceil M^2B^2\delta^{-2}\rceil$, for accuracy $\delta > 0$. Consider the update $\mathbf{A}_{t+1} \leftarrow \mathrm{proj}_B(\mathbf{A}_t - \lambda\mathbf{G}_t)$, where $\mathbf{A}_1 = \mathbf{0}$, $\mathbf{G}_t \in \partial G(\mathbf{A}_t)$, $\lambda \coloneqq B/(M\sqrt{T})$, and proj_B is the Euclidean projection onto the Frobenius ball of radius B centered at $\mathbf{0}$. Then, the averaged iterate $\bar{\mathbf{A}}_T = T^{-1}\sum_{t=1}^T \mathbf{A}_t$ achieves $G(\bar{\mathbf{A}}_T) \le \min_{\mathbf{A}} G(\mathbf{A}) + \delta$.*

The proof in Section G.2 is routine. Hence, we can solve SNR to accuracy δ with $O(\delta^{-2})$ iterations, each computing a top singular vector. To accelerate this rate, we use that our nonsmooth penalty admits an efficiently computable proximal operator, enabling the application of FISTA (Beck and Teboulle, 2009). In Section G.3, we prove an improved rate for the resulting Algorithm 1.[4]

Proposition 4.3 (SNR FISTA convergence). *Fix $\tilde{\mathbf{\Sigma}} \in \mathbb{S}_+^d$, $\varepsilon > 0$, and $\delta > 0$. Then Algorithm 1 achieves $G(\mathbf{A}_T) \le \min_{\mathbf{A}} G(\mathbf{A}) + \delta$ after $T = \lceil(4d\lambda_{\max}(\tilde{\mathbf{\Sigma}}_X)\,\mathrm{tr}(\tilde{\mathbf{\Sigma}}_Y))^{1/2}\varepsilon^{-1}\cdot\delta^{-1/2}\rceil$ iterations.*

In particular, FISTA achieves accuracy δ using $O(\delta^{-1/2})$ full singular value decompositions.

[3] The first two terms for this subgradient representative are obtained directly by differentiating $L(\cdot, \tilde{\mathbf{\Sigma}})$. The last term follows, for example, from the characterization of the spectral-norm subdifferential set in Example 2.8 of Lewis (1995).

[4] Faster linear rates are available if $L(\cdot, \tilde{\mathbf{\Sigma}})$ is strongly convex, but this requires $\lambda_{\min}(\tilde{\mathbf{\Sigma}}_X)$ to be bounded away from 0, which is not guaranteed.

Algorithm 1 FISTA for SNR

Require: Observed second moments $\tilde{\mathbf{\Sigma}}_X \in \mathbb{R}^{p\times p}$, $\tilde{\mathbf{\Sigma}}_{YX} \in \mathbb{R}^{m\times p}$, penalty $\varepsilon > 0$, iteration count T

Compute $M = 2\lambda_{\max}(\tilde{\mathbf{\Sigma}}_X)$
if $M = 0$ **then return** $\mathbf{A}_T = \mathbf{0}$
Set $\alpha = \varepsilon^2/M$, initialize $\mathbf{A}_0 = \mathbf{0}_{m\times p}$, $\mathbf{Y}_0 = \mathbf{A}_0$, $t_0 = 1$
for $k = 0, 1, \ldots, T-1$ **do**
  $\mathbf{B}_k = \mathbf{Y}_k - \frac{2}{M}\big(\mathbf{Y}_k\tilde{\mathbf{\Sigma}}_X - \tilde{\mathbf{\Sigma}}_{YX}\big)$ ▷ gradient step for the smooth loss
  $\mathbf{A}_{k+1} = \texttt{ProxSpecSq}(\mathbf{B}_k, \alpha)$ ▷ proximal step
  $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$
  $\mathbf{Y}_{k+1} = \mathbf{A}_{k+1} + \frac{t_k - 1}{t_{k+1}}(\mathbf{A}_{k+1} - \mathbf{A}_k)$ ▷ Nesterov update
return $\mathbf{A}_T$

Algorithm 2 `ProxSpecSq`: proximal map of $\alpha\|\cdot\|_{\mathrm{op}}^2$

Require: Matrix $\mathbf{B} \in \mathbb{R}^{m\times p}$, parameter $\alpha > 0$

Compute thin SVD $\mathbf{B} = \mathbf{P}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_q)\mathbf{Q}^\top$, $\sigma_1 \geq \cdots \geq \sigma_q \geq 0$, $q = \min\{m, p\}$
if $\sigma_1 = 0$ **then return** $\mathbf{B}$
Define $\sigma_{q+1} = 0$. Find the unique $k_\star \in \{1, \ldots, q\}$ such that

$$\sigma_{k_\star} > \tau_{k_\star} \geq \sigma_{k_\star+1}, \qquad \tau_k = \textstyle\sum_{i=1}^k \sigma_i/(k + 2\alpha).$$

Set $\tau = \tau_{k_\star}$ and define clipped singular values $s_i = \min\{\sigma_i, \tau\}$ for $i \in \{1, \ldots, q\}$
return $\mathbf{P}\,\mathrm{diag}(s_1, \ldots, s_q)\mathbf{Q}^\top$

5 Finite-Sample Statistics

We now turn to the finite-sample setting. Let $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P$ and $P_n := n^{-1}\sum_{i=1}^n \delta_{Z_i}$. We treat two models, each specifying an observed empirical source measure $P_{\mathrm{s},n}$ and an unknown population target P_{t}. In the *sample-corruption model*, $P_{\mathrm{t}} = P$ and $P_{\mathrm{s},n} = n^{-1}\sum_{i=1}^n \delta_{\tilde{Z}_i}$ for an adversarially perturbed sample $\{\tilde{Z}_i\}_{i=1}^n$ only guaranteed to satisfy $n^{-1}\sum_{i=1}^n \|\tilde{Z}_i - Z_i\|^2 \leq \varepsilon^2$. Naturally, this implies $\mathsf{W}_2(P_{\mathrm{s},n}, P_n) \leq \varepsilon$. In the *distribution-shift model*, the learner instead observes $P_{\mathrm{s},n} = P_n$, and $P_{\mathrm{t}} = Q$ for a perturbed population law $Q \in \mathcal{B}_\varepsilon(P)$. Let $\mathbb{M}_{\varepsilon,n}(P)$ denote the family of joint laws of $(P_{\mathrm{s},n}, P_{\mathrm{t}})$ arising from either model, over all admissible, possibly randomized, sample-corruption rules and all fixed $Q \in \mathcal{B}_\varepsilon(P)$. Seeking robust, high-probability regret bounds, we thus define

$$R_{\varepsilon,n,\delta}(\hat{\mathbf{A}}_n, P) := \sup_{\mathbb{P}\in\mathbb{M}_{\varepsilon,n}(P)} \inf\Big\{\tau \geq 0 : \mathbb{P}\Big(R(\hat{\mathbf{A}}_n[P_{\mathrm{s},n}], P_{\mathrm{t}}) \leq \tau\Big) \geq 1 - \delta\Big\},$$

where $\delta \in (0, 1)$ and $\hat{\mathbf{A}}_n$ is any (measurable) estimator mapping $P_{\mathrm{s},n}$ to $\hat{\mathbf{A}}_n[P_{\mathrm{s},n}] \in \mathcal{A}$.

To control empirical convergence of second-moment matrices, we introduce the *Bures variance proxy*, a rescaled fourth-moment bound given by

$$\mathsf{V}_{\mathsf{B}}(P) := \mathbb{E}_P\left[\left\|(ZZ^\top - \mathbf{\Sigma}_P)\mathbf{\Sigma}_P^{\dagger/2}\right\|_{\mathrm{F}}^2\right] = \mathbb{E}_P\big[\|Z\|^2 Z^\top \mathbf{\Sigma}_P^\dagger Z\big] - \mathrm{tr}(\mathbf{\Sigma}_P). \tag{3}$$

Writing $\mathcal{P}_4(\mathbb{R}^d)$ for the set of laws with finite fourth moments, we prove the following in Section H.1.

Proposition 5.1 (V_{B} properties)**.** *Let $P \in \mathcal{P}_4(\mathbb{R}^d)$. Then $\mathsf{V}_{\mathsf{B}}(P) < \infty$, and $\mathsf{V}_{\mathsf{B}}(P) = 0$ if and only if $ZZ^\top = \mathbf{\Sigma}_P$ for almost all $Z \sim P$. If $\mathbf{\Sigma}_P \succ 0$, then $\mathsf{V}_{\mathsf{B}}(P) \leq \mathbb{E}_P[\|Z\|^4]/\lambda_{\min}(\mathbf{\Sigma}_P)$. For Gaussian $P = \mathcal{N}(0, \mathbf{\Sigma}_P)$, we have $\mathsf{V}_{\mathsf{B}}(P) = (\mathrm{rank}(\mathbf{\Sigma}_P) + 1)\,\mathrm{tr}(\mathbf{\Sigma}_P)$, and if further $P \in \mathcal{P}^*$ is non-degenerate, then $\mathsf{V}_{\mathsf{B}}(P) \geq p\,\mathrm{tr}(\mathbf{\Sigma}_E^P)$. Crucially, we have $\mathbb{E}[\mathsf{B}(\mathbf{\Sigma}_P, \mathbf{\Sigma}_{P_n})^2] \leq \mathsf{V}_{\mathsf{B}}(P)/n$.*

Hence, we can extend population-limit results to the finite-sample setting by treating sampling as an adversarial Bures perturbation and slightly inflating ε. However, this approach uses Markov's inequality, leading to sample complexity with inverse polynomial dependence on the failure probability. Alternatively, sub-Gaussian concentration controls the empirical condition number directly,

avoiding radius inflation. For $K > 0$, recall that a random vector W on $\mathbb{R}^d$ is K-sub-Gaussian if $\mathbb{E}[\exp((u^\top W)^2/K^2)] \le 2$ for each unit vector $u \in \mathbb{R}^d$. We prove the following in Section H.2.

Theorem 5.2 (Finite-sample regret bounds)**.** *Let $P \in \mathcal{P}^*$, $\varepsilon > 0$, and $\delta \in (0,1)$. Suppose first that $P \in \mathcal{P}_4(\mathbb{R}^d)$, and fix $V \ge \mathsf{V}_\mathsf{B}(P)$. Then, for sample size $n \ge V\delta^{-1}\varepsilon^{-2}$ and expanded radius $\varepsilon_n \coloneqq \varepsilon + \sqrt{V/(n\delta)}$, we have*

$$R_{\varepsilon,n,\delta}(\mathsf{MRO}_{\varepsilon_n}, P), R_{\varepsilon,n,\delta}(\mathsf{SNR}_{\varepsilon_n}, P) \lesssim \kappa_P \varepsilon^2.$$

Alternatively, let $(X,Y) \sim P$, write $E = Y - \mathbf{A}_P X$, and suppose that the whitened random vector $((\mathbf{\Sigma}_X^P)^{-1/2}X, (\mathbf{\Sigma}_E^P)^{\dagger/2}E)$ is K-sub-Gaussian. Then, for $n \gtrsim K^4(d + \log\frac{1}{\delta})(1 + \mathrm{tr}(\mathbf{\Sigma}_E^P)\varepsilon^{-2})$,

$$R_{\varepsilon,n,\delta}(\mathsf{MRO}_{\varepsilon}, P), R_{\varepsilon,n,\delta}(\mathsf{SNR}_{\varepsilon}, P) \lesssim \kappa_P \varepsilon^2.$$

Notably, the sub-Gaussian guarantee improves the polynomial dependence on δ^{-1} of the first sample-size bound to logarithmic, and it requires no radius inflation. Still, these sample complexities are not comparable in general, and we defer the question of optimal sample complexity to future work.

6 Experiments

Finally, we implement our estimators and compare their performance on synthetic finite-sample instances. All methods use the empirical source second-moment matrix $\mathbf{\Sigma}_{P_{\mathrm{s},n}}$. For ERM, we compute the closed-form solution using NumPy. For MRO, we implement the lifted objective $D(\mathbf{A}, \eta)$ from Theorem 4.1 in PyTorch, evaluating its tr_p term by eigendecomposition and computing subgradients by automatic differentiation. We use the same template for DRO, replacing $D(\mathbf{A}, \eta)$ with the simpler dual objective in (2). For both methods, we use the normalized subgradient method with kth iteration step size $k^{-1/2}$, initialize $\mathbf{A}$ at the ERM solution and $\eta = 2 + \|\mathbf{A}\|_{\mathrm{op}}^2$, and repair feasibility after each step by taking $\eta \leftarrow \max\{\eta, 1 + \|\mathbf{A}\|_{\mathrm{op}}^2 + \zeta\}$ for a numerical buffer $\zeta = 0.001$. We return the final iterate after $T = 800$ iterations. For SNR, we implement the FISTA method in Algorithm 1 using NumPy, with the same ERM initialization and $T = 60$ iterations. The population benchmarks discussed below served as numerical diagnostics for selecting iteration counts and tolerances.

In Fig. 1, we consider two experimental settings based on the Lemma F.10 construction. Both use corruption budget $\varepsilon = 1$ and dimensions $p = m = d/2 \ge 2$. We take target measure $P = P_{\mathrm{t}}$ as the law of (X,Y), where $X \sim \mathcal{N}(0, \mathbf{I}_m)$ and $Y = X + U$, and $U \sim \mathcal{N}(0, \mathrm{diag}(0, \mathbf{I}_{m-1}))$ is independent of X. Then $\mathbf{A}_P = \mathbf{\Sigma}_X^P = \mathbf{I}_m$, $\mathbf{\Sigma}_E^P = \mathrm{diag}(0, \mathbf{I}_{m-1})$, and hence $\kappa_P = 3$. Target population regret is therefore $R(\mathbf{A}, P) = \|\mathbf{A} - \mathbf{I}_m\|_{\mathrm{F}}^2$. Drawing $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n$ i.i.d. from P, we replace each $X_1^{(i)}$ by $m^{-1/4}X_1^{(i)}$, stopping if needed before the first replacement that would exceed a total squared perturbation budget of $n\varepsilon^2$. All other feature coordinates, responses, and unselected observations remain unchanged. Writing $\tilde{X}^{(i)}$ for the resulting features, the observed empirical source $P_{\mathrm{s},n} = n^{-1}\sum_{i=1}^n \delta_{(\tilde{X}^{(i)}, Y^{(i)})}$ satisfies $\mathsf{W}_2(P_{\mathrm{s},n}, P_n)^2 \le \frac{1}{n}\sum_{i=1}^n \|\tilde{X}^{(i)} - X^{(i)}\|^2 \le \varepsilon^2$, obeying our Section 5 sample-corruption model. We use sample size $n = 10m^2$ to match, for fixed confidence, the $d(1 + \mathrm{tr}(\mathbf{\Sigma}_E^P)) \asymp m^2$ dimensional dependence in the sub-Gaussian bound of Theorem 5.2.

We also compute population benchmarks using the source law P_{s} of $(\mathrm{diag}(m^{-1/4}, \mathbf{I}_{m-1})X, Y)$ for $(X,Y) \sim P$, corresponding to contracting every observation. This choice is justified because the average cost of contracting every observation converges almost surely to $(1 - m^{-1/4})^2 < \varepsilon^2$ as $n \to \infty$. The estimators then receive $\mathbf{\Sigma}_{P_{\mathrm{s}}}$ instead of $\mathbf{\Sigma}_{P_{\mathrm{s},n}}$. Appendix I uses rotational symmetry to reduce the population problems to optimization over the two-parameter family $\mathbf{A}(\alpha, \beta) = \mathrm{diag}(\alpha, \beta\mathbf{I}_{m-1})$. Under this reduction, ERM and SNR admit closed-form solutions, while DRO and MRO admit constant-size semidefinite programs which we solve using CVXPY (Diamond and Boyd, 2016; Agrawal et al., 2018). These computations benchmark the estimators at exact population second moments and provide diagnostics for the finite-sample optimization routines. For the population MRO benchmark, we break ties by minimizing the source risk $L(\cdot, P_{\mathrm{s}})$.

For **Setting A**, we supply radius $\rho = \varepsilon = 1$, vary d between 20 and 400, and plot the target population regret $R(\hat{\mathbf{A}}[P_{\mathrm{s},n}], P)$ for $\hat{\mathbf{A}} \in \{\mathsf{ERM}, \mathsf{DRO}_\rho, \mathsf{MRO}_\rho, \mathsf{SNR}_\rho\}$. This setting illustrates the dimension-free behavior of MRO and SNR and $\sqrt{d}$ growth of DRO regret. For **Setting B**, we fix $d = 400$ and vary the supplied radius $\rho \in [0.1\varepsilon, 20\varepsilon]$. For small ρ, DRO, MRO, and SNR perform similarly to

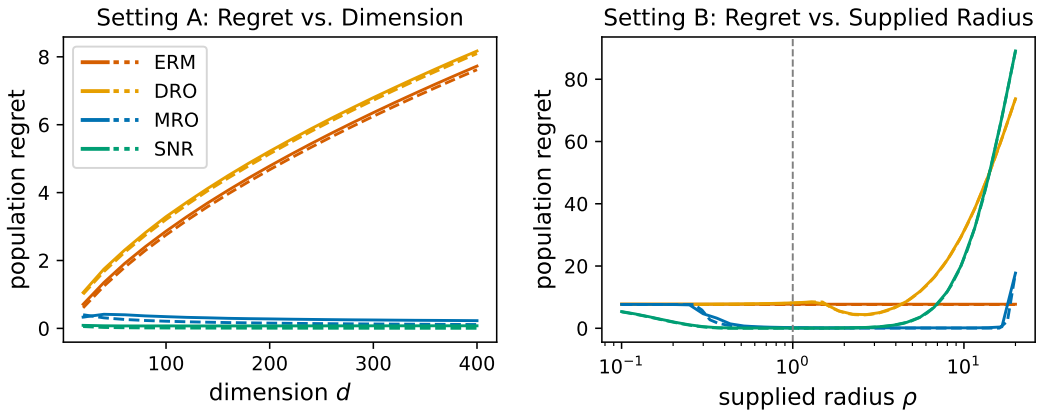


Figure 1: Target population regret $R(\hat{\mathbf{A}}[P_{\mathrm{s},n}], P)$ for Settings A (left) and B (right). Solid curves show finite-sample averages, while dashed curves show population benchmarks. For Setting B, the dashed vertical line marks the corruption budget $\varepsilon = 1$.

ERM, and their relative ranking at $\rho = \varepsilon$ matches Setting A. As the radius ρ varies, the best method depends on its scale; SNR is strongest for radii below the budget ε, MRO is strongest for most larger radii, and for very large radii excessive regularization or robustification causes ERM to dominate. This precise behavior is beyond the scope of our current theoretical results and warrants further study.

For both settings, the plots report target population regret averaged over 10 independent trials, with error bars showing the minimum and maximum over trials, and dashed curves showing the population benchmarks. Error bars are nearly imperceptible due to low variance at the chosen sample size.

See Appendix J for additional details and code, an alternative to Setting A where d is fixed but the effective rank of $\mathbf{\Sigma}_E^P$ is varied, and a final experiment with real-world network traffic data (Koumar et al., 2025). We emphasize that the observed separations between MRO/SNR and ERM/DRO need not hold universally, with ample room remaining for finer-grained instance-dependent analysis.

7 Concluding Remarks

In this paper, we introduced a new framework for regret-optimal Wasserstein-robust decision-making, focused on the setting of least-squares regression. We fully characterized the complexity of robust estimation in the $\varepsilon \to 0$ limit, via the condition number, and extended these asymptotic guarantees to all $\varepsilon > 0$ using regret stability. While DRO is a natural robust algorithm, we proved and verified empirically that it can be outperformed by MRO and our new procedure, SNR, for high dimensional tasks. Moreover, we developed a new convex reformulation for MRO and an efficient proximal gradient method for SNR. Moving forward, we hope that this initial work spurs more research into characterizing and efficiently achieving regret-optimal performance for robust stochastic optimization.

Acknowledgements and AI Usage Statement

This work was supported as a part of NCCR Automation, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40 225155). We also thank the Optimal Transport and Optimization Workshop (OT+OPT '26) attendees for their helpful feedback on an early version of this work.

We used ChatGPT (versions 5-6 Pro) for editing, checking and refining mathematical arguments, and tuning lower bound constructions. Version 5.5 suggested an improvement to Lemma 3.1 that allowed us to extend our analysis from MRO to SNR. It also suggested the argument used in Step 1 of the proof of Theorem 4.1, enabling our convex reformulation of MRO. We have independently verified all mathematical arguments and take full responsibility for the content and correctness of this work.

References

Agarwal, A. and Zhang, T. (2022). Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory (COLT)*.

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. (2018). A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Bhatia, R. (2009). *Positive Definite Matrices*. Princeton University Press.

Bhatia, R., Jain, T., and Lim, Y. (2019). On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191.

Blanchet, J. and Murthy, K. (2019). Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357.

Chao, P. and Dobriban, E. (2023). Minimax statistical estimation under Wasserstein contamination. *arXiv preprint arXiv:2308.01853*.

Chen, Z. and Xie, W. (2021). Regret in the newsvendor model with demand and yield randomness. *Production and Operations Management*, 30(11):4176–4197.

Cho, Y. and Yang, I. (2024). Wasserstein distributionally robust regret minimization. *IEEE Control Systems Letters*, 8:820–825.

Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Fiechtner, L.-B. and Blanchet, J. (2025). Wasserstein distributionally robust regret optimization. *arXiv preprint arXiv:2504.10796*.

Fiechtner, L.-B. and Blanchet, J. (2026). Distributionally robust regret optimal LQR with common stage-law ambiguity. In *Proceedings of the 65th IEEE Conference on Decision and Control (CDC)*. To appear; arXiv:2604.06158, version 3.

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

Gao, R. and Kleywegt, A. (2022). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655.

Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. Cambridge University Press.

Jacobsson, S., Durisi, G., Coldrey, M., Goldstein, T., and Studer, C. (2017). Quantized precoding for massive MU-MIMO. *IEEE Transactions on Communications*, 65(11):4670–4684.

Koumar, J., Hynek, K., Čejka, T., and Šiška, P. (2025). CESNET-TimeSeries24: Time series dataset for network traffic anomaly detection and forecasting. *Scientific Data*, 12(1):338.

Kuhn, D., Mohajerin Esfahani, P., Nguyen, V. A., and Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, INFORMS TutORials in Operations Research, pages 130–166. INFORMS.

Kuhn, D., Shafiee, S., and Wiesemann, W. (2025). Distributionally robust optimization. *Acta Numerica*, 34:579–804.

Lee, J. and Raginsky, M. (2018). Minimax statistical learning with Wasserstein distances. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lewis, A. S. (1995). The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1–2):173–183.

Lin, F., Fang, X., and Gao, Z. (2022). Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization*, 12(1):159–212.

Liu, Z. and Loh, P.-L. (2023). Robust W-GAN-based estimation under Wasserstein contamination. *Information and Inference*, 12(1):312–362.

Nguyen, V. A., Kuhn, D., and Mohajerin Esfahani, P. (2022). Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *Operations Research*, 70(1):490–515.

Nguyen, V. A., Shafiee, S., Filipović, D., and Kuhn, D. (2021a). Mean-covariance robust risk measurement. *arXiv preprint arXiv:2112.09959*.

Nguyen, V. A., Shafieezadeh-Abadeh, S., Kuhn, D., and Mohajerin Esfahani, P. (2021b). Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization. *Mathematics of Operations Research*, 48(1):1–37.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239.

Perakis, G. and Roels, G. (2008). Regret in the newsvendor model with partial information. *Operations Research*, 56(1):188–203.

Petersen, K. B. and Pedersen, M. S. (2012). The Matrix Cookbook. Technical report, Technical University of Denmark.

Pólik, I. and Terlaky, T. (2007). A survey of the S-lemma. *SIAM Review*, 49(3):371–418.

Rahimian, H. and Mehrotra, S. (2022). Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85. Article 4.

Ricceri, B. (2019). Minimax theorems in a fully non-convex setting. *Journal of Nonlinear and Variational Analysis*, 3(1):45–52.

Shafieezadeh-Abadeh, S., Nguyen, V. A., Kuhn, D., and Mohajerin Esfahani, P. (2018). Wasserstein distributionally robust Kalman filtering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Sinha, A., Namkoong, H., and Duchi, J. C. (2018). Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*.

Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176.

Taşkesen, B., Iancu, D. A., Koçyiğit, Ç., and Kuhn, D. (2023). Distributionally robust linear quadratic control. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Vershynin, R. (2026). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2 edition.

Wedin, P.-Å. (1972). Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111.

Zhu, B., Jiao, J., and Steinhardt, J. (2022). Generalized resilience and robust statistics. *The Annals of Statistics*, 50(4):2256–2283.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We accuracy summarized our contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, our work is theoretical and we carefully specify all needed assumptions.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Our appendix provides complete proofs for all results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our implementation follows the provided pseudo-code and we fully specify our experimental settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All code is provided in the supplement.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All such details are given in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All plots include error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: These details are specified in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the ethics code.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a theoretical work, and we envision no major societal impact beyond those for general statistics theory.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the paper associated with the one external implementation we use. All data is synthetic.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our supplement code directly follows pseudo-code and experimental details specified in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: No human subjects were involved.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No LLM usage in this capacity.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Appendix Contents

A Linear Algebra Preliminaries **21**

A.1 Basic Matrix Inequalities and Identities 21

A.2 Scalar, Tensor, and Spectral Variational Facts 23

B Bures–Wasserstein Geometry and Regression **24**

B.1 Common Block Decompositions 24

B.2 Regression in Bures Space 24

B.3 Bures–Wasserstein Geometry 26

C DRO Duality and Symmetry Reductions **28**

C.1 Strong Duality for the Inner DRO Problem 28

C.2 Averaging over Symmetries 30

D Remaining Proofs for Section 1 **32**

D.1 Existence of Minimizers (Footnote 1) 32

E Remaining Proofs for Section 2 **32**

E.1 Condition Number Formula (Proof of Theorem 2.2) 32

F Remaining Proofs for Section 3 **36**

F.1 Regret Transfer (Proof of Lemma 3.1) 36

F.2 Stability Implications for ERM and DRO (Remaining Proof for Lemma 3.2) 38

F.3 Rank-Adaptive DRO Bound (Proof of Lemma 3.3) 41

F.4 Main Population-Limit Regret Bounds (Proof of Theorem 3.4) 42

G Remaining Proofs for Section 4 **48**

G.1 Convex Reformulation for MRO (Proof of Theorem 4.1) 48

G.2 Subgradient Method Convergence for SNR (Proof of Proposition 4.2) 52

G.3 FISTA Convergence for SNR (Proof of Proposition 4.3) 53

H Remaining Proofs for Section 5 **53**

H.1 Properties of the Bures Variance Proxy (Proof of Proposition 5.1) 53

H.2 Finite-Sample Regret Bounds (Proof of Theorem 5.2) 55

I Exact Reductions for Population Benchmarks in Section 6 **59**

I.1 Reduction to Scalar Optimization Using Rotational Symmetry 59

I.2 Constant-Size SDP Reformulations for DRO and MRO 59

I.3 Closed form for SNR 62

J Full Experimental Details and Additional Experiments **63**

J.1 Additional Synthetic Experiment with Varied Residual Rank 63

J.2 Network-Traffic Experiment with CESNET Dataset 64

A Linear Algebra Preliminaries

This appendix collects linear algebraic tools used repeatedly in the proofs. We include short derivations when they are more convenient than a citation, and otherwise point to standard references.

Additional notation. Write $\mathbb{S}^n$ for the space of symmetric $n \times n$ matrices. For a matrix $\mathbf{M} \in \mathbb{R}^{a\times b}$, let $\mathrm{range}(\mathbf{M})$ denote its column space, so that $\mathrm{range}(\mathbf{M}^\top)$ is its row space, and write $\mathrm{rank}(\mathbf{M}) \coloneqq \dim \mathrm{range}(\mathbf{M})$. Write $\mathrm{vec}(\mathbf{M}) \in \mathbb{R}^{ab}$ for its column-stacked vectorization and $\otimes$ for the Kronecker product. For matrices of the same size, we write $\langle \mathbf{M}, \mathbf{N}\rangle_{\mathrm{F}} \coloneqq \mathrm{tr}(\mathbf{M}^\top\mathbf{N})$ for their Frobenius inner product. We order the singular values of $\mathbf{M} \in \mathbb{R}^{a\times b}$ as $\sigma_1(\mathbf{M}) \geq \cdots \geq \sigma_{\min\{a,b\}}(\mathbf{M}) \geq 0$, including zeros, and write $\sigma_{\min}(\mathbf{M}) \coloneqq \sigma_{\min\{a,b\}}(\mathbf{M})$. For $\mathbf{H} \in \mathbb{S}^d$ and $k \in \{1,\dots,d\}$, let $\mathrm{tr}_k(\mathbf{H})$ denote the sum of its k largest eigenvalues and write $\mathbf{H}_+$ for its positive semidefinite part, obtained by replacing negative eigenvalues by zero. Write $(t)_+ \coloneqq \max\{t, 0\}$ for scalars. We write 0_n for the zero vector in $\mathbb{R}^n$ and e_i for the ith standard basis vector, with its ambient dimension specified when needed.

A.1 Basic Matrix Inequalities and Identities

Lemma A.1 (Basic norm and projection inequalities)**.** *For dimension-compatible matrices,*

$$\begin{aligned}\|\mathbf{ABC}\|_{\mathrm{F}} &\leq \|\mathbf{A}\|_{\mathrm{op}}\,\|\mathbf{B}\|_{\mathrm{F}}\,\|\mathbf{C}\|_{\mathrm{op}},\\ \|\mathbf{AB}\|_{\mathrm{op}} &\leq \|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{B}\|_{\mathrm{op}},\\ \|\mathbf{AB}\|_{\mathrm{F}} &\leq \|\mathbf{A}\|_{\mathrm{op}}\|\mathbf{B}\|_{\mathrm{F}},\\ \|\mathbf{A}\|_{\mathrm{op}} \leq \|\mathbf{A}\|_{\mathrm{F}}, &\qquad \|\mathbf{A}\|_{\mathrm{F}} \leq \sqrt{\mathrm{rank}(\mathbf{A})}\,\|\mathbf{A}\|_{\mathrm{op}}.\end{aligned}$$

If $\mathbf{\Pi}$ is an orthogonal projection, then

$$\|\mathbf{M\Pi}\|_{\mathrm{F}} \leq \|\mathbf{M}\|_{\mathrm{F}}, \qquad \|\mathbf{M}\|_{\mathrm{F}}^2 = \|\mathbf{M\Pi}\|_{\mathrm{F}}^2 + \|\mathbf{M}(\mathbf{I}-\mathbf{\Pi})\|_{\mathrm{F}}^2.$$

Finally,

$$|\langle \mathbf{A}, \mathbf{B}\rangle_{\mathrm{F}}| \leq \|\mathbf{A}\|_{\mathrm{F}}\|\mathbf{B}\|_{\mathrm{F}}.$$

Proof. The first inequality holds for all unitarily invariant matrix norms, including the Frobenius norm; see, e.g., Horn and Johnson, 1991, Section 3.5, Exercise 4. The second is a standard submultiplicativity property; see, e.g., Horn and Johnson, 1991, Section 3.5, Exercise 3. The third inequality follows from the first by taking $\mathbf{C} = \mathbf{I}$. The comparisons between Frobenius and operator norms follow from the singular-value identities $\|\mathbf{A}\|_{\mathrm{op}} = \sigma_1(\mathbf{A})$ and $\|\mathbf{A}\|_{\mathrm{F}}^2 = \sum_i \sigma_i(\mathbf{A})^2$. The projection identities follow from $\mathbf{\Pi}(\mathbf{I}-\mathbf{\Pi}) = \mathbf{0}$ and the Pythagorean theorem for the Frobenius inner product. The last inequality is Cauchy–Schwarz for the Frobenius inner product. □

Lemma A.2 (Pseudoinverse and projection identities)**.** *Let $\mathbf{M} \in \mathbb{R}^{r\times d}$. Then $\mathbf{M}^\dagger\mathbf{M}$ is the orthogonal projection onto the row space of $\mathbf{M}$, and $\mathbf{MM}^\dagger$ is the orthogonal projection onto the column space of $\mathbf{M}$. We always have $\mathbf{M}^\dagger = \mathbf{M}^\top(\mathbf{MM}^\top)^\dagger$. If $\mathbf{M}$ has full row rank, then*

$$\mathbf{M}^\dagger = \mathbf{M}^\top(\mathbf{MM}^\top)^{-1}, \qquad \mathbf{MM}^\dagger = \mathbf{I}_r, \qquad \|\mathbf{M}^\dagger\|_{\mathrm{op}} = \sigma_{\min}(\mathbf{M})^{-1}.$$

If $\mathbf{B} \succeq \mathbf{0}$, then $\mathbf{B}^\dagger\mathbf{B}^2 = \mathbf{B}$.

Proof. Take a compact singular value decomposition $\mathbf{M} = \mathbf{UDV}^\top$ over its nonzero singular values. Then $\mathbf{M}^\dagger = \mathbf{VD}^{-1}\mathbf{U}^\top$, so $\mathbf{M}^\dagger\mathbf{M} = \mathbf{VV}^\top$, $\mathbf{MM}^\dagger = \mathbf{UU}^\top$. These are exactly the claimed row-space and column-space projections. Further, $\mathbf{MM}^\top = \mathbf{UD}^2\mathbf{U}^\top$, and so

$$\mathbf{M}^\top(\mathbf{MM}^\top)^\dagger = \mathbf{VDU}^\top\big(\mathbf{UD}^{-2}\mathbf{U}^\top\big) = \mathbf{VD}^{-1}\mathbf{U}^\top = \mathbf{M}^\dagger$$

If $\mathbf{M}$ has full row rank, then $\mathbf{MM}^\top$ is invertible and the displayed formulas follow directly. The claim for $\mathbf{B} \succeq \mathbf{0}$ follows from the eigenvalue decomposition of $\mathbf{B}$. □

Lemma A.3 (Schur complement)**.** *Let*

$$\mathbf{Q} = \begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^\top & \mathbf{C}\end{pmatrix}$$

be symmetric. If $\mathbf{A} \succ \mathbf{0}$*, then*

$$\mathbf{Q} \succeq \mathbf{0} \quad \iff \quad \mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1} \mathbf{B} \succeq \mathbf{0}.$$

The analogous statement holds with $\mathbf{C} \succ \mathbf{0}$ *and Schur complement* $\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^\top$.

Proof. Use the congruence

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{B}^\top \mathbf{A}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\mathbf{A}^{-1}\mathbf{B} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} - \mathbf{B}^\top \mathbf{A}^{-1} \mathbf{B} \end{pmatrix}.$$

Congruence by an invertible matrix preserves positive (semi)definiteness. □

Lemma A.4 (Rectangular resolvent identities)**.** *Let* $\mathbf{M} \in \mathbb{R}^{r \times d}$ *and* $\eta > \|\mathbf{M}\|_{\mathrm{op}}^2$*. Then*

$$\begin{aligned} (\eta \mathbf{I}_d - \mathbf{M}^\top \mathbf{M})^{-1} \mathbf{M}^\top &= \mathbf{M}^\top (\eta \mathbf{I}_r - \mathbf{M}\mathbf{M}^\top)^{-1}, \\ \mathbf{M}(\eta \mathbf{I}_d - \mathbf{M}^\top \mathbf{M})^{-1} &= (\eta \mathbf{I}_r - \mathbf{M}\mathbf{M}^\top)^{-1} \mathbf{M}, \\ \mathbf{M}^\top \mathbf{M}(\eta \mathbf{I}_d - \mathbf{M}^\top \mathbf{M})^{-1} &= \mathbf{M}^\top (\eta \mathbf{I}_r - \mathbf{M}\mathbf{M}^\top)^{-1} \mathbf{M}, \\ \eta(\eta \mathbf{I}_d - \mathbf{M}^\top \mathbf{M})^{-1} &= \mathbf{I}_d + \mathbf{M}^\top (\eta \mathbf{I}_r - \mathbf{M}\mathbf{M}^\top)^{-1} \mathbf{M}. \end{aligned}$$

Proof. We start by computing

$$\mathbf{M}^\top (\eta \mathbf{I}_r - \mathbf{M}\mathbf{M}^\top) = (\eta \mathbf{I}_d - \mathbf{M}^\top \mathbf{M}) \mathbf{M}^\top.$$

The terms in parentheses are both invertible due to our lower bound on η, so suitable left/right multiplication gives the first identity. The second identity follows by taking transposes, and the third follows by left multiplication of the second identity by $\mathbf{M}^\top$. The fourth identity follows by combining the third with the identity $\eta(\eta \mathbf{I} - \mathbf{C})^{-1} = \mathbf{I} + \mathbf{C}(\eta \mathbf{I} - \mathbf{C})^{-1}$, which can be confirmed by multiplying both expressions on the right by $\eta \mathbf{I} - \mathbf{C}$. □

Lemma A.5 (Lipschitz continuity of the smallest singular value)**.** *Let* $\mathbf{S}, \bar{\mathbf{S}} \in \mathbb{R}^{r \times d}$*. Then*

$$|\sigma_{\min}(\bar{\mathbf{S}}) - \sigma_{\min}(\mathbf{S})| \leq \|\bar{\mathbf{S}} - \mathbf{S}\|_{\mathrm{op}}.$$

Proof. This is the standard singular-value perturbation inequality; see, e.g., Theorem 3.3.16(c) of Horn and Johnson (1991). □

Lemma A.6 (Row-space projection perturbation)**.** *Let* $\mathbf{M}_0 \in \mathbb{R}^{r \times d}$ *have full row rank, and let* $\mathbf{M}_1 = \mathbf{M}_0 + \mathbf{\Delta}$ *for some* $\mathbf{\Delta} \in \mathbb{R}^{r \times d}$*. Let* $\mathbf{\Pi}_0, \mathbf{\Pi}_1 \in \mathbb{R}^{d \times d}$ *denote orthogonal projections onto the row spaces of* $\mathbf{M}_0, \mathbf{M}_1$*, respectively. Then,*

$$\|(\mathbf{I}_d - \mathbf{\Pi}_0)\mathbf{\Pi}_1\|_{\mathrm{F}} \leq \|(\mathbf{I}_d - \mathbf{\Pi}_1)\mathbf{\Pi}_0\|_{\mathrm{F}} \leq \frac{\|\mathbf{\Delta}\|_{\mathrm{F}}}{\sigma_{\min}(\mathbf{M}_0)}.$$

Proof. Although we have not seen this precise statement in the literature, the left-hand side connects to the principal angles between row spaces and has been well examined, with the closely related Davis–Kahan–Wedin sine theorem providing a similar bound (see, e.g., Section 3 of Wedin, 1972).

For the proof, define the complementary projections $\mathbf{\Pi}_i^\perp := \mathbf{I}_d - \mathbf{\Pi}_i$ for $i \in \{0, 1\}$. To prove the first inequality, we compute

$$\begin{aligned} \|\mathbf{\Pi}_0^\perp \mathbf{\Pi}_1\|_{\mathrm{F}}^2 &= \mathrm{tr}((\mathbf{I}_d - \mathbf{\Pi}_0)\mathbf{\Pi}_1) && (\mathbf{\Pi}_0, \mathbf{\Pi}_1 \text{ projections}) \\ &= \mathrm{tr}(\mathbf{\Pi}_1) - \mathrm{tr}(\mathbf{\Pi}_0 \mathbf{\Pi}_1) \\ &\leq \mathrm{tr}(\mathbf{\Pi}_0) - \mathrm{tr}(\mathbf{\Pi}_0 \mathbf{\Pi}_1) \\ &= \|\mathbf{\Pi}_1^\perp \mathbf{\Pi}_0\|_{\mathrm{F}}^2. \end{aligned}$$

The inequality uses that $\mathrm{tr}(\mathbf{\Pi}_1) \leq r = \mathrm{tr}(\mathbf{\Pi}_0)$, since $\mathbf{M}_0$ has full row rank. Thus, it suffices to bound $\|\mathbf{\Pi}_1^\perp \mathbf{\Pi}_0\|_{\mathrm{F}}^2$.

Next, since $\mathbf{M}_0$ is full row rank, we can take a thin singular value decomposition

$$\mathbf{M}_0 = \mathbf{U}\mathbf{D}\mathbf{V}^\top,$$

where $\mathbf{U} \in \mathbb{O}^r$ is orthogonal, $\mathbf{D} \in \mathbb{R}^{r\times r}$ is diagonal with strictly positive entries, and $\mathbf{V} \in \mathbb{R}^{d\times r}$ has orthonormal columns. By this choice, $\mathbf{V} = \mathbf{M}_0^\top \mathbf{U}\mathbf{D}^{-1}$ and $\mathbf{\Pi}_0 = \mathbf{V}\mathbf{V}^\top$, implying that

$$\|\mathbf{\Pi}_1^\perp \mathbf{\Pi}_0\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{V}^\top \mathbf{\Pi}_1^\perp \mathbf{V}) = \|\mathbf{\Pi}_1^\perp \mathbf{V}\|_{\mathrm{F}}^2.$$

Since $\mathbf{\Pi}_1^\perp$ is the orthogonal projection onto the orthogonal complement of the row space $R_1 \subseteq \mathbb{R}^d$ of $\mathbf{M}_1$, we have

$$\|\mathbf{\Pi}_1^\perp \mathbf{V}\|_{\mathrm{F}} = \min\big\{\|\mathbf{V} - \mathbf{W}\|_{\mathrm{F}} : \text{columns of } \mathbf{W} \text{ lie in } R_1\big\}.$$

Consider the specific competitor $\mathbf{W} \coloneqq \mathbf{M}_1^\top \mathbf{U}\mathbf{D}^{-1}$, whose columns lie in R_1. We then have

$$\mathbf{V} - \mathbf{W} = (\mathbf{M}_0^\top - \mathbf{M}_1^\top)\mathbf{U}\mathbf{D}^{-1} = -\mathbf{\Delta}^\top \mathbf{U}\mathbf{D}^{-1}.$$

Therefore,

$$\|\mathbf{\Pi}_1^\perp \mathbf{V}\|_{\mathrm{F}} \le \|\mathbf{\Delta}^\top \mathbf{U}\mathbf{D}^{-1}\|_{\mathrm{F}} \le \|\mathbf{\Delta}\|_{\mathrm{F}} \|\mathbf{D}^{-1}\|_{\mathrm{op}} = \frac{\|\mathbf{\Delta}\|_{\mathrm{F}}}{\sigma_{\min}(\mathbf{M}_0)}.$$

Combining the steps above, we obtain

$$\|(\mathbf{I}_d - \mathbf{\Pi}_0)\mathbf{\Pi}_1\|_{\mathrm{F}} = \|\mathbf{\Pi}_0^\perp \mathbf{\Pi}_1\|_{\mathrm{F}} \le \|\mathbf{\Pi}_1^\perp \mathbf{\Pi}_0\|_{\mathrm{F}} = \|\mathbf{\Pi}_1^\perp \mathbf{V}\|_{\mathrm{F}} \le \frac{\|\mathbf{\Delta}\|_{\mathrm{F}}}{\sigma_{\min}(\mathbf{M}_0)},$$

as desired. □

A.2 Scalar, Tensor, and Spectral Variational Facts

Lemma A.7 (Elementary scalar identities)**.** *For $a_1, \dots, a_k \ge 0$,*

$$\left(\sum_{i=1}^k a_i\right)^2 \le k \sum_{i=1}^k a_i^2.$$

For $A, B \ge 0$,

$$\inf_{\delta>0} (A+\delta)\left(1 + \frac{B}{\delta}\right) = (\sqrt{A} + \sqrt{B})^2,$$

with optimizer $\delta = \sqrt{AB}$ when $AB > 0$. Finally, for any norm $\|\cdot\|$, we have

$$\inf_x \max\{\|x - x_0\|^2, \|x - x_1\|^2\} = \frac{1}{4}\|x_0 - x_1\|^2.$$

Proof. The first inequality is Cauchy–Schwarz. For the second identity, expand

$$(A+\delta)\left(1 + \frac{B}{\delta}\right) = A + B + \delta + \frac{AB}{\delta}.$$

Basic calculus confirms that this is minimized at $\delta = \sqrt{AB}$ if $AB > 0$ or otherwise in the limit as $\delta \to 0$. In both cases, the infimum value is $(\sqrt{A} + \sqrt{B})^2$, as claimed.

The last identity follows because any x with both distances at most r must satisfy $\|x_0 - x_1\| \le 2r$, while the midpoint achieves radius $\|x_0 - x_1\|/2$. □

Lemma A.8 (Kronecker trace identity)**.** *For matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ with dimensions such that the following products are well-defined,*

$$\mathrm{tr}(\mathbf{A}\mathbf{C}\mathbf{B}\mathbf{C}^\top) = \mathrm{vec}(\mathbf{C})^\top (\mathbf{B}^\top \otimes \mathbf{A})\,\mathrm{vec}(\mathbf{C}).$$

Proof. This follows from the matrix vectorization identities $\mathrm{vec}(\mathbf{A}\mathbf{C}\mathbf{B}) = (\mathbf{B}^\top \otimes \mathbf{A})\,\mathrm{vec}(\mathbf{C})$ and $\mathrm{tr}(\mathbf{U}\mathbf{V}^\top) = \mathrm{vec}(\mathbf{V})^\top \mathrm{vec}(\mathbf{U})$; see, e.g., Petersen and Pedersen, 2012, Section 10.2. □

Lemma A.9 (Variational matrix formulas)**.** *For any matrix* $\mathbf{A}$*,*

$$\|\mathbf{A}\|_{\mathrm{F}} = \sup_{\|\mathbf{B}\|_{\mathrm{F}} \le 1} \langle \mathbf{A}, \mathbf{B} \rangle_{\mathrm{F}}.$$

For any $\mathbf{H} \in \mathbb{S}^d$ *and* $k \in \{1, \dots, d\}$*, we have the Ky Fan variational formula*

$$\mathrm{tr}_k(\mathbf{H}) = \max_{\substack{\mathbf{\Pi}^2 = \mathbf{\Pi} = \mathbf{\Pi}^\top \\ \mathrm{tr}(\mathbf{\Pi}) = k}} \mathrm{tr}(\mathbf{H}\mathbf{\Pi}),$$

and equivalently

$$\mathrm{tr}_k(\mathbf{H}) = \min_{\tau \in \mathbb{R},\, \mathbf{Y} \succeq \mathbf{0},\, \mathbf{Y} \succeq \mathbf{H} - \tau \mathbf{I}_d} k\tau + \mathrm{tr}(\mathbf{Y}).$$

Proof. The Frobenius formula simply is Hilbert space duality. The projection formula is Ky Fan's maximum principle; see, e.g., Horn and Johnson, 1991, Section 3.4. For the epigraph formula, diagonalize $\mathbf{H} = \sum_{i=1}^d \lambda_i u_i u_i^\top$, for an orthonormal eigenbasis $\{u_i\}_{i=1}^d$. For fixed τ, any feasible $\mathbf{Y}$ satisfies $u_i^\top \mathbf{Y} u_i \ge \max\{\lambda_i - \tau, 0\}$ for all i. Hence,

$$\mathrm{tr}(\mathbf{Y}) \ge \sum_{i=1}^d (\lambda_i - \tau)_+.$$

Equality is attained by $\mathbf{Y} = (\mathbf{H} - \tau \mathbf{I}_d)_+$. Thus the inner minimum equals

$$k\tau + \sum_{i=1}^d (\lambda_i - \tau)_+,$$

whose minimum over τ is the sum of the k largest eigenvalues of $\mathbf{H}$. □

B Bures–Wasserstein Geometry and Regression

This appendix develops the second-moment formulation of least-squares regression in Bures space and collects some basic facts of Bures–Wasserstein geometry used throughout the proofs.

B.1 Common Block Decompositions

Throughout the proofs, we frequently apply two types of matrix block decompositions. Recall that $d = p + m$. If $\mathbf{\Sigma} \in \mathbb{S}^d$ is symmetric, we write

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_X & \mathbf{\Sigma}_{XY} \\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_Y \end{pmatrix}, \qquad \mathbf{\Sigma}_X \in \mathbb{S}^p, \quad \mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_{YX}^\top \in \mathbb{R}^{p \times m}, \quad \mathbf{\Sigma}_Y \in \mathbb{S}^m.$$

This matches our existing notation for second moments when $\mathbf{\Sigma} = \mathbf{\Sigma}_P$. We employ this decomposition when $\mathbf{\Sigma}$ is a second-moment matrix, a nearby symmetric matrix, or a tangent direction to the positive semidefinite cone.

If $\mathbf{S} \in \mathbb{R}^{d \times k}$ is introduced as a square-root factor for a matrix $\mathbf{\Sigma} \in \mathbb{S}_+^d$, i.e., $\mathbf{\Sigma} = \mathbf{S}\mathbf{S}^\top$, or as the difference between two such factors, then we often decompose

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_X \\ \mathbf{S}_Y \end{pmatrix}, \qquad \mathbf{S}_X \in \mathbb{R}^{p \times k}, \quad \mathbf{S}_Y \in \mathbb{R}^{m \times k}.$$

This decomposition arises naturally for Bures distance computations. For intuition when working with such blocks, note that if $Z = (X, Y) = (\mathbf{S}_X W, \mathbf{S}_Y W)$ for $W \sim \mathcal{N}(0, \mathbf{I}_k)$, then $\mathbb{E}[ZZ^\top] = \mathbf{S}\mathbf{S}^\top$.

B.2 Regression in Bures Space

Since all risks considered in the paper depend only on second moments, we extend the basic definitions from the main body to Bures space. For $\mathbf{\Sigma} \in \mathbb{S}_+^d$ and $\mathbf{A} \in \mathcal{A} = \mathbb{R}^{m \times p}$, define

$$\begin{aligned} \mathbf{A}_{\mathbf{\Sigma}} &\coloneqq \mathbf{\Sigma}_{YX} \mathbf{\Sigma}_X^\dagger, \\ L(\mathbf{A}, \mathbf{\Sigma}) &\coloneqq \mathrm{tr}(\mathbf{A}^\top \mathbf{A} \mathbf{\Sigma}_X) - 2\,\mathrm{tr}(\mathbf{A}\mathbf{\Sigma}_{XY}) + \mathrm{tr}(\mathbf{\Sigma}_Y), \\ R(\mathbf{A}, \mathbf{\Sigma}) &\coloneqq L(\mathbf{A}, \mathbf{\Sigma}) - L(\mathbf{A}_{\mathbf{\Sigma}}, \mathbf{\Sigma}), \\ \mathbf{\Sigma}_E &\coloneqq \begin{pmatrix} \mathbf{A}_{\mathbf{\Sigma}} & -\mathbf{I}_m \end{pmatrix} \mathbf{\Sigma} \begin{pmatrix} \mathbf{A}_{\mathbf{\Sigma}} & -\mathbf{I}_m \end{pmatrix}^\top. \end{aligned}$$

For $P \in \mathcal{P}^*$, these agree with the distributional definitions in the main body when $\mathbf{\Sigma} = \mathbf{\Sigma}_P$. We note that the residual second moment can be simplified as follows:

$$\begin{aligned}\mathbf{\Sigma}_E &= (\mathbf{A}_{\mathbf{\Sigma}} \quad -\mathbf{I}_m)\begin{pmatrix}\mathbf{\Sigma}_X & \mathbf{\Sigma}_{XY}\\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_Y\end{pmatrix}(\mathbf{A}_{\mathbf{\Sigma}} \quad -\mathbf{I}_m)^\top \\ &= \mathbf{A}_{\mathbf{\Sigma}}\mathbf{\Sigma}_X\mathbf{A}_{\mathbf{\Sigma}}^\top - \mathbf{A}_{\mathbf{\Sigma}}\mathbf{\Sigma}_{XY} - \mathbf{\Sigma}_{YX}\mathbf{A}_{\mathbf{\Sigma}}^\top + \mathbf{\Sigma}_Y \\ &= \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_X^\dagger\mathbf{\Sigma}_X\mathbf{\Sigma}_X^\dagger\mathbf{\Sigma}_{XY} - 2\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_X^\dagger\mathbf{\Sigma}_{XY} + \mathbf{\Sigma}_Y \\ &= \mathbf{\Sigma}_Y - \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_X^\dagger\mathbf{\Sigma}_{XY},\end{aligned}$$

where the last equality uses that $\mathbf{\Sigma}_X^\dagger\mathbf{\Sigma}_X\mathbf{\Sigma}_X^\dagger = \mathbf{\Sigma}_X^\dagger$.

The next lemma records several useful identities for these Bures-space regression quantities in terms of a square-root factor for $\mathbf{\Sigma}$.

Lemma B.1 (Regression factor identities and regret formula)**.** *Let $\mathbf{\Sigma} \in \mathbb{S}_+^d$, and fix any $\mathbf{S} \in \mathbb{R}^{d\times r}$ such that $\mathbf{\Sigma} = \mathbf{S}\mathbf{S}^\top$. Let $\mathbf{\Pi}_X = \mathbf{S}_X^\dagger\mathbf{S}_X$ be the orthogonal projection onto the row space of $\mathbf{S}_X$. Then*

$$\mathbf{A}_{\mathbf{\Sigma}} = \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_X^\dagger = \mathbf{S}_Y\mathbf{S}_X^\dagger.$$

For the chosen factor $\mathbf{S}$, the corresponding residual factor

$$\mathbf{E} \coloneqq \mathbf{S}_Y - \mathbf{A}_{\mathbf{\Sigma}}\mathbf{S}_X = \mathbf{S}_Y(\mathbf{I} - \mathbf{\Pi}_X)$$

satisfies

$$\mathbf{E}\mathbf{S}_X^\top = \mathbf{0}, \qquad \mathbf{E}\mathbf{\Pi}_X = \mathbf{0}, \qquad \mathbf{E}\mathbf{E}^\top = \mathbf{\Sigma}_E.$$

Finally, for every $\mathbf{A} \in \mathcal{A}$, we have $L(\mathbf{A}, \mathbf{\Sigma}) = \|\mathbf{A}\mathbf{S}_X - \mathbf{S}_Y\|_{\mathrm{F}}^2$ and

$$R(\mathbf{A}, \mathbf{\Sigma}) = \|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}_X\|_{\mathrm{F}}^2 = \|(\mathbf{A} - \mathbf{A}_{\mathbf{\Sigma}})\mathbf{S}_X\|_{\mathrm{F}}^2.$$

Proof. By Lemma A.2, $\mathbf{S}_X^\dagger = \mathbf{S}_X^\top(\mathbf{S}_X\mathbf{S}_X^\top)^\dagger$, and so

$$\mathbf{S}_Y\mathbf{S}_X^\dagger = \mathbf{S}_Y\mathbf{S}_X^\top(\mathbf{S}_X\mathbf{S}_X^\top)^\dagger = \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_X^\dagger = \mathbf{A}_{\mathbf{\Sigma}}.$$

Thus, $\mathbf{A}_{\mathbf{\Sigma}}\mathbf{S}_X = \mathbf{S}_Y\mathbf{S}_X^\dagger\mathbf{S}_X = \mathbf{S}_Y\mathbf{\Pi}_X$, and hence $\mathbf{E} = \mathbf{S}_Y(\mathbf{I} - \mathbf{\Pi}_X)$. Since $\mathbf{\Pi}_X\mathbf{S}_X^\top = \mathbf{S}_X^\top$ and $(\mathbf{I} - \mathbf{\Pi}_X)\mathbf{\Pi}_X = \mathbf{0}$, we obtain

$$\mathbf{E}\mathbf{S}_X^\top = \mathbf{0}, \qquad \mathbf{E}\mathbf{\Pi}_X = \mathbf{0}.$$

Moreover, $(\mathbf{A}_{\mathbf{\Sigma}} \quad -\mathbf{I}_m)\,\mathbf{S} = -\mathbf{E}$, and so

$$\mathbf{E}\mathbf{E}^\top = (\mathbf{A}_{\mathbf{\Sigma}} \quad -\mathbf{I}_m)\mathbf{\Sigma}(\mathbf{A}_{\mathbf{\Sigma}} \quad -\mathbf{I}_m)^\top = \mathbf{\Sigma}_E.$$

Finally,

$$L(\mathbf{A}, \mathbf{\Sigma}) = \mathrm{tr}((\mathbf{A}\mathbf{S}_X - \mathbf{S}_Y)(\mathbf{A}\mathbf{S}_X - \mathbf{S}_Y)^\top) = \|\mathbf{A}\mathbf{S}_X - \mathbf{S}_Y\|_{\mathrm{F}}^2.$$

We decompose this norm using $\mathbf{\Pi}_X$ and $\mathbf{I} - \mathbf{\Pi}_X$:

$$\|\mathbf{A}\mathbf{S}_X - \mathbf{S}_Y\|_{\mathrm{F}}^2 = \|(\mathbf{A}\mathbf{S}_X - \mathbf{S}_Y)\mathbf{\Pi}_X\|_{\mathrm{F}}^2 + \|\mathbf{S}_Y(\mathbf{I} - \mathbf{\Pi}_X)\|_{\mathrm{F}}^2.$$

The second term is independent of $\mathbf{A}$ and the first is zero for $\mathbf{A} = \mathbf{A}_{\mathbf{\Sigma}}$. Thus,

$$R(\mathbf{A}, \mathbf{\Sigma}) = \|(\mathbf{A}\mathbf{S}_X - \mathbf{S}_Y)\mathbf{\Pi}_X\|_{\mathrm{F}}^2.$$

Finally, since $\mathbf{S}_X\mathbf{\Pi}_X = \mathbf{S}_X$ and $\mathbf{A}_{\mathbf{\Sigma}}\mathbf{S}_X = \mathbf{S}_Y\mathbf{\Pi}_X$,

$$(\mathbf{A}\mathbf{S}_X - \mathbf{S}_Y)\mathbf{\Pi}_X = (\mathbf{A} - \mathbf{A}_{\mathbf{\Sigma}})\mathbf{S}_X,$$

which gives the final equality. □

Next, set

$$\mathcal{C}^* \coloneqq \{\mathbf{\Sigma}_P : P \in \mathcal{P}^*\} = \{\mathbf{\Sigma} \in \mathbb{S}_+^d : \mathbf{\Sigma}_X \succ \mathbf{0}\}.$$

Using the second-moment formulation of the risk together with Lemma 1.2, the condition number and our main-body estimators admit equivalent Bures-space formulations. For $\mathbf{\Sigma} \in \mathcal{C}^*$, define

$$R_\varepsilon^\star(\mathbf{\Sigma}) \coloneqq \inf_{\mathbf{A}\in\mathcal{A}} \sup_{\mathbf{\Sigma}'\in\mathcal{B}_\varepsilon(\mathbf{\Sigma})} R(\mathbf{A}, \mathbf{\Sigma}'), \qquad \kappa(\mathbf{\Sigma}) \coloneqq \limsup_{\varepsilon\downarrow 0} \varepsilon^{-2} R_\varepsilon^\star(\mathbf{\Sigma}). \tag{4}$$

Then, by (1) and Lemma 1.2, $\kappa_P = \kappa(\mathbf{\Sigma}_P)$ for $P \in \mathcal{P}^*$. Likewise, the estimators can be written as

$$\begin{aligned}
\mathsf{ERM}[\mathbf{\Sigma}] &\in \operatorname*{arg\,min}_{\mathbf{A}\in\mathcal{A}} L(\mathbf{A}, \mathbf{\Sigma}),\\
\mathsf{DRO}_\varepsilon[\mathbf{\Sigma}] &\in \operatorname*{arg\,min}_{\mathbf{A}\in\mathcal{A}} \sup_{\mathbf{\Sigma}'\in\mathcal{B}_\varepsilon(\mathbf{\Sigma})} L(\mathbf{A}, \mathbf{\Sigma}'),\\
\mathsf{MRO}_\varepsilon[\mathbf{\Sigma}] &\in \operatorname*{arg\,min}_{\mathbf{A}\in\mathcal{A}} \sup_{\mathbf{\Sigma}'\in\mathcal{B}_\varepsilon(\mathbf{\Sigma})} R(\mathbf{A}, \mathbf{\Sigma}'),\\
\mathsf{SNR}_\varepsilon[\mathbf{\Sigma}] &\in \operatorname*{arg\,min}_{\mathbf{A}\in\mathcal{A}} L(\mathbf{A}, \mathbf{\Sigma}) + \varepsilon^2 \|\mathbf{A}\|_{\mathrm{op}}^2.
\end{aligned}$$

B.3 Bures–Wasserstein Geometry

We now turn to proofs of the Section 1.2 lemmas on the Bures metric.

Proof of Lemma 1.1. Given $\mathbf{A}, \mathbf{B} \in \mathbb{S}_+^d$, the identity

$$\mathsf{B}(\mathbf{A}, \mathbf{B}) = \min_{\mathbf{U}\in\mathbb{O}^d} \|\mathbf{A}^{\frac{1}{2}} - \mathbf{B}^{\frac{1}{2}}\mathbf{U}\|_{\mathrm{F}} \tag{5}$$

$$= \min_{\substack{\mathbf{S},\mathbf{T}\in\mathbb{R}^{d\times d}\\ \mathbf{A}=\mathbf{S}\mathbf{S}^\top,\, \mathbf{B}=\mathbf{T}\mathbf{T}^\top}} \|\mathbf{S} - \mathbf{T}\|_{\mathrm{F}}. \tag{6}$$

is well-known; see, e.g., Bhatia et al. (2019, Theorem 1).

Suppose $\mathsf{B}(\mathbf{A}, \mathbf{B}) \le \varepsilon$. Fix $\mathbf{U} \in \mathbb{O}^d$ minimizing (5), so that $\|\mathbf{A}^{1/2} - \mathbf{B}^{1/2}\mathbf{U}\|_{\mathrm{F}} \le \varepsilon$. Setting $\mathbf{\Delta} \coloneqq \mathbf{A}^{1/2}\mathbf{U}^\top - \mathbf{B}^{1/2}$, we then have $\|\mathbf{\Delta}\|_{\mathrm{F}} = \|\mathbf{\Delta}\mathbf{U}\|_{\mathrm{F}} \le \varepsilon$ and

$$(\mathbf{B}^{\frac{1}{2}} + \mathbf{\Delta})(\mathbf{B}^{\frac{1}{2}} + \mathbf{\Delta})^\top = \mathbf{A}^{\frac{1}{2}}\mathbf{U}^\top\mathbf{U}\mathbf{A}^{\frac{1}{2}} = \mathbf{A},$$

as desired. On the other hand, if such $\mathbf{\Delta}$ exists, then $\mathbf{S} \coloneqq \mathbf{B}^{\frac{1}{2}} + \mathbf{\Delta}$ and $\mathbf{T} \coloneqq \mathbf{B}^{\frac{1}{2}}$ are feasible square-root factors for (6), and so

$$\mathsf{B}(\mathbf{A}, \mathbf{B}) \le \|\mathbf{S} - \mathbf{T}\|_{\mathrm{F}} = \|\mathbf{\Delta}\|_{\mathrm{F}} \le \varepsilon.$$

If $\mathbf{A}$ and $\mathbf{B}$ are simultaneously diagonalizable, then $\mathbf{A}^{1/2}$ and $\mathbf{B}^{1/2}$ commute. In this case, we have

$$\begin{aligned}
\mathsf{B}(\mathbf{A}, \mathbf{B})^2 &= \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B}) - 2\operatorname{tr}\big((\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\big)\\
&= \operatorname{tr}(\mathbf{A}) + \operatorname{tr}(\mathbf{B}) - 2\operatorname{tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}})\\
&= \|\mathbf{A}^{\frac{1}{2}} - \mathbf{B}^{\frac{1}{2}}\|_{\mathrm{F}}^2,
\end{aligned}$$

implying that the minimum in (5) is attained by $\mathbf{U} = \mathbf{I}_d$, as claimed. □

Proof of Lemma 1.2. The second equality follows directly from Lemma 1.1. We now prove the first.

Proposition 1 of Nguyen et al. (2021a) establishes a related claim for the augmented Gelbrich distance over mean–covariance pairs, assuming non-singular covariance matrices. Here we give a direct argument for the Bures case with uncentered and potentially singular second-moment matrices.

$\subseteq$ direction. Fix $Q \in \mathcal{B}_\varepsilon(P)$ and any coupling of random variables (Z, Z') with $Z \sim P$ and $Z' \sim Q$. Setting $\mathbf{C} \coloneqq \mathbb{E}[Z(Z')^\top]$, the block second-moment matrix

$$\begin{pmatrix} \mathbf{\Sigma}_P & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{\Sigma}_Q \end{pmatrix}$$

is PSD. Hence, there exists a contraction $\mathbf{K} \in \mathbb{R}^{d\times d}$ with $\|\mathbf{K}\|_{\mathrm{op}} \le 1$ such that $\mathbf{C} = \mathbf{\Sigma}_P^{1/2}\mathbf{K}\mathbf{\Sigma}_Q^{1/2}$ (see, e.g., Bhatia, 2009, Proposition 1.3.2). Therefore, by duality between the operator norm $\|\cdot\|_{\mathrm{op}}$ and nuclear norm $\|\cdot\|_*$ (the sum of singular values), we have

$$\operatorname{tr}(\mathbf{C}) = \operatorname{tr}\big(\mathbf{K}\mathbf{\Sigma}_Q^{1/2}\mathbf{\Sigma}_P^{1/2}\big) \le \big\|\mathbf{\Sigma}_Q^{1/2}\mathbf{\Sigma}_P^{1/2}\big\|_* = \operatorname{tr}\Big((\mathbf{\Sigma}_P^{1/2}\mathbf{\Sigma}_Q\mathbf{\Sigma}_P^{1/2})^{\frac{1}{2}}\Big).$$

It then follows from definition of B that

$$\mathbb{E}\,\|Z - Z'\|^2 = \operatorname{tr}(\mathbf{\Sigma}_P) + \operatorname{tr}(\mathbf{\Sigma}_Q) - 2\operatorname{tr}(\mathbf{C}) \geq \mathsf{B}(\mathbf{\Sigma}_P, \mathbf{\Sigma}_Q)^2.$$

Infimizing over couplings gives $\mathsf{W}_2(P, Q) \geq \mathsf{B}(\mathbf{\Sigma}_P, \mathbf{\Sigma}_Q)$. Thus $\mathbf{\Sigma}_Q \in \mathcal{B}_\varepsilon(\mathbf{\Sigma}_P)$.

$\supseteq$ **direction.** Fix $\mathbf{\Sigma} \in \mathcal{B}_\varepsilon(\mathbf{\Sigma}_P)$. It suffices to identify $Q \in \mathcal{B}_\varepsilon(P)$ such that $\mathbf{\Sigma} = \mathbf{\Sigma}_Q$. Abbreviate $\mathbf{B}_P \coloneqq \mathbf{\Sigma}_P^{1/2}$. By the Bures bound and Lemma 1.1, there exists a factor $\mathbf{B} \in \mathbb{R}^{d\times d}$ such that $\|\mathbf{B}_P - \mathbf{B}\|_{\mathrm{F}} \leq \varepsilon$ and $\mathbf{B}\mathbf{B}^\top = \mathbf{\Sigma}$. Let $\mathbf{\Pi} \in \mathbb{R}^{d\times d}$ be the orthogonal projection onto $\operatorname{range}(\mathbf{\Sigma}_P) = \operatorname{range}(\mathbf{B}_P)$. By Lemma A.2, we have $\mathbf{\Pi} = \mathbf{B}_P\mathbf{B}_P^\dagger = \mathbf{B}_P^\dagger\mathbf{\Sigma}_P\mathbf{B}_P^\dagger$, where $\dagger$ denotes the Moore–Penrose pseudoinverse. Let $Z \sim P$ and $W \sim \mathcal{N}(0, \mathbf{I}_d - \mathbf{\Pi})$ be independent. Now consider the random variable $U \coloneqq \mathbf{B}_P^\dagger Z + W$, constructed so that

$$\begin{aligned}\mathbb{E}\big[UU^\top\big] &= \mathbf{B}_P^\dagger\mathbf{\Sigma}_P\mathbf{B}_P^\dagger + \mathbf{I}_d - \mathbf{\Pi} = \mathbf{\Pi} + \mathbf{I}_d - \mathbf{\Pi} = \mathbf{I}_d\\ \mathbf{B}_P U &= \mathbf{\Pi}Z + \mathbf{B}_P W = Z.\end{aligned}$$

Indeed, Z is supported on the range of $\mathbf{\Sigma}_P$ and W is supported on the null space of $\mathbf{B}_P$. We take Q as the law of $Z' \coloneqq \mathbf{B}U$. Indeed, since U has identity second moment, we have $\mathbf{\Sigma}_Q = \mathbf{B}\mathbf{B}^\top = \mathbf{\Sigma}$. Moreover, the coupling (Z, Z') certifies that

$$\mathsf{W}_2(P, Q)^2 \leq \mathbb{E}[\|Z - Z'\|^2] = \mathbb{E}[\|(\mathbf{B}_P - \mathbf{B})U\|^2] = \|\mathbf{B}_P - \mathbf{B}\|_{\mathrm{F}}^2 \leq \varepsilon^2.$$

Thus, $Q \in \mathcal{B}_\varepsilon(P)$, as desired. □

Lastly, we apply Lemma 1.1 to prove a simple upper bound on the Bures distance.

Lemma B.2 (Bures distance upper bound)**.** *For $\mathbf{A}, \mathbf{B} \in \mathbb{S}_+^d$ with $\operatorname{range}(\mathbf{B}) \subseteq \operatorname{range}(\mathbf{A})$, we have*

$$\mathsf{B}(\mathbf{A}, \mathbf{B})^2 \leq \operatorname{tr}\big((\mathbf{B} - \mathbf{A})\mathbf{A}^\dagger(\mathbf{B} - \mathbf{A})\big).$$

Proof. We assume without loss of generality that $\mathbf{A} \succ \mathbf{0}$. Indeed, if $\mathbf{A} = \mathbf{0}$, then the range constraint forces $\mathbf{B} = \mathbf{0}$ and the claim holds trivially. Otherwise, since $\operatorname{range}(\mathbf{B}) \subseteq \operatorname{range}(\mathbf{A})$, both matrices vanish on $\operatorname{range}(\mathbf{A})^\perp$. We may therefore restrict to $\operatorname{range}(\mathbf{A})$, where $\mathbf{A}$ is positive definite and $\mathbf{A}^\dagger$ agrees with its inverse.

Set $\mathbf{C} \coloneqq \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} \succeq \mathbf{0}$. By construction, we have

$$\mathbf{B} = \mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{1/2} = \big(\mathbf{A}^{1/2}\mathbf{C}^{1/2}\big)\big(\mathbf{A}^{1/2}\mathbf{C}^{1/2}\big)^\top.$$

Using the square-root characterization of the Bures distance (Lemma 1.1) with the particular square-root factors $\mathbf{A}^{1/2}$ and $\mathbf{A}^{1/2}\mathbf{C}^{1/2}$, we obtain

$$\mathsf{B}(\mathbf{A}, \mathbf{B})^2 \leq \big\|\mathbf{A}^{1/2}(\mathbf{C}^{1/2} - \mathbf{I}_d)\big\|_{\mathrm{F}}^2.$$

Since $\mathbf{C}^{1/2} - \mathbf{I}_d = (\mathbf{C} - \mathbf{I}_d)(\mathbf{C}^{1/2} + \mathbf{I}_d)^{-1}$ and $\|(\mathbf{C}^{1/2} + \mathbf{I}_d)^{-1}\|_{\mathrm{op}} \leq 1$, we further bound

$$\begin{aligned}\big\|\mathbf{A}^{1/2}(\mathbf{C}^{1/2} - \mathbf{I}_d)\big\|_{\mathrm{F}} &\leq \big\|\mathbf{A}^{1/2}(\mathbf{C} - \mathbf{I}_d)\big\|_{\mathrm{F}}\|(\mathbf{C}^{1/2} + \mathbf{I}_d)^{-1}\|_{\mathrm{op}} && \text{(Lemma A.1)}\\ &\leq \big\|\mathbf{A}^{1/2}(\mathbf{C} - \mathbf{I}_d)\big\|_{\mathrm{F}}.\end{aligned}$$

By our choice of $\mathbf{C}$, we have $\mathbf{A}^{1/2}(\mathbf{C} - \mathbf{I}_d) = (\mathbf{B} - \mathbf{A})\mathbf{A}^{-1/2}$, and thus

$$\big\|\mathbf{A}^{1/2}(\mathbf{C} - \mathbf{I}_d)\big\|_{\mathrm{F}}^2 = \operatorname{tr}\big((\mathbf{B} - \mathbf{A})\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\big).$$

Combining the above gives the claimed bound on $\mathsf{B}(\mathbf{A}, \mathbf{B})$. □

Finally, we establish Lipschitz continuity of the square root of the smallest feature eigenvalue.

Lemma B.3 (Bures continuity of the smallest feature eigenvalue)**.** *For $\mathbf{\Sigma}, \bar{\mathbf{\Sigma}} \in \mathbb{S}_+^d$, we have*

$$\Big|\sqrt{\lambda_{\min}(\bar{\mathbf{\Sigma}}_X)} - \sqrt{\lambda_{\min}(\mathbf{\Sigma}_X)}\Big| \leq \mathsf{B}(\bar{\mathbf{\Sigma}}, \mathbf{\Sigma}).$$

In particular, $\lambda_{\min}(\mathbf{\Sigma}_X)$ is continuous with respect to the Bures metric.

Proof. By Lemma 1.1, there exist square-root factors $\mathbf{S}, \bar{\mathbf{S}} \in \mathbb{R}^{d\times d}$ of $\mathbf{\Sigma}$ and $\bar{\mathbf{\Sigma}}$ such that

$$\|\bar{\mathbf{S}} - \mathbf{S}\|_{\mathrm{F}} = \mathsf{B}(\bar{\mathbf{\Sigma}}, \mathbf{\Sigma}).$$

Using the Section B.1 block decompositions, we have $\mathbf{\Sigma}_X = \mathbf{S}_X\mathbf{S}_X^\top$ and $\bar{\mathbf{\Sigma}}_X = \bar{\mathbf{S}}_X\bar{\mathbf{S}}_X^\top$. Hence, by Lemma A.5,

$$\begin{aligned}\left|\sqrt{\lambda_{\min}(\bar{\mathbf{\Sigma}}_X)} - \sqrt{\lambda_{\min}(\mathbf{\Sigma}_X)}\right| &= \left|\sigma_{\min}(\bar{\mathbf{S}}_X) - \sigma_{\min}(\mathbf{S}_X)\right| \\ &\le \|\bar{\mathbf{S}}_X - \mathbf{S}_X\|_{\mathrm{op}} \\ &\le \|\bar{\mathbf{S}} - \mathbf{S}\|_{\mathrm{F}} \\ &= \mathsf{B}(\bar{\mathbf{\Sigma}}, \mathbf{\Sigma}).\end{aligned}$$

The continuity claim follows immediately. □

C DRO Duality and Symmetry Reductions

This appendix establishes a dual formulation for the inner DRO maximization problem and an averaging principle for DRO, MRO, and SNR under orthogonal symmetries. We use $\sup \varnothing = -\infty$.

C.1 Strong Duality for the Inner DRO Problem

We first establish a helper lemma, which will be applied both for the inner DRO problem below and for the fixed-projection problem in Section G.1. Very similar arguments appear in Shafieezadeh-Abadeh et al. (2018); Nguyen et al. (2021b), but we provide a self-contained proof to account for the case where $r < d$ (untreated in those works).

Lemma C.1 (Quadratic-ball duality)**.** *Let $\mathbf{M} \in \mathbb{R}^{k\times d}$ and $\tilde{\mathbf{W}} \in \mathbb{R}^{d\times r}$, where $k, r \ge 1$. Set $b \coloneqq \|\mathbf{M}\|_{\mathrm{op}}^2$. For every $\eta > b$, we have*

$$\sup_{\mathbf{W}\in\mathbb{R}^{d\times r}} \Big\{\|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 - \eta\|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2\Big\} = \eta\,\mathrm{tr}\Big((\eta\mathbf{I}_k - \mathbf{M}\mathbf{M}^\top)^{-1}\mathbf{M}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^\top\mathbf{M}^\top\Big). \tag{7}$$

Moreover, for every $t \in \mathbb{R}$, with $\sup\varnothing = -\infty$, we have

$$\begin{aligned}\sup_{\substack{\mathbf{W}\in\mathbb{R}^{d\times r}\\ \|\mathbf{W}-\tilde{\mathbf{W}}\|_{\mathrm{F}}^2\le t}} \|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 &= \sup_{\mathbf{W}\in\mathbb{R}^{d\times r}} \inf_{\eta\ge 0}\Big\{\|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 + \eta\big(t - \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2\big)\Big\} \\ &= \inf_{\eta\ge 0}\sup_{\mathbf{W}\in\mathbb{R}^{d\times r}} \Big\{\|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 + \eta\big(t - \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2\big)\Big\} \\ &= \inf_{\eta> b} \eta\Big(t + \mathrm{tr}\Big((\eta\mathbf{I}_k - \mathbf{M}\mathbf{M}^\top)^{-1}\mathbf{M}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^\top\mathbf{M}^\top\Big)\Big).\end{aligned} \tag{8}$$

Proof. Fix $t \in \mathbb{R}$. For $\eta \ge 0$, define

$$\Phi_t(\eta) \coloneqq \sup_{\mathbf{W}\in\mathbb{R}^{d\times r}} \Big\{\|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 + \eta\big(t - \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2\big)\Big\}.$$

We first compute this supremum for $\eta > b$. Expanding the quadratic expression gives

$$\|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 - \eta\|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2 = -\,\mathrm{tr}\big(\mathbf{W}^\top(\eta\mathbf{I}_d - \mathbf{M}^\top\mathbf{M})\mathbf{W}\big) + 2\eta\langle\mathbf{W}, \tilde{\mathbf{W}}\rangle_{\mathrm{F}} - \eta\|\tilde{\mathbf{W}}\|_{\mathrm{F}}^2.$$

Since $\eta\mathbf{I}_d - \mathbf{M}^\top\mathbf{M} \succ \mathbf{0}$, this expression is strictly concave in $\mathbf{W}$. Its unique maximizer satisfies

$$(\eta\mathbf{I}_d - \mathbf{M}^\top\mathbf{M})\mathbf{W} = \eta\tilde{\mathbf{W}},$$

and therefore equals $\eta(\eta\mathbf{I}_d - \mathbf{M}^\top\mathbf{M})^{-1}\tilde{\mathbf{W}}$. Substitution yields

$$\begin{aligned}\Phi_t(\eta) &= \eta t + \eta^2\,\mathrm{tr}\Big(\tilde{\mathbf{W}}^\top(\eta\mathbf{I}_d - \mathbf{M}^\top\mathbf{M})^{-1}\tilde{\mathbf{W}}\Big) - \eta\|\tilde{\mathbf{W}}\|_{\mathrm{F}}^2 \\ &= \eta t + \eta\,\mathrm{tr}\Big(\tilde{\mathbf{W}}^\top\big[\eta(\eta\mathbf{I}_d - \mathbf{M}^\top\mathbf{M})^{-1} - \mathbf{I}_d\big]\tilde{\mathbf{W}}\Big) \\ &= \eta t + \eta\,\mathrm{tr}\Big((\eta\mathbf{I}_k - \mathbf{M}\mathbf{M}^\top)^{-1}\mathbf{M}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^\top\mathbf{M}^\top\Big),\end{aligned} \tag{9}$$

where the last equality uses Lemma A.4 and cyclicity of trace. Subtracting ηt proves (7).

We next prove the final equality in (8). Given (9), it suffices to show

$$\inf_{\eta \geq 0} \Phi_t(\eta) = \inf_{\eta > b} \Phi_t(\eta).$$

For $0 \leq \eta < b$, varying one column of $\mathbf{W}$ along a top eigenvector of $\mathbf{M}^\top \mathbf{M}$ makes the expression defining $\Phi_t(\eta)$ unbounded above, since its leading quadratic coefficient is $b - \eta > 0$. Thus $\Phi_t(\eta) = +\infty$. At the endpoint $\eta = b$, observe that as $\eta \downarrow b$, the functions $\mathbf{W} \mapsto \|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 - \eta\|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2$ increase pointwise to their values at b, so their suprema converge to the supremum at b. Further, $\eta \mapsto \eta t$ is continuous, so we obtain

$$\lim_{\eta \downarrow b} \Phi_t(\eta) = \Phi_t(b),$$

allowing $+\infty$. Hence excluding all $\eta \leq b$ does not change the infimum.

We now return to the initial constrained supremum in (8). For each fixed $\mathbf{W}$, we have

$$\inf_{\eta \geq 0} \Big\{ \|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 + \eta\big(t - \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2\big) \Big\} = \begin{cases} \|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2, & \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2 \leq t, \\ -\infty, & \text{otherwise.} \end{cases}$$

This proves the first equality in (8).

For the remaining equality, suppose first that $t > 0$ and set

$$\gamma \coloneqq \max_{\substack{\mathbf{W} \in \mathbb{R}^{d \times r} \\ \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2 \leq t}} \|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2.$$

This maximum exists and is finite because the feasible set is compact and the objective is continuous. By definition of γ, the quadratic implication

$$t - \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2 \geq 0 \quad \Longrightarrow \quad \gamma - \|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 \geq 0$$

holds for every $\mathbf{W}$. The first inequality is strictly satisfied at $\mathbf{W} = \tilde{\mathbf{W}}$. After vectorizing $\mathbf{W}$, the S-lemma (Pólik and Terlaky, 2007, Theorem 2.2 and Section 8.1) thus gives some $\eta \geq 0$ such that

$$\gamma - \|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 - \eta\big(t - \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2\big) \geq 0 \qquad \forall\, \mathbf{W} \in \mathbb{R}^{d \times r}.$$

Equivalently, $\Phi_t(\eta) \leq \gamma$. Conversely, $\Phi_t(\eta) \geq \gamma$ for every $\eta \geq 0$, since the additional term in the definition of Φ_t is nonnegative on the feasible set. Consequently,

$$\gamma = \inf_{\eta \geq 0} \Phi_t(\eta).$$

This yields the second equality in (8) when $t > 0$.

If $t = 0$, the only feasible point is $\mathbf{W} = \tilde{\mathbf{W}}$, so the constrained supremum is $\|\mathbf{M}\tilde{\mathbf{W}}\|_{\mathrm{F}}^2$. Moreover, the formula (9) established for $\Phi_t(\eta)$ implies that

$$\inf_{\eta \geq 0} \Phi_0(\eta) \leq \lim_{\eta \to \infty} \Phi_0(\eta) = \|\mathbf{M}\tilde{\mathbf{W}}\|_{\mathrm{F}}^2,$$

using that $\eta(\eta \mathbf{I}_k - \mathbf{M}\mathbf{M}^\top)^{-1} \to \mathbf{I}_k$. Since $\Phi_0(\eta)$ is at least the constrained supremum for every $\eta \geq 0$, equality follows.

Finally, if $t < 0$, the constrained feasible set is empty and its supremum equals $-\infty$. Meanwhile, the formula (9) for Φ_t gives

$$\Phi_t(\eta) = \eta t + O(1) \to -\infty \qquad \text{as } \eta \to \infty,$$

so $\inf_{\eta \geq 0} \Phi_t(\eta) = -\infty$ as well. □

We now apply this to the inner maximization problem in the DRO formulation. For $\mathbf{A} \in \mathcal{A}$, write

$$\mathbf{M}_{\mathbf{A}} \coloneqq \begin{pmatrix} \mathbf{A} & -\mathbf{I}_m \end{pmatrix} \in \mathbb{R}^{m \times d},$$

so that $L(\mathbf{A}, P) = \mathbb{E}_P \|\mathbf{M}_{\mathbf{A}} Z\|^2$. Direct multiplication gives

$$\mathbf{M}_{\mathbf{A}} \mathbf{M}_{\mathbf{A}}^\top = \mathbf{A}\mathbf{A}^\top + \mathbf{I}_m, \qquad \|\mathbf{M}_{\mathbf{A}}\|_{\mathrm{op}}^2 = 1 + \|\mathbf{A}\|_{\mathrm{op}}^2.$$

We state the dual result for a generic quadratic loss before specializing to regression.

Lemma C.2 (Inner DRO strong duality)**.** *Let $P \in \mathcal{P}_2(\mathbb{R}^d)$, $\varepsilon > 0$, and $\mathbf{M} \in \mathbb{R}^{k\times d}$ for $1 \le k \le d$. Then*

$$\sup_{Q\in\mathcal{B}_\varepsilon(P)} \mathbb{E}_Q \|\mathbf{M}Z\|^2 = \inf_{\eta>\|\mathbf{M}\|_{\mathrm{op}}^2} \eta\big(\varepsilon^2 + \mathrm{tr}\big((\eta\mathbf{I}_k - \mathbf{M}\mathbf{M}^\top)^{-1}\mathbf{M}\boldsymbol{\Sigma}_P\mathbf{M}^\top\big)\big).$$

For regression, taking $k = m$ and $\mathbf{M} = \mathbf{M}_\mathbf{A}$ gives

$$\sup_{Q\in\mathcal{B}_\varepsilon(P)} L(\mathbf{A}, Q) = \inf_{\eta>1+\|\mathbf{A}\|_{\mathrm{op}}^2} \eta\big(\varepsilon^2 + \mathrm{tr}\big(((\eta-1)\mathbf{I}_m - \mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{M}_\mathbf{A}\boldsymbol{\Sigma}_P\mathbf{M}_\mathbf{A}^\top\big)\big).$$

Proof. By Lemma 1.2, the second moments of laws in $\mathcal{B}_\varepsilon(P)$ are precisely the matrices $\mathbf{W}\mathbf{W}^\top$, where $\mathbf{W} \in \mathbb{R}^{d\times d}$ satisfies $\|\mathbf{W} - \boldsymbol{\Sigma}_P^{1/2}\|_{\mathrm{F}}^2 \le \varepsilon^2$. Since $\mathbb{E}_Q \|\mathbf{M}Z\|^2 = \mathrm{tr}(\mathbf{M}\boldsymbol{\Sigma}_Q\mathbf{M}^\top)$, we obtain

$$\begin{aligned}\sup_{Q\in\mathcal{B}_\varepsilon(P)} \mathbb{E}_Q \|\mathbf{M}Z\|^2 &= \sup_{\substack{\mathbf{W}\in\mathbb{R}^{d\times d}\\ \|\mathbf{W}-\boldsymbol{\Sigma}_P^{1/2}\|_{\mathrm{F}}^2\le\varepsilon^2}} \|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 \\ &= \inf_{\eta>\|\mathbf{M}\|_{\mathrm{op}}^2} \eta\big(\varepsilon^2 + \mathrm{tr}\big((\eta\mathbf{I}_k - \mathbf{M}\mathbf{M}^\top)^{-1}\mathbf{M}\boldsymbol{\Sigma}_P\mathbf{M}^\top\big)\big),\end{aligned}$$

where the second equality applies Lemma C.1 with $r = d$, $\tilde{\mathbf{W}} = \boldsymbol{\Sigma}_P^{1/2}$, and $t = \varepsilon^2$.

Finally, the regression specialization follows by taking $\mathbf{M} = \mathbf{M}_\mathbf{A}$ and using $\mathbf{M}_\mathbf{A}\mathbf{M}_\mathbf{A}^\top = \mathbf{A}\mathbf{A}^\top + \mathbf{I}_m$ together with $L(\mathbf{A}, Q) = \mathbb{E}_Q \|\mathbf{M}_\mathbf{A}Z\|^2$. □

We note that the infimum over the open dual domain need not be attained; its value may be approached at the left endpoint. We next specialize to the case where $\mathbf{A}\mathbf{A}^\top$ is diagonal.

Corollary C.3 (Diagonal case)**.** *Let $P \in \mathcal{P}_2(\mathbb{R}^d)$, $\varepsilon > 0$, and $\mathbf{A} \in \mathcal{A}$ satisfy*

$$\mathbf{A}\mathbf{A}^\top = \mathrm{diag}(a_1^2, \dots, a_m^2).$$

For $i = 1, \dots, m$, define the ith coordinate residual second moment

$$r_i^2 \coloneqq \mathbb{E}_P\big[(e_i^\top(\mathbf{A}X - Y))^2\big] = e_i^\top\mathbf{M}_\mathbf{A}\boldsymbol{\Sigma}_P\mathbf{M}_\mathbf{A}^\top e_i.$$

We then have

$$\sup_{Q\in\mathcal{B}_\varepsilon(P)} L(\mathbf{A}, Q) = \inf_{\eta>1+\max_i a_i^2} \eta\left(\varepsilon^2 + \sum_{i=1}^m \frac{r_i^2}{\eta - 1 - a_i^2}\right).$$

Proof. In the regression formula of Lemma C.2,

$$((\eta-1)\mathbf{I}_m - \mathbf{A}\mathbf{A}^\top)^{-1} = \mathrm{diag}\left(\frac{1}{\eta-1-a_1^2}, \dots, \frac{1}{\eta-1-a_m^2}\right).$$

Taking the trace of its product with $\mathbf{M}_\mathbf{A}\boldsymbol{\Sigma}_P\mathbf{M}_\mathbf{A}^\top$ gives the stated sum. □

C.2 Averaging over Symmetries

All symmetry reductions used below preserve the feature and response subspaces. We therefore formulate the averaging principle directly for block-diagonal orthogonal transformations. Recall that for a measurable map T, $T_\sharp P$ denotes the pushforward law of $T(Z)$ when $Z \sim P$.

Lemma C.4 (Averaging over symmetries)**.** *Let U be a compact subgroup of*

$$\{\mathrm{diag}(\mathbf{U}_X, \mathbf{U}_Y) : \mathbf{U}_X \in \mathbb{O}^p,\ \mathbf{U}_Y \in \mathbb{O}^m\}.$$

Suppose $P, \tilde{P} \in \mathcal{P}_2(\mathbb{R}^d)$ satisfy $\mathbf{U}_\sharp P = P$ and $\mathbf{U}_\sharp\tilde{P} = \tilde{P}$ for every $\mathbf{U} \in U$. Let $\varepsilon > 0$, and fix any minimizer $\hat{\mathbf{A}} \in \mathcal{A}$ for the $\mathsf{DRO}_\varepsilon[\tilde{P}]$, $\mathsf{MRO}_\varepsilon[\tilde{P}]$, or $\mathsf{SNR}_\varepsilon[\tilde{P}]$ optimization problem.

Let μ be normalized Haar measure on U and define

$$\bar{\mathbf{A}} \coloneqq \int_U \mathbf{U}_Y^\top\hat{\mathbf{A}}\mathbf{U}_X \,\mathrm{d}\mu(\mathbf{U}).$$

Then $\bar{\mathbf{A}}$ is a minimizer of the same optimization problem, is invariant in the sense that

$$\mathbf{U}_Y^\top \bar{\mathbf{A}} \mathbf{U}_X = \bar{\mathbf{A}} \qquad \forall\, \mathbf{U} \in U,$$

and satisfies $R(\bar{\mathbf{A}}, P) \le R(\hat{\mathbf{A}}, P)$. Consequently, restricting any of these optimization problems to invariant coefficient matrices preserves its optimal value. Moreover, to lower bound the target regret of every optimizer, it suffices to lower bound the target regret of every invariant optimizer.

Proof. For $\mathbf{U} = \operatorname{diag}(\mathbf{U}_X, \mathbf{U}_Y) \in U$, define the linear coefficient transformation

$$T_{\mathbf{U}}(\mathbf{A}) \coloneqq \mathbf{U}_Y^\top \mathbf{A} \mathbf{U}_X.$$

Orthogonality gives

$$\begin{aligned} L(\mathbf{A}, \mathbf{U}_\sharp Q) &= \mathbb{E}_Q \left\| \mathbf{A}\mathbf{U}_X X - \mathbf{U}_Y Y \right\|^2 \\ &= \mathbb{E}_Q \left\| \mathbf{U}_Y^\top \mathbf{A}\mathbf{U}_X X - Y \right\|^2 \\ &= L(T_{\mathbf{U}}(\mathbf{A}), Q). \end{aligned}$$

The map $T_{\mathbf{U}}$ is a bijection from $\mathcal{A}$ to itself, so

$$\inf_{\mathbf{B} \in \mathcal{A}} L(\mathbf{B}, \mathbf{U}_\sharp Q) = \inf_{\mathbf{B} \in \mathcal{A}} L(\mathbf{B}, Q).$$

Subtracting this identity from the preceding risk identity yields

$$R(\mathbf{A}, \mathbf{U}_\sharp Q) = R(T_{\mathbf{U}}(\mathbf{A}), Q).$$

Also, left and right multiplication by orthogonal matrices gives $\|T_{\mathbf{U}}(\mathbf{A})\|_{\mathrm{op}} = \|\mathbf{A}\|_{\mathrm{op}}$.

The source ambiguity set is invariant under each $\mathbf{U} \in U$. Indeed,

$$\mathsf{W}_2(\mathbf{U}_\sharp Q, \tilde{P}) = \mathsf{W}_2(\mathbf{U}_\sharp Q, \mathbf{U}_\sharp \tilde{P}) = \mathsf{W}_2(Q, \tilde{P}).$$

Applying this also to $\mathbf{U}^{-1}$ proves $\mathbf{U}_\sharp \mathcal{B}_\varepsilon(\tilde{P}) = \mathcal{B}_\varepsilon(\tilde{P})$. It follows that each of the three objectives

$$\sup_{Q \in \mathcal{B}_\varepsilon(\tilde{P})} L(\mathbf{A}, Q), \qquad \sup_{Q \in \mathcal{B}_\varepsilon(\tilde{P})} R(\mathbf{A}, Q), \qquad L(\mathbf{A}, \tilde{P}) + \varepsilon^2 \|\mathbf{A}\|_{\mathrm{op}}^2$$

is invariant under $T_{\mathbf{U}}$. Each objective is also convex in $\mathbf{A}$.

The integral defining $\bar{\mathbf{A}}$ is well-defined because its integrand is continuous on the compact group U. For any fixed $\mathbf{V} = \operatorname{diag}(\mathbf{V}_X, \mathbf{V}_Y) \in U$,

$$\begin{aligned} \mathbf{V}_Y^\top \bar{\mathbf{A}} \mathbf{V}_X &= \int_U (\mathbf{U}_Y \mathbf{V}_Y)^\top \hat{\mathbf{A}} (\mathbf{U}_X \mathbf{V}_X) \, \mathrm{d}\mu(\mathbf{U}) \\ &= \bar{\mathbf{A}}, \end{aligned}$$

where the last equality uses right invariance of Haar measure under $\mathbf{U} \mapsto \mathbf{U}\mathbf{V}$. Thus $\bar{\mathbf{A}}$ is invariant.

Let J denote whichever of the three objectives is minimized by $\hat{\mathbf{A}}$. By convexity, invariance, and Jensen's inequality,

$$J(\bar{\mathbf{A}}) \le \int_U J(T_{\mathbf{U}}(\hat{\mathbf{A}})) \, \mathrm{d}\mu(\mathbf{U}) = J(\hat{\mathbf{A}}).$$

Since $\hat{\mathbf{A}}$ is a minimizer, $\bar{\mathbf{A}}$ is also a minimizer.

Finally, invariance of P gives $R(T_{\mathbf{U}}(\hat{\mathbf{A}}), P) = R(\hat{\mathbf{A}}, P)$. The regret is convex in its coefficient argument, so another application of Jensen's inequality yields

$$R(\bar{\mathbf{A}}, P) \le \int_U R(T_{\mathbf{U}}(\hat{\mathbf{A}}), P) \, \mathrm{d}\mu(\mathbf{U}) = R(\hat{\mathbf{A}}, P).$$

This proves all claims. □

D Remaining Proofs for Section 1

D.1 Existence of Minimizers (Footnote 1)

We treat the estimators individually.

ERM. Solving the normal equation $\nabla_{\mathbf{A}} L(\mathbf{A}, P) = \mathbf{0}$ reveals that $\mathbf{\Sigma}_{YX}^P (\mathbf{\Sigma}_X^P)^\dagger$ is always a minimizer of $L(\cdot, P)$.

DRO. Since $L(\cdot, Q)$ is continuous and convex, the DRO objective $F(\mathbf{A}) = \sup_{Q \in \mathcal{B}_\varepsilon(P)} L(\mathbf{A}, Q)$ is convex and lower semicontinuous as the supremum of continuous convex functions. It thus suffices to show that F is coercive. To prove this, consider the feasible measure $Q = P * \mathcal{N}(0, \varepsilon^2 \operatorname{diag}(\mathbf{I}_p, \mathbf{0}_{m\times m})/p)$, which certifies that

$$F(\mathbf{A}) \geq L(\mathbf{A}, Q) = L(\mathbf{A}, P) + \frac{\varepsilon^2}{p} \mathbb{E}_{W \sim \mathcal{N}(0, \mathbf{I}_p)} \|\mathbf{A}W\|^2 \geq \frac{\varepsilon^2}{p} \|\mathbf{A}\|_{\mathrm{F}}^2.$$

Thus $\|\mathbf{A}\|_{\mathrm{F}} \to \infty$ implies $F(\mathbf{A}) \to \infty$, and so F is coercive, as desired.

MRO. Similar to the above, continuity and convexity of $R(\cdot, Q)$ imply that the MRO objective $G(\mathbf{A}) = \sup_{Q \in \mathcal{B}_\varepsilon(P)} R(\mathbf{A}, Q)$ is convex and lower semicontinuous. It thus remains to show that G is coercive. To prove this, consider the same feasible measure Q as for DRO. Since Q is non-degenerate, we have

$$G(\mathbf{A}) \geq R(\mathbf{A}, Q) = \left\|(\mathbf{A} - \mathbf{A}_Q)(\mathbf{\Sigma}_X^Q)^{1/2}\right\|_{\mathrm{F}}^2 \geq \frac{\varepsilon^2}{p} \|\mathbf{A} - \mathbf{A}_Q\|_{\mathrm{F}}^2.$$

Thus $\|\mathbf{A}\|_{\mathrm{F}} \to \infty$ implies $G(\mathbf{A}) \to \infty$, and so G is coercive, as desired.

SNR. The regularized objective $H(\mathbf{A}) = L(\mathbf{A}, P) + \varepsilon^2 \|\mathbf{A}\|_{\mathrm{op}}^2$ is clearly convex and lower semicontinuous. Moreover, since the risk is non-negative, it is also coercive.

E Remaining Proofs for Section 2

E.1 Condition Number Formula (Proof of Theorem 2.2)

We formulate the proof in Bures space, adopting the conventions of Appendix B. Initially, we find it simpler to analyze the alternative condition number $\tilde{\kappa} : \mathcal{C}^* \to [0, \infty]$, defined for $\mathbf{\Sigma} \in \mathcal{C}^*$ by

$$\tilde{\kappa}(\mathbf{\Sigma}) \coloneqq \limsup_{\varepsilon \downarrow 0} \varepsilon^{-2} \sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} R(\mathbf{A}_{\mathbf{\Sigma}'}, \mathbf{\Sigma}), \tag{10}$$

which we will later show coincides with κ.

Our first step is to differentiate the parameter map $g(\mathbf{\Sigma}) \coloneqq \mathbf{\Sigma}_{YX} \mathbf{\Sigma}_X^{-1}$. To avoid boundary issues, we extend this map to the open domain $\mathcal{C} \coloneqq \{\mathbf{\Sigma} \in \mathbb{S}^d : \mathbf{\Sigma}_X \succ 0\}$.

Lemma E.1 (Parameter map derivative)**.** *The map $g : \mathcal{C} \to \mathcal{A}$ is continuously differentiable, with*

$$Dg(\mathbf{\Sigma})[\mathbf{H}] = (\mathbf{H}_{YX} - g(\mathbf{\Sigma})\mathbf{H}_X)\mathbf{\Sigma}_X^{-1} \quad \forall \mathbf{H} \in \mathbb{S}^d.$$

Proof. The maps $\mathbf{\Sigma} \mapsto \mathbf{\Sigma}_{YX}$ and $\mathbf{\Sigma} \mapsto \mathbf{\Sigma}_X$ are linear, while matrix inversion is smooth on the positive definite cone. Hence g is continuously differentiable on $\mathcal{C}$. By the product rule and the standard derivative of the inverse map,

$$\begin{aligned} Dg(\mathbf{\Sigma})[\mathbf{H}] &= \mathbf{H}_{YX} \mathbf{\Sigma}_X^{-1} - \mathbf{\Sigma}_{YX} \mathbf{\Sigma}_X^{-1} \mathbf{H}_X \mathbf{\Sigma}_X^{-1} \\ &= \big(\mathbf{H}_{YX} - g(\mathbf{\Sigma})\mathbf{H}_X\big)\mathbf{\Sigma}_X^{-1}, \end{aligned}$$

as claimed. □

We now characterize $\tilde{\kappa}$ in terms of this derivative. We prove a slightly strengthened, uniform claim.

Lemma E.2 (Derivative expression for $\tilde{\kappa}$)**.** *Fixing $\mathbf{\Sigma} \in \mathcal{C}^*$ with square root $\mathbf{S} \coloneqq \mathbf{\Sigma}^{1/2}$, we have*

$$\tilde{\kappa}(\mathbf{\Sigma}) = \sup_{\mathbf{\Delta} \in \mathbb{R}^{d \times d} : \|\mathbf{\Delta}\|_{\mathrm{F}} \leq 1} \left\| Dg(\mathbf{\Sigma})[\mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S}]\mathbf{\Sigma}_X^{1/2} \right\|_{\mathrm{F}}^2. \tag{11}$$

Moreover, the expansion

$$\sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} R(g(\mathbf{\Sigma}'), \mathbf{\Sigma}) = \tilde{\kappa}(\mathbf{\Sigma})\varepsilon^2 + o(\varepsilon^2) \quad \varepsilon \to 0 \tag{12}$$

holds uniformly as $\mathbf{\Sigma}$ varies in any compact subset of $\mathcal{C}^*$.

Proof. Write $\mathbf{\Sigma}(\mathbf{\Delta}) \coloneqq (\mathbf{S} + \mathbf{\Delta})(\mathbf{S} + \mathbf{\Delta})^\top = \mathbf{\Sigma} + \mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S} + \mathbf{\Delta}\mathbf{\Delta}^\top$, and let K be any compact subset of $\mathcal{C}^*$. By Lemma E.1, g is continuously differentiable on $\mathcal{C} \supset K$, and so

$$g(\mathbf{\Sigma} + \mathbf{H}) - g(\mathbf{\Sigma}) = Dg(\mathbf{\Sigma})[\mathbf{H}] + o(\|\mathbf{H}\|_{\mathrm{F}})$$

as $\mathbf{H} \to \mathbf{0}$, uniformly over $\mathbf{\Sigma} \in K$. Further,

$$\|Dg(\mathbf{\Sigma})\|_{\mathrm{op}} \coloneqq \sup_{\|\mathbf{H}\|_{\mathrm{F}} \leq 1} \|Dg(\mathbf{\Sigma})[\mathbf{H}]\|_{\mathrm{F}}$$

is bounded for $\mathbf{\Sigma} \in K$. Further

$$\|\mathbf{\Sigma}(\mathbf{\Delta}) - \mathbf{\Sigma}\|_{\mathrm{F}} \leq 2\|\mathbf{S}\|_{\mathrm{F}}\|\mathbf{\Delta}\|_{\mathrm{F}} + \|\mathbf{\Delta}\|_{\mathrm{F}}^2 = 2\sqrt{\mathrm{tr}(\mathbf{\Sigma})}\|\mathbf{\Delta}\|_{\mathrm{F}} + \|\mathbf{\Delta}\|_{\mathrm{F}}^2 = O(\|\mathbf{\Delta}\|_{\mathrm{F}})$$

as $\mathbf{\Delta} \to \mathbf{0}$, uniformly for $\mathbf{\Sigma} \in K$. Combining gives

$$\begin{aligned} g(\mathbf{\Sigma}(\mathbf{\Delta})) - g(\mathbf{\Sigma}) &= Dg(\mathbf{\Sigma})[\mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S} + \mathbf{\Delta}\mathbf{\Delta}^\top] + o(\|\mathbf{\Delta}\|_{\mathrm{F}}) \\ &= Dg(\mathbf{\Sigma})[\mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S}] + (1 + \|Dg(\mathbf{\Sigma})\|_{\mathrm{op}})o(\|\mathbf{\Delta}\|_{\mathrm{F}}) \\ &= Dg(\mathbf{\Sigma})[\mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S}] + o(\|\mathbf{\Delta}\|_{\mathrm{F}}) \end{aligned}$$

as $\mathbf{\Delta} \to \mathbf{0}$, uniformly over $\mathbf{\Sigma} \in K$. Hence,

$$\begin{aligned} \left\| (g(\mathbf{\Sigma}(\mathbf{\Delta})) - g(\mathbf{\Sigma}))\mathbf{\Sigma}_X^{1/2} \right\|_{\mathrm{F}}^2 &= \left\| \big(Dg(\mathbf{\Sigma})[\mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S}] + o(\|\mathbf{\Delta}\|_{\mathrm{F}}) \big)\mathbf{\Sigma}_X^{1/2} \right\|_{\mathrm{F}}^2 \\ &= \left\| Dg(\mathbf{\Sigma})[\mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S}]\mathbf{\Sigma}_X^{1/2} \right\|_{\mathrm{F}}^2 + E(\mathbf{\Sigma}, \mathbf{\Delta}) \end{aligned}$$

where, by Cauchy–Schwarz and submultiplicativity of the Frobenius norm

$$\begin{aligned} |E(\mathbf{\Sigma}, \mathbf{\Delta})| &\leq 2\|Dg(\mathbf{\Sigma})\|_{\mathrm{op}}\|\mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S}\|_{\mathrm{F}} o(\|\mathbf{\Delta}\|_{\mathrm{F}})\|\mathbf{\Sigma}_X\|_{\mathrm{F}} + o(\|\mathbf{\Delta}\|_{\mathrm{F}}^2)\|\mathbf{\Sigma}_X\|_{\mathrm{F}} \\ &\leq 4\|Dg(\mathbf{\Sigma})\|_{\mathrm{op}}\sqrt{\mathrm{tr}(\mathbf{\Sigma})}\|\mathbf{\Delta}\|_{\mathrm{F}} o(\|\mathbf{\Delta}\|_{\mathrm{F}})\|\mathbf{\Sigma}_X\|_{\mathrm{F}} + o(\|\mathbf{\Delta}\|_{\mathrm{F}}^2)\|\mathbf{\Sigma}_X\|_{\mathrm{F}} \\ &= o(\|\mathbf{\Delta}\|_{\mathrm{F}}^2) \end{aligned}$$

as $\mathbf{\Delta} \to \mathbf{0}$, uniformly over $\mathbf{\Sigma} \in K$. Combining this with Lemma 1.1 gives

$$\begin{aligned} \sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} R(g(\mathbf{\Sigma}'), \mathbf{\Sigma}) &= \sup_{\mathbf{\Delta} \in \mathbb{R}^{d \times d} : \|\mathbf{\Delta}\|_{\mathrm{F}} \leq \varepsilon} R\big(g(\mathbf{\Sigma}(\mathbf{\Delta})), \mathbf{\Sigma}\big) \\ &= \sup_{\mathbf{\Delta} \in \mathbb{R}^{d \times d} : \|\mathbf{\Delta}\|_{\mathrm{F}} \leq \varepsilon} \left\| (g(\mathbf{\Sigma}(\mathbf{\Delta})) - g(\mathbf{\Sigma}))\mathbf{\Sigma}_X^{1/2} \right\|_{\mathrm{F}}^2 \\ &= \sup_{\mathbf{\Delta} \in \mathbb{R}^{d \times d} : \|\mathbf{\Delta}\|_{\mathrm{F}} \leq \varepsilon} \left\| \big(Dg(\mathbf{\Sigma})[\mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S}] + o(\|\mathbf{\Delta}\|_{\mathrm{F}}) \big)\mathbf{\Sigma}_X^{1/2} \right\|_{\mathrm{F}}^2 \\ &= \sup_{\mathbf{\Delta} \in \mathbb{R}^{d \times d} : \|\mathbf{\Delta}\|_{\mathrm{F}} \leq 1} \left\| Dg(\mathbf{\Sigma})[\mathbf{S}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{S}]\mathbf{\Sigma}_X^{1/2} \right\|_{\mathrm{F}}^2 \varepsilon^2 + o(\varepsilon^2), \end{aligned}$$

as $\varepsilon \to 0$, uniformly over $\mathbf{\Sigma} \in K$. The lemma then follows by the definition of $\tilde{\kappa}$. □

Next, we substitute the derivative formula from Lemma E.1 into (11) and simplify. We use the variational form of the Frobenius norm $\|\mathbf{X}\|_{\mathrm{F}} = \sup_{\mathbf{Y} : \|\mathbf{Y}\|_{\mathrm{F}} \leq 1} \langle \mathbf{X}, \mathbf{Y} \rangle_{\mathrm{F}}$ to streamline calculations.

Lemma E.3 (Closed form for $\tilde{\kappa}$)**.** *For $\boldsymbol{\Sigma} \in \mathcal{C}^*$, we have*

$$\tilde{\kappa}(\boldsymbol{\Sigma}) = 1 + \lambda_{\max}\big(\mathbf{A}_{\boldsymbol{\Sigma}}\mathbf{A}_{\boldsymbol{\Sigma}}^\top + \lambda_{\min}(\boldsymbol{\Sigma}_X)^{-1}\boldsymbol{\Sigma}_E\big).$$

In particular, we have

$$1 + \max\left\{\|\mathbf{A}_{\boldsymbol{\Sigma}}\|_{\mathrm{op}}^2, \frac{\lambda_{\max}(\boldsymbol{\Sigma}_E)}{\lambda_{\min}(\boldsymbol{\Sigma}_X)}\right\} \le \tilde{\kappa}(\boldsymbol{\Sigma}) \le 1 + \|\mathbf{A}_{\boldsymbol{\Sigma}}\|_{\mathrm{op}}^2 + \frac{\lambda_{\max}(\boldsymbol{\Sigma}_E)}{\lambda_{\min}(\boldsymbol{\Sigma}_X)}.$$

Proof. Abbreviate $\mathbf{A} \coloneqq g(\boldsymbol{\Sigma}) = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}$ and $\mathbf{S} \coloneqq \boldsymbol{\Sigma}^{1/2}$. Combining Lemmas E.1 and E.2 gives

$$\begin{aligned}
\tilde{\kappa}(\boldsymbol{\Sigma}) &= \sup_{\boldsymbol{\Delta}\in\mathbb{R}^{d\times d}:\|\boldsymbol{\Delta}\|_{\mathrm{F}}\le 1} \|Dg(\boldsymbol{\Sigma})[\mathbf{S}\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\mathbf{S}]\boldsymbol{\Sigma}_X^{1/2}\|_{\mathrm{F}}^2 \\
&= \sup_{\boldsymbol{\Delta}\in\mathbb{R}^{d\times d}:\|\boldsymbol{\Delta}\|_{\mathrm{F}}\le 1} \left\|\big(\mathbf{S}_Y\boldsymbol{\Delta}_X^\top + \boldsymbol{\Delta}_Y\mathbf{S}_X^\top - \mathbf{A}(\mathbf{S}_X\boldsymbol{\Delta}_X^\top + \boldsymbol{\Delta}_X\mathbf{S}_X^\top)\big)\boldsymbol{\Sigma}_X^{-1/2}\right\|_{\mathrm{F}}^2 \\
&= \sup_{\boldsymbol{\Delta}\in\mathbb{R}^{d\times d}:\|\boldsymbol{\Delta}\|_{\mathrm{F}}\le 1} \left\|\big((\boldsymbol{\Delta}_Y - \mathbf{A}\boldsymbol{\Delta}_X)\mathbf{S}_X^\top + (\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\boldsymbol{\Delta}_X^\top\big)\boldsymbol{\Sigma}_X^{-1/2}\right\|_{\mathrm{F}}^2.
\end{aligned}$$

To simplify the objective, write $\mathbf{Q} \coloneqq \mathbf{S}_X^\top\boldsymbol{\Sigma}_X^{-1/2} \in \mathbb{R}^{d\times p}$, noting that $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_p$ and that $\mathbf{Q}\mathbf{Q}^\top \in \mathbb{R}^{d\times d}$ is the orthogonal projection onto the row space of $\mathbf{S}_X$. Further define the residual factor $\mathbf{E} \coloneqq \mathbf{S}_Y - \mathbf{A}\mathbf{S}_X \in \mathbb{R}^{m\times d}$, which by Lemma B.1 satisfies $\mathbf{E}\mathbf{E}^\top = \boldsymbol{\Sigma}_E$ and $\mathbf{E}\mathbf{S}_X^\top = \mathbf{0}$ (implying $\mathbf{E}\mathbf{Q} = \mathbf{0}$). Substituting these definitions into the equation above and rearranging, we obtain

$$\begin{aligned}
\tilde{\kappa}(\boldsymbol{\Sigma}) &= \sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}\le 1} \left\|(\boldsymbol{\Delta}_Y - \mathbf{A}\boldsymbol{\Delta}_X)\mathbf{Q} + \mathbf{E}\boldsymbol{\Delta}_X^\top\boldsymbol{\Sigma}_X^{-1/2}\right\|_{\mathrm{F}}^2. \\
&= \sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}\le 1} \Big\|\underbrace{\mathbf{E}\boldsymbol{\Delta}_X^\top\boldsymbol{\Sigma}_X^{-1/2} - \mathbf{A}\boldsymbol{\Delta}_X\mathbf{Q}}_{\text{feature perturbations}} + \underbrace{\boldsymbol{\Delta}_Y\mathbf{Q}}_{\text{response perturbations}}\Big\|_{\mathrm{F}}^2.
\end{aligned}$$

The first component depends on the perturbation $\boldsymbol{\Delta}_X$ to the feature factor $\mathbf{S}_X$, while the second component depends on the perturbation $\boldsymbol{\Delta}_Y$ to the response factor $\mathbf{S}_Y$.

To simplify further, we use the variational form of the Frobenius norm. For $\mathbf{B} \in \mathbb{R}^{m\times p}$, define

$$\phi(\mathbf{B};\boldsymbol{\Delta}) \coloneqq \big\langle \mathbf{B}, \mathbf{E}\boldsymbol{\Delta}_X^\top\boldsymbol{\Sigma}_X^{-1/2} - \mathbf{A}\boldsymbol{\Delta}_X\mathbf{Q} + \boldsymbol{\Delta}_Y\mathbf{Q}\big\rangle_{\mathrm{F}}.$$

Since the Frobenius unit ball is symmetric, we have

$$\tilde{\kappa}(\boldsymbol{\Sigma}) = \sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}\le 1}\left(\sup_{\|\mathbf{B}\|_{\mathrm{F}}\le 1}\phi(\mathbf{B};\boldsymbol{\Delta})\right)^2 = \sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}\le 1}\sup_{\|\mathbf{B}\|_{\mathrm{F}}\le 1}\phi(\mathbf{B};\boldsymbol{\Delta})^2.$$

We now swap these suprema and maximize over $\boldsymbol{\Delta}$ for fixed $\mathbf{B}$. Grouping terms according to the perturbed row block and using cyclicity of the trace gives

$$\phi(\mathbf{B};\boldsymbol{\Delta}) = \big\langle \boldsymbol{\Sigma}_X^{-1/2}\mathbf{B}^\top\mathbf{E} - \mathbf{A}^\top\mathbf{B}\mathbf{Q}^\top,\, \boldsymbol{\Delta}_X\big\rangle_{\mathrm{F}} + \big\langle \mathbf{B}\mathbf{Q}^\top, \boldsymbol{\Delta}_Y\big\rangle_{\mathrm{F}}.$$

Since the Frobenius norm is self-dual,

$$\begin{aligned}
\sup_{\|\boldsymbol{\Delta}\|_{\mathrm{F}}\le 1}\phi(\mathbf{B};\boldsymbol{\Delta})^2 &= \left\|\begin{pmatrix}\boldsymbol{\Sigma}_X^{-1/2}\mathbf{B}^\top\mathbf{E} - \mathbf{A}^\top\mathbf{B}\mathbf{Q}^\top \\ \mathbf{B}\mathbf{Q}^\top\end{pmatrix}\right\|_{\mathrm{F}}^2 \\
&= \left\|\boldsymbol{\Sigma}_X^{-1/2}\mathbf{B}^\top\mathbf{E} - \mathbf{A}^\top\mathbf{B}\mathbf{Q}^\top\right\|_{\mathrm{F}}^2 + \|\mathbf{B}\mathbf{Q}^\top\|_{\mathrm{F}}^2 \\
&= \underbrace{\|\boldsymbol{\Sigma}_X^{-1/2}\mathbf{B}^\top\mathbf{E}\|_{\mathrm{F}}^2 + \|\mathbf{A}^\top\mathbf{B}\|_{\mathrm{F}}^2}_{\text{feature contribution}} + \underbrace{\|\mathbf{B}\|_{\mathrm{F}}^2}_{\text{response contribution}}.
\end{aligned}$$

The last equality uses $\mathbf{E}\mathbf{Q} = \mathbf{0}$ to eliminate the cross term within the feature contribution and applies $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_p$ to simplify the other norms. For fixed $\mathbf{B}$, restricting to $\boldsymbol{\Delta}_Y = \mathbf{0}$ leaves exactly the first bracket, while restricting to $\boldsymbol{\Delta}_X = \mathbf{0}$ leaves the final term.

By homogeneity, the outer supremum may be restricted to $\|\mathbf{B}\|_{\mathrm{F}} = 1$. The response contribution is then 1 for every $\mathbf{B}$. Using $\mathbf{E}\mathbf{E}^\top = \mathbf{\Sigma}_E$ and Lemma A.8, we obtain

$$\begin{aligned}\tilde{\kappa}(\mathbf{\Sigma}) &= 1 + \sup_{\|\mathbf{B}\|_{\mathrm{F}}=1} \left\{\mathrm{tr}(\mathbf{\Sigma}_X^{-1}\mathbf{B}^\top\mathbf{\Sigma}_E\mathbf{B}) + \mathrm{tr}(\mathbf{A}\mathbf{A}^\top\mathbf{B}\mathbf{B}^\top)\right\} \\ &= 1 + \sup_{\|\mathbf{B}\|_{\mathrm{F}}=1} \mathrm{vec}(\mathbf{B})^\top \left(\mathbf{\Sigma}_X^{-1} \otimes \mathbf{\Sigma}_E + \mathbf{I}_p \otimes \mathbf{A}\mathbf{A}^\top\right) \mathrm{vec}(\mathbf{B}) \\ &= 1 + \left\|\mathbf{\Sigma}_X^{-1} \otimes \mathbf{\Sigma}_E + \mathbf{I}_p \otimes \mathbf{A}\mathbf{A}^\top\right\|_{\mathrm{op}}.\end{aligned}$$

The remaining operator norm measures worst-case sensitivity to feature perturbations alone.

To compute it, eigendecompose $\mathbf{\Sigma}_X = \mathbf{U}\,\mathrm{diag}(s_1, \ldots, s_p)\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{O}^p$ and $0 < s_1 \leq \cdots \leq s_p$. Conjugating by $\mathbf{U} \otimes \mathbf{I}_m$ gives a block-diagonal matrix with blocks $s_i^{-1}\mathbf{\Sigma}_E + \mathbf{A}\mathbf{A}^\top$. Since $s_i^{-1} \leq s_1^{-1}$ and $\mathbf{\Sigma}_E \succeq \mathbf{0}$, the first block dominates all others in Loewner order. Consequently,

$$\tilde{\kappa}(\mathbf{\Sigma}) = \underbrace{1}_{\text{response contribution}} + \underbrace{\lambda_{\max}\left(\mathbf{A}\mathbf{A}^\top + s_1^{-1}\mathbf{\Sigma}_E\right)}_{\text{feature contribution}}.$$

This proves the formula since $s_1 = \lambda_{\min}(\mathbf{\Sigma}_X)$.

Finally, positive semidefiniteness of the two matrices inside $\lambda_{\max}$ and subadditivity of $\lambda_{\max}$ give

$$\max\left\{\|\mathbf{A}\|_{\mathrm{op}}^2,\, s_1^{-1}\lambda_{\max}(\mathbf{\Sigma}_E)\right\} \leq \tilde{\kappa}(\mathbf{\Sigma}) - 1 \leq \|\mathbf{A}\|_{\mathrm{op}}^2 + s_1^{-1}\lambda_{\max}(\mathbf{\Sigma}_E),$$

establishing the remaining bounds. □

We next show that κ and $\tilde{\kappa}$ coincide. The upper bound is relatively direct in light of (12).

Lemma E.4 (Condition number comparison, upper bound)**.** *For all $\mathbf{\Sigma} \in \mathcal{C}^*$, $\kappa(\mathbf{\Sigma}) \leq \tilde{\kappa}(\mathbf{\Sigma})$.*

Proof. We upper bound $\kappa(\mathbf{\Sigma})$ in (4) by considering the feasible parameter $\mathbf{A} = \mathbf{A}_{\mathbf{\Sigma}} = g(\mathbf{\Sigma})$. This gives

$$\begin{aligned}\kappa(\mathbf{\Sigma}) &\leq \limsup_{\varepsilon\downarrow 0} \varepsilon^{-2} \sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} R(g(\mathbf{\Sigma}), \mathbf{\Sigma}') \\ &\leq \limsup_{\varepsilon\downarrow 0} \varepsilon^{-2} \sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} \sup_{\mathbf{\Sigma}'' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma}')} R(g(\mathbf{\Sigma}''), \mathbf{\Sigma}').\end{aligned}$$

Noting that the convergence in (12) is uniform over $\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})$, that expansion implies

$$\begin{aligned}\kappa(\mathbf{\Sigma}) &\leq \limsup_{\varepsilon\downarrow 0} \varepsilon^{-2} \left(\sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} \tilde{\kappa}(\mathbf{\Sigma}')\varepsilon^2 + o(\varepsilon^2) \right) \\ &= \limsup_{\varepsilon\downarrow 0} \sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} \tilde{\kappa}(\mathbf{\Sigma}') \\ &= \tilde{\kappa}(\mathbf{\Sigma}).\end{aligned}$$

The final equality uses continuity of $\tilde{\kappa}$, which is evident from our closed form in Lemma E.3. □

We prove a slightly stronger result in the other direction (using $\liminf$ instead of $\limsup$ below).

Lemma E.5 (Condition number comparison, lower bound)**.** *For all $\mathbf{\Sigma} \in \mathcal{C}^*$, we have*

$$\kappa(\mathbf{\Sigma}) \geq J_{\mathbf{\Sigma}} \coloneqq \liminf_{\varepsilon\downarrow 0} \varepsilon^{-2} \inf_{\mathbf{A} \in \mathcal{A}} \sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} R(\mathbf{A}, \mathbf{\Sigma}') \geq \tilde{\kappa}(\mathbf{\Sigma}).$$

Proof. The first inequality is immediate from (4), since $\limsup \geq \liminf$.

For the second inequality, fix $\mathbf{\Delta}_\star \in \mathbb{R}^{d\times d}$ with $\|\mathbf{\Delta}_\star\|_{\mathrm{F}} = 1$ achieving the supremum in formula (11) for $\tilde{\kappa}(\mathbf{\Sigma})$, noting that a maximizer exists by compactness of the Frobenius ball and continuity

of the objective. Now, for each $t \in \mathbb{R}$, define $\mathbf{\Sigma}^t := (\mathbf{\Sigma}^{1/2} + t\mathbf{\Delta}_\star)(\mathbf{\Sigma}^{1/2} + t\mathbf{\Delta}_\star)^\top$, which satisfies $\mathsf{B}(\mathbf{\Sigma}^t, \mathbf{\Sigma}) \le |t|$ by Lemma 1.1. In particular, $\mathbf{\Sigma}^{\pm\varepsilon} \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})$. We thus have

$$\begin{aligned}
J_{\mathbf{\Sigma}} &:= \liminf_{\varepsilon \downarrow 0} \varepsilon^{-2} \inf_{\mathbf{A} \in \mathcal{A}} \sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} R(\mathbf{A}, \mathbf{\Sigma}') \\
&\ge \liminf_{\varepsilon \downarrow 0} \varepsilon^{-2} \inf_{\mathbf{A} \in \mathcal{A}} \max\big\{R(\mathbf{A}, \mathbf{\Sigma}^{+\varepsilon}), R(\mathbf{A}, \mathbf{\Sigma}^{-\varepsilon})\big\} \\
&= \liminf_{\varepsilon \downarrow 0} \varepsilon^{-2} \inf_{\mathbf{A} \in \mathcal{A}} \max\Big\{\big\|(\mathbf{A} - \mathbf{A}_{+\varepsilon})(\mathbf{\Sigma}_X^{+\varepsilon})^{1/2}\big\|_{\mathrm{F}}^2, \big\|(\mathbf{A} - \mathbf{A}_{-\varepsilon})(\mathbf{\Sigma}_X^{-\varepsilon})^{1/2}\big\|_{\mathrm{F}}^2\Big\}, \qquad (13)
\end{aligned}$$

where $\mathbf{A}_t := g(\mathbf{\Sigma}^t)$. Since $\mathbf{\Sigma}_X^t \to \mathbf{\Sigma}_X$ as $t \to 0$, we have for fixed $\lambda \in (0,1)$ that $\mathbf{\Sigma}_X^t \succ \lambda \mathbf{\Sigma}_X$ for sufficiently small t. Consequently, we have

$$\begin{aligned}
J_{\mathbf{\Sigma}} &\ge \sup_{\lambda \in (0,1)} \lambda \liminf_{\varepsilon \downarrow 0} \varepsilon^{-2} \inf_{\mathbf{A} \in \mathcal{A}} \max\Big\{\big\|(\mathbf{A} - \mathbf{A}_{+\varepsilon})\mathbf{\Sigma}_X^{1/2}\big\|_{\mathrm{F}}^2, \big\|(\mathbf{A} - \mathbf{A}_{-\varepsilon})\mathbf{\Sigma}_X^{1/2}\big\|_{\mathrm{F}}^2\Big\} \\
&= \liminf_{\varepsilon \downarrow 0} \varepsilon^{-2} \inf_{\mathbf{A} \in \mathcal{A}} \max\Big\{\big\|(\mathbf{A} - \mathbf{A}_{+\varepsilon})\mathbf{\Sigma}_X^{1/2}\big\|_{\mathrm{F}}^2, \big\|(\mathbf{A} - \mathbf{A}_{-\varepsilon})\mathbf{\Sigma}_X^{1/2}\big\|_{\mathrm{F}}^2\Big\}.
\end{aligned}$$

By Lemma A.7, the infimum is attained by the midpoint $\bar{\mathbf{A}}_\varepsilon = (\mathbf{A}_{+\varepsilon} + \mathbf{A}_{-\varepsilon})/2$. This gives,

$$J_{\mathbf{\Sigma}} \ge \frac{1}{4} \liminf_{\varepsilon \downarrow 0} \varepsilon^{-2} \big\|(\mathbf{A}_{+\varepsilon} - \mathbf{A}_{-\varepsilon})\mathbf{\Sigma}_X^{1/2}\big\|_{\mathrm{F}}^2$$

By construction, we have

$$\mathbf{\Sigma}^{+\varepsilon} - \mathbf{\Sigma}^{-\varepsilon} = 2\varepsilon\big(\mathbf{\Sigma}^{1/2}\mathbf{\Delta}_\star^\top + \mathbf{\Delta}_\star\mathbf{\Sigma}^{1/2}\big),$$

which, since g is continuously differentiable, gives

$$\begin{aligned}
\mathbf{A}_{+\varepsilon} - \mathbf{A}_{-\varepsilon} &= g(\mathbf{\Sigma}^{+\varepsilon}) - g(\mathbf{\Sigma}^{-\varepsilon}) \\
&= Dg(\mathbf{\Sigma})[\mathbf{\Sigma}^{+\varepsilon} - \mathbf{\Sigma}] - Dg(\mathbf{\Sigma})[\mathbf{\Sigma}^{-\varepsilon} - \mathbf{\Sigma}] + o(\varepsilon) \\
&= Dg(\mathbf{\Sigma})[\mathbf{\Sigma}^{+\varepsilon} - \mathbf{\Sigma}^{-\varepsilon}] + o(\varepsilon) \\
&= 2\varepsilon Dg(\mathbf{\Sigma})\big[\mathbf{\Sigma}^{1/2}\mathbf{\Delta}_\star^\top + \mathbf{\Delta}_\star\mathbf{\Sigma}^{1/2}\big] + o(\varepsilon)
\end{aligned}$$

as $\varepsilon \to 0$. Substituting this expansion into the preceding lower bound for $J_{\mathbf{\Sigma}}$ and using optimality of $\mathbf{\Delta}_\star$, we obtain

$$\begin{aligned}
J_{\mathbf{\Sigma}} &\ge \big\|Dg(\mathbf{\Sigma})\big[\mathbf{\Sigma}^{1/2}\mathbf{\Delta}_\star^\top + \mathbf{\Delta}_\star\mathbf{\Sigma}^{1/2}\big]\mathbf{\Sigma}_X^{1/2}\big\|_{\mathrm{F}}^2 \\
&= \sup_{\mathbf{\Delta} \in \mathbb{R}^{d \times d} : \|\mathbf{\Delta}\|_{\mathrm{F}} \le 1} \|Dg(\mathbf{\Sigma})[\mathbf{\Sigma}^{1/2}\mathbf{\Delta}^\top + \mathbf{\Delta}\mathbf{\Sigma}^{1/2}]\mathbf{\Sigma}_X^{1/2}\|_{\mathrm{F}}^2 \\
&= \tilde{\kappa}(\mathbf{\Sigma}),
\end{aligned}$$

as desired. □

Combining Lemma E.4 and Lemma E.5 with (4) gives

$$\kappa(\mathbf{\Sigma}) = \tilde{\kappa}(\mathbf{\Sigma}) = \lim_{\varepsilon \downarrow 0} \varepsilon^{-2} \inf_{\mathbf{A} \in \mathcal{A}} \sup_{\mathbf{\Sigma}' \in \mathcal{B}_\varepsilon(\mathbf{\Sigma})} R(\mathbf{A}, \mathbf{\Sigma}') = \lim_{\varepsilon \downarrow 0} \varepsilon^{-2} R_\varepsilon^\star(\mathbf{\Sigma})$$

for $\mathbf{\Sigma} \in \mathcal{C}^*$. The theorem then follows by the Lemma E.3 formula and inequalities. □

F Remaining Proofs for Section 3

F.1 Regret Transfer (Proof of Lemma 3.1)

We first establish a refined bound by comparing projected residual factors. The argument separates the direct effect of the perturbation from the change in the feature row space, and it allows the perturbed feature second-moment matrix to be singular.

Lemma F.1 (Regret stability, refined)**.** *For all $\varepsilon \ge 0$, $\mathbf{A} \in \mathcal{A}$, $P \in \mathcal{P}^*$, and $Q \in \mathcal{B}_\varepsilon(P)$, we have*

$$\Big|\sqrt{R(\mathbf{A}, P)} - \sqrt{R(\mathbf{A}, Q)}\Big| \le \left(\max\Big\{\|\mathbf{A} - \mathbf{A}_P\|_{\mathrm{op}}, \sqrt{1 + \|\mathbf{A}\|_{\mathrm{op}}^2}\Big\} + \sqrt{\frac{\lambda_{\max}(\mathbf{\Sigma}_E^P)}{\lambda_{\min}(\mathbf{\Sigma}_X^P)}}\right)\varepsilon.$$

Proof. We adopt the Bures-space conventions and block decompositions of Appendix B. Set $\mathbf{\Sigma} \coloneqq \mathbf{\Sigma}_P$ and $\bar{\mathbf{\Sigma}} \coloneqq \mathbf{\Sigma}_Q$. By Lemmas 1.1 and 1.2, there are square-root factors $\mathbf{S}, \bar{\mathbf{S}} \in \mathbb{R}^{d\times d}$ such that

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{S}^\top, \qquad \bar{\mathbf{\Sigma}} = \bar{\mathbf{S}}\bar{\mathbf{S}}^\top, \qquad \mathbf{\Delta} \coloneqq \bar{\mathbf{S}} - \mathbf{S}, \qquad \|\mathbf{\Delta}\|_{\mathrm{F}} \le \varepsilon.$$

Let $\mathbf{\Pi}_X$ and $\bar{\mathbf{\Pi}}_X$ denote the orthogonal projections onto the row spaces of $\mathbf{S}_X$ and $\bar{\mathbf{S}}_X$, respectively, and define the residual factors

$$\mathbf{E}_{\mathbf{A}} \coloneqq \mathbf{S}_Y - \mathbf{A}\mathbf{S}_X, \qquad \bar{\mathbf{E}}_{\mathbf{A}} \coloneqq \bar{\mathbf{S}}_Y - \mathbf{A}\bar{\mathbf{S}}_X, \qquad \mathbf{E} \coloneqq \mathbf{E}_{\mathbf{A}_{\mathbf{\Sigma}}}.$$

By Lemma B.1, we have

$$\sqrt{R(\mathbf{A}, \mathbf{\Sigma})} = \|\mathbf{E}_{\mathbf{A}}\mathbf{\Pi}_X\|_{\mathrm{F}}, \qquad \sqrt{R(\mathbf{A}, \bar{\mathbf{\Sigma}})} = \|\bar{\mathbf{E}}_{\mathbf{A}}\bar{\mathbf{\Pi}}_X\|_{\mathrm{F}},$$

and the optimal residual factor $\mathbf{E}$ satisfies $\mathbf{E}\mathbf{\Pi}_X = \mathbf{0}$ and $\mathbf{E}\mathbf{E}^\top = \mathbf{\Sigma}_E$.

We compare these projected residuals directly. Since $\mathbf{S}_X\mathbf{\Pi}_X = \mathbf{S}_X$ and $\bar{\mathbf{S}}_X\bar{\mathbf{\Pi}}_X = \bar{\mathbf{S}}_X$, we have

$$\begin{aligned}\mathbf{E}_{\mathbf{A}}\mathbf{\Pi}_X - \bar{\mathbf{E}}_{\mathbf{A}}\bar{\mathbf{\Pi}}_X &= (\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}_X - (\bar{\mathbf{S}}_Y - \mathbf{A}\bar{\mathbf{S}}_X)\bar{\mathbf{\Pi}}_X \\ &= \mathbf{S}_Y(\mathbf{\Pi}_X - \bar{\mathbf{\Pi}}_X) - \mathbf{\Delta}_Y\bar{\mathbf{\Pi}}_X + \mathbf{A}\mathbf{\Delta}_X.\end{aligned}$$

Since our desired bound does not depend on $\mathbf{S}_Y$ or $\mathbf{\Sigma}_Y$ directly, we substitute $\mathbf{S}_Y = \mathbf{A}_{\mathbf{\Sigma}}\mathbf{S}_X + \mathbf{E}$ for the first term to obtain

$$\begin{aligned}\mathbf{S}_Y(\mathbf{\Pi}_X - \bar{\mathbf{\Pi}}_X) &= (\mathbf{A}_{\mathbf{\Sigma}}\mathbf{S}_X + \mathbf{E})(\mathbf{\Pi}_X - \bar{\mathbf{\Pi}}_X) \\ &= \mathbf{A}_{\mathbf{\Sigma}}\mathbf{S}_X - \mathbf{A}_{\mathbf{\Sigma}}\mathbf{S}_X\bar{\mathbf{\Pi}}_X - \mathbf{E}\bar{\mathbf{\Pi}}_X \qquad (\mathbf{S}_X\mathbf{\Pi}_X = \mathbf{S}_X, \mathbf{E}\mathbf{\Pi}_X = \mathbf{0}) \\ &= \mathbf{A}_{\mathbf{\Sigma}}\mathbf{S}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) - \mathbf{E}\bar{\mathbf{\Pi}}_X \\ &= \mathbf{A}_{\mathbf{\Sigma}}\bar{\mathbf{S}}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) - \mathbf{A}_{\mathbf{\Sigma}}\mathbf{\Delta}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) - \mathbf{E}\bar{\mathbf{\Pi}}_X \\ &= -\mathbf{A}_{\mathbf{\Sigma}}\mathbf{\Delta}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) - \mathbf{E}\bar{\mathbf{\Pi}}_X,\end{aligned}$$

where the last equality uses $\bar{\mathbf{S}}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) = \mathbf{0}$. Substituting this into the previous equation gives

$$\begin{aligned}\mathbf{E}_{\mathbf{A}}\mathbf{\Pi}_X - \bar{\mathbf{E}}_{\mathbf{A}}\bar{\mathbf{\Pi}}_X &= -\mathbf{A}_{\mathbf{\Sigma}}\mathbf{\Delta}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) - \mathbf{E}\bar{\mathbf{\Pi}}_X - \mathbf{\Delta}_Y\bar{\mathbf{\Pi}}_X + \mathbf{A}\mathbf{\Delta}_X \\ &= (\mathbf{A} - \mathbf{A}_{\mathbf{\Sigma}})\mathbf{\Delta}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) + (\mathbf{A}\mathbf{\Delta}_X - \mathbf{\Delta}_Y)\bar{\mathbf{\Pi}}_X - \mathbf{E}\bar{\mathbf{\Pi}}_X.\end{aligned}$$

Let $\mathbf{T}$ denote the sum of the first two terms in the last line, and write

$$c \coloneqq \max\Big\{\|\mathbf{A} - \mathbf{A}_{\mathbf{\Sigma}}\|_{\mathrm{op}},\ \sqrt{1 + \|\mathbf{A}\|_{\mathrm{op}}^2}\Big\}.$$

Note that $1 + \|\mathbf{A}\|_{\mathrm{op}}^2 = \lambda_{\max}(\mathbf{I}_m + \mathbf{A}\mathbf{A}^\top) = \|\begin{pmatrix}\mathbf{A} & -\mathbf{I}_m\end{pmatrix}\|_{\mathrm{op}}^2$. Moreover, the two terms of $\mathbf{T}$ have orthogonal right projections, so we obtain

$$\begin{aligned}\|\mathbf{T}\|_{\mathrm{F}}^2 &= \|(\mathbf{A} - \mathbf{A}_{\mathbf{\Sigma}})\mathbf{\Delta}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X)\|_{\mathrm{F}}^2 + \|(\mathbf{A}\mathbf{\Delta}_X - \mathbf{\Delta}_Y)\bar{\mathbf{\Pi}}_X\|_{\mathrm{F}}^2 \\ &\le \|\mathbf{A} - \mathbf{A}_{\mathbf{\Sigma}}\|_{\mathrm{op}}^2\|\mathbf{\Delta}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X)\|_{\mathrm{F}}^2 + \|\begin{pmatrix}\mathbf{A} & -\mathbf{I}_m\end{pmatrix}\|_{\mathrm{op}}^2\|\mathbf{\Delta}\bar{\mathbf{\Pi}}_X\|_{\mathrm{F}}^2 \qquad \text{(Lemma A.1)} \\ &\le c^2\big(\|\mathbf{\Delta}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X)\|_{\mathrm{F}}^2 + \|\mathbf{\Delta}_X\bar{\mathbf{\Pi}}_X\|_{\mathrm{F}}^2 + \|\mathbf{\Delta}_Y\bar{\mathbf{\Pi}}_X\|_{\mathrm{F}}^2\big) \\ &\le c^2\big(\|\mathbf{\Delta}_X\|_{\mathrm{F}}^2 + \|\mathbf{\Delta}_Y\|_{\mathrm{F}}^2\big) = c^2\|\mathbf{\Delta}\|_{\mathrm{F}}^2.\end{aligned}$$

The last inequality uses the orthogonal decomposition of $\mathbf{\Delta}_X$ and the contraction property of $\bar{\mathbf{\Pi}}_X$.

It remains to bound $\mathbf{E}\bar{\mathbf{\Pi}}_X$, the component of the original optimal residual in the perturbed feature row space. Since $\mathbf{E}\mathbf{\Pi}_X = \mathbf{0}$, the row-space perturbation bound (Lemma A.6) gives

$$\begin{aligned}\|\mathbf{E}\bar{\mathbf{\Pi}}_X\|_{\mathrm{F}} &= \|\mathbf{E}(\mathbf{I}_d - \mathbf{\Pi}_X)\bar{\mathbf{\Pi}}_X\|_{\mathrm{F}} \\ &\le \|\mathbf{E}\|_{\mathrm{op}}\|(\mathbf{I}_d - \mathbf{\Pi}_X)\bar{\mathbf{\Pi}}_X\|_{\mathrm{F}} \\ &\le \frac{\|\mathbf{E}\|_{\mathrm{op}}}{\sigma_{\min}(\mathbf{S}_X)}\|\mathbf{\Delta}_X\|_{\mathrm{F}} \\ &\le \sqrt{\frac{\lambda_{\max}(\mathbf{\Sigma}_E)}{\lambda_{\min}(\mathbf{\Sigma}_X)}}\,\varepsilon.\end{aligned}$$

Here $\mathbf{S}_X$ has full row rank because $\mathbf{\Sigma}_X \succ \mathbf{0}$; no rank assumption on $\bar{\mathbf{S}}_X$ is needed.

Finally, the reverse triangle inequality yields

$$\begin{aligned}\left|\sqrt{R(\mathbf{A},\boldsymbol{\Sigma})}-\sqrt{R(\mathbf{A},\bar{\boldsymbol{\Sigma}})}\right| &\le \|\mathbf{E}_{\mathbf{A}}\boldsymbol{\Pi}_X-\bar{\mathbf{E}}_{\mathbf{A}}\bar{\boldsymbol{\Pi}}_X\|_{\mathrm{F}}\\ &\le \|\mathbf{T}\|_{\mathrm{F}}+\|\mathbf{E}\bar{\boldsymbol{\Pi}}_X\|_{\mathrm{F}}\\ &\le \left(c+\sqrt{\frac{\lambda_{\max}(\boldsymbol{\Sigma}_E)}{\lambda_{\min}(\boldsymbol{\Sigma}_X)}}\right)\varepsilon.\end{aligned}$$

Translating back to P and Q proves the claim. □

Proof of Lemma 3.1. By Theorem 2.2,

$$\sqrt{\kappa_P}\ge\max\left\{\sqrt{1+\|\mathbf{A}_P\|_{\mathrm{op}}^2},\sqrt{\frac{\lambda_{\max}(\boldsymbol{\Sigma}_E^P)}{\lambda_{\min}(\boldsymbol{\Sigma}_X^P)}}\right\}.$$

Moreover, the triangle inequality for $\|\cdot\|_{\mathrm{op}}$ and subadditivity of the square root imply

$$\begin{aligned}\max\left\{\|\mathbf{A}-\mathbf{A}_P\|_{\mathrm{op}},\sqrt{1+\|\mathbf{A}\|_{\mathrm{op}}^2}\right\} &\le \|\mathbf{A}\|_{\mathrm{op}}+\max\{\|\mathbf{A}_P\|_{\mathrm{op}},1\}\\ &\le \|\mathbf{A}\|_{\mathrm{op}}+\sqrt{\kappa_P}.\end{aligned}$$

Applying Lemma F.1 with $\varepsilon\coloneqq\mathsf{W}_2(P,Q)$ thus gives

$$\left|\sqrt{R(\mathbf{A},P)}-\sqrt{R(\mathbf{A},Q)}\right|\le\big(\|\mathbf{A}\|_{\mathrm{op}}+2\sqrt{\kappa_P}\big)\mathsf{W}_2(P,Q).$$

For $\mathbf{A}=\mathbf{A}_P$, we have $R(\mathbf{A}_P,P)=0$, and so Lemma F.1 gives

$$\begin{aligned}R(\mathbf{A}_P,Q) &\le \left(\sqrt{1+\|\mathbf{A}_P\|_{\mathrm{op}}^2}+\sqrt{\frac{\lambda_{\max}(\boldsymbol{\Sigma}_E^P)}{\lambda_{\min}(\boldsymbol{\Sigma}_X^P)}}\right)^2\mathsf{W}_2(P,Q)^2\\ &\le 4\kappa_P\mathsf{W}_2(P,Q)^2,\end{aligned}$$

proving the second claim. □

F.2 Stability Implications for ERM and DRO (Remaining Proof for Lemma 3.2)

We employ a few helper lemmas, adopting the notation and block decompositions from Appendix B.

Lemma F.2 (Perturbed condition number)**.** *Fix $\boldsymbol{\Sigma}\in\mathcal{C}^*$ and $\bar{\boldsymbol{\Sigma}}\in\mathbb{S}_+^d$ with $\mathsf{B}(\bar{\boldsymbol{\Sigma}},\boldsymbol{\Sigma})\le\sqrt{\lambda_{\min}(\boldsymbol{\Sigma}_X)}/2$. Then $\bar{\boldsymbol{\Sigma}}\in\mathcal{C}^*$ and $\kappa(\bar{\boldsymbol{\Sigma}})\le 19\kappa(\boldsymbol{\Sigma})$.*

Proof. By Theorem 2.2, the condition number is determined up to constants by the operator norm of the optimal coefficient matrix, the maximum eigenvalue of the residual second-moment matrix, and the minimum eigenvalue of the feature second-moment matrix. We now control each of these quantities for $\bar{\boldsymbol{\Sigma}}$ in terms of those for $\boldsymbol{\Sigma}$, beginning with the minimum feature eigenvalue.

Set $\varepsilon\coloneqq\mathsf{B}(\bar{\boldsymbol{\Sigma}},\boldsymbol{\Sigma})$, $\lambda\coloneqq\lambda_{\min}(\boldsymbol{\Sigma}_X)>0$ and $\bar{\lambda}\coloneqq\lambda_{\min}(\bar{\boldsymbol{\Sigma}}_X)$. By Lemma B.3,

$$\sqrt{\bar{\lambda}}\ge\sqrt{\lambda}-\varepsilon\ge\sqrt{\lambda}/2,$$

since $\varepsilon\le\sqrt{\lambda}/2$. Thus $\bar{\lambda}\ge\lambda/4$ and $\bar{\boldsymbol{\Sigma}}\in\mathcal{C}^*$.

Next, let $\mathbf{A}\coloneqq\mathbf{A}_{\boldsymbol{\Sigma}}$, $\bar{\mathbf{A}}\coloneqq\mathbf{A}_{\bar{\boldsymbol{\Sigma}}}$, and $k\coloneqq\kappa(\boldsymbol{\Sigma})$. We then bound

$$\begin{aligned}\sqrt{\bar{\lambda}}\,\|\bar{\mathbf{A}}-\mathbf{A}\|_{\mathrm{op}} &\le \sqrt{\bar{\lambda}}\,\|\bar{\mathbf{A}}-\mathbf{A}\|_{\mathrm{F}} && \text{(Lemma A.1)}\\ &\le \big\|(\bar{\mathbf{A}}-\mathbf{A})\bar{\boldsymbol{\Sigma}}_X^{1/2}\big\|_{\mathrm{F}}\\ &= \sqrt{R(\mathbf{A},\bar{\boldsymbol{\Sigma}})} && \text{(regret formula)}\\ &\le 2\sqrt{k}\,\varepsilon. && \text{(Lemma 3.1)}\end{aligned}$$

Since $\varepsilon/\sqrt{\bar{\lambda}} \le 1$ and $\|\mathbf{A}\|_{\mathrm{op}} \le \sqrt{k}$ by Theorem 2.2, it follows that

$$\|\bar{\mathbf{A}}\|_{\mathrm{op}} \le \|\mathbf{A}\|_{\mathrm{op}} + \frac{2\sqrt{k}\varepsilon}{\sqrt{\bar{\lambda}}} \le 3\sqrt{k}.$$

We now bound the residual relative to the smallest feature eigenvalue. By Lemma 1.1, there are square-root factors $\mathbf{S}, \bar{\mathbf{S}} \in \mathbb{R}^{d\times d}$ of $\mathbf{\Sigma}, \bar{\mathbf{\Sigma}}$ with $\mathbf{\Delta} \coloneqq \bar{\mathbf{S}} - \mathbf{S}$ satisfying $\|\mathbf{\Delta}\|_{\mathrm{F}} \le \varepsilon$. Define the corresponding residual factors and projection

$$\mathbf{E} \coloneqq \mathbf{S}_Y - \mathbf{A}\mathbf{S}_X, \qquad \bar{\mathbf{E}} \coloneqq \bar{\mathbf{S}}_Y - \bar{\mathbf{A}}\bar{\mathbf{S}}_X, \qquad \bar{\mathbf{\Pi}}_X \coloneqq \bar{\mathbf{S}}_X^{\dagger}\bar{\mathbf{S}}_X.$$

By Lemma B.1, $\mathbf{E}\mathbf{E}^\top = \mathbf{\Sigma}_E$, $\bar{\mathbf{E}}\bar{\mathbf{E}}^\top = \bar{\mathbf{\Sigma}}_E$, $\bar{\mathbf{E}} = \bar{\mathbf{S}}_Y(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X)$, and $\bar{\mathbf{S}}_X(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) = \mathbf{0}$. Further,

$$\bar{\mathbf{S}}_Y = \mathbf{S}_Y + \mathbf{\Delta}_Y = \mathbf{E} + \mathbf{A}\mathbf{S}_X + \mathbf{\Delta}_Y = \mathbf{E} + \mathbf{A}\bar{\mathbf{S}}_X - \mathbf{A}\mathbf{\Delta}_X + \mathbf{\Delta}_Y.$$

We thus compute

$$\begin{aligned}
\bar{\mathbf{E}} &= \bar{\mathbf{S}}_Y(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) \\
&= (\mathbf{E} + \mathbf{A}\bar{\mathbf{S}}_X - \mathbf{A}\mathbf{\Delta}_X + \mathbf{\Delta}_Y)(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X) \\
&= (\mathbf{E} + \mathbf{\Delta}_Y - \mathbf{A}\mathbf{\Delta}_X)(\mathbf{I}_d - \bar{\mathbf{\Pi}}_X).
\end{aligned}$$

Since the right-hand projection is a contraction, we have

$$\begin{aligned}
\|\bar{\mathbf{E}}\|_{\mathrm{op}} &\le \|\mathbf{E}\|_{\mathrm{op}} + \|\mathbf{\Delta}_Y - \mathbf{A}\mathbf{\Delta}_X\|_{\mathrm{op}} \\
&\le \sqrt{\|\mathbf{\Sigma}_E\|_{\mathrm{op}}} + \sqrt{1 + \|\mathbf{A}\|_{\mathrm{op}}^2}\,\varepsilon \\
&\le \sqrt{k}\,(\sqrt{\lambda} + \varepsilon).
\end{aligned}$$

Here the second inequality uses $\|(-\mathbf{A} \quad \mathbf{I}_m)\|_{\mathrm{op}} = \sqrt{1 + \|\mathbf{A}\|_{\mathrm{op}}^2}$, and the last uses Theorem 2.2. Therefore, we bound

$$\frac{\|\bar{\mathbf{\Sigma}}_E\|_{\mathrm{op}}}{\bar{\lambda}} = \frac{\|\bar{\mathbf{E}}\|_{\mathrm{op}}^2}{\bar{\lambda}} \le k\frac{(\sqrt{\lambda} + \varepsilon)^2}{(\sqrt{\lambda} - \varepsilon)^2} \le 9k,$$

where the last inequality uses $\varepsilon \le \sqrt{\lambda}/2$.

Finally, Theorem 2.2 gives

$$\kappa(\bar{\mathbf{\Sigma}}) \le 1 + \|\bar{\mathbf{A}}\|_{\mathrm{op}}^2 + \frac{\|\bar{\mathbf{\Sigma}}_E\|_{\mathrm{op}}}{\bar{\lambda}} \le 1 + 18k \le 19k$$

using $k \ge 1$. □

Lemma F.3 (DRO strong duality)**.** *Let $\varepsilon \ge 0$. For all $P \in \mathcal{P}^*$ and $\hat{P} \in \mathcal{B}_\varepsilon(P)$, we have*

$$\inf_{\mathbf{A}\in\mathcal{A}} \sup_{Q\in\mathcal{B}_\varepsilon(\hat{P})} L(\mathbf{A}, Q) = \max_{Q\in\mathcal{B}_\varepsilon(\hat{P})} \inf_{\mathbf{A}\in\mathcal{A}} L(\mathbf{A}, Q).$$

In particular, a maximizer exists for the right-hand side.

Proof. By Lemma 1.2, it suffices to prove

$$\inf_{\mathbf{A}\in\mathcal{A}} \max_{\mathbf{\Sigma}\in K} L(\mathbf{A}, \mathbf{\Sigma}) = \max_{\mathbf{\Sigma}\in K} \inf_{\mathbf{A}\in\mathcal{A}} L(\mathbf{A}, \mathbf{\Sigma}), \tag{14}$$

where $K \coloneqq \mathcal{B}_\varepsilon(\mathbf{\Sigma}_{\hat{P}})$. The factor representation in Lemma 1.2 shows that K is the continuous image of a compact Frobenius ball and hence compact. It is convex because Wasserstein balls are convex under mixtures, as follows by mixing transport couplings, and the second-moment map is affine. Since L is continuous, convex in $\mathbf{A}$, and affine in $\mathbf{\Sigma}$, Sion's minimax theorem (Sion, 1958) gives equality of the infimum–supremum values. Finally, $\mathbf{\Sigma} \mapsto \inf_{\mathbf{A}\in\mathcal{A}} L(\mathbf{A}, \mathbf{\Sigma})$ is upper semicontinuous as an infimum of continuous functions, so its supremum on K is attained. □

We now prove the remaining excess risk bounds for ERM and DRO.

Lemma F.4 (Stability implications, ERM)**.** *Let $P \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varepsilon > 0$. We then have*

$$R_\varepsilon(\mathsf{ERM}, P) \leq 4\varepsilon^2 \sup_{Q \in \mathcal{B}_\varepsilon(P)} \kappa_Q.$$

If additionally $\varepsilon^2 \leq \lambda_{\min}(\mathbf{\Sigma}_X^P)/4$, then $R_\varepsilon(\mathsf{ERM}, P) \leq 76\kappa_P\varepsilon^2$.

Proof. The first bound is trivial if $\mathcal{B}_\varepsilon(P)$ contains a law outside of $\mathcal{P}^*$, since that law has infinite condition number. We may therefore assume $\mathcal{B}_\varepsilon(P) \subseteq \mathcal{P}^*$.

For any $Q \in \mathcal{B}_\varepsilon(P)$, the ERM solutions at P and Q are unique. Applying Lemma 3.1 then gives

$$R(\mathsf{ERM}[P], Q) = R(\mathbf{A}_P, Q) \leq 4\kappa_P\varepsilon^2,$$
$$R(\mathsf{ERM}[Q], P) = R(\mathbf{A}_Q, P) \leq 4\kappa_Q\varepsilon^2.$$

Since $P \in \mathcal{B}_\varepsilon(P)$, both bounds are at most $4\varepsilon^2 \sup_{Q' \in \mathcal{B}_\varepsilon(P)} \kappa_{Q'}$. Taking the maximum over the two orientations and then the supremum over $Q \in \mathcal{B}_\varepsilon(P)$ proves the first bound.

If $\varepsilon^2 \leq \lambda_{\min}(\mathbf{\Sigma}_X^P)/4$, then Lemmas 1.2 and F.2 give $\kappa_Q \leq 19\kappa_P$ for every $Q \in \mathcal{B}_\varepsilon(P)$. Substitution yields the second bound. □

Lemma F.5 (Stability implications, DRO)**.** *Let $P \in \mathcal{P}_2(\mathbb{R}^d)$ and $\varepsilon > 0$. We then have*

$$R_\varepsilon(\mathsf{DRO}_\varepsilon, P) \leq 16\varepsilon^2 \sup_{Q \in \mathcal{B}_{2\varepsilon}(P)} \kappa_Q.$$

If $\varepsilon^2 \leq \lambda_{\min}(\mathbf{\Sigma}_X^P)/16$, this implies that $R_\varepsilon(\mathsf{DRO}_\varepsilon, P) \leq 304\kappa_P\varepsilon^2$.

Proof. The first bound is trivial if $\mathcal{B}_{2\varepsilon}(P)$ contains a law outside of $\mathcal{P}^*$, since that law has infinite condition number. We may therefore assume $\mathcal{B}_{2\varepsilon}(P) \subseteq \mathcal{P}^*$.

Fix $Q \in \mathcal{B}_\varepsilon(P)$ and let $(P_{\mathrm{s}}, P_{\mathrm{t}})$ be either (P, Q) or (Q, P). Set $\hat{\mathbf{A}} \coloneqq \mathsf{DRO}_\varepsilon[P_{\mathrm{s}}]$. We first identify $\hat{\mathbf{A}}$ as the ERM solution at a least-favorable law, then transfer its zero regret there to the target.

By Lemma F.3, there exists a maximizer $\hat{Q}$ of

$$\max_{Q' \in \mathcal{B}_\varepsilon(P_{\mathrm{s}})} \inf_{\mathbf{A} \in \mathcal{A}} L(\mathbf{A}, Q').$$

DRO optimality and the minimax identity (Lemma F.3) give

$$\begin{aligned} L(\hat{\mathbf{A}}, \hat{Q}) &\leq \sup_{Q' \in \mathcal{B}_\varepsilon(P_{\mathrm{s}})} L(\hat{\mathbf{A}}, Q') \\ &= \inf_{\mathbf{A} \in \mathcal{A}} \sup_{Q' \in \mathcal{B}_\varepsilon(P_{\mathrm{s}})} L(\mathbf{A}, Q') \\ &= \max_{Q' \in \mathcal{B}_\varepsilon(P_{\mathrm{s}})} \inf_{\mathbf{A} \in \mathcal{A}} L(\mathbf{A}, Q') \\ &= \inf_{\mathbf{A} \in \mathcal{A}} L(\mathbf{A}, \hat{Q}). \end{aligned}$$

The reverse inequality is automatic, so $\hat{\mathbf{A}}$ minimizes $L(\cdot, \hat{Q})$. Moreover, $\hat{Q} \in \mathcal{B}_\varepsilon(P_{\mathrm{s}}) \subseteq \mathcal{B}_{2\varepsilon}(P) \subseteq \mathcal{P}^*$. Its risk minimizer is therefore unique, implying $\hat{\mathbf{A}} = \mathbf{A}_{\hat{Q}}$.

Since both $\hat{Q}$ and P_{t} belong to $\mathcal{B}_\varepsilon(P_{\mathrm{s}})$, we have $\mathsf{W}_2(\hat{Q}, P_{\mathrm{t}}) \leq 2\varepsilon$. Applying Lemma 3.1 with reference law $\hat{Q}$ therefore gives

$$R(\hat{\mathbf{A}}, P_{\mathrm{t}}) = R(\mathbf{A}_{\hat{Q}}, P_{\mathrm{t}}) \leq 4\kappa_{\hat{Q}}\mathsf{W}_2(\hat{Q}, P_{\mathrm{t}})^2 \leq 16\varepsilon^2 \sup_{Q' \in \mathcal{B}_{2\varepsilon}(P)} \kappa_{Q'}.$$

Supremizing over source–target pairs $Q \in \mathcal{B}_\varepsilon(P)$ yields the claimed bound for $R_\varepsilon(\mathsf{DRO}_\varepsilon, P)$.

Finally, if $\varepsilon^2 \leq \lambda_{\min}(\mathbf{\Sigma}_X^P)/16$, then $2\varepsilon \leq \sqrt{\lambda_{\min}(\mathbf{\Sigma}_X^P)}/2$. By Lemmas 1.2 and F.2, $\kappa_{Q'} \leq 19\kappa_P$ for all $Q' \in \mathcal{B}_{2\varepsilon}(P)$. Substitution then gives $R_\varepsilon(\mathsf{DRO}_\varepsilon, P) \leq 304\kappa_P\varepsilon^2$. □

F.3 Rank-Adaptive DRO Bound (Proof of Lemma 3.3)

To start, we establish an additional robust regret bound for DRO by more directly exploiting its min–max problem formulation. Similar bounds exist in the literature, typically stated for more generic loss functions (Gao et al., 2024; Lee and Raginsky, 2018).

Lemma F.6 (DRO regularizer bound)**.** *For $P \in \mathcal{P}^*$, define*

$$\Gamma_P \coloneqq \sqrt{\operatorname{tr}\big((\mathbf{I}_m + \mathbf{A}_P\mathbf{A}_P^\top)\mathbf{\Sigma}_E^P\big)}.$$

Then, for every $\varepsilon \geq 0$,

$$R_\varepsilon(\mathsf{DRO}_\varepsilon, P) \leq 4\Gamma_P\varepsilon + 5\kappa_P\varepsilon^2. \tag{15}$$

Proof. Fix $P \in \mathcal{P}^*$, and abbreviate

$$\mathbf{A} \coloneqq \mathbf{A}_P, \qquad \mathbf{S} \coloneqq \mathbf{\Sigma}_P^{1/2}, \qquad \mathbf{M} \coloneqq (\mathbf{A} \quad -\mathbf{I}_m), \qquad \mathbf{E} \coloneqq \mathbf{S}_Y - \mathbf{A}\mathbf{S}_X.$$

By Lemma B.1, $\mathbf{E}\mathbf{E}^\top = \mathbf{\Sigma}_E^P$. Hence

$$\|\mathbf{M}^\top\mathbf{E}\|_{\mathrm{F}}^2 = \operatorname{tr}\big((\mathbf{I}_m + \mathbf{A}\mathbf{A}^\top)\mathbf{\Sigma}_E^P\big) = \Gamma_P^2.$$

Also, $\|\mathbf{M}\|_{\mathrm{op}}^2 = 1 + \|\mathbf{A}\|_{\mathrm{op}}^2 \leq \kappa_P$ by Theorem 2.2.

We first bound changes in the risk of the fixed predictor $\mathbf{A}$ in the neighborhood of P. For any $\delta \geq 0$ and $Q \in \mathcal{B}_\delta(P)$, Lemma 1.2 provides $\mathbf{\Delta} \in \mathbb{R}^{d\times d}$ such that

$$\mathbf{\Sigma}_Q = (\mathbf{S} + \mathbf{\Delta})(\mathbf{S} + \mathbf{\Delta})^\top, \qquad \|\mathbf{\Delta}\|_{\mathrm{F}} \leq \delta.$$

Since the residual factor at Q for $\mathbf{A}$ is $\mathbf{E} - \mathbf{M}\mathbf{\Delta}$, we have by Lemma B.1 that

$$\begin{aligned} L(\mathbf{A}, Q) - L(\mathbf{A}, P) &= \|\mathbf{E} - \mathbf{M}\mathbf{\Delta}\|_{\mathrm{F}}^2 - \|\mathbf{E}\|_{\mathrm{F}}^2 \\ &= -2\langle\mathbf{M}^\top\mathbf{E}, \mathbf{\Delta}\rangle_{\mathrm{F}} + \|\mathbf{M}\mathbf{\Delta}\|_{\mathrm{F}}^2. \end{aligned}$$

Cauchy–Schwarz and $\|\mathbf{M}\mathbf{\Delta}\|_{\mathrm{F}}^2 \leq \|\mathbf{M}\|_{\mathrm{op}}^2\|\mathbf{\Delta}\|_{\mathrm{F}}^2 \leq \kappa_P\delta^2$ then give the two-sided bound

$$-2\Gamma_P\delta \leq L(\mathbf{A}, Q) - L(\mathbf{A}, P) \leq 2\Gamma_P\delta + \kappa_P\delta^2. \tag{16}$$

For the lower bound, we have simply discarded the nonnegative quadratic term.

Target role. We now bound regret when P is the target law. Fix $Q \in \mathcal{B}_\varepsilon(P)$ and set $\hat{\mathbf{A}} \coloneqq \mathsf{DRO}_\varepsilon[Q]$. Since $P \in \mathcal{B}_\varepsilon(Q)$, DRO optimality and a comparison with $\mathbf{A}$ imply that

$$\begin{aligned} L(\hat{\mathbf{A}}, P) &\leq \sup_{Q' \in \mathcal{B}_\varepsilon(Q)} L(\hat{\mathbf{A}}, Q') \\ &\leq \sup_{Q' \in \mathcal{B}_\varepsilon(Q)} L(\mathbf{A}, Q') \\ &\leq \sup_{Q' \in \mathcal{B}_{2\varepsilon}(P)} L(\mathbf{A}, Q') \\ &\leq L(\mathbf{A}, P) + 4\Gamma_P\varepsilon + 4\kappa_P\varepsilon^2. \end{aligned}$$

The third inequality uses $\mathcal{B}_\varepsilon(Q) \subseteq \mathcal{B}_{2\varepsilon}(P)$, and the last applies (16) with $\delta = 2\varepsilon$. Subtracting $L(\mathbf{A}, P) = \inf_{\mathbf{B}\in\mathcal{A}} L(\mathbf{B}, P)$ gives

$$R(\mathsf{DRO}_\varepsilon[Q], P) \leq 4\Gamma_P\varepsilon + 4\kappa_P\varepsilon^2.$$

We note that, using the terminology of Gao et al. (2024), the RHS is an upper bound on the Wasserstein regularizer of the loss $(x, y) \mapsto \|\mathbf{A}_P x - y\|^2$ at P, evaluated at radius 2ε.

Source role. Now set $\hat{\mathbf{A}} \coloneqq \mathsf{DRO}_\varepsilon[P]$. Comparing with $\mathbf{A}$, we bound regret at $Q \in \mathcal{B}_\varepsilon(P)$ by

$$\begin{aligned} R(\hat{\mathbf{A}}, Q) &= L(\hat{\mathbf{A}}, Q) - L(\mathbf{A}, Q) + R(\mathbf{A}, Q) \\ &\leq \sup_{Q' \in \mathcal{B}_\varepsilon(P)} L(\mathbf{A}, Q') - L(\mathbf{A}, Q) + R(\mathbf{A}, Q) \\ &\leq 4\Gamma_P\varepsilon + \kappa_P\varepsilon^2 + R(\mathbf{A}, Q) \\ &\leq 4\Gamma_P\varepsilon + 5\kappa_P\varepsilon^2. \end{aligned}$$

The first inequality uses feasibility of Q and DRO optimality. The second applies the upper bound in (16) to Q' and its lower bound to Q, both with $\delta = \varepsilon$. The last uses $R(\mathbf{A}_P, Q) \le 4\kappa_P \mathsf{W}_2(P,Q)^2$ from Lemma 3.1.

Taking the maximum of these two role-specific bounds and then the supremum over $Q \in \mathcal{B}_\varepsilon(P)$ gives (15). □

Next, we combine Lemma F.6 with Lemma F.5 to obtain the claimed result.

Proof of Lemma 3.3. Fixing $P \in \mathcal{P}^*$, we abbreviate

$$\lambda \coloneqq \lambda_{\min}(\mathbf{\Sigma}_X^P), \qquad \mu \coloneqq \lambda_{\max}(\mathbf{\Sigma}_E^P), \qquad r \coloneqq r_{\text{eff}}(\mathbf{\Sigma}_E^P), \qquad s \coloneqq \kappa_P - 1.$$

The condition number formula (Theorem 2.2) gives

$$\|\mathbf{A}_P\|_{\text{op}}^2 \le s, \qquad \mu \le s\lambda.$$

Also $\operatorname{tr}(\mathbf{\Sigma}_E^P) = r\mu$, including when $\mathbf{\Sigma}_E^P = \mathbf{0}$ under the convention $r = 0$.

If $\varepsilon^2 \le \lambda/16$, the small-radius bound in Lemma F.5 gives

$$R_\varepsilon(\mathsf{DRO}_\varepsilon, P) \lesssim \kappa_P \varepsilon^2,$$

which suffices. We therefore assume $\varepsilon^2 > \lambda/16$ below.

To bound Γ_P, we estimate its coefficient-dependent trace term in two ways:

$$\begin{aligned}
\operatorname{tr}(\mathbf{A}_P \mathbf{A}_P^\top \mathbf{\Sigma}_E^P) &\le \|\mathbf{A}_P\|_{\text{op}}^2 \operatorname{tr}(\mathbf{\Sigma}_E^P) = r\mu \|\mathbf{A}_P\|_{\text{op}}^2, \\
\operatorname{tr}(\mathbf{A}_P \mathbf{A}_P^\top \mathbf{\Sigma}_E^P) &\le \mu \|\mathbf{A}_P\|_{\text{F}}^2 \le p\mu \|\mathbf{A}_P\|_{\text{op}}^2,
\end{aligned}$$

with the inequalities using Lemma A.1. Taking the smaller bound and using $\|\mathbf{A}_P\|_{\text{op}}^2 \le s$ yields

$$\begin{aligned}
\Gamma_P^2 &= \operatorname{tr}(\mathbf{\Sigma}_E^P) + \operatorname{tr}(\mathbf{A}_P \mathbf{A}_P^\top \mathbf{\Sigma}_E^P) \\
&\le \mu\big(r + s \min\{p, r\}\big) \\
&\le \lambda\big(rs + s^2 \min\{p, r\}\big).
\end{aligned}$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$, we further bound

$$\Gamma_P \le \sqrt{\lambda}\Big(\sqrt{rs} + s\sqrt{\min\{p, r\}}\Big).$$

Since $\sqrt{\lambda} < 4\varepsilon$, Lemma F.6 now gives

$$\begin{aligned}
R_\varepsilon(\mathsf{DRO}_\varepsilon, P) &\le 4\Gamma_P \varepsilon + 5\kappa_P \varepsilon^2 \\
&\le \Big(16s\sqrt{\min\{p, r\}} + 16\sqrt{rs} + 5\kappa_P\Big)\varepsilon^2 \\
&\lesssim \Big(\sqrt{\min\{p, r\}}\,(\kappa_P - 1) + \sqrt{r(\kappa_P - 1)} + \kappa_P\Big)\varepsilon^2,
\end{aligned}$$

as desired. □

F.4 Main Population-Limit Regret Bounds (Proof of Theorem 3.4)

We start with upper bounds, which follow from our existing results. Indeed, Lemma 3.2 implies that $R_\varepsilon(\mathsf{SNR}_\varepsilon, \mathcal{G}), R_\varepsilon(\mathsf{MRO}_\varepsilon, \mathcal{G}) \lesssim \kappa\varepsilon^2$. Moreover, Lemma 3.3 and the basic effective rank bound $r_{\text{eff}}(\mathbf{\Sigma}_E^P) \le m$ imply that

$$R_\varepsilon(\mathsf{DRO}_\varepsilon, \mathcal{G}) \lesssim \Big(\sqrt{\min\{p, m\}}\,(\kappa - 1) + \sqrt{m(\kappa - 1)} + \kappa\Big)\varepsilon^2,$$

as desired. Of course, no upper bound was claimed for ERM.

For the lower bounds, we scale up the local minimax result to obtain the minimax lower bound for $\mathcal{G}$. Then, we give an explicit construction establishing unbounded regret for ERM. Two further constructions, whose analysis is more subtle, yield the dimension-dependent terms in the DRO rate.

We start with a helper lemma that demonstrates a certain invariance to rescaling.

Lemma F.7 (Rescaling invariance)**.** *Let $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathbf{A} \in \mathcal{A}$. For $\lambda > 0$, let P_λ and Q_λ denote the pushforwards of P and Q, respectively, under $z \mapsto \lambda z$. We then have*

$$\begin{aligned} \mathsf{W}_2(P_\lambda, Q_\lambda) &= \lambda \mathsf{W}_2(P, Q), \\ L(\mathbf{A}, P_\lambda) &= \lambda^2 L(\mathbf{A}, P), \\ R(\mathbf{A}, P_\lambda) &= \lambda^2 R(\mathbf{A}, P), \\ \kappa_{P_\lambda} &= \kappa_P. \end{aligned}$$

Moreover, for $\varepsilon \geq 0$, $R^\star_{\lambda\varepsilon}(P_\lambda) = \lambda^2 R^\star_\varepsilon(P)$, and the optimization problems defining $\mathsf{DRO}_{\lambda\varepsilon}[P_\lambda]$ and $\mathsf{DRO}_\varepsilon[P]$ have the same minimizer set.

Proof. Scaling both coordinates of a transport coupling multiplies its expected squared cost by λ^2. Applying the same argument to the inverse scaling proves the Wasserstein identity. The risk and regret identities follow from $\mathbf{\Sigma}_{P_\lambda} = \lambda^2 \mathbf{\Sigma}_P$. The condition number formula is invariant under this scaling, as is membership in $\mathcal{P}^*$.

The Wasserstein identity implies that $\mathcal{B}_{\lambda\varepsilon}(P_\lambda) = \{Q_\lambda : Q \in \mathcal{B}_\varepsilon(P)\}$. Combining this with regret homogeneity and (1) proves the local minimax identity. Likewise,

$$\sup_{Q' \in \mathcal{B}_{\lambda\varepsilon}(P_\lambda)} L(\mathbf{A}, Q') = \lambda^2 \sup_{Q \in \mathcal{B}_\varepsilon(P)} L(\mathbf{A}, Q),$$

so the DRO minimizer sets coincide. □

We now turn to the minimax lower bound for $\mathcal{G}(\kappa)$.

Lemma F.8 (Minimax lower bound)**.** *For every $\varepsilon > 0$ and $\kappa \geq 1$, we have $R^\star_\varepsilon(\mathcal{G}(\kappa)) \geq \kappa\varepsilon^2$.*

Proof. Let $P \in \mathcal{P}_2(\mathbb{R}^d)$ be the law of (X, Y) where

$$X \sim \mathcal{N}(0, \mathbf{I}_p), \qquad Y = \sqrt{\kappa - 1}\, X_1 e_1, \qquad e_1 \in \mathbb{R}^m.$$

By construction, $\mathbf{A}_P$ has only one potentially nonzero entry, $\sqrt{\kappa - 1}$ in position $(1, 1)$, and the optimal residual is zero. Thus, Theorem 2.2 gives $\kappa_P = \kappa$.

For $t > 0$, let P_t be the pushforward of P under $z \mapsto tz$. By Lemma F.7, $P_t \in \mathcal{G}(\kappa)$ and

$$R^\star_\varepsilon(\mathcal{G}(\kappa)) \geq R^\star_\varepsilon(P_t) = t^2 R^\star_{\varepsilon/t}(P).$$

Further, the local expansion in Theorem 2.2 gives

$$\lim_{t \to \infty} t^2 R^\star_{\varepsilon/t}(P) = \kappa\varepsilon^2.$$

Since the preceding lower bound holds for every $t > 0$, the claim follows. □

Next, we prove the lower bound for ERM.

Lemma F.9 (ERM lower bound)**.** *For every $\varepsilon > 0$ and $\kappa \geq 1$, $R_\varepsilon(\mathsf{ERM}, \mathcal{G}(\kappa)) = \infty$, independently of the choice of ERM minimizer.*

Proof. Consider target $P \in \mathcal{P}_2(\mathbb{R}^d)$ defined as the law of $(\varepsilon W/2, 0_m)$ where $W \sim \mathcal{N}(0, \mathbf{I}_p)$. Then $\mathbf{\Sigma}^P_X = (\varepsilon^2/4)\mathbf{I}_p$, $\mathbf{A}_P = \mathbf{0}$, and $\mathbf{\Sigma}^P_E = \mathbf{0}$. Hence $\kappa_P = 1$ and $P \in \mathcal{G}(\kappa)$.

For $\delta \in (0, 1)$, consider source $Q_\delta \in \mathcal{P}_2(\mathbb{R}^d)$ taken as the law of

$$\tilde{X} = \frac{\varepsilon}{2}(\delta W_1, W_2, \ldots, W_p), \qquad \tilde{Y} = \frac{\varepsilon}{2} W_1 e_1, \qquad e_1 \in \mathbb{R}^m.$$

The coupling using the same W gives

$$\mathsf{W}_2(P, Q_\delta)^2 \leq \frac{\varepsilon^2}{4}\big((1 - \delta)^2 + 1\big) \leq \varepsilon^2.$$

The source feature second-moment matrix is positive definite, and its unique ERM solution has entry $1/\delta$ in position $(1, 1)$ and zeros elsewhere. Therefore,

$$R_\varepsilon(\mathsf{ERM}, \mathcal{G}(\kappa)) \geq R(\mathsf{ERM}[Q_\delta], P) = \frac{\varepsilon^2}{4\delta^2}.$$

Letting $\delta \downarrow 0$ proves the claim. □

For the DRO result, we combine two lower bounds in addition to the minimax lower bound. Although DRO solutions are generally difficult to characterize, we are able to streamline analysis in both cases by exploiting rotational symmetry. We fix $\varepsilon = 1$ for now to simplify presentation.

Lemma F.10 (DRO lower bound, A)**.** *For every $\kappa \geq 1$,*

$$R_1(\mathsf{DRO}_1, \mathcal{G}(\kappa)) \gtrsim \sqrt{\min\{p, m\}}\,(\kappa - 1),$$

independently of the choice of DRO minimizer.

Proof. The claim is immediate for $\kappa = 1$. Write $q \coloneqq \min\{p, m\}$. If $q < 65$, then Lemma F.8 gives

$$R_1(\mathsf{DRO}_1, \mathcal{G}(\kappa)) \geq R_1^\star(\mathcal{G}(\kappa)) \gtrsim \kappa \gtrsim \sqrt{q}\,(\kappa - 1).$$

We may therefore assume $\kappa > 1$ and $q \geq 65$.

Construction. Set $b \coloneqq \sqrt{(\kappa - 1)/2} > 0$. Let $X \sim \mathcal{N}(0, \mathbf{I}_p)$ and

$$E \sim \mathcal{N}\big(0, \operatorname{diag}(0, b^2\mathbf{I}_{q-1}, \mathbf{0}_{m-q})\big)$$

be independent. Take target P as the law of (X, Y) where

$$Y = \begin{pmatrix} bX_{1:q} \\ 0_{m-q} \end{pmatrix} + E$$

Here and below, empty blocks of size zero should be omitted. By this construction,

$$\mathbf{A}_P = \begin{pmatrix} b\mathbf{I}_q & \mathbf{0}_{q\times(p-q)} \\ \mathbf{0}_{(m-q)\times q} & \mathbf{0}_{(m-q)\times(p-q)} \end{pmatrix}, \qquad \mathbf{\Sigma}_X^P = \mathbf{I}_p,$$

and $\mathbf{\Sigma}_E^P = \operatorname{diag}(0, b^2\mathbf{I}_{q-1}, \mathbf{0}_{m-q})$. Thus,

$$\mathbf{A}_P\mathbf{A}_P^\top + \lambda_{\min}(\mathbf{\Sigma}_X^P)^{-1}\mathbf{\Sigma}_E^P = \operatorname{diag}(b^2, 2b^2\mathbf{I}_{q-1}, \mathbf{0}_{m-q}),$$

and Theorem 2.2 gives $\kappa_P = 1 + 2b^2 = \kappa$.

For the perturbed source, take $\tilde{P}$ as the law of $(\tilde{X}, Y)$, where $\tilde{X}_1 = cX_1$ and $\tilde{X}_{2:p} = X_{2:p}$, for $c \in (0, 1]$ to be tuned. This construction ensures $\mathsf{W}_2(\tilde{P}, P)^2 \leq \mathbb{E}[\|X - \tilde{X}\|^2] \leq 1$.

Average over symmetries. Consider the compact subgroup

$$U \coloneqq \big\{\operatorname{diag}(1, \mathbf{H}, s\mathbf{I}_{p-q}, 1, \mathbf{H}, t\mathbf{I}_{m-q}) : \mathbf{H} \in \mathbb{O}^{q-1},\ s, t \in \{-1, 1\}\big\} \subseteq \mathbb{O}^d.$$

Both P and $\tilde{P}$ are invariant under pushforwards by each $\mathbf{U} \in U$. Moreover, the invariant coefficient matrices (in the sense of Lemma C.4) are exactly

$$\mathbf{A}(\alpha, \beta) \coloneqq \begin{pmatrix} \operatorname{diag}(\alpha, \beta\mathbf{I}_{q-1}) & \mathbf{0}_{q\times(p-q)} \\ \mathbf{0}_{(m-q)\times q} & \mathbf{0}_{(m-q)\times(p-q)} \end{pmatrix}, \qquad \alpha, \beta \in \mathbb{R}.$$

Indeed, varying s and t eliminates the extra feature columns and response rows. Within the leading $q \times q$ block, taking $\mathbf{H} = -\mathbf{I}_{q-1}$ eliminates the off-diagonal blocks, and commuting with every $\mathbf{H} \in \mathbb{O}^{q-1}$ forces the remaining $(q-1) \times (q-1)$ block to be a scalar multiple of the identity. By Lemma C.4, it therefore suffices to lower bound the target regret of invariant DRO optimizers.

Construct candidate DRO optimizer. Define the invariant objective by

$$J(\alpha, \beta) \coloneqq \sup_{Q \in \mathcal{B}_1(\tilde{P})} L(\mathbf{A}(\alpha, \beta), Q).$$

We have

$$\mathbf{A}(\alpha, \beta)\mathbf{A}(\alpha, \beta)^\top = \operatorname{diag}(\alpha^2, \beta^2\mathbf{I}_{q-1}, \mathbf{0}_{m-q}),$$

and the coordinate residual second moments at the source are

$$\mathbb{E}\big[(\mathbf{A}(\alpha, \beta)\tilde{X} - Y)_i^2\big] = \begin{cases} (\alpha c - b)^2, & i = 1, \\ (\beta - b)^2 + b^2, & 2 \leq i \leq q, \\ 0, & q < i \leq m. \end{cases}$$

Thus Corollary C.3 with $\varepsilon = 1$ gives

$$J(\alpha, \beta) = \inf_{\eta > 1 + \max\{\alpha^2, \beta^2\}} \eta \left(1 + \frac{(\alpha c - b)^2}{\eta - 1 - \alpha^2} + \frac{(q-1)[(\beta - b)^2 + b^2]}{\eta - 1 - \beta^2} \right). \tag{17}$$

Now consider the related problem associated with coordinates $2, \ldots, q$:

$$V(\beta) \coloneqq \inf_{\eta > 1 + \beta^2} \eta \left(1 + \frac{(q-1)[(\beta - b)^2 + b^2]}{\eta - 1 - \beta^2} \right).$$

By Lemma A.7, with $\delta = \eta - (1 + \beta^2)$, this infimum is attained at

$$\eta(\beta) \coloneqq 1 + \beta^2 + \sqrt{(1 + \beta^2)(q-1)[(\beta - b)^2 + b^2]},$$

with value

$$V(\beta) = \left(\sqrt{1 + \beta^2} + \sqrt{(q-1)[(\beta - b)^2 + b^2]} \right)^2.$$

This function is strictly convex and coercive. Let $\beta_\star$ denote its unique minimizer, set $\eta_\star \coloneqq \eta(\beta_\star)$, and let $\alpha_\star \coloneqq \sqrt{\eta_\star - 1} > 0$. We set the free scaling parameter c to $b/\alpha_\star > 0$. Below we shall verify that $c \leq 1$, ensuring that $\tilde{P} \in \mathcal{B}_1(P)$.

Certify optimality. We next show that J is uniquely minimized at $(\alpha_\star, \beta_\star)$. For every feasible η for (17), $\eta/(\eta - 1 - \alpha^2) \geq 1$. The remaining terms are bounded from below by $V(\beta)$, so

$$J(\alpha, \beta) \geq V(\beta) + (\alpha c - b)^2 \geq V(\beta_\star).$$

At $(\alpha_\star, \beta_\star)$, the first-coordinate residual vanishes because $\alpha_\star c = b$. Also $\alpha_\star^2 > \beta_\star^2$, so

$$J(\alpha_\star, \beta_\star) = \inf_{\eta > \eta_\star} \eta \left(1 + \frac{(q-1)[(\beta_\star - b)^2 + b^2]}{\eta - 1 - \beta_\star^2} \right).$$

When minimizing instead over the larger range $\eta > 1 + \beta_\star^2$, our construction ensures that the minimum is attained at $\eta = \eta_\star$. Although $\eta_\star$ is excluded from the current feasible range, the objective is continuous there, so the infimum is approached as $\eta \downarrow \eta_\star$ and equals $V(\beta_\star)$. The preceding lower bound then forces every minimizing pair (α, β) to satisfy $\beta = \beta_\star$ and $\alpha c = b$. Since $c = b/\alpha_\star > 0$, this implies that $(\alpha_\star, \beta_\star)$ is the unique minimizing pair.

Bound $\alpha_\star$ and $\beta_\star$ from below, verify c is admissible. Define

$$h(\beta) \coloneqq \sqrt{V(\beta)} = \sqrt{1 + \beta^2} + \sqrt{q-1}\sqrt{(\beta - b)^2 + b^2},$$

so that $\beta_\star \in \arg\min_\beta h(\beta)$. This function is strictly convex and coercive. For $\beta \leq b/2$,

$$h'(\beta) = \frac{\beta}{\sqrt{1 + \beta^2}} + \sqrt{q-1}\,\frac{\beta - b}{\sqrt{(\beta - b)^2 + b^2}} \leq 1 - \sqrt{\frac{q-1}{5}} < 0.$$

The inequality uses $\beta - b \leq -b/2$. Hence $\beta_\star > b/2$. It follows that

$$\begin{aligned} \alpha_\star^2 &= \eta_\star - 1 \\ &\geq \sqrt{(1 + \beta_\star^2)(q-1)[(\beta_\star - b)^2 + b^2]} \\ &\geq \frac{b^2}{2}\sqrt{q-1}. \end{aligned} \tag{18}$$

Since $q \geq 65$, this implies $\alpha_\star \geq 2b$ and hence $c = b/\alpha_\star \leq 1/2$.

Lower bound regret. Since $c \leq 1$, $\tilde{P} \in \mathcal{B}_1(P)$, and its unique invariant DRO optimizer satisfies

$$\begin{aligned} R(\mathbf{A}(\alpha_\star, \beta_\star), P) &= (\alpha_\star - b)^2 + (q-1)(\beta_\star - b)^2 \\ &\geq (\alpha_\star - b)^2 \\ &\geq \frac{1}{4}\alpha_\star^2 \\ &\geq \frac{b^2}{8}\sqrt{q-1} \gtrsim \sqrt{q}\,(\kappa - 1). \end{aligned}$$

The first equality uses $\mathbf{\Sigma}_X^P = \mathbf{I}_p$ and the regret formula, the second inequality uses that $\alpha_\star \geq 2b$, the third inequality uses (18), and the last inequality uses $b^2 = (\kappa - 1)/2$.

By Lemma C.4, every DRO optimizer at $\tilde{P}$ has target regret at least that of $\mathbf{A}(\alpha_\star, \beta_\star)$. Since $P \in \mathcal{G}(\kappa)$ and $\tilde{P} \in \mathcal{B}_1(P)$,

$$R_1(\mathsf{DRO}_1, \mathcal{G}(\kappa)) \geq R(\mathsf{DRO}_1[\tilde{P}], P) \geq R(\mathbf{A}(\alpha_\star, \beta_\star), P) \gtrsim \sqrt{q}\,(\kappa - 1).$$

This proves the claim for every choice of DRO minimizer. □

Lemma F.11 (DRO lower bound, B)**.** *For every $\kappa \geq 1$,*

$$R_1(\mathsf{DRO}_1, \mathcal{G}(\kappa)) \gtrsim \sqrt{m(\kappa - 1)},$$

independently of the choice of DRO minimizer.

Proof. The claim is immediate for $\kappa = 1$. If $m = 1$, then Lemma F.8 gives

$$R_1(\mathsf{DRO}_1, \mathcal{G}(\kappa)) \geq R_1^\star(\mathcal{G}(\kappa)) \gtrsim \kappa \geq \sqrt{\kappa - 1}.$$

We may therefore assume $\kappa > 1$ and $m \geq 2$.

Construction. Set

$$s \coloneqq \kappa - 1 > 0, \qquad N \coloneqq (m-1)s > 0, \qquad a \coloneqq \frac{1}{2}\min\{\sqrt{s}, N^{1/4}\} > 0.$$

Let $X \sim \mathcal{N}(0, \mathbf{I}_p)$ and $E \sim \mathcal{N}(0, \mathrm{diag}(0, s\mathbf{I}_{m-1}))$ be independent. Fix target $P \in \mathcal{P}_2(\mathbb{R}^d)$ as the law of (X, Y), where

$$Y = aX_1 e_1 + E, \qquad e_1 \in \mathbb{R}^m.$$

By this construction,

$$\mathbf{A}_P = \begin{pmatrix} ae_1 & \mathbf{0}_{m\times(p-1)} \end{pmatrix}, \qquad \mathbf{\Sigma}_X^P = \mathbf{I}_p, \qquad \mathbf{\Sigma}_E^P = \mathrm{diag}(0, s\mathbf{I}_{m-1}).$$

Thus

$$\mathbf{A}_P\mathbf{A}_P^\top + \lambda_{\min}(\mathbf{\Sigma}_X^P)^{-1}\mathbf{\Sigma}_E^P = \mathrm{diag}(a^2, s\mathbf{I}_{m-1}).$$

Since $a^2 \leq s$ and $m \geq 2$, Theorem 2.2 gives $\kappa_P = 1 + s = \kappa$.

For the perturbed source, take $\tilde{P}$ as the law of $(\tilde{X}, Y)$, where $\tilde{X}_1 = cX_1$ and $\tilde{X}_{2:p} = X_{2:p}$, for $c \in (0, 1]$ to be specified. Coupling the same X, E gives

$$\mathsf{W}_2(\tilde{P}, P)^2 \leq \mathbb{E}\,\|X - \tilde{X}\|^2 = (1 - c)^2 \leq 1.$$

Average over symmetries. Consider the compact subgroup

$$U \coloneqq \left\{\mathrm{diag}(1, t\mathbf{I}_{p-1}, 1, \mathbf{H}) : t \in \{-1, 1\},\ \mathbf{H} \in \mathbb{O}^{m-1}\right\} \subseteq \mathbb{O}^d.$$

Both P and $\tilde{P}$ are invariant under pushforwards by each $\mathbf{U} \in U$. The invariant coefficient matrices (in the sense of Lemma C.4) are exactly

$$\mathbf{A}(\alpha) \coloneqq \begin{pmatrix} \alpha e_1 & \mathbf{0}_{m\times(p-1)} \end{pmatrix}, \qquad \alpha \in \mathbb{R}.$$

Indeed, varying t eliminates columns $2, \ldots, p$, and taking $\mathbf{H} = -\mathbf{I}_{m-1}$ eliminates rows $2, \ldots, m$. By Lemma C.4, it therefore suffices to lower bound the target regret of invariant DRO optimizers.

Construct candidate DRO optimizer. Define the invariant objective by

$$J(\alpha) \coloneqq \sup_{Q \in \mathcal{B}_1(\tilde{P})} L(\mathbf{A}(\alpha), Q).$$

We have

$$\mathbf{A}(\alpha)\mathbf{A}(\alpha)^\top = \mathrm{diag}(\alpha^2, \mathbf{0}_{m-1}),$$

and the coordinate residual second moments at the source are

$$\mathbb{E}\big[(\mathbf{A}(\alpha)\tilde{X} - Y)_i^2\big] = \begin{cases} (\alpha c - a)^2, & i = 1, \\ s, & 2 \leq i \leq m. \end{cases}$$

Thus Corollary C.3 with $\varepsilon = 1$ gives

$$J(\alpha) = \inf_{\eta > 1+\alpha^2} \eta\left(1 + \frac{(\alpha c - a)^2}{\eta - 1 - \alpha^2} + \frac{N}{\eta - 1}\right). \tag{19}$$

Now consider the related problem associated with coordinates $2, \dots, m$:

$$V \coloneqq \inf_{\eta > 1} \eta\left(1 + \frac{N}{\eta - 1}\right).$$

By Lemma A.7, with $\delta = \eta - 1$, this infimum is attained at $\eta_\star \coloneqq 1 + \sqrt{N}$ and $V = (1 + \sqrt{N})^2$. We now set $\alpha_\star \coloneqq \sqrt{\eta_\star - 1} = N^{1/4}$ and $c \coloneqq a/\alpha_\star$. Our choice of a gives $0 < a \le \alpha_\star/2$ and hence $0 < c \le 1/2$. Thus the source construction is admissible.

Certify optimality. We next show that J is uniquely minimized at $\alpha_\star$. For every feasible η in (19), $\eta/(\eta - 1 - \alpha^2) \ge 1$. The remaining terms are bounded from below by V, so

$$J(\alpha) \ge V + (\alpha c - a)^2 \ge V.$$

At $\alpha = \alpha_\star$, the first-coordinate residual vanishes because $\alpha_\star c = a$. Since $1 + \alpha_\star^2 = \eta_\star$, we obtain

$$J(\alpha_\star) = \inf_{\eta > \eta_\star} \eta\left(1 + \frac{N}{\eta - 1}\right).$$

When minimizing instead over the larger range $\eta > 1$, the bulk minimum is attained at $\eta = \eta_\star$. Although $\eta_\star$ is excluded from the restricted range, the objective is continuous there, so the infimum is approached as $\eta \downarrow \eta_\star$ and equals V. Thus $J(\alpha_\star) = V$.

If α is any minimizer of J, the preceding lower bound gives $0 = J(\alpha) - V \ge (\alpha c - a)^2$. Hence, $\alpha c = a$ and, since $c > 0$, $\alpha = a/c = \alpha_\star$. Thus, $\mathbf{A}(\alpha_\star)$ is the unique invariant DRO optimizer.

Lower bound regret. Since $\mathbf{\Sigma}_X^P = \mathbf{I}_p$, the regret formula and our choice of $c = a/\alpha_\star$ give

$$\begin{aligned} R(\mathbf{A}(\alpha_\star), P) &= \|\mathbf{A}(\alpha_\star) - \mathbf{A}_P\|_{\mathrm{F}}^2 \\ &= (\alpha_\star - a)^2 \\ &= \alpha_\star^2 (1 - c)^2 \\ &\ge \frac{1}{4}\alpha_\star^2 = \frac{1}{4}\sqrt{(m-1)(\kappa - 1)} \gtrsim \sqrt{m(\kappa - 1)}. \end{aligned}$$

The inequality uses $c \le 1/2$, and the final comparison uses $m - 1 \ge m/2$ for $m \ge 2$.

By Lemma C.4, every DRO optimizer at $\tilde{P}$ has target regret at least that of $\mathbf{A}(\alpha_\star)$. Since $P \in \mathcal{G}(\kappa)$ and $\tilde{P} \in \mathcal{B}_1(P)$,

$$R_1(\mathsf{DRO}_1, \mathcal{G}(\kappa)) \ge R(\mathsf{DRO}_1[\tilde{P}], P) \ge R(\mathbf{A}(\alpha_\star), P) \gtrsim \sqrt{m(\kappa - 1)}.$$

This proves the claim for every choice of DRO minimizer. □

Combining the results above, we obtain the following unified lower bound.

Lemma F.12 (DRO lower bound)**.** *For all $\varepsilon > 0$ and $\kappa \ge 1$, we have*

$$R_\varepsilon(\mathsf{DRO}_\varepsilon, \mathcal{G}(\kappa)) \gtrsim \big(\sqrt{\min\{p, m\}}(\kappa - 1) + \sqrt{m(\kappa - 1)} + \kappa\big)\varepsilon^2,$$

no matter how ties are broken to select DRO solutions.

Proof. Scale the hard source–target pairs from Lemmas F.10 and F.11 by ε. By Lemma F.7, their condition numbers are unchanged, their Wasserstein distances are multiplied by ε, and the DRO minimizer sets are unchanged when the supplied radius changes from 1 to ε. Their target regrets are multiplied by ε^2. Combining these bounds with Lemma F.8 and using $\max\{a, b, c\} \ge (a + b + c)/3$ gives the claim. □

G Remaining Proofs for Section 4

G.1 Convex Reformulation for MRO (Proof of Theorem 4.1)

Throughout, we use the notation from Theorem 4.1 and the block decompositions from Section B.1. The observed matrix $\tilde{\mathbf{\Sigma}}$, its square-root factor $\tilde{\mathbf{S}}$, and the radius $\varepsilon > 0$ are fixed, and their dependence is suppressed. We use $\sup \varnothing = -\infty$. Recalling $\tilde{\mathbf{E}}_{\mathbf{A}} = \tilde{\mathbf{S}}_Y - \mathbf{A}\tilde{\mathbf{S}}_X$, our goal is to prove

$$F(\mathbf{A}) = \inf_{\eta > b_{\mathbf{A}}} \eta\Big(\varepsilon^2 - \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2 + \mathrm{tr}_p(\mathbf{H}_{\mathbf{A},\eta})\Big) \qquad \forall\, \mathbf{A} \in \mathcal{A},$$

where $b_{\mathbf{A}} \coloneqq 1 + \|\mathbf{A}\|_{\mathrm{op}}^2$ and, for $\eta > b_{\mathbf{A}}$,

$$\mathbf{H}_{\mathbf{A},\eta} \coloneqq \tilde{\mathbf{S}}_X^\top \tilde{\mathbf{S}}_X + \tilde{\mathbf{E}}_{\mathbf{A}}^\top \big((\eta - 1)\mathbf{I}_m - \mathbf{A}\mathbf{A}^\top\big)^{-1} \tilde{\mathbf{E}}_{\mathbf{A}}.$$

We must also show that the displayed objective, namely $D(\mathbf{A}, \eta)$, is jointly convex in $(\mathbf{A}, \eta)$ on its convex domain $\eta > b_{\mathbf{A}}$. Infimizing over $\mathbf{A}$ then gives the claimed convex reformulation of MRO.

By Lemmas 1.2 and B.1, the worst-case regret in the ε-neighborhood of $\tilde{\mathbf{\Sigma}}$ is

$$F(\mathbf{A}) = \sup_{\mathbf{\Sigma} \in \mathcal{B}_\varepsilon(\tilde{\mathbf{\Sigma}})} R(\mathbf{A}, \mathbf{\Sigma}) = \sup_{\substack{\mathbf{S} \in \mathbb{R}^{d \times d} \\ \|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}} \le \varepsilon}} \|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}_{\mathbf{S}_X}\|_{\mathrm{F}}^2,$$

where $\mathbf{\Pi}_{\mathbf{S}_X} \in \mathbb{R}^{d \times d}$ denotes the orthogonal projection onto the row space of $\mathbf{S}_X \in \mathbb{R}^{p \times d}$. This projection has rank at most p and depends on $\mathbf{S}_X$. We will relax this dependence by introducing an independent projection from the family

$$\Pi_p \coloneqq \{\mathbf{\Pi} \in \mathbb{S}_+^d : \mathbf{\Pi}^2 = \mathbf{\Pi},\ \mathrm{tr}(\mathbf{\Pi}) = p\}, \tag{20}$$

while retaining the constraint $\mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi}$. This means that the row space of $\mathbf{S}_X$ is contained in $\mathrm{range}(\mathbf{\Pi})$, a condition satisfied by $\mathbf{\Pi}_{\mathbf{S}_X}$ by definition.

For $\mathbf{A} \in \mathcal{A}$, a squared perturbation budget $t > 0$, and $\mathbf{\Pi} \in \Pi_p$, define

$$f_{\mathbf{\Pi}}(\mathbf{A}, t) \coloneqq \sup_{\substack{\mathbf{S} \in \mathbb{R}^{d \times d} \\ \mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi} \\ \|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}}^2 \le t}} \|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}\|_{\mathrm{F}}^2, \tag{21}$$

$$V_{\mathbf{A}}(t) \coloneqq \sup_{\mathbf{\Pi} \in \Pi_p} f_{\mathbf{\Pi}}(\mathbf{A}, t). \tag{22}$$

Every $\mathbf{\Pi}_{\mathbf{S}_X}$ is dominated in Loewner order by some $\mathbf{\Pi} \in \Pi_p$. For such an extension, $\mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi}$, and the projected squared residual can only increase. Consequently,

$$F(\mathbf{A}) \le V_{\mathbf{A}}(\varepsilon^2).$$

The remainder of the proof proceeds in three steps.

1. **Fixed-projection duality.** Consider the Lagrangian reformulation

$$f_{\mathbf{\Pi}}(\mathbf{A}, t) = \sup_{\substack{\mathbf{S} \in \mathbb{R}^{d \times d} \\ \mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi}}} \inf_{\eta \ge 0} \Big\{ \|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}\|_{\mathrm{F}}^2 + \eta\big(t - \|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}}^2\big) \Big\}. \tag{23}$$

   We first prove that the supremum and infimum can be exchanged. Specifically, for $\eta \ge 0$, define the fixed-projection Lagrangian envelope

$$\Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) \coloneqq \sup_{\substack{\mathbf{S} \in \mathbb{R}^{d \times d} \\ \mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi}}} \Big\{ \|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}\|_{\mathrm{F}}^2 + \eta\big(t - \|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}}^2\big) \Big\}. \tag{24}$$

   We show that

$$f_{\mathbf{\Pi}}(\mathbf{A}, t) = \inf_{\eta > b_{\mathbf{A}}} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta)$$

   via a change of coordinates and Lemma C.1. The same lemma also gives the closed form

$$\Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) = \eta\Big(t - \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2 + \mathrm{tr}(\mathbf{H}_{\mathbf{A},\eta}\mathbf{\Pi})\Big), \qquad \eta > b_{\mathbf{A}}. \tag{25}$$

2. **Projection–multiplier minimax exchange.** Applying the non-convex minimax theorem Ricceri (2019, Theorem 1.1), we obtain

$$\begin{aligned} V_{\mathbf{A}}(t) &= \sup_{\mathbf{\Pi}\in\Pi_p} \inf_{\eta>b_{\mathbf{A}}} \Lambda_{\mathbf{\Pi}}(\mathbf{A},t,\eta) \\ &= \inf_{\eta>b_{\mathbf{A}}} \sup_{\mathbf{\Pi}\in\Pi_p} \Lambda_{\mathbf{\Pi}}(\mathbf{A},t,\eta) \\ &= \inf_{\eta>b_{\mathbf{A}}} \eta\Big(t - \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2 + \mathrm{tr}_p(\mathbf{H}_{\mathbf{A},\eta})\Big). \end{aligned}$$

The last equality uses the Ky Fan variational formula from Lemma A.9. At $t = \varepsilon^2$, this matches our claimed expression for $F(\mathbf{A})$. It remains to show that $F(\mathbf{A}) \geq V_{\mathbf{A}}(\varepsilon^2)$.

3. **Exactness and joint convexity.** For fixed $\mathbf{A}$, the final expression above reveals that $V_{\mathbf{A}}$ is concave in the squared budget t, and its definition implies that it is finite for every $t > 0$. Hence, $V_{\mathbf{A}}$ is continuous on $(0,\infty)$. For $0 < s < \varepsilon^2$, fix $\mathbf{\Pi} \in \Pi_p$ and any $\mathbf{S}$ feasible for the supremum in (21) defining $f_{\mathbf{\Pi}}(\mathbf{A}, s)$. Using the spare perturbation budget, we approximate $\mathbf{S}$ by factors within distance ε of $\tilde{\mathbf{S}}$ whose feature row-space projection is exactly $\mathbf{\Pi}$. Their actual regrets approach $\|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}\|_{\mathrm{F}}^2$. Supremizing over $\mathbf{S}$ and then $\mathbf{\Pi}$ gives

$$V_{\mathbf{A}}(s) \leq F(\mathbf{A}) \leq V_{\mathbf{A}}(\varepsilon^2).$$

Letting $s \uparrow \varepsilon^2$ and using continuity proves $F(\mathbf{A}) = V_{\mathbf{A}}(\varepsilon^2)$. Joint convexity follows by expressing D as the supremum over $\mathbf{\Pi} \in \Pi_p$ of the envelopes $\Lambda_{\mathbf{\Pi}}(\mathbf{A}, \varepsilon^2, \eta)$, each jointly convex in $(\mathbf{A}, \eta)$.

We now establish these three steps in order.

Lemma G.1 (Fixed-projection duality)**.** *For every $\mathbf{A} \in \mathcal{A}$, $t > 0$, and $\mathbf{\Pi} \in \Pi_p$, the Lagrangian envelope $\Lambda_{\mathbf{\Pi}}$ satisfies* (25) *whenever $\eta > b_{\mathbf{A}}$. Moreover,*

$$f_{\mathbf{\Pi}}(\mathbf{A}, t) = \inf_{\eta\geq 0} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) = \inf_{\eta > b_{\mathbf{A}}} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta). \tag{26}$$

Proof. Fix $\mathbf{A} \in \mathcal{A}$, $t > 0$, and $\mathbf{\Pi} \in \Pi_p$, and write $\mathbf{M} \coloneqq \begin{pmatrix} \mathbf{A} & -\mathbf{I}_m \end{pmatrix}$ so that $\|\mathbf{M}\|_{\mathrm{op}}^2 = b_{\mathbf{A}}$.

We first reparameterize the supremum in (21) defining $f_{\mathbf{\Pi}}$ using reduced coordinates which eliminate the row-space constraint. We then apply Lemma C.1 to establish strong duality and derive (25).

Step 1: Reduced coordinates. For any $\mathbf{S} \in \mathbb{R}^{d\times d}$ satisfying $\mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi}$, feasible for the supremum in (21), consider substituting $\mathbf{S}_Y$ with $\mathbf{S}_Y\mathbf{\Pi} + \tilde{\mathbf{S}}_Y(\mathbf{I}_d - \mathbf{\Pi})$. This preserves the objective and cannot increase $\|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}}^2$, so it preserves feasibility. Thus, we may freely restrict the feasible set to those $\mathbf{S}$ which further satisfy $\mathbf{S}_Y(\mathbf{I}_d - \mathbf{\Pi}) = \tilde{\mathbf{S}}_Y(\mathbf{I}_d - \mathbf{\Pi})$. Now fix any $\mathbf{U} \in \mathbb{R}^{d\times p}$ whose columns form an orthonormal basis of $\mathrm{range}(\mathbf{\Pi})$, so that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$ and $\mathbf{U}\mathbf{U}^\top = \mathbf{\Pi}$. Define the reduced coordinates of feasible $\mathbf{S}$ by

$$\mathbf{W}_X \coloneqq \mathbf{S}_X\mathbf{U} \in \mathbb{R}^{p\times p}, \qquad \mathbf{W}_Y \coloneqq \mathbf{S}_Y\mathbf{U} \in \mathbb{R}^{m\times p}, \qquad \mathbf{W} \coloneqq \begin{pmatrix} \mathbf{W}_X \\ \mathbf{W}_Y \end{pmatrix} = \mathbf{S}\mathbf{U} \in \mathbb{R}^{d\times p}.$$

On the restricted feasible set, the correspondence between $\mathbf{S}$ and $\mathbf{W}$ is bijective, since $\mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi} = \mathbf{W}_X\mathbf{U}^\top$, $\mathbf{S}_Y\mathbf{\Pi} = \mathbf{W}_Y\mathbf{U}^\top$, and $\mathbf{S}_Y(\mathbf{I}_d - \mathbf{\Pi})$ is fixed. Further define

$$\tilde{\mathbf{W}}_X \coloneqq \tilde{\mathbf{S}}_X\mathbf{U}, \qquad \tilde{\mathbf{W}}_Y \coloneqq \tilde{\mathbf{S}}_Y\mathbf{U}, \qquad \tilde{\mathbf{W}} \coloneqq \begin{pmatrix} \tilde{\mathbf{W}}_X \\ \tilde{\mathbf{W}}_Y \end{pmatrix} = \tilde{\mathbf{S}}\mathbf{U}.$$

We then have

$$\begin{aligned} \|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}}^2 &= \|(\mathbf{S}_X - \tilde{\mathbf{S}}_X)\mathbf{\Pi}\|_{\mathrm{F}}^2 + \|(\mathbf{S}_X - \tilde{\mathbf{S}}_X)(\mathbf{I}_d - \mathbf{\Pi})\|_{\mathrm{F}}^2 + \|(\mathbf{S}_Y - \tilde{\mathbf{S}}_Y)\mathbf{\Pi}\|_{\mathrm{F}}^2 \\ &= \|(\mathbf{W}_X - \tilde{\mathbf{W}}_X)\mathbf{U}^\top\|_{\mathrm{F}}^2 + \|\tilde{\mathbf{S}}_X(\mathbf{I}_d - \mathbf{\Pi})\|_{\mathrm{F}}^2 + \|(\mathbf{W}_Y - \tilde{\mathbf{W}}_Y)\mathbf{U}^\top\|_{\mathrm{F}}^2 \\ &= \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2 + \underbrace{\|\tilde{\mathbf{S}}_X(\mathbf{I}_d - \mathbf{\Pi})\|_{\mathrm{F}}^2}_{\coloneqq c_{\mathbf{\Pi}}}, \end{aligned}$$

using $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_p$. Returning to (21), this yields

$$\begin{aligned} f_{\mathbf{\Pi}}(\mathbf{A}, t) &= \sup_{\substack{\mathbf{S} \in \mathbb{R}^{d \times d} \\ \mathbf{S}_X = \mathbf{S}_X \mathbf{\Pi} \\ \mathbf{S}_Y(\mathbf{I}_d - \mathbf{\Pi}) = \tilde{\mathbf{S}}_Y(\mathbf{I}_d - \mathbf{\Pi}) \\ \|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}}^2 \le t}} \|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}\|_{\mathrm{F}}^2 \\ &= \sup_{\substack{\mathbf{W} \in \mathbb{R}^{d \times p} \\ \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2 \le t - c_{\mathbf{\Pi}}}} \|(\mathbf{W}_Y - \mathbf{A}\mathbf{W}_X)\mathbf{U}^\top\|_{\mathrm{F}}^2 \\ &= \sup_{\substack{\mathbf{W} \in \mathbb{R}^{d \times p} \\ \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2 \le t - c_{\mathbf{\Pi}}}} \|\mathbf{W}_Y - \mathbf{A}\mathbf{W}_X\|_{\mathrm{F}}^2. \end{aligned} \tag{27}$$

Likewise, for all $\eta \ge 0$, an analogous argument gives

$$\Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) = \sup_{\mathbf{W} \in \mathbb{R}^{d \times p}} \Big\{ \|\mathbf{M}\mathbf{W}\|_{\mathrm{F}}^2 + \eta\big(t - c_{\mathbf{\Pi}} - \|\mathbf{W} - \tilde{\mathbf{W}}\|_{\mathrm{F}}^2\big) \Big\}. \tag{28}$$

Step 2: Apply quadratic-ball duality. Since $\mathbf{W}_Y - \mathbf{A}\mathbf{W}_X = -\mathbf{M}\mathbf{W}$, the reduced formulations in (27) and (28) match the constrained and dual problems in Lemma C.1, with $k = m$, $r = p$, center $\tilde{\mathbf{W}}$, and budget parameter $t - c_{\mathbf{\Pi}}$. Since $\|\mathbf{M}\|_{\mathrm{op}}^2 = b_{\mathbf{A}}$, that lemma gives

$$f_{\mathbf{\Pi}}(\mathbf{A}, t) = \inf_{\eta \ge 0} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) = \inf_{\eta > b_{\mathbf{A}}} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta).$$

For $\eta > b_{\mathbf{A}}$, (7) also gives

$$\Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) = \eta(t - c_{\mathbf{\Pi}}) + \eta \operatorname{tr}\Big((\eta \mathbf{I}_m - \mathbf{M}\mathbf{M}^\top)^{-1} \mathbf{M}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^\top \mathbf{M}^\top \Big).$$

Recall that

$$\mathbf{M}\mathbf{M}^\top = \mathbf{A}\mathbf{A}^\top + \mathbf{I}_m, \qquad \mathbf{M}\tilde{\mathbf{W}} = (\mathbf{A}\tilde{\mathbf{S}}_X - \tilde{\mathbf{S}}_Y)\mathbf{U} = -\tilde{\mathbf{E}}_{\mathbf{A}}\mathbf{U}.$$

Since $\mathbf{U}\mathbf{U}^\top = \mathbf{\Pi}$, substitution yields

$$\Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) = \eta(t - c_{\mathbf{\Pi}}) + \eta \operatorname{tr}\Big(((\eta - 1)\mathbf{I}_m - \mathbf{A}\mathbf{A}^\top)^{-1} \tilde{\mathbf{E}}_{\mathbf{A}} \mathbf{\Pi} \tilde{\mathbf{E}}_{\mathbf{A}}^\top \Big), \qquad \eta > b_{\mathbf{A}}. \tag{29}$$

Combining (29) with the identity $c_{\mathbf{\Pi}} = \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2 - \operatorname{tr}(\tilde{\mathbf{S}}_X^\top \tilde{\mathbf{S}}_X \mathbf{\Pi})$, cyclicity of the trace, and the definition of $\mathbf{H}_{\mathbf{A},\eta}$, we obtain (25). □

Lemma G.2 (Projection–multiplier minimax exchange)**.** *For every $\mathbf{A} \in \mathcal{A}$ and $t > 0$,*

$$\begin{aligned} V_{\mathbf{A}}(t) &= \inf_{\eta > b_{\mathbf{A}}} \sup_{\mathbf{\Pi} \in \Pi_p} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) \\ &= \inf_{\eta > b_{\mathbf{A}}} \eta\Big(t - \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2 + \operatorname{tr}_p(\mathbf{H}_{\mathbf{A},\eta}) \Big). \end{aligned} \tag{30}$$

Proof. Fix $\mathbf{A} \in \mathcal{A}$ and $t > 0$, and write

$$I \coloneqq (b_{\mathbf{A}}, \infty), \qquad I_j \coloneqq [b_{\mathbf{A}} + 1/j, b_{\mathbf{A}} + j], \quad j \ge 2.$$

These nested compact intervals exhaust I. By Lemma G.1 and the definition of $V_{\mathbf{A}}$,

$$V_{\mathbf{A}}(t) = \sup_{\mathbf{\Pi} \in \Pi_p} \inf_{\eta \in I} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta).$$

We now apply the minimax theorem of Ricceri (2019, Theorem 1.1) to the joint objective $g(\mathbf{\Pi}, \eta) \coloneqq -\Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta)$, with exhaustion sequence $\{I_j\}_{j \ge 2}$. This requires checking the following assumptions.

Compactness, continuity, and convexity. The set Π_p is closed, since its defining constraints in (20) are continuous. Moreover, $\|\mathbf{\Pi}\|_{\mathrm{F}}^2 = \operatorname{tr}(\mathbf{\Pi}) = p$ for each $\mathbf{\Pi} \in \Pi_p$, so it is in fact compact. This implies that each $\mathbf{\Pi} \mapsto \inf_{\eta \in I_j} g(\mathbf{\Pi}, \eta)$ is relatively inf-compact (see the definition in Ricceri, 2019) because the domain Π_p is compact. Equation (25), established in Lemma G.1, shows that $(\mathbf{\Pi}, \eta) \mapsto \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta)$ is jointly continuous on $\Pi_p \times I$. Finally, for fixed $\mathbf{\Pi}$, $\eta \mapsto \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta)$ is convex as a supremum of affine functions in (24).

Connectedness of best-response sets. Fixing $\eta \in I$, we next prove that the set of projections $\mathbf{\Pi} \in \Pi_p$ maximizing $\Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta)$ is connected. Since $\eta > 0$, (25) gives

$$\arg\max_{\mathbf{\Pi}\in\Pi_p} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) = \arg\max_{\mathbf{\Pi}\in\Pi_p} \operatorname{tr}(\mathbf{H}_{\mathbf{A},\eta}\mathbf{\Pi}).$$

List the eigenvalues of $\mathbf{H}_{\mathbf{A},\eta} \in \mathbb{S}^d_+$ in decreasing order, counted with multiplicity, and let α be the pth eigenvalue. Let $\mathbf{P}_>$ project onto the direct sum of eigenspaces with eigenvalues strictly greater than α, set $r_> \coloneqq \operatorname{rank}(\mathbf{P}_>) < p$, and write $\mathcal{V} \coloneqq \operatorname{null}(\mathbf{H}_{\mathbf{A},\eta} - \alpha\mathbf{I}_d)$. The equality conditions in the Ky Fan principle (Lemma A.9) show that every maximizing projection is $\mathbf{P}_> + \mathbf{Q}$, where $\mathbf{Q}$ projects onto an r-dimensional subspace of $\mathcal{V}$, with $r \coloneqq p - r_>$.

Set $n \coloneqq \dim \mathcal{V}$. If $r = n$, the maximizing set is a singleton. Otherwise, choose $\mathbf{V} \in \mathbb{R}^{d\times n}$ whose columns form an orthonormal basis of $\mathcal{V}$, and set $\mathbf{Q}_0 \coloneqq \operatorname{diag}(\mathbf{I}_r, \mathbf{0}_{n-r})$. The maximizing set is then

$$\big\{\mathbf{P}_> + \mathbf{V}\mathbf{O}\mathbf{Q}_0\mathbf{O}^\top\mathbf{V}^\top : \mathbf{O} \in \mathbb{O}^n,\ \det(\mathbf{O}) = 1\big\}.$$

Indeed, allowing all $\mathbf{O} \in \mathbb{O}^n$ generates every projection onto an r-dimensional subspace of $\mathcal{V}$. The determinant restriction loses none of them, since negating one column of $\mathbf{O}$ changes its determinant sign but leaves $\mathbf{O}\mathbf{Q}_0\mathbf{O}^\top$ unchanged because $\mathbf{Q}_0$ is diagonal. The displayed set is therefore a continuous image of the special orthogonal group in dimension n, which is path-connected. Hence the maximizing set is connected.

Lastly, fixing $\mathbf{\Pi} \in \Pi_p$ and $j \geq 2$, we must show that the set of $\eta \in I_j$ minimizing $\Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta)$ is nonempty and connected. This follows directly by continuity and convexity of the objective in η.

To conclude, Ricceri's theorem then gives

$$V_{\mathbf{A}}(t) = \inf_{\eta > b_{\mathbf{A}}} \sup_{\mathbf{\Pi}\in\Pi_p} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta).$$

Substituting (25) and applying Lemma A.9 gives

$$\sup_{\mathbf{\Pi}\in\Pi_p} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, t, \eta) = \eta\Big(t - \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2 + \operatorname{tr}_p(\mathbf{H}_{\mathbf{A},\eta})\Big),$$

which proves the second equality in (30). □

Lemma G.3 (Exactness and joint convexity)**.** *For each $\mathbf{A} \in \mathcal{A}$, the function $t \mapsto V_{\mathbf{A}}(t)$ is finite, concave, and continuous on $(0, \infty)$, and*

$$F(\mathbf{A}) = V_{\mathbf{A}}(\varepsilon^2).$$

Moreover, D, extended by $+\infty$ outside $\eta > b_{\mathbf{A}}$, is jointly convex on $\mathcal{A} \times \mathbb{R}$.

Proof. We prove the claims separately.

Finiteness, concavity, and continuity. Fixing $\mathbf{A} \in \mathcal{A}$, the supremum in (22) defining $V_{\mathbf{A}}(t)$ is feasible for each $t > 0$. Indeed, we may take $\mathbf{S} = \tilde{\mathbf{S}}$ and let $\mathbf{\Pi} \in \Pi_p$ project onto a p-dimensional subspace containing the row space of $\tilde{\mathbf{S}}_X$. Moreover, since each $\mathbf{\Pi} \in \Pi_p$ is dominated in Loewner order by $\mathbf{I}_d$, we have

$$\begin{aligned}
0 &\leq V_{\mathbf{A}}(t) \\
&\leq \sup_{\substack{\mathbf{S}\in\mathbb{R}^{d\times d} \\ \|\mathbf{S}-\tilde{\mathbf{S}}\|_{\mathrm{F}}^2 \leq t}} \|\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X\|_{\mathrm{F}}^2 \\
&\leq b_{\mathbf{A}} \sup_{\substack{\mathbf{S}\in\mathbb{R}^{d\times d} \\ \|\mathbf{S}-\tilde{\mathbf{S}}\|_{\mathrm{F}}^2 \leq t}} \|\mathbf{S}\|_{\mathrm{F}}^2 && \text{(Lemma A.1)} \\
&\leq b_{\mathbf{A}}\big(\|\tilde{\mathbf{S}}\|_{\mathrm{F}} + \sqrt{t}\big)^2 < \infty.
\end{aligned}$$

By Lemma G.2, $t \mapsto V_{\mathbf{A}}(t)$ is the infimum of affine functions of t, hence concave. Being finite and concave on the open interval $(0, \infty)$, this function is continuous there.

Exactness of the projection relaxation. The inequality $F(\mathbf{A}) \le V_{\mathbf{A}}(\varepsilon^2)$ was established above. For the reverse inequality, fix $0 < s < \varepsilon^2$ and a pair $(\mathbf{\Pi}, \mathbf{S}) \in \Pi_p \times \mathbb{R}^{d\times d}$ satisfying

$$\mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi}, \qquad \|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}}^2 \le s.$$

Now choose $\mathbf{U} \in \mathbb{R}^{d\times p}$ with $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$ and $\mathbf{\Pi} = \mathbf{U}\mathbf{U}^\top$. Since $\mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi}$, we may write $\mathbf{S}_X = \mathbf{C}\mathbf{U}^\top$ with $\mathbf{C} \coloneqq \mathbf{S}_X\mathbf{U} \in \mathbb{R}^{p\times p}$. Then, for $\tau > 0$, define

$$\mathbf{S}_X^{(\tau)} \coloneqq (\mathbf{C} + \tau\mathbf{I}_p)\mathbf{U}^\top, \qquad \mathbf{S}^{(\tau)} \coloneqq \begin{pmatrix} \mathbf{S}_X^{(\tau)} \\ \mathbf{S}_Y \end{pmatrix}.$$

The polynomial $\det(\mathbf{C} + \tau\mathbf{I}_p)$ has leading coefficient 1, so it is nonzero for sufficiently small $\tau > 0$. For such τ, $\mathbf{S}_X^{(\tau)}$ has full row rank and $\mathbf{\Pi}_{\mathbf{S}_X^{(\tau)}} = \mathbf{\Pi}$. Moreover,

$$\|\mathbf{S}^{(\tau)} - \tilde{\mathbf{S}}\|_{\mathrm{F}} \le \|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}} + \|\mathbf{S}^{(\tau)} - \mathbf{S}\|_{\mathrm{F}} \le \sqrt{s} + \tau\sqrt{p} < \varepsilon$$

for sufficiently small τ. By Lemma 1.2, $\mathbf{\Sigma}^{(\tau)} \coloneqq \mathbf{S}^{(\tau)}(\mathbf{S}^{(\tau)})^\top$ thus belongs to $\mathcal{B}_\varepsilon(\tilde{\mathbf{\Sigma}})$ for sufficiently small τ. Hence, Lemma B.1 gives

$$F(\mathbf{A}) \ge \lim_{\tau\downarrow 0} R(\mathbf{A}, \mathbf{\Sigma}^{(\tau)}) = \lim_{\tau\downarrow 0} \|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X^{(\tau)})\mathbf{\Pi}\|_{\mathrm{F}}^2 = \|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}\|_{\mathrm{F}}^2.$$

Taking the supremum over $(\mathbf{\Pi}, \mathbf{S})$ gives

$$V_{\mathbf{A}}(s) \le F(\mathbf{A}) \le V_{\mathbf{A}}(\varepsilon^2).$$

Letting $s \uparrow \varepsilon^2$ and using continuity proves $F(\mathbf{A}) = V_{\mathbf{A}}(\varepsilon^2)$.

Joint convexity. For fixed $\mathbf{S}, \mathbf{\Pi}$, the expression

$$\|(\mathbf{S}_Y - \mathbf{A}\mathbf{S}_X)\mathbf{\Pi}\|_{\mathrm{F}}^2 + \eta\big(\varepsilon^2 - \|\mathbf{S} - \tilde{\mathbf{S}}\|_{\mathrm{F}}^2\big)$$

is convex in $\mathbf{A}$ and affine in η, with no mixed term. It is therefore jointly convex in $(\mathbf{A}, \eta)$. The constraint $\mathbf{S}_X = \mathbf{S}_X\mathbf{\Pi}$ is independent of these variables, so taking the supremum in (24) preserves joint convexity. Thus $\Lambda_{\mathbf{\Pi}}(\cdot, \varepsilon^2, \cdot)$ is jointly convex on $\mathcal{A} \times [0, \infty)$ for every fixed $\mathbf{\Pi}$. By (25) and the Ky Fan formula,

$$D(\mathbf{A}, \eta) = \sup_{\mathbf{\Pi}\in\Pi_p} \Lambda_{\mathbf{\Pi}}(\mathbf{A}, \varepsilon^2, \eta) \qquad \text{when } \eta > b_{\mathbf{A}}.$$

Hence D is jointly convex on this domain. The domain itself is convex, being the strict epigraph of $\mathbf{A} \mapsto 1 + \|\mathbf{A}\|_{\mathrm{op}}^2$. Extending D by $+\infty$ outside this domain preserves convexity. □

Proof of Theorem 4.1. By Lemmas G.2 and G.3,

$$F(\mathbf{A}) = V_{\mathbf{A}}(\varepsilon^2) = \inf_{\eta > b_{\mathbf{A}}} D(\mathbf{A}, \eta) \qquad \forall\, \mathbf{A} \in \mathcal{A},$$

and D is jointly convex as an extended-value function. Taking the infimum over $\mathbf{A}$ gives

$$\inf_{\mathbf{A}\in\mathcal{A}} F(\mathbf{A}) = \inf_{\mathbf{A}\in\mathcal{A},\ \eta\in\mathbb{R}} D(\mathbf{A}, \eta),$$

as claimed. □

G.2 Subgradient Method Convergence for SNR (Proof of Proposition 4.2)

If $\tilde{\mathbf{\Sigma}}_Y = \mathbf{0}$, then $G(\mathbf{0}) = 0$. Since G is always non-negative, $\mathbf{0}$ is a minimizer, as claimed.

Otherwise, recall that

$$\begin{aligned} B &\coloneqq (d \operatorname{tr} \tilde{\mathbf{\Sigma}}_Y)^{1/2}/\varepsilon \\ M &\coloneqq 2(B\|\tilde{\mathbf{\Sigma}}_X\|_{\mathrm{op}} + \|\tilde{\mathbf{\Sigma}}_{YX}\|_{\mathrm{F}} + \varepsilon^2 B) \\ T &\coloneqq \lceil M^2 B^2 \delta^{-2} \rceil. \end{aligned}$$

First, we note a simple bound on the norm of a minimizer of $G(\mathbf{A}) = L(\mathbf{A}, \tilde{\mathbf{\Sigma}}) + \varepsilon^2\|\mathbf{A}\|_{\mathrm{op}}^2$.

Lemma G.4 (SNR solution bound)**.** *Each minimizer $\mathbf{A}_\star$ of G satisfies $\|\mathbf{A}_\star\|_{\mathrm{F}} \le B$.*

Proof. By optimality, we have

$$\varepsilon^2 \|\mathbf{A}_\star\|_{\mathrm{op}}^2 \le G(\mathbf{A}_\star) \le G(\mathbf{0}) = \mathrm{tr}(\tilde{\mathbf{\Sigma}}_Y),$$

and so $\|\mathbf{A}_\star\|_{\mathrm{F}} \le \sqrt{d}\|\mathbf{A}_\star\|_{\mathrm{op}} \le (d\,\mathrm{tr}(\tilde{\mathbf{\Sigma}}_Y))^{1/2}/\varepsilon = B$. □

Thus, writing $\mathcal{K}$ for the Frobenius ball of radius B centered at $\mathbf{A}_1 = \mathbf{0}$, we have $G(\mathbf{A}_\star) = \min_{\mathbf{A}\in\mathcal{K}} G(\mathbf{A})$. Moreover, the projection operator appearing in the proposition is exactly the Frobenius projection onto $\mathcal{K}$.

Next, we claim that G is M-Lipschitz under the Frobenius norm over the domain $\mathcal{K}$. Indeed, we can write the subdifferential set $\partial G(\mathbf{A})$ as the convex hull

$$\mathrm{conv}\Big\{2\mathbf{A}\tilde{\mathbf{\Sigma}}_X - 2\tilde{\mathbf{\Sigma}}_{YX} + 2\varepsilon^2\|\mathbf{A}\|_{\mathrm{op}} uv^\top : u \in \mathbb{R}^m, v \in \mathbb{R}^p, \|u\| = \|v\| = 1, \mathbf{A}v = \|\mathbf{A}\|_{\mathrm{op}} u\Big\},$$

Here, the operator-norm subgradient is handled via the characterization of the spectral-norm subdifferential set in Example 2.8 of Lewis (1995). Thus the Lipschitz constant of G over $\mathcal{K}$ is at most

$$\sup_{\substack{\mathbf{A}\in\mathcal{K} \\ u\in\mathbb{R}^m, \|u\|\le 1 \\ v\in\mathbb{R}^p, \|v\|\le 1}} \Big\| 2\mathbf{A}\tilde{\mathbf{\Sigma}}_X - 2\tilde{\mathbf{\Sigma}}_{YX} + 2\varepsilon^2\|\mathbf{A}\|_{\mathrm{op}} uv^\top \Big\|_{\mathrm{F}} \le M,$$

using the triangle inequality and Lemma A.1.

Noting that the Frobenius geometry is Euclidean, Proposition 4.2 then follows by standard guarantees for the projected subgradient method (e.g., Theorem 3.2 of Bubeck, 2015). □

G.3 FISTA Convergence for SNR (Proof of Proposition 4.3)

First, we observe that Algorithm 2, given input $\mathbf{B}$ and parameter α, evaluates the proximal operator of $\alpha\|\cdot\|_{\mathrm{op}}^2$ at $\mathbf{B}$. In other words, it solves the minimization problem

$$\min_{\mathbf{C}\in\mathbb{R}^{m\times p}} \alpha\|\mathbf{C}\|_{\mathrm{op}}^2 + \frac{1}{2}\|\mathbf{C} - \mathbf{B}\|_{\mathrm{F}}^2.$$

Indeed, by the discussion of orthogonally invariant matrix functions in Parikh and Boyd (2014, Sections 6.7.2–6.7.3), the solution to the proximal problem above can be obtained by first computing an SVD of $\mathbf{B}$ and then applying the ℓ_∞^2 proximal operator to the vector of its singular values. The clipping procedure of Algorithm 2 is the squared-ℓ_∞ proximal operator derived in Jacobsson et al. (2017, Theorem 4).

Given that Algorithm 2 correctly implements the proximal operator for $\alpha\|\cdot\|_{\mathrm{op}}^2$, Algorithm 1 is simply a standard application of the FISTA algorithm (Beck and Teboulle, 2009), applied to the objective G, decomposed as the sum of the smooth term $L(\cdot, \tilde{\mathbf{\Sigma}})$ and the proximable term $\varepsilon^2\|\cdot\|_{\mathrm{op}}^2$. In particular, letting $\mathbf{A}_\star$ denote any minimizer of G, Theorem 4.4 of Beck and Teboulle (2009) gives

$$G(\mathbf{A}_T) - G(\mathbf{A}_\star) \le \frac{2M\|\mathbf{A}_\star\|_{\mathrm{F}}^2}{(T+1)^2},$$

using that $L(\cdot, \tilde{\mathbf{\Sigma}})$ is $M = 2\lambda_{\max}(\tilde{\mathbf{\Sigma}}_X)$ smooth. By Lemma G.4, $\|\mathbf{A}_\star\|_{\mathrm{F}}^2 \le d\,\mathrm{tr}(\tilde{\mathbf{\Sigma}}_Y)/\varepsilon^2$, and so $\mathbf{A}_T$ has additive approximation error δ once $T \ge (2Md\,\mathrm{tr}(\tilde{\mathbf{\Sigma}}_Y))^{1/2}\varepsilon^{-1}\cdot\delta^{-1/2}$, as claimed.

H Remaining Proofs for Section 5

H.1 Properties of the Bures Variance Proxy (Proof of Proposition 5.1)

Let $Z \sim P$ and abbreviate $\mathbf{\Sigma} \coloneqq \mathbf{\Sigma}_P$. Since $\mathsf{V}_{\mathsf{B}}(P)$ is the expectation of a quartic in the entries of Z, it is finite for $P \in \mathcal{P}_4(\mathbb{R}^d)$. By non-negativity of the Frobenius norm, $\mathsf{V}_{\mathsf{B}}(P)$ vanishes if and only if

$$(ZZ^\top - \mathbf{\Sigma})\mathbf{\Sigma}_P^{\dagger/2} = 0 \iff ZZ^\top = \mathbf{\Sigma}$$

almost surely. The equivalence holds because Z lies in the range of $\mathbf{\Sigma}$ almost surely.

Expanding the squared norm in the definition of V_B and using $\mathbf{\Sigma}\mathbf{\Sigma}^\dagger Z = Z$ almost surely gives

$$\left\|(ZZ^\top - \mathbf{\Sigma})\mathbf{\Sigma}^{\dagger/2}\right\|_{\mathrm{F}}^2 = \|Z\|^2 Z^\top \mathbf{\Sigma}^\dagger Z - 2\|Z\|^2 + \operatorname{tr}(\mathbf{\Sigma}).$$

Taking expectations and using $\mathbb{E}_P \|Z\|^2 = \operatorname{tr}(\mathbf{\Sigma})$ proves the equality in (3).

For $\mathbf{\Sigma} \neq \mathbf{0}$, let $\lambda_{\min}^+(\mathbf{\Sigma})$ denote its smallest positive eigenvalue. The fourth-moment bound follows from the inequality

$$Z^\top \mathbf{\Sigma}^\dagger Z \leq \frac{\|Z\|^2}{\lambda_{\min}^+(\mathbf{\Sigma})}$$

on $\operatorname{range}(\mathbf{\Sigma})$, implying that

$$\mathsf{V}_\mathsf{B}(P) \leq \mathbb{E}_P\big[\|Z\|^2 Z^\top \mathbf{\Sigma}^\dagger Z\big] \leq \frac{\mathbb{E}\,\|Z\|^4}{\lambda_{\min}^+(\mathbf{\Sigma})}.$$

This is finite whenever $P \in \mathcal{P}_4(\mathbb{R}^d)$ and $\mathbf{\Sigma} \neq 0$. The $\mathbf{\Sigma} = 0$ case requires $Z = 0$ almost surely and hence $\mathsf{V}_\mathsf{B}(P) = 0$. When $\mathbf{\Sigma} \succ \mathbf{0}$, $\lambda_{\min}^+(\mathbf{\Sigma}) = \lambda_{\min}(\mathbf{\Sigma})$, giving the stated bound.

Now suppose that $P = \mathcal{N}(0, \mathbf{\Sigma})$ is Gaussian, and assume without loss of generality that $\mathbf{\Sigma}$ is diagonalizable in the standard basis, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Restricting to $\operatorname{range}(\mathbf{\Sigma})$, let $r \coloneqq \operatorname{rank}(\mathbf{\Sigma})$ and write

$$Z = \sum_{j=1}^{r} \sqrt{\lambda_j} G_j e_j, \qquad G_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

This gives

$$Z^\top \mathbf{\Sigma}^\dagger Z = \sum_{j=1}^{r} G_j^2, \qquad \|Z\|^2 = \sum_{j=1}^{r} \lambda_j G_j^2,$$

and so we compute

$$\begin{aligned}
\mathbb{E}_P\big[\|Z\|^2 Z^\top \mathbf{\Sigma}^\dagger Z\big] &= \sum_{i=1}^{r} \lambda_i\, \mathbb{E}\left[G_i^2 \sum_{j=1}^{r} G_j^2\right] \\
&= \sum_{i=1}^{r} \lambda_i \big(3 + (r-1)\big) \\
&= (r+2)\operatorname{tr}(\mathbf{\Sigma}).
\end{aligned}$$

Therefore, $\mathsf{V}_\mathsf{B}(P) = (r+2)\operatorname{tr}(\mathbf{\Sigma}) - \operatorname{tr}(\mathbf{\Sigma}) = (r+1)\operatorname{tr}(\mathbf{\Sigma})$.

If further $P \in \mathcal{P}^*$ is non-degenerate, we then have $r = \operatorname{rank}(\mathbf{\Sigma}) \geq \operatorname{rank}(\mathbf{\Sigma}_X) = p$. Moreover,

$$\operatorname{tr}(\mathbf{\Sigma}) \geq \operatorname{tr}(\mathbf{\Sigma}_Y) = \operatorname{tr}(\mathbf{A}_P \mathbf{\Sigma}_X \mathbf{A}_P^\top) + \operatorname{tr}(\mathbf{\Sigma}_E) \geq \operatorname{tr}(\mathbf{\Sigma}_E).$$

Combining with the above gives $\mathsf{V}_\mathsf{B}(P) \geq (p+1)\operatorname{tr}(\mathbf{\Sigma}_E) \geq p \operatorname{tr}(\mathbf{\Sigma}_E)$, as claimed.

It remains to prove the empirical Bures bound. Since each $Z_i \in \operatorname{range}(\mathbf{\Sigma})$ almost surely, we have $\operatorname{range}(\mathbf{\Sigma}_{P_n}) \subseteq \operatorname{range}(\mathbf{\Sigma})$ almost surely. Therefore Lemma B.2 gives

$$\mathsf{B}(\mathbf{\Sigma}, \mathbf{\Sigma}_{P_n})^2 \leq \operatorname{tr}\big((\mathbf{\Sigma}_{P_n} - \mathbf{\Sigma})\mathbf{\Sigma}^\dagger(\mathbf{\Sigma}_{P_n} - \mathbf{\Sigma})\big).$$

Next, set $\mathbf{H}_i \coloneqq Z_i Z_i^\top - \mathbf{\Sigma}$, so that $\mathbf{\Sigma}_{P_n} - \mathbf{\Sigma} = \sum_{i=1}^n \mathbf{H}_i / n$. The $\mathbf{H}_i$ matrices are independent and mean zero, so the cross terms vanish in expectation:

$$\mathbb{E}\operatorname{tr}\big((\mathbf{\Sigma}_{P_n} - \mathbf{\Sigma})\mathbf{\Sigma}^\dagger(\mathbf{\Sigma}_{P_n} - \mathbf{\Sigma})\big) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}\operatorname{tr}(\mathbf{H}_i \mathbf{\Sigma}^\dagger \mathbf{H}_i) = \frac{1}{n}\mathsf{V}_\mathsf{B}(P),$$

where the last equality uses (3). This proves the claim. □

H.2 Finite-Sample Regret Bounds (Proof of Theorem 5.2)

We first establish some empirical estimates needed for the sub-Gaussian guarantee.

Lemma H.1 (Empirical conditioning and ERM error)**.** *Fix $P \in \mathcal{P}^*$ and $\delta \in (0,1)$. Let $(X,Y) \sim P$ and $E = Y - \mathbf{A}_P X$. Suppose that the random vector $((\mathbf{\Sigma}_X^P)^{-1/2}X, (\mathbf{\Sigma}_E^P)^{\dagger/2}E)$ is K-sub-Gaussian. Then for $n \gtrsim K^4(d + \log(1/\delta))$, the following hold simultaneously with probability at least $1-\delta$:*

$$\begin{aligned} \tfrac{3}{4}\mathbf{\Sigma}_X^P \preceq \mathbf{\Sigma}_X^{P_n} &\preceq \tfrac{5}{4}\mathbf{\Sigma}_X^P, \\ \kappa_{P_n} &\le 4\kappa_P, \\ R(\mathsf{ERM}[P_n], P) &\lesssim \frac{K^4 \operatorname{tr}(\mathbf{\Sigma}_E^P)(d + \log(1/\delta))}{n}. \end{aligned}$$

In particular, $P_n \in \mathcal{P}^*$ on this event.

Proof. Define the random vectors

$$\bar{X} \coloneqq (\mathbf{\Sigma}_X^P)^{-1/2}X, \qquad \bar{E} \coloneqq (\mathbf{\Sigma}_E^P)^{\dagger/2}E, \qquad W \coloneqq (\bar{X}, \bar{E}).$$

We will consider both P and P_n as their base probability law. Let $\mathbf{\Pi}_E$ project onto $\operatorname{range}(\mathbf{\Sigma}_E^P)$. Then $E \in \operatorname{range}(\mathbf{\Sigma}_E^P)$ almost surely, and the population normal equation gives $\mathbb{E}_P[EX^\top] = \mathbf{0}$. Therefore,

$$\mathbf{\Sigma}_W^P \coloneqq \mathbb{E}_P[WW^\top] = \operatorname{diag}(\mathbf{I}_p, \mathbf{\Pi}_E).$$

The vector W is supported on $\mathbb{R}^p \times \operatorname{range}(\mathbf{\Sigma}_E^P)$ and has identity second moment on this subspace, whose dimension is at most d. The feature directions have unit second moment, so $K \ge 1$. Applying Vershynin (2026, Theorem 4.7.1 and Remark 4.7.3) on this subspace gives an event $\mathcal{E}$ of probability at least $1-\delta$ on which

$$\left\|\mathbb{E}_{P_n}[WW^\top] - \mathbf{\Sigma}_W^P\right\|_{\mathrm{op}} \le \alpha_n \coloneqq CK^2\sqrt{\frac{d + \log(1/\delta)}{n}} \le \frac{1}{4}.$$

Here C is an absolute constant, and the sample-size assumption absorbs the linear term in the cited concentration bound. The second-moment formulation does not require W to be centered.

Define the empirical blocks

$$\begin{aligned} \mathbf{H}_n &\coloneqq \mathbb{E}_{P_n}[\bar{X}\bar{X}^\top] \in \mathbb{R}^{p\times p}, \\ \mathbf{C}_n &\coloneqq \mathbb{E}_{P_n}[\bar{E}\bar{X}^\top] \in \mathbb{R}^{m\times p}, \\ \mathbf{J}_n &\coloneqq \mathbb{E}_{P_n}[\bar{E}\bar{E}^\top] \in \mathbb{R}^{m\times m}. \end{aligned}$$

On $\mathcal{E}$ we have

$$\|\mathbf{H}_n - \mathbf{I}_p\|_{\mathrm{op}}, \|\mathbf{C}_n\|_{\mathrm{op}}, \|\mathbf{J}_n - \mathbf{\Pi}_E\|_{\mathrm{op}} \le \alpha_n \le \frac{1}{4}. \tag{31}$$

Condition on this event for the remainder of the proof. Since

$$\mathbf{\Sigma}_X^{P_n} = (\mathbf{\Sigma}_X^P)^{1/2}\mathbf{H}_n(\mathbf{\Sigma}_X^P)^{1/2},$$

we obtain the claimed bounds for the empirical feature second-moment matrix, with the lower bound implying that $P_n \in \mathcal{P}^*$. Also, $\|\mathbf{H}_n^{-1}\|_{\mathrm{op}} \le (1-\alpha_n)^{-1} \le 4/3$.

The identities $Y = \mathbf{A}_P X + E$ and $E = (\mathbf{\Sigma}_E^P)^{1/2}\bar{E}$ imply that

$$\mathbf{\Sigma}_{YX}^{P_n} = \mathbf{A}_P\mathbf{\Sigma}_X^{P_n} + \mathbb{E}_{P_n}[EX^\top] = \mathbf{A}_P\mathbf{\Sigma}_X^{P_n} + (\mathbf{\Sigma}_E^P)^{1/2}\mathbf{C}_n(\mathbf{\Sigma}_X^P)^{1/2}.$$

Combining the above with the formula for $\mathbf{A}_{P_n}$ gives

$$\begin{aligned} \mathbf{A}_{P_n} - \mathbf{A}_P &= \mathbf{\Sigma}_{YX}^{P_n}(\mathbf{\Sigma}_X^{P_n})^{-1} - \mathbf{A}_P \\ &= (\mathbf{\Sigma}_E^P)^{1/2}\mathbf{C}_n(\mathbf{\Sigma}_X^P)^{1/2}(\mathbf{\Sigma}_X^{P_n})^{-1} \\ &= (\mathbf{\Sigma}_E^P)^{1/2}\mathbf{C}_n\mathbf{H}_n^{-1}(\mathbf{\Sigma}_X^P)^{-1/2}. \end{aligned} \tag{32}$$

Consequently,

$$\|\mathbf{A}_{P_n} - \mathbf{A}_P\|_{\mathrm{op}} \le \frac{\alpha_n}{1-\alpha_n}\sqrt{\frac{\lambda_{\max}(\mathbf{\Sigma}_E^P)}{\lambda_{\min}(\mathbf{\Sigma}_X^P)}} \le \frac{1}{3}\sqrt{\frac{\lambda_{\max}(\mathbf{\Sigma}_E^P)}{\lambda_{\min}(\mathbf{\Sigma}_X^P)}}.$$

Using (32), the empirical fitted residual can be written as

$$Y - \mathbf{A}_{P_n} X = E + (\mathbf{A}_P - \mathbf{A}_{P_n})X = (\mathbf{\Sigma}_E^P)^{1/2}(\bar{E} - \mathbf{C}_n \mathbf{H}_n^{-1} \bar{X}).$$

Using the definitions of $\mathbf{C}_n$, $\mathbf{H}_n$ and $\mathbf{J}_n$, along with symmetry of $\mathbf{H}_n^{-1}$, we compute

$$\begin{aligned}
&\mathbb{E}_{P_n}\big[(\bar{E} - \mathbf{C}_n \mathbf{H}_n^{-1} \bar{X})(\bar{E} - \mathbf{C}_n \mathbf{H}_n^{-1} \bar{X})^\top\big] \\
&\quad = \mathbf{J}_n - \mathbb{E}_{P_n}[\bar{E}\bar{X}^\top]\mathbf{H}_n^{-1}\mathbf{C}_n^\top - \mathbf{C}_n\mathbf{H}_n^{-1}\,\mathbb{E}_{P_n}[\bar{X}\bar{E}^\top] + \mathbf{C}_n\mathbf{H}_n^{-1}\,\mathbb{E}_{P_n}[\bar{X}\bar{X}^\top]\mathbf{H}_n^{-1}\mathbf{C}_n^\top \\
&\quad = \mathbf{J}_n - 2\mathbf{C}_n\mathbf{H}_n^{-1}\mathbf{C}_n^\top + \mathbf{C}_n\mathbf{H}_n^{-1}\mathbf{H}_n\mathbf{H}_n^{-1}\mathbf{C}_n^\top \\
&\quad = \mathbf{J}_n - \mathbf{C}_n\mathbf{H}_n^{-1}\mathbf{C}_n^\top.
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
\mathbf{\Sigma}_E^{P_n} &= \mathbb{E}_{P_n}\big[(Y - \mathbf{A}_{P_n}X)(Y - \mathbf{A}_{P_n}X)^\top\big] \\
&= (\mathbf{\Sigma}_E^P)^{1/2}(\mathbf{J}_n - \mathbf{C}_n\mathbf{H}_n^{-1}\mathbf{C}_n^\top)(\mathbf{\Sigma}_E^P)^{1/2} \\
&\preceq (\mathbf{\Sigma}_E^P)^{1/2}\mathbf{J}_n(\mathbf{\Sigma}_E^P)^{1/2} \\
&\preceq \frac{5}{4}\mathbf{\Sigma}_E^P.
\end{aligned}$$

The last inequality follows from $\mathbf{J}_n \preceq \mathbf{\Pi}_E + \alpha_n \mathbf{I}_m$ and $(\mathbf{\Sigma}_E^P)^{1/2}\mathbf{\Pi}_E = (\mathbf{\Sigma}_E^P)^{1/2}$. Combining the above with Theorem 2.2 gives

$$\begin{aligned}
\kappa_{P_n} &\le 1 + \|\mathbf{A}_{P_n}\|_{\mathrm{op}}^2 + \frac{\lambda_{\max}(\mathbf{\Sigma}_E^{P_n})}{\lambda_{\min}(\mathbf{\Sigma}_X^{P_n})} \\
&\le 1 + 2\|\mathbf{A}_P\|_{\mathrm{op}}^2 + 2\|\mathbf{A}_{P_n} - \mathbf{A}_P\|_{\mathrm{op}}^2 + \frac{\frac{5}{4}\lambda_{\max}(\mathbf{\Sigma}_E^P)}{\frac{3}{4}\lambda_{\min}(\mathbf{\Sigma}_X^P)} \\
&\le 1 + 2\|\mathbf{A}_P\|_{\mathrm{op}}^2 + \left(\frac{2}{9} + \frac{5}{3}\right)\frac{\lambda_{\max}(\mathbf{\Sigma}_E^P)}{\lambda_{\min}(\mathbf{\Sigma}_X^P)} \\
&\le 4\kappa_P.
\end{aligned}$$

Finally, (32) gives

$$\begin{aligned}
R(\mathsf{ERM}[P_n], P) &= \|(\mathbf{A}_{P_n} - \mathbf{A}_P)(\mathbf{\Sigma}_X^P)^{1/2}\|_{\mathrm{F}}^2 \\
&= \|(\mathbf{\Sigma}_E^P)^{1/2}\mathbf{C}_n\mathbf{H}_n^{-1}\|_{\mathrm{F}}^2 \\
&\le \frac{\alpha_n^2}{(1-\alpha_n)^2}\operatorname{tr}(\mathbf{\Sigma}_E^P) \\
&\lesssim \frac{K^4 \operatorname{tr}(\mathbf{\Sigma}_E^P)(d + \log(1/\delta))}{n},
\end{aligned}$$

as claimed. □

We now prove the two guarantees separately. For each, we establish a high-probability event depending only on the i.i.d. sample from P, under which the claimed bounds hold uniformly over the admissible choices in both perturbation models.

Fourth-moment guarantee. Abbreviate $\rho \coloneqq \varepsilon_n$. By Proposition 5.1, $\mathbb{E}[\mathsf{B}(\mathbf{\Sigma}_{P_n}, \mathbf{\Sigma}_P)^2] \le V/n$, so Markov's inequality gives $\mathsf{B}(\mathbf{\Sigma}_{P_n}, \mathbf{\Sigma}_P) \le \sqrt{V/(\delta n)}$ with probability at least $1 - \delta$. On this event,

$$\mathsf{B}(\mathbf{\Sigma}_{P_{\mathrm{s},n}}, \mathbf{\Sigma}_P) + \mathsf{B}(\mathbf{\Sigma}_{P_{\mathrm{t}}}, \mathbf{\Sigma}_P) \le \varepsilon + \sqrt{\frac{V}{\delta n}} = \rho \tag{33}$$

under either adversary model. Indeed, under the sample-corruption model, we have $P_{\mathrm{t}} = P$ and $\mathsf{W}_2(P_{\mathrm{s},n}, P_n) \le \varepsilon$, so the claim follows by the W_2–Bures equivalence (Lemma 1.2). Likewise, under the distribution-shift model, we have $P_{\mathrm{s},n} = P_n$ and $\mathsf{B}(\mathbf{\Sigma}_{P_t}, \mathbf{\Sigma}_P) \le \mathsf{W}_2(P_t, P) \le \varepsilon$.

Equation (33) implies that $\mathsf{B}(\mathbf{\Sigma}_{P_{\mathrm{s},n}}, \mathbf{\Sigma}_P), \mathsf{B}(\mathbf{\Sigma}_{P_{\mathrm{t}}}, \mathbf{\Sigma}_P), \mathsf{B}(\mathbf{\Sigma}_{P_{\mathrm{s},n}}, \mathbf{\Sigma}_{P_{\mathrm{t}}}) \le \rho$, with the last case holding by the triangle inequality. Note that $\rho \le 2\varepsilon$ under our assumed sample complexity. We

now proceed as in the population-limit proof. In particular, we use the equivalent second-moment formulations of the estimators and regret stability, justified by Lemma 1.2.

For the MRO guarantee, set $\hat{\mathbf{A}} \coloneqq \mathsf{MRO}_\rho[P_{\mathrm{s},n}]$ and bound

$$\begin{aligned}
R(\hat{\mathbf{A}}, P_{\mathrm{t}}) &\le \sup_{\boldsymbol{\Sigma} \in \mathcal{B}_\rho(\boldsymbol{\Sigma}_{P_{\mathrm{s},n}})} R(\hat{\mathbf{A}}, \boldsymbol{\Sigma}) && (\mathsf{B}(\boldsymbol{\Sigma}_{P_{\mathrm{s},n}}, \boldsymbol{\Sigma}_{P_{\mathrm{t}}}) \le \rho) \\
&\le \sup_{\boldsymbol{\Sigma} \in \mathcal{B}_\rho(\boldsymbol{\Sigma}_{P_{\mathrm{s},n}})} R(\mathbf{A}_P, \boldsymbol{\Sigma}) && (\text{MRO optimality}) \\
&\le \sup_{\boldsymbol{\Sigma} \in \mathcal{B}_{2\rho}(\boldsymbol{\Sigma}_P)} R(\mathbf{A}_P, \boldsymbol{\Sigma}) && (\mathsf{B}(\boldsymbol{\Sigma}_{P_{\mathrm{s},n}}, \boldsymbol{\Sigma}_P) \le \rho) \\
&\le 4\kappa_P (2\rho)^2 && (\text{Lemma 3.1}) \\
&\le 64 \kappa_P \varepsilon^2, && (\rho \le 2\varepsilon)
\end{aligned}$$

as claimed.

For SNR, set $\hat{\mathbf{A}} \coloneqq \mathsf{SNR}_\rho[P_{\mathrm{s},n}]$. Since risk and regret differ by a constant independent of the coefficient matrix, SNR optimality implies

$$R(\hat{\mathbf{A}}, P_{\mathrm{s},n}) + \rho^2 \|\hat{\mathbf{A}}\|_{\mathrm{op}}^2 \le R(\mathbf{A}_P, P_{\mathrm{s},n}) + \rho^2 \|\mathbf{A}_P\|_{\mathrm{op}}^2 \le 5\kappa_P \rho^2.$$

The last inequality follows from $\mathsf{B}(\boldsymbol{\Sigma}_{P_{\mathrm{s},n}}, \boldsymbol{\Sigma}_P) \le \rho$, Lemma 3.1, and $\|\mathbf{A}_P\|_{\mathrm{op}}^2 \le \kappa_P$ from Theorem 2.2. To convert this to a regret bound under P_{t}, we apply regret stability twice, using P as the reference law in both applications:

$$\begin{aligned}
\sqrt{R(\hat{\mathbf{A}}, P_{\mathrm{t}})} &\le \sqrt{R(\hat{\mathbf{A}}, P)} + \big(2\sqrt{\kappa_P} + \|\hat{\mathbf{A}}\|_{\mathrm{op}}\big) \mathsf{B}(\boldsymbol{\Sigma}_{P_{\mathrm{t}}}, \boldsymbol{\Sigma}_P) \\
&\le \sqrt{R(\hat{\mathbf{A}}, P_{\mathrm{s},n})} + \big(2\sqrt{\kappa_P} + \|\hat{\mathbf{A}}\|_{\mathrm{op}}\big) \rho,
\end{aligned}$$

where the last inequality uses (33). By Cauchy–Schwarz and the SNR objective comparison above,

$$\sqrt{R(\hat{\mathbf{A}}, P_{\mathrm{t}})} \le \sqrt{2} \sqrt{R(\hat{\mathbf{A}}, P_{\mathrm{s},n}) + \rho^2 \|\hat{\mathbf{A}}\|_{\mathrm{op}}^2} + 2\sqrt{\kappa_P}\rho \le 6\sqrt{\kappa_P}\rho \le 12\sqrt{\kappa_P}\varepsilon.$$

Squaring gives the desired SNR bound.

Thus, both estimators incur target regret $O(\kappa_P \varepsilon^2)$ under the specified event. The bounds hold uniformly over all admissible sample corruptions and population targets in the respective models, including randomized, sample-dependent choices. The event has probability at least $1 - \delta$ under every admissible rule, since the reference sample has law $P^{\otimes n}$. Taking the supremum over $\mathbb{M}_{\varepsilon,n}(P)$ therefore proves the first assertion.

Sub-Gaussian guarantee. Our assumption on the sample size allows us to apply Lemma H.1. Fix a realization on its event of probability at least $1 - \delta$. Then $P_n \in \mathcal{P}^*$, $\kappa_{P_n} \le 4\kappa_P$, and $\boldsymbol{\Sigma}_X^P \preceq \frac{4}{3} \boldsymbol{\Sigma}_X^{P_n}$. Moreover, we have

$$e_n \coloneqq R(\mathbf{A}_{P_n}, P) \lesssim \frac{K^4 \operatorname{tr}(\boldsymbol{\Sigma}_E^P)(d + \log(1/\delta))}{n} \lesssim \varepsilon^2,$$

where the last inequality uses the theorem's sample-size assumption.

We first establish a comparison that will be used in both perturbation models. For each $\mathbf{A} \in \mathcal{A}$, the regret formula and the squared triangle inequality imply that

$$\begin{aligned}
R(\mathbf{A}, P) &= \|(\mathbf{A} - \mathbf{A}_P)(\boldsymbol{\Sigma}_X^P)^{1/2}\|_{\mathrm{F}}^2 \\
&\le 2\|(\mathbf{A} - \mathbf{A}_{P_n})(\boldsymbol{\Sigma}_X^P)^{1/2}\|_{\mathrm{F}}^2 + 2e_n \\
&\le \frac{8}{3} \|(\mathbf{A} - \mathbf{A}_{P_n})(\boldsymbol{\Sigma}_X^{P_n})^{1/2}\|_{\mathrm{F}}^2 + 2e_n \\
&= \frac{8}{3} R(\mathbf{A}, P_n) + 2e_n, \qquad (34)
\end{aligned}$$

where the last inequality uses $\boldsymbol{\Sigma}_X^P \preceq \frac{4}{3} \boldsymbol{\Sigma}_X^{P_n}$.

In the sample-corruption model, fix $\hat{\mathbf{A}}_n \in \{\mathsf{MRO}_\varepsilon, \mathsf{SNR}_\varepsilon\}$ and any admissible observed source $P_{\mathrm{s},n}$. Set $\hat{\mathbf{A}} \coloneqq \hat{\mathbf{A}}_n[P_{\mathrm{s},n}]$. Since $\mathsf{W}_2(P_{\mathrm{s},n}, P_n) \le \varepsilon$ and $P_n \in \mathcal{P}^*$, Lemma 3.2 gives

$$R(\hat{\mathbf{A}}, P_n) \le R_\varepsilon(\hat{\mathbf{A}}_n, P_n) \lesssim \kappa_{P_n}\varepsilon^2 \lesssim \kappa_P\varepsilon^2.$$

Applying (34), we obtain

$$R(\hat{\mathbf{A}}, P_{\mathrm{t}}) = R(\hat{\mathbf{A}}, P) \le \frac{8}{3}R(\hat{\mathbf{A}}, P_n) + 2e_n \lesssim \kappa_P\varepsilon^2,$$

using $e_n \lesssim \varepsilon^2$ and $\kappa_P \ge 1$. This proves the desired target regret bound uniformly over admissible corruptions.

In the distribution-shift model, $P_{\mathrm{s},n} = P_n$ and $P_{\mathrm{t}} = Q \in \mathcal{B}_\varepsilon(P)$. Since Q need not belong to $\mathcal{B}_\varepsilon(P_n)$, we will first transfer regret from P_n to P, then from P to Q. For the second transfer, we also need to control the coefficient norm.

For $\hat{\mathbf{A}} = \mathsf{SNR}_\varepsilon[P_n]$, comparison with the empirical risk minimizer $\mathbf{A}_{P_n}$ gives

$$R(\hat{\mathbf{A}}, P_n) + \varepsilon^2\|\hat{\mathbf{A}}\|_{\mathrm{op}}^2 \le R(\mathbf{A}_{P_n}, P_n) + \varepsilon^2\|\mathbf{A}_{P_n}\|_{\mathrm{op}}^2 = \varepsilon^2\|\mathbf{A}_{P_n}\|_{\mathrm{op}}^2 \le \kappa_{P_n}\varepsilon^2.$$

Here the first inequality uses SNR optimality, and the last uses Theorem 2.2. Since both terms on the left are nonnegative, this bounds the empirical regret and coefficient norm separately.

For MRO, define the objective $F_n(\mathbf{A}) \coloneqq \sup_{Q' \in \mathcal{B}_\varepsilon(P_n)} R(\mathbf{A}, Q')$. We first show that

$$F_n(\mathbf{A}) \ge \varepsilon^2(1 + \|\mathbf{A}\|_{\mathrm{op}}^2) \qquad \forall\, \mathbf{A} \in \mathcal{A}. \tag{35}$$

By the W_2–Bures equivalence (Lemma 1.2), we may apply Theorem 4.1 with observed second moment $\mathbf{\Sigma}_{P_n}$. Define the square-root factor $\mathbf{S}_n \coloneqq \mathbf{\Sigma}_{P_n}^{1/2}$ and let $\mathbf{S}_{n,X} \in \mathbb{R}^{p\times d}$ be its feature row block. For every $\eta > 1 + \|\mathbf{A}\|_{\mathrm{op}}^2$, the matrix inside the tr_p term in that theorem is $\mathbf{S}_{n,X}^\top\mathbf{S}_{n,X}$ plus a positive semidefinite matrix. The tr_p term is therefore at least

$$\mathrm{tr}_p(\mathbf{S}_{n,X}^\top\mathbf{S}_{n,X}) = \mathrm{tr}(\mathbf{S}_{n,X}^\top\mathbf{S}_{n,X}) = \|\mathbf{S}_{n,X}\|_{\mathrm{F}}^2,$$

where the first equality uses $\mathrm{rank}(\mathbf{S}_{n,X}^\top\mathbf{S}_{n,X}) \le p$. Consequently, the dual objective is at least $\eta\varepsilon^2$, and the reformulation gives

$$F_n(\mathbf{A}) \ge \inf_{\eta > 1+\|\mathbf{A}\|_{\mathrm{op}}^2} \eta\varepsilon^2 = \varepsilon^2(1 + \|\mathbf{A}\|_{\mathrm{op}}^2).$$

This proves (35).

Now take $\hat{\mathbf{A}} = \mathsf{MRO}_\varepsilon[P_n]$. Since $P_n \in \mathcal{B}_\varepsilon(P_n)$, MRO optimality gives

$$R(\hat{\mathbf{A}}, P_n) \le F_n(\hat{\mathbf{A}}) \le F_n(\mathbf{A}_{P_n}) \le 4\kappa_{P_n}\varepsilon^2.$$

The last inequality follows from Lemma 3.1 with reference law P_n and coefficient $\mathbf{A}_{P_n}$. Combining this upper bound with (35) also gives

$$\varepsilon^2(1 + \|\hat{\mathbf{A}}\|_{\mathrm{op}}^2) \le F_n(\hat{\mathbf{A}}) \le 4\kappa_{P_n}\varepsilon^2,$$

and hence $\|\hat{\mathbf{A}}\|_{\mathrm{op}}^2 \le 4\kappa_{P_n}$.

Thus, for either choice of $\hat{\mathbf{A}}_n \in \{\mathsf{MRO}_\varepsilon, \mathsf{SNR}_\varepsilon\}$ and $\hat{\mathbf{A}} \coloneqq \hat{\mathbf{A}}_n[P_n]$, we have

$$R(\hat{\mathbf{A}}, P_n) \lesssim \kappa_P\varepsilon^2, \qquad \|\hat{\mathbf{A}}\|_{\mathrm{op}}^2 \lesssim \kappa_P,$$

using $\kappa_{P_n} \le 4\kappa_P$. The empirical-to-population comparison (34) thus yields

$$R(\hat{\mathbf{A}}, P) \le \tfrac{8}{3}R(\hat{\mathbf{A}}, P_n) + 2e_n \lesssim \kappa_P\varepsilon^2.$$

Finally, for every $Q \in \mathcal{B}_\varepsilon(P)$, Lemma 3.1 gives

$$\begin{aligned}\sqrt{R(\hat{\mathbf{A}}, Q)} &\le \sqrt{R(\hat{\mathbf{A}}, P)} + \big(2\sqrt{\kappa_P} + \|\hat{\mathbf{A}}\|_{\mathrm{op}}\big)\mathsf{W}_2(P, Q) \\ &\le \sqrt{R(\hat{\mathbf{A}}, P)} + \big(2\sqrt{\kappa_P} + \|\hat{\mathbf{A}}\|_{\mathrm{op}}\big)\varepsilon \\ &\lesssim \sqrt{\kappa_P}\varepsilon.\end{aligned}$$

Squaring proves that $R(\hat{\mathbf{A}}, P_{\mathrm{t}}) \lesssim \kappa_P\varepsilon^2$ in the distribution-shift model.

On the chosen sample event, both estimators satisfy the claimed bounds for every admissible sample corruption and every population target $Q \in \mathcal{B}_\varepsilon(P)$ in the respective models. The event has probability at least $1 - \delta$ under every admissible rule, since the reference sample has law $P^{\otimes n}$. Taking the supremum over $\mathbb{M}_{\varepsilon,n}(P)$ therefore proves the second assertion. □

I Exact Reductions for Population Benchmarks in Section 6

In this section, we derive closed forms and exact constant-size semidefinite program (SDP) reformulations for the population benchmarks in Section 6. Write $P \coloneqq P_{\mathrm{t}}$ and $\tilde{P} \coloneqq P_{\mathrm{s}}$ for the target and source laws, respectively, and let $\rho > 0$ denote the supplied radius, which may differ from the corruption budget $\varepsilon = 1$.

Set $q \coloneqq m - 1$. Recall that $m = p = d/2 \geq 2$ and that our P is the law of (X, Y), where

$$X \sim \mathcal{N}(0, \mathbf{I}_m), \qquad Y = X + U, \qquad U \sim \mathcal{N}\big(0, \mathrm{diag}(0, \mathbf{I}_q)\big),$$

and X and U are independent. For $\alpha, \beta \in \mathbb{R}$, consider the diagonal coefficient matrix given by

$$\mathbf{A}(\alpha, \beta) \coloneqq \mathrm{diag}(\alpha, \beta \mathbf{I}_q).$$

With this notation, our source measure $\tilde{P}$ is the law of $(\mathbf{A}(c, 1)X, Y)$, where $c \coloneqq m^{-1/4}$. Since $\mathbf{A}_P = \mathbf{I}_m$ and $\mathbf{\Sigma}_X^P = \mathbf{I}_m$, the target regret of $\mathbf{A}(\alpha, \beta)$ is given by

$$R(\mathbf{A}(\alpha, \beta), P) = (\alpha - 1)^2 + q(\beta - 1)^2. \tag{36}$$

Moreover, the source risk is

$$L\big(\mathbf{A}(\alpha, \beta), \tilde{P}\big) = (c\alpha - 1)^2 + q\big((\beta - 1)^2 + 1\big). \tag{37}$$

Finally, since $\mathbf{\Sigma}_X^{\tilde{P}} = \mathrm{diag}(c^2, \mathbf{I}_q) \succ \mathbf{0}$, the population ERM solution is unique and satisfies

$$\mathsf{ERM}[\tilde{P}] = \mathbf{A}(c^{-1}, 1), \qquad R(\mathsf{ERM}[\tilde{P}], P) = (c^{-1} - 1)^2.$$

I.1 Reduction to Scalar Optimization Using Rotational Symmetry

As in Lemma F.10, both P and $\tilde{P}$ are invariant under the compact group of orthogonal transformations

$$\{\mathrm{diag}(1, \mathbf{H}, 1, \mathbf{H}) : \mathbf{H} \in \mathbb{O}^q\}.$$

Its invariant coefficient matrices (in the sense of Lemma C.4) are exactly those of the form $\mathbf{A}(\alpha, \beta)$ for some $\alpha, \beta \in \mathbb{R}$. Indeed, taking $\mathbf{H} = -\mathbf{I}_q$ eliminates the off-diagonal blocks, and commuting with every $\mathbf{H} \in \mathbb{O}^q$ forces the remaining $q \times q$ block to be a scalar multiple of $\mathbf{I}_q$. Hence, by Lemma C.4, each objective

$$\begin{aligned} J_{\mathrm{DRO}}(\mathbf{A}) &\coloneqq \sup_{Q \in \mathcal{B}_\rho(\tilde{P})} L(\mathbf{A}, Q), \\ J_{\mathrm{MRO}}(\mathbf{A}) &\coloneqq \sup_{Q \in \mathcal{B}_\rho(\tilde{P})} R(\mathbf{A}, Q), \\ J_{\mathrm{SNR}}(\mathbf{A}) &\coloneqq L(\mathbf{A}, \tilde{P}) + \rho^2 \|\mathbf{A}\|_{\mathrm{op}}^2 \end{aligned}$$

admits a minimizer of the form $\mathbf{A}(\alpha, \beta)$.

I.2 Constant-Size SDP Reformulations for DRO and MRO

Using this symmetry reduction, we now derive SDP reformulations for the DRO and MRO problems. Fix $\mathbf{A} = \mathbf{A}(\alpha, \beta)$ for now and write $\gamma \coloneqq 1 + \max\{\alpha^2, \beta^2\}$.

Our first goal is to derive a clean dual formulation for $J_{\mathrm{DRO}}(\mathbf{A})$ and $J_{\mathrm{MRO}}(\mathbf{A})$. We then introduce epigraph variables and simplify the resulting convex programs using Schur complements and a variational formula for tr_p.

Step 1: Initial dual problems. A convenient square-root factor of the source second moment is

$$\tilde{\mathbf{S}} \coloneqq \begin{pmatrix} \mathrm{diag}(c, \mathbf{I}_q) & \mathbf{0}_{m \times m} \\ \mathbf{I}_m & \mathrm{diag}(0, \mathbf{I}_q) \end{pmatrix}, \qquad \tilde{\mathbf{S}}\tilde{\mathbf{S}}^\top = \mathbf{\Sigma}_{\tilde{P}}.$$

Set $\tilde{\mathbf{E}}_{\mathbf{A}} \coloneqq \tilde{\mathbf{S}}_Y - \mathbf{A}\tilde{\mathbf{S}}_X$ and $\mathbf{B}_{\mathbf{A},\eta} \coloneqq (\eta - 1)\mathbf{I}_m - \mathbf{A}\mathbf{A}^\top$. Then, the W_2–Bures equivalence (Lemma 1.2) and the DRO and MRO dual formulas in (2) and Theorem 4.1 give

$$\begin{aligned} J_{\mathrm{DRO}}(\mathbf{A}) &= \inf_{\eta > \gamma} \Big\{ \eta\rho^2 + \mathrm{tr}\big(\eta \tilde{\mathbf{E}}_{\mathbf{A}}^\top \mathbf{B}_{\mathbf{A},\eta}^{-1} \tilde{\mathbf{E}}_{\mathbf{A}}\big) \Big\}, \\ J_{\mathrm{MRO}}(\mathbf{A}) &= \inf_{\eta > \gamma} \Big\{ \eta(\rho^2 - \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2) + \mathrm{tr}_m\big(\eta \tilde{\mathbf{S}}_X^\top \tilde{\mathbf{S}}_X + \eta \tilde{\mathbf{E}}_{\mathbf{A}}^\top \mathbf{B}_{\mathbf{A},\eta}^{-1} \tilde{\mathbf{E}}_{\mathbf{A}}\big) \Big\}. \end{aligned}$$

Here, we used $p = m$ and positive homogeneity of tr_m. To unify these formulas, define

$$\mathbf{K}_\eta \coloneqq \eta\tilde{\mathbf{S}}^\top\tilde{\mathbf{S}} + \eta\tilde{\mathbf{E}}_{\mathbf{A}}^\top\mathbf{B}_{\mathbf{A},\eta}^{-1}\tilde{\mathbf{E}}_{\mathbf{A}}.$$

Substituting into the above gives

$$J_{\mathrm{DRO}}(\mathbf{A}) = \inf_{\eta>\gamma}\Big\{\eta(\rho^2 - \|\tilde{\mathbf{S}}\|_{\mathrm{F}}^2) + \mathrm{tr}(\mathbf{K}_\eta)\Big\}, \tag{38}$$

$$J_{\mathrm{MRO}}(\mathbf{A}) = \inf_{\eta>\gamma}\Big\{\eta(\rho^2 - \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2) + \mathrm{tr}_m(\mathbf{K}_\eta - \eta\tilde{\mathbf{S}}_Y^\top\tilde{\mathbf{S}}_Y)\Big\}. \tag{39}$$

The matrices here are structured, due to our choice of source law, leading to further simplifications. First, writing $\mathbf{M}_{\mathbf{A}} \coloneqq (\mathbf{A} \quad -\mathbf{I}_m)$, Lemma A.4 implies that

$$\mathbf{K}_\eta = \eta^2\tilde{\mathbf{S}}^\top(\eta\mathbf{I}_d - \mathbf{M}_{\mathbf{A}}^\top\mathbf{M}_{\mathbf{A}})^{-1}\tilde{\mathbf{S}}.$$

Next, note that the source random variables can be written as

$$(\tilde{X}_1, Y_1) = (cG_1, G_1), \qquad (\tilde{X}_i, Y_i) = (G_i, G_i + H_i), \quad i = 2, \dots, m,$$

with independent standard Gaussian latent variables G_i and H_i. The corresponding factors are

$$\mathbf{S}_1 \coloneqq \begin{pmatrix} c \\ 1 \end{pmatrix}, \qquad \mathbf{S}_b \coloneqq \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

with the subscript "b" denoting the "bulk" of the coordinates. For a scalar coefficient $a \in \mathbb{R}$ and a fixed block factor $\mathbf{S} \in \mathbb{R}^{2\times k}$, define

$$\mathbf{M}(a) \coloneqq (a \quad -1), \qquad \mathbf{K}(a, \mathbf{S}, \eta) \coloneqq \eta^2\mathbf{S}^\top(\eta\mathbf{I}_2 - \mathbf{M}(a)^\top\mathbf{M}(a))^{-1}\mathbf{S}, \quad \eta > 1 + a^2.$$

Write $k_1 \coloneqq \mathbf{K}(\alpha, \mathbf{S}_1, \eta)$, identifying the 1×1 matrix with its entry, and $\mathbf{K}_b \coloneqq \mathbf{K}(\beta, \mathbf{S}_b, \eta)$. In latent-coordinate order $G_1, (G_2, H_2), \dots, (G_m, H_m), H_1$, the matrices $\mathbf{K}_\eta$ and $\tilde{\mathbf{S}}_Y^\top\tilde{\mathbf{S}}_Y$ become

$$\mathrm{diag}(k_1, \underbrace{\mathbf{K}_b, \dots, \mathbf{K}_b}_{q \text{ times}}, 0), \qquad \mathrm{diag}(1, \underbrace{1_2 1_2^\top, \dots, 1_2 1_2^\top}_{q \text{ times}}, 0), \quad 1_2 = (1, 1)^\top.$$

The final zero comes from the unused latent coordinate H_1. We also compute

$$\|\tilde{\mathbf{S}}\|_{\mathrm{F}}^2 = c^2 + 1 + 3q, \qquad \|\tilde{\mathbf{S}}_X\|_{\mathrm{F}}^2 = c^2 + q.$$

Substituting these decompositions into (38) and (39), we arrive at

$$J_{\mathrm{DRO}}(\mathbf{A}) = \inf_{\eta>\gamma}\Big\{\eta(\rho^2 - c^2 - 1 - 3q) + k_1 + q\,\mathrm{tr}(\mathbf{K}_b)\Big\}, \tag{40}$$

$$J_{\mathrm{MRO}}(\mathbf{A}) = \inf_{\eta>\gamma}\Big\{\eta(\rho^2 - c^2 - q) + \mathrm{tr}_m\Big(\mathrm{diag}\big(k_1 - \eta, \underbrace{\mathbf{K}_b - \eta 1_2 1_2^\top, \dots, \mathbf{K}_b - \eta 1_2 1_2^\top}_{q \text{ copies}}\big)\Big)\Big\}. \tag{41}$$

Here, we removed the final zero from the matrix inside tr_m. This is permissible because this matrix coincides up to a permutation of rows and columns with $\mathbf{K}_\eta - \eta\tilde{\mathbf{S}}_Y^\top\tilde{\mathbf{S}}_Y$. By the definition of $\mathbf{K}_\eta$, that matrix is greater in Loewner order than $\eta\tilde{\mathbf{S}}_X^\top\tilde{\mathbf{S}}_X \succeq \mathbf{0}$ and is thus positive semidefinite. The remaining blocks have dimension $2m - 1 \geq m$, so removing the zero does not change the eigenvalue sum.

Step 2: Introduce epigraph variables and simplify. We now introduce epigraph variables $t_1 \geq k_1$ and $\mathbf{T}_b \succeq \mathbf{K}_b$. To make these lower bound constraints affine in $\mathbf{A}$ and η, we use that for $\eta > 1 + a^2$, a Schur complement (Lemma A.3) gives

$$\mathbf{T} \succeq \mathbf{K}(a, \mathbf{S}, \eta) \quad \Longleftrightarrow \quad \begin{pmatrix} \eta\mathbf{I}_2 - \mathbf{M}(a)^\top\mathbf{M}(a) & \eta\mathbf{S} \\ \eta\mathbf{S}^\top & \mathbf{T} \end{pmatrix} \succeq \mathbf{0}.$$

The upper-left block is quadratic in a. A second Schur complement replaces this quadratic term by an additional scalar block, yielding the affine constraint $\mathcal{Q}(a, \mathbf{S}, \eta, \mathbf{T}) \succeq \mathbf{0}$, where

$$\mathcal{Q}(a, \mathbf{S}, \eta, \mathbf{T}) \coloneqq \begin{pmatrix} \eta\mathbf{I}_2 & \mathbf{M}(a)^\top & \eta\mathbf{S} \\ \mathbf{M}(a) & 1 & \mathbf{0}_{1\times k} \\ \eta\mathbf{S}^\top & \mathbf{0}_{k\times 1} & \mathbf{T} \end{pmatrix} \in \mathbb{S}^{k+3}.$$

In particular, for $\eta > 1 + a^2$,

$$\mathcal{Q}(a, \mathbf{S}, \eta, \mathbf{T}) \succeq \mathbf{0} \quad \Longleftrightarrow \quad \mathbf{T} \succeq \mathbf{K}(a, \mathbf{S}, \eta).$$

For general $\eta \in \mathbb{R}$, its Schur complement at the scalar block 1 also shows that

$$\mathcal{Q}(a, \mathbf{S}, \eta, \mathbf{T}) \succeq \mathbf{0} \quad \Longrightarrow \quad \eta \geq 1 + a^2. \tag{42}$$

Define the two instances of this constraint by

$$\mathcal{Q}_1(\alpha, \eta, t_1) \coloneqq \mathcal{Q}(\alpha, \mathbf{S}_1, \eta, t_1) \in \mathbb{S}^4,$$
$$\mathcal{Q}_b(\beta, \eta, \mathbf{T}_b) \coloneqq \mathcal{Q}(\beta, \mathbf{S}_b, \eta, \mathbf{T}_b) \in \mathbb{S}^5.$$

Retaining the constraint that $\eta > \gamma$ in (40) and (41), we replace $k_1, \mathbf{K}_b$ by the epigraph variables $t_1, \mathbf{T}_b$ and minimize over them subject to $\mathcal{Q}_1(\alpha, \eta, t_1) \succeq \mathbf{0}$ and $\mathcal{Q}_b(\beta, \eta, \mathbf{T}_b) \succeq \mathbf{0}$. Both traces are nondecreasing in Loewner order, and equality is feasible at $t_1 = k_1$, $\mathbf{T}_b = \mathbf{K}_b$. Thus this substitution preserves the values of (40) and (41). For MRO, we additionally reparameterize using variables

$$z_1 \coloneqq t_1 - \eta, \qquad \mathbf{Z}_b \coloneqq \mathbf{T}_b - \eta 1_2 1_2^\top, \qquad \mathbf{Z} \coloneqq \operatorname{diag}(z_1, \underbrace{\mathbf{Z}_b, \ldots, \mathbf{Z}_b}_{q \text{ times}}).$$

By Lemma A.9, we have

$$\operatorname{tr}_m(\mathbf{Z}) = \min_{\substack{\tau \in \mathbb{R},\ y_1 \geq 0,\ y_1 \geq z_1 - \tau \\ \mathbf{Y}_b \in \mathbb{S}^2,\ \mathbf{Y}_b \succeq \mathbf{0},\ \mathbf{Y}_b \succeq \mathbf{Z}_b - \tau \mathbf{I}_2}} m\tau + y_1 + q \operatorname{tr}(\mathbf{Y}_b). \tag{43}$$

Substituting these representations into (40) and (41) and taking the infimum over (α, β) results in the programs below with the additional constraint that $\eta > \gamma$. We show next that this constraint can be dropped and that the resulting minima are attained.

Proposition I.1 (Constant-size SDP formulations)**.** *A population DRO optimizer is obtained by extracting coefficient variables (α, β) from any solution to*

$$\begin{aligned} \min_{\alpha, \beta, \eta, t_1, \mathbf{T}_b} \quad & t_1 + q \operatorname{tr}(\mathbf{T}_b) + \eta(\rho^2 - c^2 - 1 - 3q) \\ \text{s.t.} \quad & \mathcal{Q}_1(\alpha, \eta, t_1) \succeq \mathbf{0}, \\ & \mathcal{Q}_b(\beta, \eta, \mathbf{T}_b) \succeq \mathbf{0}, \\ & \eta \geq 0, \qquad \mathbf{T}_b \in \mathbb{S}^2. \end{aligned} \tag{44}$$

Likewise, a population MRO optimizer is obtained by taking (α, β) from any solution to

$$\begin{aligned} \min_{\alpha, \beta, \eta, t_1, \mathbf{T}_b, \tau, y_1, \mathbf{Y}_b} \quad & \eta(\rho^2 - c^2 - q) + m\tau + y_1 + q \operatorname{tr}(\mathbf{Y}_b) \\ \text{s.t.} \quad & \mathcal{Q}_1(\alpha, \eta, t_1) \succeq \mathbf{0}, \\ & \mathcal{Q}_b(\beta, \eta, \mathbf{T}_b) \succeq \mathbf{0}, \\ & \eta \geq 0, \\ & y_1 \geq 0, \qquad y_1 \geq z_1 - \tau, \\ & \mathbf{Y}_b \succeq \mathbf{0}, \qquad \mathbf{Y}_b \succeq \mathbf{Z}_b - \tau \mathbf{I}_2, \\ & \mathbf{T}_b, \mathbf{Y}_b \in \mathbb{S}^2. \end{aligned} \tag{45}$$

Both minima are attained, and the regret of the returned coefficient pairs at the target law P can be computed via (36)*.*

Proof. By (42), both SDPs impose $\eta \geq \gamma = \gamma(\alpha, \beta)$. The preceding argument establishes the desired optimal value identities with the additional restriction $\eta > \gamma$. It remains to remove this strict constraint and prove that minimizers are attained.

For fixed (α, β), let $\eta > \gamma$. By taking the remaining variables sufficiently large, we can obtain a feasible point with this strictly feasible value of η. Moreover, the feasible set is convex, and the objective is continuous. Thus, the presence of the strict constraint does not affect the infimum.

For attainment, fix an optimal coefficient pair (α, β), which exists by Section I.1. Choose a feasible objective value C for either SDP in the statement, and restrict to feasible points with objective at most C. This does not change the infimum. We show that the restricted feasible set is compact.

First, $\eta \geq \gamma \geq 1$. Extracting principal submatrices from the $\mathcal{Q}_1, \mathcal{Q}_b \succeq \mathbf{0}$ constraints gives

$$\begin{pmatrix} \eta \mathbf{I}_2 & \eta \mathbf{S}_1 \\ \eta \mathbf{S}_1^\top & t_1 \end{pmatrix} \succeq \mathbf{0}, \qquad \begin{pmatrix} \eta \mathbf{I}_2 & \eta \mathbf{S}_b \\ \eta \mathbf{S}_b^\top & \mathbf{T}_b \end{pmatrix} \succeq \mathbf{0}.$$

Since $\eta > 0$, Lemma A.3 implies that

$$t_1 \geq \eta \mathbf{S}_1^\top \mathbf{S}_1 = \eta(c^2 + 1), \qquad \mathbf{T}_b \succeq \eta \mathbf{S}_b^\top \mathbf{S}_b = \eta \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

For MRO, we therefore have

$$z_1 = t_1 - \eta \geq \eta c^2, \qquad \mathbf{Z}_b = \mathbf{T}_b - \eta 1_2 1_2^\top \succeq \eta \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

implying that the combined matrix $\mathbf{Z} = \operatorname{diag}(z_1, \mathbf{Z}_b, \dots \mathbf{Z}_b)$, with q copies of $\mathbf{Z}_b$, satisfies $\mathbf{Z} \succeq \mathbf{0}$ and $\operatorname{tr}_m(\mathbf{Z}) \geq \eta(c^2 + q)$. The SDP objective formulas and (43) now show that both objectives are at least $\eta\rho^2$. By our choice of C, the multiplier η must therefore lie in the bounded interval $[\gamma, C/\rho^2]$.

For DRO, boundedness of the objective and multiplier bounds $t_1 + q \operatorname{tr}(\mathbf{T}_b)$. Since $t_1 \geq 0$, $\mathbf{T}_b \succeq \mathbf{0}$, and $q \geq 1$, both t_1 and $\mathbf{T}_b$ must be bounded.

For MRO, boundedness of the objective and multiplier bounds $m\tau + y_1 + q \operatorname{tr}(\mathbf{Y}_b)$ from above. In the other direction, the constraints $y_1 \geq 0$ and $\mathbf{Y}_b \succeq \mathbf{0}$ imply that

$$m\tau + y_1 + q \operatorname{tr}(\mathbf{Y}_b) \geq m\tau.$$

Further, the SDP constraints for y_1 and $\mathbf{Y}_b$ and the early observation that $\mathbf{Z} \succeq \mathbf{0}$ imply that

$$\begin{aligned} m\tau + y_1 + q \operatorname{tr}(\mathbf{Y}_b) &\geq m\tau + z_1 - \tau + q \operatorname{tr}(\mathbf{Z}_b - \tau \mathbf{I}_2) \\ &\geq \operatorname{tr}(\mathbf{Z}) - (1 + 2q - m)\tau \\ &\geq -q\tau. \end{aligned}$$

Since $m, q > 0$, these bounds control τ in both directions. They then bound y_1 and $\operatorname{tr}(\mathbf{Y}_b)$ separately, since $y_1 \geq 0$ and $\mathbf{Y}_b \succeq \mathbf{0}$. Moreover,

$$0 \leq z_1 \leq y_1 + \tau, \qquad \mathbf{0} \preceq \mathbf{Z}_b \preceq \mathbf{Y}_b + \tau \mathbf{I}_2,$$

so the shifted blocks z_1 and $\mathbf{Z}_b$ are bounded. The identities $t_1 = z_1 + \eta$ and $\mathbf{T}_b = \mathbf{Z}_b + \eta 1_2 1_2^\top$ then bound the remaining variables t_1 and $\mathbf{T}_b$.

Thus, for either SDP, the restricted feasible set is bounded. It is also nonempty and closed, because the constraints are closed and the objective is continuous. It is thus compact, and so the continuous objective attains its minimum on this set. □

I.3 Closed form for SNR

For SNR, we obtain an explicit solution and closed form for the regret under P.

Lemma I.2 (Closed-form SNR solution)**.** *Let $\lambda := \rho^2$. Then the unique minimizer of $\mathbf{A} \mapsto L(\mathbf{A}, \tilde{P}) + \lambda \|\mathbf{A}\|_{\mathrm{op}}^2$ is $\mathbf{A}(\alpha_\lambda, \beta_\lambda)$, where*

$$(\alpha_\lambda, \beta_\lambda) = \begin{cases} \left(\dfrac{c}{c^2 + \lambda}, 1 \right), & 0 \leq \lambda \leq c(1 - c), \\ (t_\lambda, t_\lambda), \quad t_\lambda := \dfrac{q + c}{q + c^2 + \lambda}, & \lambda \geq c(1 - c). \end{cases} \tag{46}$$

Consequently, its exact clean regret is

$$R(\mathsf{SNR}_\rho[\tilde{P}], P) = \begin{cases} \left(\dfrac{c}{c^2 + \lambda} - 1 \right)^2, & 0 \leq \lambda \leq c(1 - c), \\ m \left(\dfrac{q + c}{q + c^2 + \lambda} - 1 \right)^2, & \lambda \geq c(1 - c). \end{cases} \tag{47}$$

The two expressions agree at $\lambda = c(1 - c)$.

Proof. By our discussion in Section I.1, it suffices to minimize over $\mathbf{A}(\alpha, \beta)$. Dropping the additive constant q from (37), the reduced objective is

$$F_\lambda(\alpha, \beta) = (c\alpha - 1)^2 + q(\beta - 1)^2 + \lambda \max\{\alpha^2, \beta^2\}. \tag{48}$$

Recall that $c = m^{-1/4} \in (0, 1]$. Because $\mathbf{\Sigma}_X^{\tilde{P}} = \operatorname{diag}(c^2, \mathbf{I}_q) \succ 0$, the risk is strongly convex in the full matrix $\mathbf{A}$. Hence the regularized objective has a unique minimizer.

Introduce an epigraph variable $t \geq 0$ with constraints $|\alpha| \leq t$ and $|\beta| \leq t$. For fixed t, the quadratic part of (48) is minimized by projecting its unconstrained minimizers $\alpha = c^{-1}$ and $\beta = 1$ onto $[-t, t]$. Since both are positive, the constrained minimizers are

$$\alpha(t) \coloneqq \min\{t, c^{-1}\}, \qquad \beta(t) \coloneqq \min\{t, 1\}.$$

Thus the epigraph problem reduces to minimizing

$$\phi_\lambda(t) \coloneqq \lambda t^2 + (1 - ct)_+^2 + q(1 - t)_+^2, \qquad t \geq 0.$$

Here $\lambda = \rho^2 > 0$ and $0 < c < 1$, so ϕ_λ is strictly convex. For $t > c^{-1}$ it equals λt^2 and increases, so its minimizer lies in $[0, c^{-1}]$. Moreover,

$$\frac{1}{2}\phi_\lambda'(t) = \lambda t - c(1 - ct)_+ - q(1 - t)_+, \qquad \frac{1}{2}\phi_\lambda'(1) = \lambda - c(1 - c).$$

If $0 < \lambda \leq c(1 - c)$, the minimizer lies in $[1, c^{-1}]$, where the derivative vanishes at

$$t = \frac{c}{c^2 + \lambda}.$$

The corresponding coefficients are $\alpha = t$ and $\beta = 1$. If $\lambda \geq c(1 - c)$, the minimizer lies in $[0, 1]$, where the derivative vanishes at

$$t = \frac{q + c}{q + c^2 + \lambda}.$$

In this case $\alpha = \beta = t$. At $\lambda = c(1 - c)$, both expressions give $t = 1$. This proves (46). Substitution into (36) proves (47). □

J Full Experimental Details and Additional Experiments

The synthetic experiments for Fig. 1 in Section 6 and Section J.1 below were run in 8 hours on an `m8g.large` Amazon EC2 instance with 2 vCPUs and 8 GB of RAM, while the experiment for Section J.2 was run in less than one hour on an M1 MacBook Air with an 8-core CPU and 16 GB RAM. Code for the experiments and plots is available at `https://github.com/sbnietert/regret-optimal-regression`.

J.1 Additional Synthetic Experiment with Varied Residual Rank

For **Setting C**, we modify Setting A by fixing $p = m = 200$ ($d = 400$), $n = 10m^2$, and $\rho = \varepsilon = 1$, and varying the residual effective rank over 10 equally spaced values $r \in [1, m - 1]$. Specifically, for each r, the target law P_r is the law of (X, Y), where

$$X \sim \mathcal{N}(0, \mathbf{I}_m), \qquad Y = X + U_r, \qquad U_r \sim \mathcal{N}\left(0, \operatorname{diag}\left(0, 1, \frac{r - 1}{m - 2}\mathbf{I}_{m-2}\right)\right),$$

and X and U_r are independent. Then $\mathbf{A}_{P_r} = \mathbf{\Sigma}_X^{P_r} = \mathbf{I}_m$, and the residual second moment satisfies

$$\operatorname{tr}(\mathbf{\Sigma}_E^{P_r}) = r, \qquad \|\mathbf{\Sigma}_E^{P_r}\|_{\mathrm{op}} = 1, \qquad r_{\mathrm{eff}}(\mathbf{\Sigma}_E^{P_r}) = r.$$

In particular, each $\kappa_{P_r} = 3$. The endpoint $r = m - 1$ recovers the target law of Setting A at $m = 200$.

For each r, we draw n i.i.d. target observations and apply the same corruption as in Setting A. We fit each estimator $\hat{\mathbf{A}} \in \{\mathsf{ERM}, \mathsf{DRO}_\rho, \mathsf{MRO}_\rho, \mathsf{SNR}_\rho\}$ to the resulting empirical source law $P_{\mathrm{s},n}$, using the implementations described in Section 6. In Fig. 2, we plot the target population regret $R(\hat{\mathbf{A}}[P_{\mathrm{s},n}], P_r) = \|\hat{\mathbf{A}}[P_{\mathrm{s},n}] - \mathbf{I}_m\|_{\mathrm{F}}^2$ against r. Curves report averages over 10 independent trials at each value of r, with error bars showing the minimum and maximum over trials (again imperceptible due to low variance).

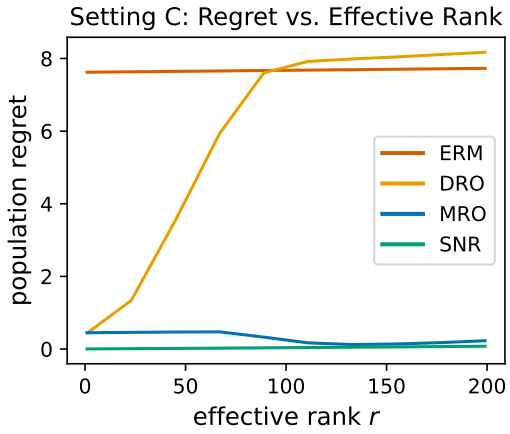


Figure 2: Target population regret $R(\hat{\mathbf{A}}[P_{\mathrm{s},n}], P_r)$ for Setting C, as a function of the target residual effective rank $r = r_{\mathrm{eff}}(\mathbf{\Sigma}_E^{P_r})$.

J.2 Network-Traffic Experiment with CESNET Dataset

Finally, we test our estimators on a real-world network-traffic dataset, albeit with synthetically introduced Wasserstein perturbations. We employ the CESNET-TimeSeries24 dataset (Koumar et al., 2025), which contains network-traffic time series data from the CESNET3 network connecting public and research institutions in the Czech Republic. Specifically, it records hourly, institution-level counts of packets and bytes transmitted over the network between October 2023 and July 2024, and we attempt to predict byte totals from packet counts. We use 268 institutions with nonconstant counts and finite, nonnegative measurements for at least 99% of the 6,717 observed hourly timestamps, and intersect their valid timestamps, obtaining $n = 6{,}228$ paired observations. For each response dimension $m \in \{8, 16, 24, 32, 48, 64, 96, 128, 192, 256, 268\}$, we take our target response vector to be the first m byte-total channels and its corresponding feature vector to be the first $q = m/2$ matching packet-count channels plus an appended intercept term of 1, so $p = q + 1$. We center and normalize each physical channel according to its mean and standard deviation under the full empirical law of the raw observations, keeping these transformations and the observation set fixed throughout. Denote the resulting standardized target law by $P_m \in \mathcal{P}_2(\mathbb{R}^{p+m})$, treated as a population measure to isolate the complexity of robust regression from that of handling sampling error.

This setup was selected so that some of the feature and response coordinates are highly correlated, while others are nearly independent, yielding a range of effective ranks. Coordinate-wise normalization is a relatively common preprocessing procedure, and it helps to keep the condition number of the target law from growing too quickly with m. In particular, the effective rank of $\mathbf{\Sigma}_E^{P_m}$ grows roughly linearly with m, while the target condition number κ_{P_m} grows more slowly, as shown in Fig. 4.

To obtain the corresponding corrupted source law $\tilde{P}_m$, we apply the frozen transformation above to the same n observations, but with the first packet-count channel first multiplied by a gain of 0.5 (all other channels are unchanged). By coupling observations according to their timestamps, one can directly upper bound $\mathsf{W}_2(P_m, \tilde{P}_m) \le \varepsilon := 1.08679$, independent of m. We then fit each estimator $\hat{\mathbf{A}} \in \{\mathsf{ERM}, \mathsf{DRO}_\varepsilon, \mathsf{MRO}_\varepsilon, \mathsf{SNR}_\varepsilon\}$ to $\tilde{P}_m$, using the implementations described in Section 6. In Fig. 3, we plot the target regret $R(\hat{\mathbf{A}}[\tilde{P}_m], P_m)$ against m. We observe qualitative similarities to the separation observed in our synthetic experiments, with MRO and SNR outperforming DRO and ERM as the dimension grows (although the gap is less pronounced). We emphasize that ample room remains for more refined empirical study, e.g., to identify the most realistic modes of data corruption for specific applications and to compare the performance of these estimators beyond the worst case.

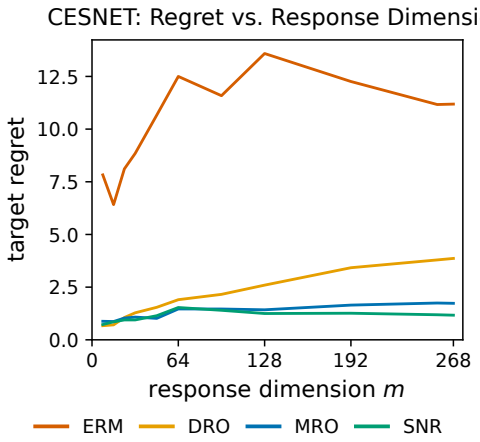


Figure 3: Target regret $R(\hat{\mathbf{A}}[\tilde{P}_m], P_m)$ for our CESNET experiment, as a function of the response dimension m.

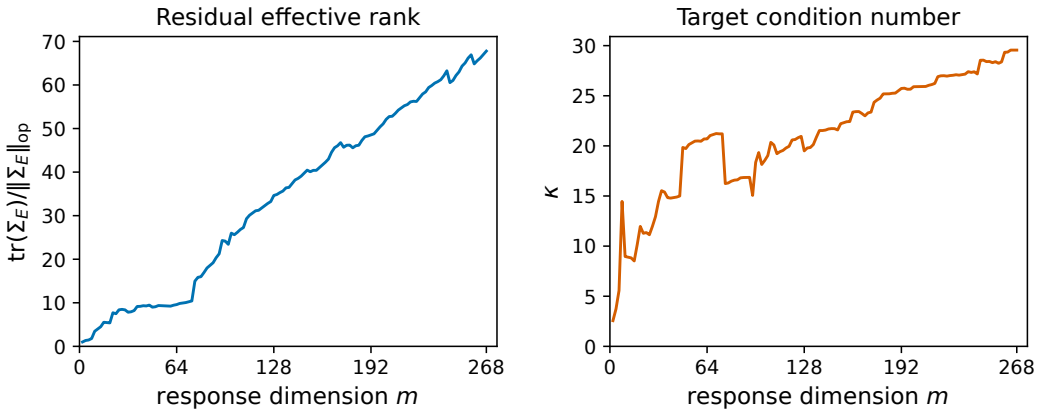


Figure 4: Effective rank $\mathrm{tr}(\mathbf{\Sigma}_E^{P_m})/\|\mathbf{\Sigma}_E^{P_m}\|_{\mathrm{op}}$ and condition number κ_{P_m} of the target law P_m in our CESNET experiment, as a function of the response dimension m.