The radius of statistical efficiency

Joshua Cutler* Mateo Díaz† Dmitriy Drusvyatskiy‡

Abstract

Classical results in asymptotic statistics show that the Fisher information matrix controls the difficulty of estimating a statistical model from observed data. In this work, we introduce a companion measure of robustness of an estimation problem: the radius of statistical efficiency (RSE) is the size of the smallest perturbation to the problem data that renders the Fisher information matrix singular. We compute RSE up to numerical constants for a variety of test bed problems, including principal component analysis, generalized linear models, phase retrieval, bilinear sensing, and matrix completion. In all cases, the RSE quantifies the compatibility between the covariance of the population data and the latent model parameter. Interestingly, we observe a precise reciprocal relationship between RSE and the intrinsic complexity/sensitivity of the problem instance, paralleling the classical Eckart–Young theorem in numerical analysis.

1 Introduction

A central theme in computational mathematics is that the numerical difficulty of solving a given problem is closely linked to both (*i*) the sensitivity of its solution to perturbations and (*ii*) the shortest distance of the problem to an ill-posed instance. As a rudimentary example, consider solving an $m \times d$ linear system $Ax = b$. The celebrated Eckart–Young theorem asserts the equality:

$$\underbrace{\min_{B \in \mathbf{R}^{m \times d}} \{\|A - B\|_F \mid B \text{ is singular}\}}_{\text{distance to ill-posedness}} \quad = \quad \underbrace{\sigma_{\min}(A)}_{\text{difficulty/sensitivity}} . \tag{1}$$

Although the proof is elementary, the conclusion is intriguing since it equates two conceptually distinct quantities. Namely, the reciprocal of the minimal singular value $1/\sigma_{\min}(A)$ is classically known to control both the numerical difficulty of solving the linear system $Ax = b$ and the Lipschitz stability of the solution to perturbations in b. In contrast, the left side of the equation (1) is geometric; it measures the smallest perturbation to the data that renders the problem ill-posed.

The *exact* equality in (1) is somewhat misleading because it is specific to linear systems. We would expect that for more sophisticated problems, the difficulty/sensitivity of the problem should be inversely proportional to the distance to ill-posedness. This is indeed the case for a wide class of problems in numerical analysis [13, 29] and optimization [31, 36, 49, 55, 56], including computing eigenvalues and eigenvectors, finding zeros of polynomials, pole assignment in control systems, conic optimization, nonlinear programming, and variational inequalities. Despite this impressive body of work, this line of research is largely unexplored in statistical contexts. Therefore, here we ask:

*Department of Mathematics, U. Washington, Seattle, WA 98195; jocutler@uw.edu.

†Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA; https://mateodd25.github.io.

‡Department of Mathematics, U. Washington, Seattle, WA 98195; www.math.washington.edu/∼ddrusv. Research of Drusvyatskiy was supported by the NSF DMS-2306322, NSF CCF 1740551, and AFOSR FA9550-24-1-0092 awards.

Is there a succinct relationship between complexity, sensitivity, and distance to ill-posedness for problems in statistical inference and learning?

We will see that in a certain precise sense the answer is indeed yes for a wide class of problems. The starting point for our development is that the statistical difficulty of estimation is tightly controlled by (quantities akin to) the Fisher information matrix for maximum likelihood estimation. This connection is made precise for example by the Cramér-Rao lower bound [19, 53] and the local asymptotic minimax theory of Hájek and Le Cam [35, 39, 65]. From an optimization viewpoint, the minimal eigenvalue of the Fisher information matrix is closely related to the quadratic growth constant of the objective function modeling the learning problem at hand. In particular, (near-) singularity of this matrix signifies that the problem is ill-conditioned. Inspired by this observation, we introduce a new measure of robustness associated to an estimation task: the *radius of statistical efficiency (RSE)* is the size of the smallest perturbation to the problem data that renders the Fisher information matrix singular. Thus large RSE signifies existence of a large neighborhood around the problem instance comprised only of well-posed problems.

We compute RSE for a variety of test bed problems, including principal component analysis (PCA), generalized linear models, phase retrieval, bilinear sensing, and rank-one matrix completion. In all cases, the RSE exhibits a precise reciprocal relationship with the statistical difficulty of solving the target problem, thereby directly paralleling the Eckart–Young theorem and its numerous extensions in numerical analysis and optimization. Moreover, we provide gradient-based conditions for general estimation problems which ensure validity of such a reciprocal relationship. Slope-based criteria for error bounds, due to Ioffe [37] and Azé-Corvellec [2], play a key role in this development.

Before delving into the technical details, we illustrate the main thread of our work with two examples—linear regression and PCA—where the conclusions are appealingly simple to state.

Linear regression. The problem of linear regression is to recover a vector $\beta_\star \in \mathbf{R}^d$ from noisy linear measurements

$$y = \langle x, \beta_\star \rangle + \varepsilon,$$

where $x \in \mathbf{R}^d$ is drawn from a probability distribution D and ε is zero-mean noise vector that is independent of x. The standard approach to this task is to minimize the mean-squared error

$$\min_{\beta \in \mathbf{R}^d} \ \tfrac{1}{2}\, \mathbb{E}_{x,y} \left(\langle x, \beta \rangle - y\right)^2. \tag{2}$$

Classical results show that the asymptotic performance of estimators for this problem is tightly controlled by Σ^{-1} where $\Sigma := \mathbb{E}_{x \sim \mathsf{D}}\, xx^\top$ is the second moment matrix of the population. The closer the matrix Σ is to being singular, the more challenging the problem (2) is to solve, requiring a higher number of samples. Seeking to estimate a neighborhood of well-posed problems around D, the RSE is defined to be the minimal Wasserstein-2 distance from D to distributions with a singular second moment matrix. We will see that for linear regression (2), RSE is simple to compute:

$$\boxed{\mathrm{RSE}(\mathsf{D}) = \sqrt{\lambda_{\min}(\Sigma)}.} \tag{3}$$

That is, the simplest measure of ill-conditioning of the target problem $1/\sqrt{\lambda_{\min}(\Sigma)}$ has a geometric interpretation as the reciprocal of the distance to the nearest ill-posed problem. The representation of RSE in (3) is not specific to linear regression and holds much more generally for (quasi) maximum likelihood estimation [46] with strongly convex cumulant functions.

Principal component analysis (PCA). Principal component analysis (PCA) seeks to find a q-dimensional subspace $\mathcal{V} \subset \mathbf{R}^d$ that captures most of the variance of a centered random vector x drawn from a probability distribution D. Analytically, this amounts to solving the problem

$$\max_{R \in \mathrm{Gr}(q,d)} \mathop{\mathbb{E}}_{x \sim \mathsf{D}} \|Rx\|_2^2,$$

where the Grassmannian manifold $\mathrm{Gr}(q,d)$ consists of all orthogonal projection matrices $R \in \mathbf{R}^{d \times d}$ onto q-dimensional subspaces of $\mathbf{R}^d$. The column space of the optimal matrix R is called the *top q principal subspace*. Intuitively, the hardness of the problem is governed by the gap $\lambda_q - \lambda_{q+1}$ between the q'th and $(q+1)$'th eigenvalues of $\Sigma = \mathbb{E}_{\mathsf{D}}\, xx^\top$: the smaller the gap is, the higher the number of samples required for estimation, and this can indeed be made rigorous. Again, RSE by definition is the minimal Wasserstein-2 distance from D to a distribution with covariance having equal q'th and $(q+1)$'st eigenvalues. We will show that RSE admits the simple form

$$\boxed{\mathrm{RSE}(\mathsf{D}) = \tfrac{1}{\sqrt{2}}\left(\sqrt{\lambda_q(\Sigma)} - \sqrt{\lambda_{q+1}(\Sigma)}\right).} \tag{4}$$

In particular, the expression (4) endows the gap $\sqrt{\lambda_q(\Sigma)} - \sqrt{\lambda_{q+1}(\Sigma)}$—the reciprocal of the problem's complexity—with a geometric meaning as the distance to a nearest ill-conditioned problem instance.

The two examples of linear regression and PCA can be understood within the broader context of stochastic optimization:

$$\min_{\beta \in \mathcal{M}} f(\beta) \quad \text{where} \quad f(\beta) = \mathop{\mathbb{E}}_{z \sim \mathsf{D}} \ell(\beta; z).$$

Here z is data drawn from a distribution D, the function $\ell(\beta; \cdot)$ is a loss parameterized by β, and $\mathcal{M} \subset \mathbf{R}^d$ is a smooth manifold of allowable model parameters. For example, $\mathtt{SP}(\mathsf{D})$ may correspond to least-squares regression or maximum likelihood estimation. In both cases, the asymptotic efficiency of estimators for finding the minimizer $\beta_\star$ of $\mathtt{SP}(\mathsf{D})$ is tightly controlled by the following matrix akin to Fisher information:

$$\mathcal{I}(\beta_\star, \mathsf{D}) = P_{\mathcal{T}} \nabla^2_{\beta\beta} \mathcal{L}(\beta_\star, z, \lambda_\star) P_{\mathcal{T}}.$$

Here, $P_{\mathcal{T}}$ is the projection onto the tangent space of $\mathcal{M}$ at $\beta_\star$ and $\mathcal{L}(\beta_\star, z, \lambda_\star)$ is the Lagrangian function for $\mathtt{SP}(\mathsf{D})$ with optimal multiplier $\lambda_\star$. For simplicity, we will abuse notation and call $\mathcal{I}(\beta_\star, \mathsf{D})$ the Fisher information matrix. The matrix $\mathcal{I}(\beta_\star, \mathsf{D})$ plays a central role in estimation, as highlighted by the lower bounds of Cramér-Rao and Hájek-Le Cam [39, 65], as well as their recent extensions to stochastic optimization of Duchi-Ruan [35]. Moreover, the minimal nonzero eigenvalue of $\mathcal{I}(\beta_\star, \mathsf{D})$ controls both the coefficient of quadratic growth of the objective function in $\mathtt{SP}(\mathsf{D})$ and the Lipschitz stability of the solution under linear perturbations. In particular, the problem $\mathtt{SP}(\mathsf{D})$ becomes ill-conditioned when the minimal eigenvalue of $\mathcal{I}(\beta_\star, \mathsf{D})$ is small.

In summary, the matrix $\mathcal{I}(\beta_\star, \mathsf{D})$ tightly controls the difficulty of solving $\mathtt{SP}(\mathsf{D})$. Consequently, it is appealing to consider as a measure of robustness of $\mathtt{SP}(\mathsf{D})$ the size of the smallest perturbation to the data D, say in the Wasserstein-2 distance $W_2(\cdot,\cdot)$, that renders the Fisher information matrix singular. This is the viewpoint we explore in the current work, and call this quantity the *radius of statistical efficiency (RSE)*. We choose to use the Wasserstein-2 distance in the definition of RSE, as opposed to other metrics on measures, because it leads to concise and easily interpretable estimates in examples. In the rest of the paper, we study basic properties of RSE and compute it up to numerical constants for a variety of test bed problems: generalized linear models, PCA, phase retrieval, blind deconvolution, and matrix completion. In all cases, the RSE translates intuitive measures of "well-posedness" into quantified neighborhoods of stable problems. Moreover, in all cases we show a reciprocal relationship between the minimal eigenvalue of $\mathcal{I}(\beta_\star, \mathsf{D})$ and RSE, thereby paralleling the Eckart–Young theorem in numerical analysis and optimization.

Outline The remainder of this section covers related work and basic notation we use. Section 2 formally describes the radius of statistical efficiency and establishes a few general-purpose results relating RSE to the minimal eigenvalue of the Fisher information matrix. The subsequent sections characterize RSE for several problems: PCA (Section 3), generalized linear models and (quasi) maximum likelihood estimation (Section 4), rank-one matrix regression (Section 5). Section 6 closes the paper with conclusions. All proofs appear in the appendix in order to streamline the reading.

1.1 Related work

Our work is closely related to a number of topics in statistics and computational mathematics.

Local minimax lower bounds in estimation. There is a rich literature on minimax lower bounds in statistical estimation problems; we refer the reader to [68, Chapter 15] for a detailed treatment. Typical results of this type lower-bound the performance of any statistical procedure on a worst-case instance for that procedure. Minimax lower bounds can be quite loose as they do not consider the complexity of the particular problem that one is trying to solve but rather that of an entire problem class to which it belongs. More precise local minimax lower bounds, as developed by Hájek and Le Cam [39, 65], provide much finer problem-specific guarantees. Simply put, a single object akin to the Fisher information matrix controls both the difficulty of estimation from finitely many samples and the stability of the model parameters to perturbation of the density. Extensions of this theory to stochastic nonlinear programming were developed by Duchi and Ruan [35] and extended to decision-dependent problems in [21] and to a wider class of (partly smooth) problems in [26]. In particular, it is known that popular algorithms such as sample average approximation [65] and stochastic gradient descent with iterate averaging [26, 51] match asymptotic local lower-bound, and are therefore asymptotically optimal. Weaker ad hoc results, based on the Cramér-Rao lower bound, have been established for a handful of problems [3, 4, 44, 48, 62].

Radius theorems. Classical numerical analysis literature emphasizes the close interplay between efficiency of numerical algorithms and their sensitivity to perturbation. Namely, problems with solutions that change rapidly due to small perturbation are typically difficult to solve. Examples of this phenomenon abound in computational mathematics; e.g. eigenvalue problems and polynomial equations [13, 29] and optimization [31, 32, 57]. Motivated by this observation, Demmel in [28] introduced a new robustness measure, called the radius of regularity, which measures the size of a neighborhood around a problem instance within which all other problem instances are stable. A larger neighborhood thereby signifies a more robust problem instance. Estimates on the radius of regularity have now been computed for a wealth of computational problems; e.g. solving polynomial systems [12, 13], linear and conic programming [36, 49, 55, 56], and nonlinear optimization [31]. The radius of statistical efficiency, introduced here, serves as a direct analogue for statistical estimation.

Conditioning and radius theorems in recovery problems. Several condition numbers—controlling the convergence of first-order methods—are closely related to notions of strong identifiability, e.g., the restricted isometry property (RIP), in the context of statistical recovery problems [8, 14, 15, 16, 30, 45]. A few works [5, 58, 71] have established connections between these notions of strong identifiability and a suitably-defined radius to ill-posed instances. In particular, [58] established a connection between Renegar's conic distance to infeasibility [56] and the null space property [18] in compressed sensing. In a similar vein, [71] linked the ℓ_1-distance to ill-posed problems and the RIP for generalized rank-one matrix completion. Finally, [5] defined a condition number for the LASSO variable selection problem via the reciprocal of the distance to ill-posedness, designed an algorithm whose complexity depends solely on this condition number, and proved an impossibility result for instances with infinite condition number. The radius of statistical efficiency, defined in this work, is distinct from and complementary to these metrics.

Error bounds. The basic question we explore is the relationship between the minimal eigenvalue of the Fisher information matrix (complexity) and the distance to the set where this eigenvalue is zero (RSE). The theory of error bounds exactly addresses questions of this type; namely when does the function's value polynomially bound the distance to the set of minimizers. See for example the authoritative monographs on the subject, [20, Chapter 8] and [38, Chapter 3]. Indeed, Demmel's original work [28] makes heavy use of this interpretation. We explore this path here as well when developing infinitesimal characterizations of RSE in Theorem 2.2. The added complication is that the functions we deal with are defined over a complete metric space, and therefore the techniques we use rely on variational principles (á la Ekeland) and computations of the slope.

1.2 Notation

Linear algebra. Throughout, we let $\mathbf{R}^d$ denote the standard d-dimensional Euclidean space with dot product $\langle x, y\rangle = x^\top y$ and norm $\|x\|_2 = \sqrt{\langle x, x\rangle}$. The unit sphere in $\mathbf{R}^d$ will be denoted by $\mathbb{S}^{d-1}$, while the set of nonnegative vectors will be written as $\mathbf{R}^d_+$. The symbol $\mathbf{R}^{m\times n}$ will denote the Euclidean space of real $m \times n$ matrices, endowed with the trace inner product $\langle X, Y\rangle = \mathrm{tr}(X^\top Y)$. The symbol $\otimes$ denotes Kronecker product. The Frobenius and operator norms will be written as $\|\cdot\|_F$ and $\|\cdot\|_{\mathrm{op}}$, respectively. The singular values of a matrix $A \in \mathbf{R}^{m\times n}$ will be arranged in nonincreasing order:

$$\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_{m\wedge n}(A).$$

The space of real symmetric $d \times d$ matrices is denoted by $\boldsymbol{S}^d$ and is equipped with the trace product as well. The cone of $d \times d$ positive semidefinite matrices will be written as $\boldsymbol{S}^d_+$. The eigenvalues of a matrix $A \in \boldsymbol{S}^d$ will be arranged in nonincreasing order:

$$\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_d(A).$$

For any subspace $\mathcal{K} \subset \mathbf{R}^d$, the symbol $\mathsf{P}_{\mathcal{K}} \colon \mathbf{R}^d \to \mathcal{K}$ will denote the orthogonal projection onto $\mathcal{K}$. The compression of any matrix $A \in \mathbf{R}^{d\times d}$ to $\mathcal{K}$, denoted $A|_{\mathcal{K}} \colon \mathcal{K} \to \mathcal{K}$, is the map $A|_{\mathcal{K}} := \mathsf{P}_{\mathcal{K}} A \mathsf{P}_{\mathcal{K}}$.

Probability theory. We will require some background on the Wasserstein geometry on the space of probability measures on $\mathbf{R}^d$. In order to streamline the reading, we record here only the most essential notation that we will need. A detailed review of Wasserstein geometry appears in Section A. To this end, we let $\mathcal{P}_p(\mathbf{R}^d)$ be the space of measures μ on $\mathbf{R}^d$ with a finite p'th moment $\mathbb{E}_\mu \|x\|_p^p < \infty$. The subset of measures of $\mathcal{P}_p(\mathbf{R}^d)$ that are centered, meaning $\mathbb{E}_\mu[x] = 0$, will be written as $\mathcal{P}_p^\circ(\mathbf{R}^d)$. When the space $\mathbf{R}^d$ is clear from context, we use the shorthand $\mathcal{P}_p$ and $\mathcal{P}_p^\circ$. A convenient metric on $\mathcal{P}_p$ is furnished by the Wasserstein distance $W_p(\mu, \nu)$; see Section A for details. The distance function to a set of measures $Q \subset \mathcal{P}_p$ is defined by $W_p(\mu, Q) = \inf_{\nu\in Q} W_p(\mu, \nu)$. The symbol $\Sigma_\mu = \mathbb{E}_\mu xx^\top$ will denote the second moment matrix of any measure $\mu \in \mathcal{P}_2$. In the rest of the paper, we will use the symbol D to denote a distinguished measure associated with the problem of interest, while we use μ as a placeholder for arbitrary measures.

2 The distance to ill-conditioned problems

In this section, we formally define the radius of statistical efficiency (RSE), and develop some techniques for computing it. Throughout, we will focus on the stochastic optimization problem

$$\min_{\beta\in\mathcal{M}} f(\beta) \quad \text{where} \quad f(\beta) = \mathop{\mathbb{E}}_{z\sim\mathsf{D}} \ell(\beta; z).$$

Here, the set $\mathcal{M} \subset \mathbf{R}^d$ is a C^2 manifold and z is drawn from a distribution $\mathsf{D} \in \mathcal{P}_2(\mathcal{Z})$, where $\mathcal{Z}$ is a finite-dimensional Euclidean space equipped with its Borel σ-algebra. We assume that the function $\ell(\beta; z)$ is measurable and twice differentiable in β for every z and that f is C^2-smooth. We also make the blanket assumption that the gradient and Hessian of $\ell(\cdot; z)$ are D-integrable.

The difficulty of solving the problem SP(D) from finitely many samples $z_1, z_2, \ldots, z_n \overset{\text{iid}}{\sim} \mathsf{D}$ is tightly controlled by a matrix akin to Fisher Information. This object, which we now describe, plays a central role in our work. Let $\beta_\star$ be a minimizer of SP(D) and define the solution map:

$$\sigma(v) = \operatorname*{argmin}_{\beta\in\mathcal{M}\cap B_\varepsilon(\beta_\star)} f(\beta) - \langle v, \beta\rangle.$$

Thus, the set $\sigma(v)$ is comprised of all solutions to a problem obtained from SP(D) by subtracting a linear/tilt perturbation $\langle v, \beta\rangle$. Clearly, a desirable property is for σ to be single-valued and smooth. With this in mind, we introduce the following notion due to Poliquin and Rockafellar [50].

Definition: (Tilt-stable minimizer)

The point $\beta_\star$ is a *tilt-stable minimizer* of SP(D) if the map $\sigma(\cdot)$ satisfies $\sigma(0) = \beta_\star$ and is single-valued and C^1-smooth on some neighborhood of the origin. Then the *regularity modulus* of the problem is defined to be

$$\mathrm{REG}(\mathsf{D}) = \|\nabla\sigma(0)\|_{\mathrm{op}}.$$

If $\beta_\star$ is not tilt-stable, we call $\beta_\star$ *unstable* and set $\mathrm{REG}(\mathsf{D}) = +\infty$.

In particular, we will regard $\mathrm{REG}(\mathsf{D})$ as the measure of difficulty of solving the problem SP(D). When $\mathcal{M}$ is the whole space, $\beta_\star$ is a tilt-stable minimizer if and only if the Hessian $\nabla^2 f(\beta_\star)$ is nonsingular, in which case equality $\nabla\sigma(0) = [\nabla^2 f(\beta_\star)]^{-1}$ holds [50, Proposition 1.2]. In particular, when SP(D) corresponds to maximum likelihood estimation, the matrix $\nabla\sigma(0)$ reduces to the inverse of the Fisher information. More generally, tilt-stability can be characterized either in terms of definiteness of the covariant Hessian or the Hessian of the Lagrangian on the tangent space to $\mathcal{M}$. Since we will use both of these viewpoints, we review them now. The reader may safely skip this discussion during the first reading since it will not be used until the appendix.

Lagrangian characterization. Let $G = 0$ be the local defining equations for $\mathcal{M}$ around $\beta_\star$. That is $G\colon \mathbf{R}^d \to \mathbf{R}^m$ is a C^2-smooth map with surjective Jacobian $\nabla G(\beta_\star)$ and such that the two sets $\mathcal{M}$ and $\{\beta : G(\beta) = 0\}$ coincide near $\beta_\star$. Then the tangent space to $\mathcal{M}$ at $\beta_\star$ is $\mathcal{T} = \mathrm{Null}(\nabla G(\beta_\star))$. Define the Lagrangian function

$$\mathcal{L}(\beta, \lambda) := f(\beta) + \langle \lambda, G(\beta)\rangle.$$

First order optimality conditions at $\beta_\star$ ensure that there exists a unique vector $\lambda_\star \in \mathbf{R}^m$ satisfying $\nabla_\beta \mathcal{L}(\beta_\star, \lambda_\star) = 0$. Define the matrix

$$\mathcal{I}(\beta_\star, \mathsf{D}) := P_{\mathcal{T}} \cdot \nabla^2_{\beta\beta}\mathcal{L}(\beta_\star, \lambda_\star) \cdot P_{\mathcal{T}}, \tag{5}$$

where $P_{\mathcal{T}}$ is the orthogonal projection onto the tangent space $\mathcal{T}$. If $\beta_\star$ is a local minimizer of the problem, then $\mathcal{I}(\beta_\star, \mathsf{D})$ is positive semidefinite. Conversely:

$$\mathcal{I}(\beta_\star, \mathsf{D}) \text{ is positive definite on } \mathcal{T} \text{ if and only if } \beta_\star \text{ is a tilt-stable minimizer.}$$

Moreover, in this case equality $\nabla\sigma(0) = \mathcal{I}(\beta_\star, \mathsf{D})^\dagger$ holds, where † denotes the Moore–Penrose inverse. In particular, one may regard $\nabla\sigma(0)$ as akin to the inverse of the Fisher information matrix for MLE. Note that when $\beta_\star$ is tilt-stable, the reciprocal of the regularity modulus $1/\mathrm{REG}(\mathsf{D})$ coincides the minimal nonzero eigenvalue of $\mathcal{I}(\beta_\star, \mathsf{D})$.

Intrinsic characterization. Often, the defining equations of the manifold are either unknown or difficult to work with. In this case, tilt-stability can be characterized through second-order expansions along curves. Namely, for any tangent vector $u \in \mathcal{T}$, there exists a C^2-smooth curve $\gamma_u \colon (-\epsilon, \epsilon) \to \mathcal{M}$ for some $\epsilon > 0$ satisfying $\gamma_u(0) = \beta_\star$ and $\dot{\gamma_u}(0) = u$. The covariant Hessian of f at $\beta_\star$ is the unique symmetric bilinear form $\nabla^2_{\mathcal{M}} f(\beta_\star) \colon \mathcal{T} \times \mathcal{T} \to \mathbf{R}$ satisfying

$$\nabla^2_{\mathcal{M}} f(\beta_\star)[u, u] = (f \circ \gamma_u)''(0).$$

It is classically known that equality holds

$$\nabla^2_{\mathcal{M}} f(\beta_\star)[u, u] = u^\top \cdot \mathcal{I}(\beta_\star, \mathsf{D}) \cdot u \qquad \forall u \in \mathcal{T},$$

where the matrix $\mathcal{I}(\beta_\star, \mathsf{D})$ is defined in (5). Consequently, $\nabla^2_{\mathcal{M}} f(\beta_\star)$ is positive semidefinite when $\beta_\star$ is a local minimizer. Conversely:

$$\nabla^2_{\mathcal{M}} f(\beta_\star) \text{ is positive definite on } \mathcal{T} \text{ if and only if } \beta_\star \text{ is a tilt-stable minimizer.}$$

In this case, identifying $\nabla^2_{\mathcal{M}} f(\beta_\star)$ with a matrix, equality $\nabla\sigma(0) = (P_{\mathcal{T}} \nabla^2_{\mathcal{M}} f(\beta_\star) P_{\mathcal{T}})^\dagger$ holds. In particular, the equality $\mathrm{REG}(\mathsf{D})^{-1} = \lambda_{\min}(\nabla^2_{\mathcal{M}} f(\beta_\star))$ holds.

The sensitivity matrix $\nabla\sigma(0)$ figures prominently in the asymptotic performance of estimation procedures. Notably, building on classical ideas due to Hájek and Le Cam, the recent paper of Duchi and Ruan [35] established a lower bound on asymptotic covariance of arbitrary estimators $\widehat{\beta}_n$ of $\beta_\star$ using n samples $z_1 \ldots, z_n$. The precise lower bound is quite technical, and we refer the interested reader to their paper. In summary, their result shows that if $\beta_\star$ is a tilt-stable minimizer, then the asymptotic covariance of $\sqrt{n}(\widehat{\beta}_n - \beta_\star)$ is lower bounded in the Loewner order by the matrix

$$\Sigma := \nabla\sigma(0) \cdot \mathrm{Cov}\left(\nabla\ell(\beta_\star, z)\right) \cdot \nabla\sigma(0), \tag{6}$$

Moreover, in typical settings the expression in (6) simplifies to the equality $\Sigma = \nabla\sigma(0)$; this is the case for example for (quasi) maximum likelihood estimation (Section 4) and rank-one matrix regression problems (Section 5). Thus, asymptotically the best error that any estimator can achieve in the direction u is on the order $n^{-1/2} \cdot u^\top \Sigma u$. The direction u with the worst error matches the top eigenvector of Σ and the number of samples necessary to find an accurate approximation of $\beta_\star$ grows with $\lambda_{\max}(\Sigma)$. Reassuringly, typical algorithms such as sample average approximation [65] and the stochastic projected gradient method [26, 51] match the lower-bound (6) and thus are asymptotically optimal.

In the rest of the paper, we analyze data distributions D', nearest to a fixed measure D, for which the problem $\mathrm{SP}(\mathsf{D}')$ admits unstable minimizers. The formal definition will depend on whether the learning problem is of supervised or unsupervised type. We now describe these two settings in turn. The goal of unsupervised learning is to learn some property of a distribution D from finitely many samples $z_1, \ldots, z_n \overset{\text{iid}}{\sim} \mathsf{D}$. Dimension reduction with principal component analysis (PCA) is a primary example. In this case, the solution $\beta_\star$ of $\mathrm{SP}(\mathsf{D})$ strongly depends on the distribution D. Therefore a natural measure of robustness of the problem is the size of the smallest perturbation in the Wasserstein-2 distance $W_2(\mathsf{D}', \mathsf{D})$ so that the problem $\mathrm{SP}(\mathsf{D}')$ has an unstable minimizer.

Definition: Radius of statistical efficiency (unsupervised)

Consider the problem SP(D) and let $\mathcal{Q} \subset \mathcal{P}_2(\mathcal{Z})$ be a distinguished set of distributions. Define the *set of ill-conditioned distributions* as

$$\mathcal{E} = \{\mathsf{D}' \in \mathcal{Q} \mid \text{There exists a minimizer of SP}(\mathsf{D}') \text{ that is unstable}\}.$$

The *radius of statistical efficiency* of D is defined to be

$$\mathrm{RSE}(\mathsf{D}) := W_2(\mathsf{D}, \mathcal{E}).$$

Problems of supervised learning are distinctly different. The data consists of pairs $z = (x, y) \sim \mathsf{D}$, where $x \sim \mathsf{D}_x$ are called feature vectors and $y \sim \mathsf{D}_{y|x}$ are the labels. We will assume that the conditional distribution $\mathsf{D}_{y|x}$ depends on the features x and a latent parameter $\beta^\star$. A typical example is the setting of regression under a model $y = g(x, \beta_\star) + \varepsilon$ where ε is a noise vector that is independent of x. The goal of the corresponding optimization problem SP(D) is to recover $\beta^\star$. In contrast to unsupervised learning, the latent parameter $\beta_\star$ is fixed a priori and is not a function of the data distribution. Therefore a natural measure of robustness of the problem is the size of the smallest perturbation to the feature vectors $W_2(\mathsf{D}'_x, \mathsf{D}_x)$ so that the problem $\mathrm{SP}(\mathsf{D}'_x \times \mathsf{D}_{y|x})$ has an unstable minimizer. Notice that the conditional distributions of y given x coincide for the two measures D and $\mathsf{D}' = \mathsf{D}'_x \times \mathsf{D}_{y|x}$. This type of shift exclusively only in the feature data appears often in the literature under the name of *covariate shift* [1, 52, 54, 61, 69].

Definition: Radius of statistical efficiency (supervised)

Consider the problem SP(D) with $\mathsf{D} = \mathsf{D}_x \times \mathsf{D}_{y|x}$ and let $\beta_\star$ be its minimizer. Let $\mathcal{Q} \subset \mathcal{P}_2$ be a distinguished set of distributions. Define the *set of ill-conditioned distributions* as

$$\mathcal{E} = \left\{\mathsf{D}'_x \in \mathcal{Q} \mid \beta_\star \text{ is not a tilt-stable minimizer of } \mathrm{SP}(\mathsf{D}'_x \times \mathsf{D}_{y|x})\right\}.$$

The *radius of statistical efficiency* of D is defined to be

$$\mathrm{RSE}(\mathsf{D}) := W_2(\mathsf{D}_x, \mathcal{E}).$$

Evidently, the quantity $\mathrm{RSE}(\mathsf{D})$ measures the robustness of the problem because it quantifies the size of a neighborhood around D for which all problem instances are stable. There is a small nuance in formalizing this statement due to a lack of compactness in the Wasserstein space. Namely, we have to impose the minor assumption that any sequence of measures ν_i for which the problem $\mathrm{SP}(\nu_i)$ becomes progressively harder ($\mathrm{REG}(\nu_i) \to \infty$) must approach the set of ill-conditioned distributions ($\mathrm{RSE}(\nu_i) \to 0$). In all examples we consider, this holds at least on bounded sets $\mathcal{Q}' \subset \mathcal{P}_2$. The proof of this elementary observation appears in Appendix B.1.

Proposition 2.1 (RSE as a robustness measure)**.** *Fix a set $\mathcal{Q}' \subset \mathcal{Q}$ and suppose that for any sequence of measures $\nu_i \in \mathcal{Q}' \setminus \mathcal{E}$ the implication holds:*

$$\mathrm{REG}(\nu_i) \to \infty \qquad \Longrightarrow \qquad \mathrm{RSE}(\nu_i) \to 0. \tag{7}$$

Then for any measure $\mu \in \mathcal{Q}' \setminus \mathcal{E}$ and any radius $0 < r < \mathrm{RSE}(\mu)$, we have

$$\sup_{\nu \in \mathcal{Q}' : W_2(\nu, \mu) \le r} \mathrm{REG}(\nu) < +\infty. \tag{8}$$

Moreover, if for some $c, q > 0$, the inequality $\mathrm{RSE}(\nu)^q \le c \cdot \mathrm{REG}(\nu)^{-1}$ holds for all $\nu \in \mathcal{Q}' \setminus \mathcal{E}$, then the supremum in (8) is upper bounded by $c \cdot (\mathrm{RSE}(\mu) - r)^{-1/q}$.

At first sight, it appears that computing $\mathrm{RSE}(\mathsf{D})$ in concrete problems is difficult. Indeed, the set of ill-conditioned distributions $\mathcal{E}$ may be quite exotic and computing $\mathrm{RSE}(\mathsf{D})$ amounts to estimating the Wasserstein-2 distance $W_2(\mathsf{D}, \mathcal{E})$. In contrast, computing the regularity modulus REG(D) should be relatively straightforward. The key observation now is that the two quantities, RSE(D) and REG(D), are closely related since $\mathcal{E}$ is the set of minimizers of the function $\mathcal{J}(\mu) := 1/\mathrm{REG}(\mu)$. Thus it would be ideal if there were a quantitative relationship of the form:

$$(W_2(\mu, \mathcal{E}))^{\ell_1} \lesssim \mathcal{J}(\mu) \lesssim (W_2(\mu, \mathcal{E}))^{\ell_2} \qquad \forall \mu \in \mathcal{P}_2. \tag{9}$$

The upper should be elementary to establish because it amounts to upper bounding the growth of the functional $\mathcal{J}(\mu)$. The lower bound is more substantial because it requires lower-bounding the growth of $\mathcal{J}(\mu)$ by a nonnegative function of the distance. Such lower-estimates are called *error bounds* in nonlinear analysis and can be checked by various "slope"-based conditions. See for example the authoritative monographs on the subject, [20, Chapter 8] and [38, Chapter 3]. Indeed, Demmel's original work [28] relies on verifying an error bound property as well, albeit in the much simpler Euclidean setting. The following theorem provides a sufficient condition (10) ensuring the relationship (9). We state the theorem loosely by compressing all multiplicative numerical constants via the symbol $\lesssim$. More precise and sharper guarantees appear in Appendix B.2.[1]

Theorem 2.2: (Infinitesimal characterization of RSE)

Consider the problem SP(D) *of supervised learning and let $\beta_\star$ be its minimizer. Set $\mathcal{Q} = \mathcal{P}_2$ and suppose that the Hessian $\nabla^2_{\mathcal{M}} f(\beta_\star)$ corresponding to a measure $\mu \in \mathcal{P}_2$ can be written as*

$$\nabla^2_{\mathcal{M}} f(\beta_\star) = \mathbb{E}_\mu F(x),$$

for some C^1-smooth map $F \colon \mathbf{R}^d \to \boldsymbol{S}^k_+$ satisfying $\|DF(x)\|_{\mathrm{op}} \lesssim 1 + \|x\|_2$ for all $x \in \mathbf{R}^d$. Suppose that there exists $q_1, q_2 \in [0, 1)$ such that for all measures $\nu \in \mathcal{P}_2 \setminus \mathcal{E}$, the estimate

$$\lambda_{\min}\left(\mathbb{E}_\nu F(x)\right)^{q_1} \lesssim \sqrt{\mathbb{E}_\nu \|DF(x)^*[uu^\top]\|_2^2} \lesssim \lambda_{\min}\left(\mathbb{E}_\nu F(x)\right)^{q_2}, \tag{10}$$

holds for some eigenvector $u \in \mathbb{S}^{d-1}$ of the matrix $\mathbb{E}_\nu F(x)$ corresponding to its minimal eigenvalue. Then for every $\mu \in \mathcal{P}_2$ the inequality holds:

$$\mathrm{REG}(\mu)^{q_2 - 1} \lesssim \mathrm{RSE}(\mu) \lesssim \mathrm{REG}(\mu)^{q_1 - 1}. \tag{11}$$

The expression (10) becomes particularly enlightening when $F(x)$ decomposes as $F(x) = g(x)g(x)^\top$ for some C^1-smooth map $g \colon \mathbf{R}^d \to \mathbf{R}^k$. This situation is typical for regression problems (see Section 4). A simple computation then shows that the sufficient condition (10) reduces to

$$\left(\mathbb{E}_\nu \langle u, g(x)\rangle^2\right)^{q_1} \lesssim \sqrt{\mathbb{E}_\nu \langle u, g(x)\rangle^2 \|\nabla g(x) u\|_2^2} \lesssim \left(\mathbb{E}_\nu \langle u, g(x)\rangle^2\right)^{q_2},$$

Observe that all three terms would match exactly with $q_1 = q_2 = \frac{1}{2}$ were it not for the term $\|\nabla g(x)u\|_2^2$ that reweighs the middle integral. It is this reweighing that may impact the values of q_1 and q_2. The salient feature of Theorem 2.2 is that it completely circumvents the need for explicitly estimating the distance to the exceptional set $\mathcal{E}$. One could apply this theorem to a number of examples that will appear in the rest of the paper. That being said, in all the upcoming examples, we will be able to compute the distance to $\mathcal{E}$ explicitly, thereby obtaining sharper estimates than would follow from Theorem 2.2. Nonetheless, we believe that Theorem 2.2 is interesting in its own right and may be useful for analyzing RSE in more complex situations.

[1] In the theorem statement, the symbol $DF(x) \colon \mathbf{R}^d \to \boldsymbol{S}^k$ denotes the differential of F, while the symbol $DF(x)^* \colon \boldsymbol{S}^k \to \mathbf{R}^d$ is the adjoint linear map of $DF(x)$.

3 Principal component analysis

Principal component analysis (PCA) is a common technique for dimension reduction. The goal of PCA is to find a low-dimensional subspace that captures the majority of the variance of the distribution. In this section, we compute the radius of statistical efficiency for PCA. Setting the stage, let x be a random vector in $\mathbf{R}^d$ drawn from a zero-mean distribution $\mathsf{D} \in \mathcal{P}_2^\circ(\mathbf{R}^d)$. A unit vector v for which the random variable $\langle v, x\rangle$ has maximal variance is called the *first principal component* of D. Thus, the first principal component is the maximizer of the problem

$$\max_{v\in\mathbb{S}^{d-1}} \ \tfrac{1}{2} \mathop{\mathbb{E}}_{x\sim\mathsf{D}} \ \langle v, x\rangle^2, \tag{12}$$

Equivalently, the first principal component is the eigenvector corresponding to the maximal eigenvalue of the covariance matrix $\Sigma_{\mathsf{D}} = \mathbb{E}_{\mathsf{D}}\, xx^\top$. Intuitively, the problem (12) is more challenging when the gap between the top two eigenvalues of Σ_{D} is small. This is the content of the following lemma, whose proof appears in Appendix C.1.

Lemma 3.1. *The set of ill-conditioned distributions for (12) is given by*

$$\mathcal{E} = \{\mu \in \mathcal{P}_2^\circ : \lambda_1(\Sigma_\mu) = \lambda_2(\Sigma_\mu)\}.$$

Moreover, for any $\mu \in \mathcal{P}_2^\circ \setminus \mathcal{E}$, equality $\mathrm{REG}(\mu)^{-1} = \lambda_1(\Sigma_\mu) - \lambda_2(\Sigma_\mu)$ holds.

Therefore, estimating RSE amounts to computing the Wasserstein-2 distance of a base measure D to the set $\mathcal{E}$. The end result is the following theorem; we defer its proof to Appendix C.2.

Theorem 3.2: (RSE for top principal component)

Consider the problem (12) and define the covariance $\Sigma_{\mathsf{D}} := \mathbb{E}_{\mathsf{D}}[xx^\top]$. Then, equality holds:

$$\mathrm{RSE}(\mathsf{D}) = \tfrac{1}{\sqrt{2}}\left(\sqrt{\lambda_1(\Sigma_{\mathsf{D}})} - \sqrt{\lambda_2(\Sigma_{\mathsf{D}})}\right).$$

In particular, we have

$$\mathrm{RSE}(\mathsf{D}) \cdot \mathrm{REG}(\mathsf{D}) = \frac{1}{\sqrt{2}\left(\sqrt{\lambda_1(\Sigma_{\mathsf{D}})} + \sqrt{\lambda_2(\Sigma_{\mathsf{D}})}\right)}. \tag{13}$$

Thus treating $\lambda_1(\Sigma_{\mathsf{D}})$ in (13) as being of constant order, we see that the hardness of the problem is inversely proportional to the distance to the nearest ill-posed problem, $\mathrm{REG}(\mathsf{D}) \propto \mathrm{RSE}(\mathsf{D})^{-1}$.

More generally still, we may be interested in finding a q-dimensional subspace $\mathcal{V} \subset \mathbf{R}^d$ that captures most of the variance. Analytically, this amounts to solving the problem

$$\max_{R\in\mathrm{Gr}(q,d)} \ f(R) = \mathop{\mathbb{E}}_{x\sim\mathsf{D}} \|Rx\|_2^2, \tag{14}$$

where the Grassmannian manifold $\mathrm{Gr}(q, d)$ consists of all orthogonal projections $R \in \boldsymbol{S}^d$ onto q dimensional subspaces of $\mathbf{R}^d$. The column space of the optimal matrix R is called the *top q principal subspace*. Equivalently, the top q principal subspace is the span of the eigenspaces of Σ_{D} corresponding to its top q eigenvalues. The following lemma is a direction extension of Lemma 3.1; see Appendix C.3 for a proof.

Lemma 3.3. *The set of ill-conditioned distributions for (14) is given by*

$$\mathcal{E} = \{\mu \in \mathcal{P}_2^\circ : \lambda_q(\Sigma_\mu) = \lambda_{q+1}(\Sigma_\mu)\}.$$

Moreover, for any $\mu \in \mathcal{P}_2^\circ \setminus \mathcal{E}$, equality $\mathrm{REG}(\mu)^{-1} = \lambda_q(\Sigma_\mu) - \lambda_{q+1}(\Sigma_\mu)$ holds.

Thus, estimating RSE amounts to computing the Wasserstein-2 distance of a base measure D to the set $\mathcal{E}$. The end result is the following theorem; the proof appears in Appendix C.4.

Theorem 3.4: (RSE for PCA)

Consider the problem (14) *and define the covariance* $\Sigma := \mathbb{E}_{\mathsf{D}}[xx^\top]$. *Then, the estimate holds:*

$$\mathrm{RSE}(\mathsf{D}) = \tfrac{1}{\sqrt{2}}\left(\sqrt{\lambda_q(\Sigma)} - \sqrt{\lambda_{q+1}(\Sigma)}\right).$$

In particular, we have

$$\mathrm{RSE}(\mathsf{D}) \cdot \mathrm{REG}(\mathsf{D}) = \frac{1}{\sqrt{2}\left(\sqrt{\lambda_q(\Sigma_{\mathsf{D}})} + \sqrt{\lambda_{q+1}(\Sigma_{\mathsf{D}})}\right)}. \tag{15}$$

Thus similarly to the case $q = 1$, treating $\lambda_q(\Sigma_{\mathsf{D}})$ in (15) as being of constant order, the hardness of the problem is inversely proportional to the distance to ill-posed problems, $\mathrm{REG}(\mathsf{D}) \propto \mathrm{RSE}(\mathsf{D})^{-1}$.

The proofs of Theorems 3.2 and 3.4 rely on estimating the distance to the exceptional sets $\mathcal{E}$. Notice that these two are defined purely in terms of the spectrum of the second-moment matrix. Although such "spectral" sets in $\mathcal{P}_2$ are quite complicated, their distance can be readily computed. Geometric properties of such sets are explored in Appendix A.2 and may be of independent interest.

4 Generalized linear models and (quasi) maximum likelihood estimation

In this section, we compute the RSE for a large class of supervised learning problems arising from (quasi) maximum likelihood estimation (QMLE). The goal is to estimate a parameter $\beta_\star \in \mathcal{M}$ where the constraint set $\mathcal{M}$ is a C^2-smooth embedded submanifold of $\mathbf{R}^d$. Setting the stage, suppose that we have an L^2 random vector x (the predictor) and an L^2 random variable y (the response) satisfying the GLM conditions

$$\mathbb{E}_{\mathsf{D}}[y\,|\,x] = h'(\langle x, \beta_\star\rangle) \qquad \text{and} \qquad \mathrm{Var}_{\mathsf{D}}[y\,|\,x] = \sigma^2 \cdot h''(\langle x, \beta_\star\rangle) \tag{16}$$

for some known C^2-smooth convex function $h\colon \mathbf{R} \to \mathbf{R}$ with $h'' > 0$ and parameter $\sigma^2 > 0$. The function h is called the *cumulant function* of the model (16), and σ^2 the dispersion parameter. Here and from now on, we make the blanket assumption that $\mathcal{Q} \subset \mathcal{P}_2$ is the space of probability measures where sufficient regularity conditions are met to take expectations and to interchange differentiation and expectation as necessary. More precisely, we assume that the function $\phi\colon \mathcal{M} \to \mathbf{R}$ given by $\phi(\beta) = \mathbb{E}[h(\langle x, \beta\rangle)]$ is well defined and has a C^2-smooth local extension to a neighborhood of $\beta_\star$ in $\mathbf{R}^d$ with $\nabla\phi(\beta_\star) = \mathbb{E}[h'(\langle x, \beta_\star\rangle)x]$ and $\nabla^2\phi(\beta_\star) = \mathbb{E}[h''(\langle x, \beta_\star\rangle)xx^\top]$.

Following the seminal work of McCullagh [46], we consider the QMLE problem

$$\min_{\beta \in \mathcal{M}} f(\beta) := \mathbb{E}_{\mathsf{D}}[h(\langle x, \beta\rangle) - y\langle x, \beta\rangle]. \tag{17}$$

The function f given in (17) is called the negative log quasi-likelihood of the GLM (16). In general f is not the negative of the log likelihood function, yet it shares many of its properties and hence the name. The motivation for this loss function comes from the canonical example of (16), where the conditional density of y given x admits an exponential-family formulation. In this case, standard maximum likelihood estimation of $\beta_\star$ coincides with (17). As illustration, Table 1 lists some common examples of QMLE. Henceforth, we let $\mathcal{T} := T_{\mathcal{M}}(\beta_\star)$ denote the tangent space of $\mathcal{M}$ at $\beta_\star$.

We begin with the following lemma that characterizes the set of ill-conditioned problem instances.

Lemma 4.1. *The set of ill-conditioned distributions for* (17) *is given by*

$$\mathcal{E} = \{\mu \in \mathcal{Q} : \mathcal{T} \cap \ker(\Sigma_\mu) \neq \{0\}\}.$$

Model	**Response variable**	**Cumulant** $h(\theta)$	**Second derivative** $h''(\theta)$
Linear	$y = \langle x, \beta_\star \rangle + \varepsilon$	$\frac{1}{2}\theta^2$	1
Logistic	$y \,|\, x \sim \mathsf{Ber}\left(\frac{\exp\langle x,\beta_\star\rangle}{1+\exp\langle x,\beta_\star\rangle}\right)$	$\log(1+\exp\theta)$	$\frac{\exp(\theta)}{(1+\exp(\theta))^2}$
Poisson	$y \,|\, x \sim \mathsf{Poi}(\exp\langle x, \beta_\star\rangle)$	$\exp\theta$	$\exp(\theta)$
Gamma	$y \,|\, x \sim \Gamma(\sigma^{-2}, -\sigma^{-2}\langle x, \beta_\star\rangle)$	$-\log(-\theta)$	θ^{-2}

Table 1: Examples of QGLM problems (16). In the linear regression model, ε is random noise satisfying $\mathbb{E}[\varepsilon \,|\, x] = 0$ and $\mathrm{Var}[\varepsilon \,|\, x] = \sigma^2$.

Moreover, for any $\mu \in \mathcal{Q} \setminus \mathcal{E}$ for which there exist $c_{lb}, c_{ub} > 0$ satisfying $c_{lb} \leq h''(\langle x, \beta_\star\rangle) \leq c_{ub}$ for μ-almost every x, we have $c_{lb} \leq \mathrm{REG}(\mu) \cdot \lambda_{\min}(\Sigma_\mu|_{\mathcal{T}}) \leq c_{ub}$.

The proof of this lemma is deferred to Appendix D.1. The RSE for the problem follows quickly by computing the distance to the set $\mathcal{E}$ in Lemma 4.1; see Appendix D.2 for a proof.

Theorem 4.2: (RSE for QMLE)

Consider a QMLE problem (17) and let $\mathsf{D}_x \in \mathcal{Q}$ be the distribution of x with covariance $\Sigma := \mathbb{E}[xx^\top]$. Then, the estimate holds:

$$\mathrm{RSE}(\mathsf{D}) = \sqrt{\lambda_{\min}(\Sigma|_{\mathcal{T}})}, \tag{18}$$

In particular, if for some $c_{lb}, c_{ub} > 0$ the inequality $c_{lb} \leq h''(\langle x, \beta_\star\rangle) \leq c_{ub}$ holds for D_x-almost every x, then we have

$$c_{lb} \leq (\mathrm{RSE}(\mathsf{D}))^2 \cdot \mathrm{REG}(\mathsf{D}) \leq c_{ub}.$$

Thus we see that under mild conditions, the hardness of the problem is inversely proportional to the square of the distance to the nearest ill-posed problem, $\mathrm{REG}(\mathsf{D}) \propto \mathrm{RSE}(\mathsf{D})^{-2}$. Note that this scaling is different from the one exhibited by PCA in the previous section, $\mathrm{REG}(\mathsf{D}) \propto \mathrm{RSE}(\mathsf{D})^{-1}$.

Aside from the examples in Table 1, an interesting problem instance occurs in sparse recovery. Namely, let $\mathcal{M}$ be the submanifold of $\mathbf{R}^d$ comprised of k-sparse vectors, i.e., those which have precisely k nonzero components. Then the tangent space of $\mathcal{M}$ at $\beta_\star$ is the k-dimensional subspace of $\mathbf{R}^d$ in which $\beta_\star$ is supported:

$$\mathcal{T} = \mathrm{span}\{e_i \mid \langle e_i, \beta_\star\rangle \neq 0\},$$

where $\{e_1, \ldots, e_d\}$ denotes the standard basis of $\mathbf{R}^d$. Any regression problem from Table 1 constrained to $\mathcal{M}$ is ill-conditioned precisely when $\mathrm{supp}(x) \subset v^\perp$ for some $v \in \mathcal{T}$. The right-side of (18) is then the square root of the minimum eigenvalue of the submatrix of Σ_{D} whose columns and rows are indexed by the nonzero coordinates of $\beta_\star$.

5 Rank-one matrix regression problems

In this section, we compute the RSE for a number of tasks in low-rank matrix recovery—a large class of problems with numerous applications in control, system identification, recommendation systems, and machine learning. See for example [14, 17, 24, 70] for an overview. Setting the stage, consider the measurement model

$$y = \langle X, M_\star\rangle + \varepsilon \tag{19}$$

where the matrix $X \in \mathbf{R}^{d_1 \times d_2}$ is drawn from some distribution D_x, the noise ε is independent of X and is mean zero with variance σ^2, and $M_\star$ is a rank r matrix. The goal of low-rank matrix recovery is to estimate the latent parameter $M_\star$ for finitely many i.i.d samples $(X_1, y_1), \ldots, (X_n, y_n)$. For the rest of the section, we focus on the simplest case of rank $r = 1$ matrix recovery.

5.1 Phase retrieval

The problem of phase retrieval corresponds to (19), where the ground truth $M_\star = \beta_\star \beta_\star^\top$ and the data $X = xx^\top$ matrices are symmetric and have rank one. We will let $\mathsf{D}_x \in \mathcal{P}_4(\mathbf{R}^d)$ denote the distribution of $x \in \mathbf{R}^d$ and let D denote the joint distribution of (x, y). There are two standard ways to write the phase retrieval problem as a problem of stochastic optimization. The first is simply to minimize the mean square error over rank one matrices:

$$\min_{M \succeq 0:\, \mathrm{rank}\, M = 1} f(M) := \tfrac{1}{2} \mathop{\mathbb{E}}_{x,y \sim \mathsf{D}} \left(x^\top M x - y\right)^2. \tag{20}$$

Alternatively, one may parameterize rank one matrices as $M = \beta\beta^\top$ and then minimize the mean square error over the factors:

$$\min_{\beta \in \mathbf{R}^d} f(\beta) := \tfrac{1}{8} \mathop{\mathbb{E}}_{x,y \sim \mathsf{D}} \left(\langle x, \beta \rangle^2 - y\right)^2. \tag{21}$$

From the viewpoint of RSE there is no significant distinction between these two formulations. The proof of the next lemma appears in Appendix E.1.

Lemma 5.1. *The set of ill-conditioned distributions for both (20) and (21) is given by*

$$\mathcal{E} = \bigcup_{v \in \mathbb{S}^{d-1}} \left\{\mu \in \mathcal{P}_4(\mathbf{R}^d) : \mathrm{supp}(\mu) \subset \beta_\star^\perp \cup v^\perp\right\}.$$

Fix any measure $\mu \in \mathcal{P}_4 \setminus \mathcal{E}$ and define the matrix $\hat{\Sigma}_\mu = \mathbb{E}_\mu \langle x, \beta_\star \rangle^2 xx^\top$. Then, the estimates hold:

$$c_{lb} \leq \mathrm{REG}(\mu) \cdot \lambda_{\min}(\hat{\Sigma}_\mu) \leq c_{ub}, \tag{22}$$

where $(c_{lb}, c_{ub}) = (2\|\beta_\star\|^2, 4\|\beta_\star\|^2)$ for problem (20) and $(c_{lb}, c_{ub}) = (1, 1)$ for problem (21).

It remains to estimate the expression for the distance to the exceptional set $\mathcal{E}$. The following lemma shows that minimizing the expected squared distance to $\{\beta_\star\}^\perp \cup \{u\}^\perp$ over $u \in \mathbb{S}^{d-1}$ yields the squared W_2 distance to $\mathcal{E}$; its proof appears in Appendix E.2.

Theorem 5.2: (RSE for phase retrieval)

Consider the problems (20) and (21) and define $\Sigma := \mathbb{E}_{\mathsf{D}_x}[xx^\top]$. Then, equality holds:

$$\mathrm{RSE}(\mathsf{D}) = \min_{v \in \mathbb{S}^{d-1}} \sqrt{\mathop{\mathbb{E}}_{x \sim \mathsf{D}_x} \left[\left\langle x, \frac{\beta_\star}{\|\beta_\star\|}\right\rangle^2 \wedge \langle x, v \rangle^2\right]}.$$

Thus, using the reasoning from Lemma 5.1 and Theorem 5.2 one easily derives the following two estimates for formulation (21):

$$\mathrm{RSE}(\mathsf{D}) = \min_{v \in \mathbb{S}^{d-1}} \sqrt{\mathop{\mathbb{E}}_{x \sim \mathsf{D}_x} \left[\left\langle x, \frac{\beta_\star}{\|\beta_\star\|}\right\rangle^2 \wedge \langle x, v \rangle^2\right]},$$

$$\mathrm{REG}(\mathsf{D})^{-1} = \min_{v \in \mathbb{S}^{d-1}} \mathop{\mathbb{E}}_{x \sim \mathsf{D}_x} \left[\left\langle x, \beta_\star \right\rangle^2 \langle x, v \rangle^2\right].$$

A moment of thought leads one to realize that for reasonable distributions, the first equation should scale as $\sqrt{\lambda_{\min}(\Sigma)}$ while the scaling of the second is at least $\lambda_{\min}(\Sigma)$. We now verify that this is indeed the case when the base distribution is Gaussian $x \sim N(0, \Sigma)$ for some covariance matrix $\Sigma \succeq 0$. To this end, define the two functions

$$h_\Sigma(u,v) := \mathop{\mathbb{E}}_{x\sim \mathsf{N}(0,\Sigma)} \left[\langle x,u\rangle^2 \wedge \langle x,v\rangle^2\right],$$
$$g_\Sigma(u,v) := \mathop{\mathbb{E}}_{x\sim \mathsf{N}(0,\Sigma)} \left[\langle x,u\rangle^2 \cdot \langle x,v\rangle^2\right],$$

where $u, v \in \mathbb{S}^{d-1}$ vary over the unit sphere. We defer the proof of the next result to Appendix E.3.

Theorem 5.3. *For any $u \in \mathbb{S}^{d-1}$, the following estimates hold:*

$$(1-\tfrac{2}{\pi})\lambda_{\min}(\Sigma) \le \min_{v\in\mathbb{S}^{d-1}} h_\Sigma(u,v) \le \lambda_{\min}(\Sigma). \tag{23}$$

$$\lambda_{\min}(\Sigma)\cdot\langle \Sigma u, u\rangle \le \min_{v\in\mathbb{S}^{d-1}} g_\Sigma(u,v) \le 3\lambda_{\min}(\Sigma)\cdot\langle \Sigma u, u\rangle \tag{24}$$

Combining Lemma 5.1, Theorem 5.2, and Theorem 5.3 directly yields the following estimate on RSE with Gaussian initial data.

Theorem 5.4: (RSE for phase retrieval with Gaussian data)

Consider problems (20) and (21) with Gaussian data $\mathsf{D}_x \sim N(0,\Sigma)$. Then, the estimate hold:

$$\sqrt{(1-\tfrac{2}{\pi})\lambda_{\min}(\Sigma)} \le \mathrm{RSE}(\mathsf{D}) \le \sqrt{\lambda_{\min}(\Sigma)}.$$

In particular, the following estimates hold:

$$c_{lb}\cdot\frac{1-\frac{2}{\pi}}{3\langle \Sigma\beta_\star, \beta_\star\rangle} \le \mathrm{RSE}(\mathsf{D})^2\cdot \mathrm{REG}(\mathsf{D}) \le c_{ub}\cdot\frac{1}{\langle \Sigma\beta_\star,\beta_\star\rangle}$$

where $(c_{lb}, c_{ub}) = (2\|\beta_\star\|^2, 4\|\beta_\star\|^2)$ for problem (20) and $(c_{lb}, c_{ub}) = (1,1)$ for problem (21).

Interestingly, we see a different scaling between $\mathrm{RSE}(\mathsf{D})$ and $\mathrm{REG}(\mathsf{D})$, depending on whether $\beta_\star$ aligns with the nullspace of Σ. In the regime $\langle \Sigma\beta_\star, \beta_\star\rangle \gg \lambda_{\min}(\Sigma)$, we observe the scaling $\mathrm{REG}(\mathsf{D}) \propto \mathrm{RSE}(\mathsf{D})^{-2}$, while in the regime $\langle \Sigma\beta_\star,\beta_\star\rangle \approx \lambda_{\min}(\Sigma)$, the scaling is much worse $\mathrm{REG}(\mathsf{D}) \propto \mathrm{RSE}(\mathsf{D})^{-4}$. Thus in the latter regime, the distance to the nearest ill-posed problem has a much stronger effect on the hardness of the problem.

5.2 Bilinear sensing

The problem of bilinear sensing is an asymmetric analogue of phase retrieval, that is, the ground truth matrix $M_\star = \beta_{1\star}\beta_{2\star}^\top$ and measurement data $X = x_1 x_2^\top$ are rank one $d_1 \times d_2$ matrices, where the factors $x_1 \sim \mathsf{D}_{x_1}$ and $x_2 \sim \mathsf{D}_{x_2}$ are independent. The standard way to write this problem as stochastic optimization is to minimize the mean square error over rank one rectangular matrices:

$$\min_{M\in\mathbf{R}^{d_1\times d_2}:\,\mathrm{rank}\,M=1} f(M) := \tfrac{1}{2}\mathop{\mathbb{E}}_{\mathsf{D}}\left(x_1^\top M x_2 - y\right)^2, \tag{25}$$

where D denotes the joint distribution over (x_1, x_2, y). Throughout, we fix as the set of allowable data distributions all product measures $\mathcal{Q} = \mathcal{P}_4(\mathbf{R}^{d_1}) \times \mathcal{P}_4(\mathbf{R}^{d_1})$. We disregard the factorized formulation with $M = \beta_1\beta_2^\top$ because it results in a continuum of minimizers that are not tilt-stable. This technical difficulty could be circumvented by introducing an additional constraint, such as $\|\beta_1\| = 1$; however, we do not pursue this approach to simplify the exposition.

The following lemma characterizes the set of ill-conditioned problems; see Appendix F.1 for a proof. For any PSD matrix Σ, the symbol $\kappa(\Sigma) = \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$ denotes its condition number.

Lemma 5.5. *The set of ill-conditioned distributions for* (25) *is given by*

$$\mathcal{E} = \Big\{\mu \times \nu \in \mathcal{Q} : \text{ either } \mathbb{E}_\mu\, xx^\top \text{ or } \mathbb{E}_\nu\, xx^\top \text{ is singular}\Big\}.$$

Moreover, for any $\mu \times \nu \in \mathcal{Q} \setminus \mathcal{E}$*, with* $\Sigma_1 = \mathbb{E}_\mu\, x_1 x_1^\top$ *and* $\Sigma_2 = \mathbb{E}_\nu\, x_2 x_2^\top$*, we have*

$$\frac{2 \cdot \min\{\gamma_2 \lambda_{\min}(\Sigma_1), \gamma_1 \lambda_{\min}(\Sigma_2)\}}{\kappa(\Sigma_1)\kappa(\Sigma_2) + 1} \le \mathrm{REG}(\mu \times \nu)^{-1} \le \min\{\gamma_2 \lambda_{\min}(\Sigma_1), \gamma_1 \lambda_{\min}(\Sigma_2)\} \tag{26}$$

where $\gamma_i = \left\langle \Sigma_i \frac{\beta_{i\star}}{\|\beta_{i\star}\|}, \frac{\beta_{i\star}}{\|\beta_{i\star}\|} \right\rangle$ *for* $i = 1, 2$.

Thus an application of Theorem A.3 immediately yields the following expression for RSE.

Theorem 5.6: (RSE for bilinear sensing)

Consider the problem (25) *with* $\Sigma_1 := \mathbb{E}_{\mathsf{D}_{x_1}}\, x_1 x_1^\top$ *and* $\Sigma_2 := \mathbb{E}_{\mathsf{D}_{x_2}}\, x_2 x_2^\top$*. Then it holds:*

$$\mathrm{RSE}(\mathsf{D}) = \min\left\{\sqrt{\lambda_{\min}(\Sigma_1)}, \sqrt{\lambda_{\min}(\Sigma_2)}\right\}.$$

In particular, the following estimate holds:

$$\frac{\min\{\lambda_{\min}(\Sigma_1), \lambda_{\min}(\Sigma_2)\}}{\min\{\gamma_2 \lambda_{\min}(\Sigma_1), \gamma_1 \lambda_{\min}(\Sigma_2)\}} \le \mathrm{RSE}(\mathsf{D})^2 \cdot \mathrm{REG}(\mathsf{D}) \le C \cdot \frac{\min\{\lambda_{\min}(\Sigma_1), \lambda_{\min}(\Sigma_2)\}}{\min\{\gamma_2 \lambda_{\min}(\Sigma_1), \gamma_1 \lambda_{\min}(\Sigma_2)\}}$$

where $C = \frac{\kappa(\Sigma_1)\kappa(\Sigma_2)+1}{2}$ *with* $\gamma_i = \left\langle \Sigma_i \frac{\beta_{i\star}}{\|\beta_{i\star}\|}, \frac{\beta_{i\star}}{\|\beta_{i\star}\|} \right\rangle$ *for* $i = 1, 2$.

Similar to phase retrieval, the scaling between RSE and REG for bilinear sensing depends on the simultaneous alignment between $\beta_{i\star}$ and the least eigenvector of Σ_i for $i = 1, 2$. In particular, when $\kappa(\Sigma_1)\kappa(\Sigma_2) \approx 1$ both upper and lower bounds are off by a constant and if, further, $\lambda_{\min}(\Sigma_1) \approx \lambda_{\min}(\Sigma_2)$, there are two regimes: (1) when $\min\{\gamma_1, \gamma_2\} \gg \lambda_{\min}(\Sigma_1)$, then $\mathrm{REG}(\mathsf{D}) \propto \mathrm{RSE}(\mathsf{D})^{-2}$, and (2) when $\min\{\gamma_1, \gamma_2\} \approx \lambda_{\min}(\Sigma_1)$, then $\mathrm{REG}(\mathsf{D}) \propto \mathrm{RSE}(\mathsf{D})^{-4}$.

5.3 Matrix completion

The problem of matrix completion corresponds to (19), where the ground truth matrix $M_\star$ has low rank and the data matrices $X \sim \mathsf{D}_x$ are drawn from some discrete distribution on matrices of the form $X = e_i e_j^\top$. We will focus on the simplified setting where $M_\star$ is rank one and positive semidefinite. The standard way to write this problem as stochastic optimization is

$$\min_{M \in \mathcal{M}} f(M) := \tfrac{1}{2} \mathop{\mathbb{E}}_{\mathsf{D}} \left[(\langle X, M\rangle - y)^2\right], \tag{27}$$

where $\mathcal{M} = \{M \succeq 0 : \operatorname{rank} M = 1\}$ is the set of rank one PSD matrices. In this section, we compute the RSE of the problem in terms of the graph induced by the support of the distribution D_x.

We begin with a few observations. First, it is straightforward to verify $\nabla f(M_\star) = 0$ and

$$\nabla f(M_\star)[\Delta, \Delta] = \mathbb{E}\langle X, \Delta\rangle^2 \qquad \forall \Delta \in \boldsymbol{S}^d. \tag{28}$$

In particular, the optimal Lagrange multipliers for $M_\star$ are zero. Forming the factorization $M_\star = \beta_\star \beta_\star^\top$, the tangent space to $\mathcal{M}$ at $M_\star$ can be written as

$$\mathcal{T} = \{\beta_\star v^\top + v \beta_\star^\top : v \in \mathbf{R}^d\}.$$

Moreover, an elementary computation shows that $\|\Delta\|_F^2 / \|\beta_\star\|^2 \|v\|^2 \in [2, 4]$. In particular, Δ is zero if and only if v is zero. Consequently, the set of ill-conditioned distributions takes the form:

$$\mathcal{E} = \bigcup_{v \in \mathbf{R}^d \setminus \{0\}} \{\mu \in \mathcal{P}_2(\boldsymbol{S}^d) : \operatorname{supp}(X) \subset (v \beta_\star^\top)^\perp\}. \tag{29}$$

Indeed, estimating $\mathrm{REG}(\mu)$ at any $\mu \in \mathcal{P}_2(\boldsymbol{S}^d) \setminus \mathcal{E}$ is straightforward, and is the content of the following lemma; see Appendix G.1 for a proof.

Lemma 5.7. *Consider any measure $\mu \in \mathcal{P}_2(\boldsymbol{S}^d) \setminus \mathcal{E}$ and define the matrices*

$$\Phi_{\beta_\star} = (I \otimes \beta_\star) + (\beta_\star \otimes I) \qquad \textit{and} \qquad \Sigma_\mu = \mathbb{E}_\mu \operatorname{vec}(X)\operatorname{vec}(X)^\top.$$

Then the estimate holds:

$$2\|\beta_\star\|^2 \leq \mathrm{REG}(\mu) \cdot \lambda_{\min}(\Phi_{\beta_\star}^\top \Sigma_\mu \Phi_{\beta_\star}) \leq 4\|\beta_\star\|^2. \tag{30}$$

Observe now that most measures $\mu \in \mathcal{E}$ do not correspond to a matrix completion problem, since they do not even need to be discrete. With this in mind, we now focus on the setting where the set of admissible distributions $\mathcal{Q}$ encodes only matrix completion problems. To this end, for a matrix completion problem, X is equal to some matrix $X = e_\ell e_k^\top$, whose distribution is induced by a random pair (ℓ, k) in $[d] \times [d]$ and can be represented by a matrix of probabilities

$$P = (p_{ij}) \quad \text{where} \quad p_{ij} = \mathbb{P}\{(\ell, k) = (i, j)\} = \mathbb{P}\{X_{ij} = 1\} = \mathbb{E}[X_{ij}].$$

Since $M_\star$ is symmetric, we assume without loss of generality that P is symmetric as well. We denote the set of all such symmetric distributions on $[d] \times [d]$ by

$$\mathcal{Q} = \{(p_{ij}) \in \boldsymbol{S}^d \mid \textstyle\sum_{ij} p_{ij} = 1 \text{ and } p_{ij} \geq 0 \text{ for all } i, j \in [d]\},$$

For each $Q \in \mathcal{Q}$, we let $\mu_Q \in \mathcal{P}_2(\mathbf{R}^{d\times d})$ be the distribution over canonical matrices $e_i e_j^\top$ where $(i, j) \sim Q$.[2] Thus, the set of ill-conditioned distributions encoding matrix completion problems is

$$\mathcal{E}^{\mathrm{mc}} := \{\mu_Q \mid Q \in \mathcal{Q}\} \cap \mathcal{E},$$

where $\mathcal{E}$ is defined in (29). We are now ready to study the Wasserstein distance between μ and $\mathcal{E}^{\mathrm{mc}}$. We will see that this distance relies on the combinatorial structure of the observations.

Consider the undirected graph $G = (V, E)$ with vertices $V = [d]$, and edges $E = \{(i, j) \mid p_{ij} > 0\}$. Thus, E corresponds to the tuples of indices that are "observed" in the problem. Let

$$G^* = (V^*, E^*) \text{ be the induced graph given by } V^* = \operatorname{supp}(\beta_\star), \tag{31}$$

meaning that E^* consists of all the edges in E between elements of V^*. Further, define

$$V^0 = \{i \notin V^* \mid \text{for all } j \in V^*, (i, j) \notin E\}. \tag{32}$$

Thus, V^0 corresponds to nodes i with $\beta_i = 0$ that are not connected to any j for which $\beta_j \neq 0$.

We will see shortly that the following assumption characterizes the set of well-posed instances of matrix completion.

Assumption 1. Consider a graph $G = ([d], E)$ with G^* and V^0 defined in (31) and (32), respectively. Suppose the following two hold.

1. (**Non-bipartite**) G^* has no connected components that are bipartite.
2. (**No isolated zeros**) The set V^0 is empty.

Given a set of edges $A \subset [d] \times [d]$, we let G_A be the graph induced by A and define

$$\Omega_{\beta_\star} = \{A \subset [d] \times [d] \mid G_A \text{ does not satisfy Assumption 1}\}.$$

We are now ready to show the main result of this section; see Appendix G.2 for a proof.

[2]With a slight abuse of notation we use Q to denote both the matrix and distribution over indices.

Theorem 5.8: (RSE for matrix completion)

Let $P = (p_{ij})$ and $M_\star = \beta_\star \beta_\star^\top$ be the data of the matrix completion problem (27), and let μ be the distribution of X induced by P. Then, μ is well-posed, i.e., $\mu \notin \mathcal{E}^{\mathrm{mc}}$, if, and only if, the graph G induced by the problem satisfies Assumption 1. Additionally, the identity holds:

$$\mathrm{RSE}(\mu) = \min_{\substack{A \in \Omega_{\beta_\star} \\ A \subset \mathrm{supp}(P)}} \sum_{ij \in \mathrm{supp}(P) \setminus A} p_{ij}. \tag{33}$$

Moreover, computing $\mathrm{RSE}(\mu)$ *is NP-hard in general.*

Thus the theorem shows that one can compute $\mathrm{RSE}(\mu)$ by enumerating over all exceptional edge sets $A \subset \mathrm{supp}(P)$, meaning that G_A^* either has a connected bipartite component or V^0 is nonempty. Then, the set A for which the mass $\sum_{ij \in \mathrm{supp}(P) \setminus A} p_{ij}$ is smallest yields $\mathrm{RSE}(\mu)$. Interestingly, computing $\mathrm{RSE}(\mu)$ is NP-hard in general, as we show by a reduction from `MAXCUT`.

6 Conclusion

In this work, we introduced a new measure of robustness—*radius of statistical efficiency (RSE)*—for problems of statistical inference and estimation. We computed RSE for a number of test-bed problems, including principal component analysis, generalized linear models, phase retrieval, bilinear sensing, and matrix completion. In all cases, we verified a precise reciprocal relationship between RSE and the intrinsic complexity/sensitivity of the problem instance, thereby paralleling the classical Eckart–Young theorem and its numerous extensions in numerical analysis and optimization. More generally, we obtained sufficient conditions for such a relationship to hold that depend only on local information (gradients, Hessians), rather than an explicit description of the set of ill-conditioned distributions. We believe that this work provides an intriguing new perspective on the interplay between problem difficulty, solution sensitivity, and robustness in statistical inference and learning.

Acknowledgments

We thank John Duchi, Jorge Garza-Vargas, Zaid Harchaoui, and Eitan Levin for insightful conversations during the development of this work.

References

[1] D. Agarwal, L. Li, and A. Smola. Linear-time estimators for propensity scores. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 93–100. JMLR Workshop and Conference Proceedings, 2011.

[2] D. Azé and J.-N. Corvellec. Characterizations of error bounds for lower semicontinuous functions on metric spaces. *ESAIM: Control, Optimisation and Calculus of Variations*, 10(3):409–425, 2004.

[3] R. Balan. Reconstruction of signals from magnitudes of redundant representations: The complex case. *Foundations of Computational Mathematics*, 16:677–721, 2016.

[4] A. S. Bandeira, J. Cahill, D. G. Mixon, and A. A. Nelson. Saving phase: Injectivity and stability for phase retrieval. *Applied and Computational Harmonic Analysis*, 37(1):106–125, 2014.

[5] A. Bastounis, F. Cucker, and A. C. Hansen. When can you trust feature selection?–i: A condition-based analysis of lasso and generalised hardness of approximation. *arXiv preprint arXiv:2312.11425*, 2023.

[6] T. Bendokat, R. Zimmermann, and P.-A. Absil. A grassmann manifold handbook: Basic geometry and computational aspects. *Advances in Computational Mathematics*, 50(1):1–51, 2024.

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

[8] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

[9] J. Borwein and A. Lewis. *Convex Analysis*. Springer, 2006.

[10] N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[11] H. Brezis. *Monotonic maximal operators and semi-groups of contractions in Hilbert spaces*. Elsevier, 1973.

[12] P. Bürgisser and F. Cucker. On a problem posed by steve smale. *Annals of Mathematics*, pages 1785–1836, 2011.

[13] P. Bürgisser and F. Cucker. *Condition: The geometry of numerical algorithms*, volume 349. Springer Science & Business Media, 2013.

[14] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021.

[15] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy. Composite optimization for robust rank one bilinear sensing. *Information and Inference: A Journal of the IMA*, 10(2):333–396, 2021.

[16] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[17] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[18] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American mathematical society*, 22(1):211–231, 2009.

[19] H. Cramér. *Mathematical methods of statistics*, volume 26. Princeton university press, 1999.

[20] Y. Cui and J.-S. Pang. *Modern nonconvex nondifferentiable optimization*. SIAM, 2021.

[21] J. Cutler, M. Díaz, and D. Drusvyatskiy. Stochastic approximation with decision-dependent distributions: asymptotic normality and optimality. *arXiv preprint arXiv:2207.04173*, 2022.

[22] A. Daniilidis, A. Lewis, J. Malick, and H. Sendov. Prox-regularity of spectral functions and spectral sets. *Journal of Convex Analysis*, 15(3):547–560, 2008.

[23] A. Daniilidis, D. Drusvyatskiy, and A. S. Lewis. Orthogonal invariance and identifiability. *SIAM Journal on Matrix Analysis and Applications*, 35(2):580–598, 2014.

[24] M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[25] C. Davis. All convex invariant functions of hermitian matrices. *Archiv der Mathematik*, 8(4): 276–278, 1957.

[26] D. Davis, D. Drusvyatskiy, and L. Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *arXiv preprint arXiv:2301.06632*, 2023.

[27] E. De Giorgi. New problems on minimizing movements. *Ennio de Giorgi: Selected Papers*, pages 699–713, 1993.

[28] J. W. Demmel. On condition numbers and the distance to the nearest ill-posed problem. *Numerische Mathematik*, 51:251–289, 1987.

[29] J. W. Demmel. *Applied numerical linear algebra.* SIAM, 1997.

[30] M. Díaz. The nonsmooth landscape of blind deconvolution. *Workshop on Optimization for Machine Learning*, 2019.

[31] A. Dontchev, A. Lewis, and R. Rockafellar. The radius of metric regularity. *Transactions of the American Mathematical Society*, 355(2):493–517, 2003.

[32] A. L. Dontchev and R. T. Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.

[33] D. Drusvyatskiy and C. Paquette. Variational analysis of spectral functions simplified. *Journal of Convex Analysis*, 25(1):119–134, 2018.

[34] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. Curves of descent. *SIAM Journal on Control and Optimization*, 53(1):114–138, 2015.

[35] J. C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21–48, 2021.

[36] R. M. Freund and J. R. Vera. Some characterizations and properties of the "distance to ill-posedness" and the condition measure of a conic linear system. *Mathematical Programming*, 86(2):225–260, 1999.

[37] A. D. Ioffe. Metric regularity and subdifferential calculus. *Russian Mathematical Surveys*, 55 (3):501, 2000.

[38] A. D. Ioffe. Variational analysis of regular mappings. *Springer Monographs in Mathematics. Springer, Cham*, 2017.

[39] L. M. Le Cam and G. L. Yang. *Asymptotics in statistics: some basic concepts.* Springer Science & Business Media, 2000.

[40] A. S. Lewis. Convex analysis on the hermitian matrices. *SIAM Journal on Optimization*, 6(1): 164–177, 1996.

[41] A. S. Lewis. Nonsmooth analysis of eigenvalues. *Mathematical Programming*, 84(1):1–24, 1999.

[42] A. S. Lewis and H. S. Sendov. Nonsmooth analysis of singular values. part i: Theory. *Set-Valued Analysis*, 13:213–241, 2005.

[43] W. V. Li and A. Wei. Gaussian integrals involving absolute value functions. In *High dimensional probability V: the Luminy volume*, volume 5, pages 43–60. Institute of Mathematical Statistics, 2009.

[44] X. Liu and N. D. Sidiropoulos. Cramér-rao lower bounds for low-rank decomposition of multidimensional arrays. *IEEE Transactions on Signal Processing*, 49(9):2074–2086, 2001.

[45] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[46] P. McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, 11(1), Mar. 1983.

[47] J.-J. Moreau. Proximité and dualityé in a hilbertian space. *Bulletin of the Mathematical Society of France*, 93:273–299, 1965.

[48] R. Niazadeh, M. Babaie-Zadeh, and C. Jutten. On the achievability of cramér–rao bound in noisy compressed sensing. *IEEE Transactions on Signal Processing*, 60(1):518–526, 2011.

[49] J. Pena. Understanding the geometry of infeasible perturbations of a conic linear system. *SIAM Journal on Optimization*, 10(2):534–550, 2000.

[50] R. Poliquin and R. T. Rockafellar. Tilt stability of a local minimum. *SIAM Journal on Optimization*, 8(2):287–299, 1998.

[51] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. ISSN 0363-0129. doi: 10.1137/0330046.

[52] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN 0262170051.

[53] C. R. Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 235–247. Springer, 1992.

[54] S. Reddi, B. Poczos, and A. Smola. Doubly robust covariate shift correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[55] J. Renegar. Some perturbation theory for linear programming. *Mathematical programming*, 65 (1):73–91, 1994.

[56] J. Renegar. Incorporating condition measures into the complexity theory of linear programming. *SIAM Journal on Optimization*, 5(3):506–524, 1995.

[57] J. Renegar. Linear programming, complexity theory and elementary functional analysis. *Mathematical Programming*, 70(1):279–351, 1995.

[58] V. Roulet, N. Boumal, and A. d'Aspremont. Computational complexity versus statistical performance on sparse recovery problems. *Information and Inference: A Journal of the IMA*, 9 (1):1–32, 2020.

[59] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[60] H. S. Sendov. The higher-order derivatives of spectral functions. *Linear algebra and its applications*, 424(1):240–281, 2007.

[61] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[62] S. T. Smith. Covariance, subspace, and intrinsic crame/spl acute/r-rao bounds. *IEEE Transactions on Signal Processing*, 53(5):1610–1630, 2005.

[63] J. Sylvester. On the differentiability of $\mathbf{O}(n)$ invariant functions of symmetric matrices. *Duke Mathematical Journal*, 52(2):475 – 483, 1985. doi: 10.1215/S0012-7094-85-05223-8. URL https://doi.org/10.1215/S0012-7094-85-05223-8.

[64] C. Theobald. An inequality for the trace of the product of two symmetric matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 77, pages 265–267. Cambridge University Press, 1975.

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

[66] C. Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009.

[67] J. Von Neumann. Some matrix inequalities and metrization of matrixspace, tomsk univ. rev.(1937), 286–300. see also: Collected works vol. iv, 1962.

[68] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[69] J. Wen, C.-N. Yu, and R. Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, pages 631–639. PMLR, 2014.

[70] J. Wright and Y. Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

[71] H. Zhang, B. Yalcin, J. Lavaei, and S. Sojoudi. A new complexity metric for nonconvex rank-one generalized matrix completion. *Mathematical Programming*, pages 1–42, 2023.

[72] M. Šilhavý. Differentiability properties of isotropic functions. *Duke Mathematical Journal*, 104 (3):367 – 373, 2000. doi: 10.1215/S0012-7094-00-10431-0. URL https://doi.org/10.1215/S0012-7094-00-10431-0.

A Geometry of the Wasserstein space and distance estimation

In this section, we introduce the necessary background of the Wasserstein geometry and prove a number of results that may be on independent interest. In the following section, we will use many of these results to prove estimates on RSE announced in the paper. We follow standard notation of optimal transport, as set out for example in the monographs of Villani [66] and Santambrogio [59].

Let $(\mathcal{X}, \mathsf{d})$ be a separable complete metric space equipped with its Borel σ-algebra $\mathcal{B}_{\mathcal{X}}$. The primary example for us will be $\mathbf{R}^d$ equipped with the ℓ_p norm. The distance of a point $x \in \mathcal{X}$ to a set $\mathcal{K} \subset \mathcal{X}$ will be denoted by $\operatorname{dist}(x, \mathcal{K}) = \inf_{x' \in \mathcal{K}} \mathsf{d}(x, x')$. The set of Borel probability measures on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$, and will be abbreviated as $\mathcal{P}$ if the space $\mathcal{X}$ is clear from context. The support of a measure $\mu \in \mathcal{P}$, written as $\operatorname{supp}(\mu)$, is the smallest closed set $C \subset \mathcal{X}$ such that the complement $\mathcal{X} \setminus C$ is of zero μ-measure. For any measurable map $T\colon (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, the pushforward measure $T_{\#}\mu$ is defined to be

$$(T_{\#}\mu)(B) = \mu(T^{-1}(B)) \qquad \text{for all } B \in \mathcal{B}_{\mathcal{X}}.$$

The support of a random variable on $\mathcal{X}$ is the support of its distribution.

For any $p \geq 1$, the symbol $\mathcal{P}_p$ denotes the set of all distributions μ on $\mathcal{X}$ with finite p^{th} moment, meaning $\mathbb{E}_{x \sim \mu} \mathsf{d}(x, x_0)^p < \infty$ for some (and hence any) $x_0 \in \mathcal{X}$. The Wasserstein-p distance between two measures $\mu, \nu \in \mathcal{P}_p$ is defined by:

$$W_p(\mu, \nu) = \min_{\pi \in \Pi(\mu, \nu)} \left(\mathop{\mathbb{E}}_{(x,y) \sim \pi} \mathsf{d}(x, y)^p \right)^{1/p}.$$

Here, the set $\Pi(\mu, \nu)$ consists of couplings between μ and ν, i.e., distributions in $\mathcal{P}_p(\mathcal{X} \times \mathcal{X})$ having μ and ν as its first and second marginals. An important fact is that the pair $(\mathcal{P}_p, W_p)$ is a separable complete metric space in its own right and is called the Wasserstein-p space on $\mathcal{X}$.

We will need a few basic estimates on the W_p distance. First, consider any measures $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$ and a measurable map $T\colon \mathcal{X} \to \mathcal{X}$ satisfying $\nu = T_{\#}\mu$. Then the law of the random variable $(x, T(x))$ is a coupling between μ and ν and therefore the estimate holds:

$$W_p^p(\mu, \nu) \leq \mathop{\mathbb{E}}_{x \sim \mu} \mathsf{d}(x, T(x))^p. \tag{34}$$

Another useful observation is that for any measures $\mu, \nu \in \mathcal{P}_p$ such that the support of ν is contained in a set $C \subset \mathcal{K}$, the estimate holds:

$$W_p^p(\mu, \nu) \geq \mathop{\mathbb{E}}_{x \sim \mu} \operatorname{dist}^p(x, C). \tag{35}$$

Consequently, equality holds in (34) when T is a projection, the content of the following lemma.

Lemma A.1 (W_p metric & projections)**.** *Consider a measure $\mu \in \mathcal{P}_p$ and a set $\mathcal{K} \subset \mathcal{X}$. Suppose that the metric projection $\mathsf{P}_{\mathcal{K}}$ admits a measurable selection $s\colon \mathcal{X} \to \mathcal{X}$. Then equality holds:*

$$W_p^p(\mu, s_{\#}\mu) = \mathop{\mathbb{E}}_{X \sim \mu} \operatorname{dist}(X, \mathcal{K})^p. \tag{36}$$

Proof. We first show that $\nu := s_{\#}\mu$ lies in $\mathcal{P}_p$. Indeed, for any $x_0 \in \mathcal{K}$ and $x \in \mathcal{X}$ we have

$$\mathsf{d}(s(x), x_0) \leq \mathsf{d}(s(x), x) + \mathsf{d}(x, x_0) \leq 2\mathsf{d}(x, x_0),$$

and therefore ν lies in $\mathcal{P}_p$. Next, taking into account that the support of ν is contained in $\mathcal{K}$, combining (34) and (35) completes the proof. □

For the rest of the section, we will focus exclusively on the setting where $\mathcal{X}$ is the Euclidean space $\mathbf{R}^d$ equipped with the inner product $\langle \cdot, \cdot \rangle$. The ℓ_p norm in $\mathbf{R}^d$ will be denoted by $\|\cdot\|_p$. Finally, we denote the second moment matrix for any measure $\mu \in \mathcal{P}_2$, by the symbol

$$\Sigma_\mu := \mathop{\mathbb{E}}_{x \sim \mu} x x^\top.$$

Given a set of measures $\mathcal{V} \subset \mathcal{P}_p$, the distance function $W_2(\cdot, \mathcal{V})$ is defined in the usual way

$$W_2(\mu, \mathcal{V}) = \inf_{\nu \in \mathcal{V}} W_2(\mu, \nu).$$

Although estimating the distance function is difficult in general, we will focus on well-structured sets $\mathcal{V}$ for which $W_2(\mu,\mathcal{V})$ can be readily computed. The following two sections study, respectively, measures constrained either by the location of its support or by the spectrum of its covariance.

A.1 Sets of measures constrained by their support

Consider a linear subspace $\mathcal{K} \subseteq \mathcal{X}$ and a mean zero measure $\mu \in \mathcal{P}_p$. Recall that the compression $\Sigma_\mu|_\mathcal{K}$ of Σ_μ to $\mathcal{K}$ is the positive semidefinite quadratic form on $\mathcal{K}$ given by:

$$\langle \Sigma_\mu v, v\rangle = \mathbb{E}_{x\sim\mu}[\langle x, v\rangle^2] \qquad \text{for all } v \in \mathcal{K}. \tag{37}$$

We will need the following lemma that provides a convenient interpretation of the trace of $\Sigma_\mu|_\mathcal{K}$.

Lemma A.2 (Trace of covariance & W_2 distance)**.** *Consider a measure $\mu \in \mathcal{P}_2$ and let $\mathcal{K}$ be a proper subspace of $\mathbf{R}^d$. Then, the following equalities hold:*

$$\operatorname{tr}\Sigma_\mu|_\mathcal{K} = \mathop{\mathbb{E}}_{x\sim\mu} \operatorname{dist}(x,\mathcal{K}^\perp)^2 = W_2^2(\mu, (\mathsf{P}_{\mathcal{K}^\perp})_\#\mu).$$

Proof. We successively compute

$$\operatorname{tr}\Sigma_\mu|_\mathcal{K} = \mathop{\mathbb{E}}_{x\sim\mu}\operatorname{tr}(P_\mathcal{K} x x^\top P_\mathcal{K}) = \mathop{\mathbb{E}}_{x\sim\mu}\|P_\mathcal{K} x\|^2 = \mathop{\mathbb{E}}_{x\sim\mu}\operatorname{dist}^2(x,\mathcal{K}^\perp) = W_2^2(\mu,(P_{\mathcal{K}^\perp})_\#\mu),$$

where the last equality follows from Lemma A.1. □

In light of (37), the matrix $\Sigma_\mu|_\mathcal{K}$ is singular if and only if the inclusion $\operatorname{supp}(\mu) \subset v^\perp$ holds for some $v \in \mathcal{K}\cap\mathbb{S}^{d-1}$. Define the set of measures for which $\Sigma_\nu|_\mathcal{K}$ is indeed singular:

$$\mathcal{V}_\mathcal{K} = \{\nu \in \mathcal{P}_2 \mid \operatorname{supp}(\nu) \subset v^\perp \text{ for some } v \in \mathcal{K}\cap\mathbb{S}^{d-1}\}. \tag{38}$$

The following theorem, the main result of the section, shows that the W_2 distance to $\mathcal{V}_\mathcal{K}$ is simply the minimal eigenvalue of $\Sigma_\mu|_\mathcal{K}$.

Theorem A.3: (Distance to $\mathcal{V}_\mathcal{K}$)

Consider a measure $\mu \in \mathcal{P}_2$ and a proper linear subspace $\mathcal{K}$ of $\mathbf{R}^d$. Then, equality holds:

$$W_2^2(\mu,\mathcal{V}_\mathcal{K}) = \lambda_{\min}(\Sigma_\mu|_\mathcal{K}),$$

where $\mathcal{V}_\mathcal{K}$ is defined in (38). Moreover the distance of μ to $\mathcal{V}_\mathcal{K}$ is attained by the measure $(\mathsf{P}_{v^\perp})_\#\mu$, where $v \in \mathcal{K}$ is an eigenvector of $\Sigma_\mu|_\mathcal{K}$ corresponding to its minimal eigenvalue.

Proof. For any vector $v \in \mathcal{K}\cap\mathbb{S}^{d-1}$, applying Lemma A.2 with $\operatorname{span}(v)$ in place of $\mathcal{K}$ yields:

$$\langle \Sigma_\mu v, v\rangle = \operatorname{tr}\Sigma_\mu|_{\operatorname{span}(v)} = W_2^2(\mu,(\mathsf{P}_{\mathcal{K}^\perp})_\#\mu). \tag{39}$$

Let us now decompose $\mathcal{V}_\mathcal{K}$ into a union of simpler sets

$$\mathcal{V}_\mathcal{K} = \bigcup_{v\in\mathcal{K}\cap\mathbb{S}^{d-1}} L_v \qquad \text{where} \qquad L_v := \{\nu \in \mathcal{P}_2 : \operatorname{supp}(\nu) \subset v^\perp\}.$$

We now estimate

$$W_2(\mu, L_v) \le W_2(\mu,(\mathsf{P}_{v^\perp})_\#\mu) = \mathop{\mathbb{E}}_{x\sim\mu}\operatorname{dist}(x,v^\perp)^2 \le W_2(\mu,L_v).$$

where the first inequality holds trivially, the equality follows from Lemma A.1, and the last inequality follows from (35) with $C = v^\perp$. Thus equality holds throughout. Using (39), we then conclude

$$W_2(\mu,\mathcal{V}_\mathcal{K}) = \inf_{v\in\mathcal{K}\cap\mathbb{S}^{d-1}} W_2(\mu,L_v) = \inf_{v\in\mathcal{K}\cap\mathbb{S}^{d-1}}\langle\Sigma_\mu v,v\rangle = \lambda_{\min}(\Sigma_\mu|_\mathcal{K}),$$

as claimed. It also follows immediately that for any minimal eigenvector v of $\Sigma_\mu|_{\mathcal{K}}$, the pushforward measure $\nu = (\mathsf{P}_{v^\perp})_{\#}\mu$ attains the minimal W_2 distance of μ to $\mathcal{V}_{\mathcal{K}}$. $\square$

A.2 Spectral sets and functions of measures

In this section, we investigate a special class of functions on $\mathcal{P}_2$—called spectral—that depend on the measure only through the eigenvalues of its second moment matrix. This function class has close analogues in existing literature in matrix analysis and eigenvalue optimization. We postpone a detailed discussion on the related literature until the end of the section.

The following is the key definition.

Definition A.4 (Spectral functions of measures)**.** A function $F\colon \mathcal{P}_2 \to \mathbf{R} \cup \{+\infty\}$ is called *spectral* if for any $\mu, \nu \in \mathcal{P}_2$ with the same second moment matrix $\lambda(\Sigma_\mu) = \lambda(\Sigma_\nu)$ equality $F(\mu) = F(\nu)$ holds.

A good example to keep in mind is the Schatten norm $F(\mu) = \|\lambda(\Sigma_\mu)\|_q$ for any $q \in (0, \infty]$. Notice that in this example F factors as a composition of the eigenvalue map $\lambda(\cdot)$ and the permutation-invariant function $f(v) = \|v\|_q$ on $\mathbf{R}^d$. Evidently, all spectral functions arise in this way.

Definition A.5 (Symmetric functions)**.** A function $f\colon \mathbf{R}^d_+ \to \mathbf{R} \cup \{+\infty\}$ is called *symmetric* if equality $f(s(x)) = f(x)$ holds for all $x \in \mathbf{R}^d_+$ and all permutations of coordinates $s(\cdot)$.

An elementary observation is that a function $F\colon \mathcal{P}_2 \to \mathbf{R} \cup \{+\infty\}$ is spectral if and only if there exists a symmetric function $f : \mathbf{R}^d_+ \to \mathbf{R} \cup \{+\infty\}$ satisfying

$$F(\mu) = f(\lambda(\Sigma_\mu)) \qquad \forall \mu \in \mathcal{P}_2.$$

Concretely, the symmetric function f can be obtained from F by restricting to Gaussian measures with diagonal covariance $f(v) = F(N(0, \mathrm{Diag}(v)))$. The definition of spectral and symmetric functions easily extends to sets through indicator functions. Namely, a set $\mathcal{G} \subset \mathcal{P}_2$ is *spectral* if the indicator function $\delta_{\mathcal{G}}$ is spectral, while a set $G \subset \mathbf{R}^d_+$ is *symmetric* if the indicator δ_G is symmetric.

An interesting example of a spectral set is given by

$$\mathcal{G}_q := \{\mu \in \mathcal{P}_2 : \lambda_q(\Sigma_\mu) = \lambda_{q+1}(\Sigma_\mu)\}. \tag{40}$$

Although complicated, we may identify it with the symmetric set

$$G_q = \{v \in \mathbf{R}^d_+ : v^{(q)} = v^{(q+1)}\},$$

where $v^{(i)}$ is the i'th largest coordinate of v. Indeed, equality $\mathcal{G} = \{\mu \in \mathcal{P}_2 : \lambda(\Sigma_\mu) \in G\}$ holds.

In this section, we will show that one can express the W_2-distance function to $\mathcal{G}$ purely in terms of the ℓ_2-distance function to the much simpler set G. Indeed, we will prove the following theorem, which specialized to example (40) yields the expression

$$W_2(\mu, \mathcal{G}_q) = \mathrm{dist}_2\left(\sqrt{\lambda(\Sigma_\mu)}, G_q\right) = \tfrac{1}{\sqrt{2}}\left(\sqrt{\lambda_q} - \sqrt{\lambda_{q+1}}\right).$$

Theorem A.6: (Distance to spectral sets in $\mathcal{P}_2$)

Let $G \subset \mathbf{R}^d_+$ be a symmetric set. Define now the set of measures

$$\mathcal{G} = \{\nu \in \mathcal{P}_2 : \lambda(\Sigma_\nu) \in G\}.$$

Then for any $\mu \in \mathcal{P}_2$, equality holds:

$$W_2(\mu, \mathcal{G}) = \min_{v \in G} \left\|\sqrt{\lambda(\Sigma_\mu)} - \sqrt{v}\right\|_2.$$

Indeed, we will prove a more general statement that applies to functions, with the distance replaced by the so-called Moreau envelope. We need some further notation to proceed. Let $(\mathcal{Y}, \mathsf{d})$ be a metric space and consider a function $f\colon \mathcal{Y} \to \mathbf{R} \cup \{+\infty\}$. Then for any parameter $\rho > 0$, the *Moreau envelope* and the *proximal map* of f [27, 47], respectively, are defined as:

$$f_\rho(y) := \inf_{y' \in \mathcal{Y}} \; f(y') + \tfrac{1}{2\rho}\mathsf{d}^2(y, y'),$$

$$\operatorname{prox}_{\rho f}(y) := \operatorname*{argmin}_{y' \in W} \; f(y') + \tfrac{1}{2\rho}\mathsf{d}^2(y, y').$$

In particular, if f is an indicator function of a set Q, then f_ρ reduces to the squared distance function to Q, while $\operatorname{prox}_{\rho f}(w)$ is the nearest point projection.

We will be interested in three metric spaces and it is important to keep the metric in mind in all results that follow.

- $(\mathcal{P}_2, W_2)$ The space $\mathcal{P}_2(\mathbf{R}^d)$ equipped with the Wasserstein-2 distance $W_2(\cdot, \cdot)$.
- $(\boldsymbol{S}^d_+, W_2)$ The cone of PSD matrices $\boldsymbol{S}^d_+$ equipped with the Bures-Wasserstein distance

  $$W_2^2(A, B) = \operatorname{tr} A + \operatorname{tr} B - 2\operatorname{tr}(A^{1/2} B A^{1/2}).$$

- $(\mathbf{R}^d_+, W_2)$ The cone of nonnegative vectors $\mathbf{R}^d_+$ equipped with the Hellinger distance

  $$W_2(x, y) = \|\sqrt{x} - \sqrt{y}\|_2,$$

  where the square root is applied elementwise.

Notice that we are abusing notation by using the same symbol W_2 to denote the metric in all three spaces. The reason we are justified in doing so is that the three metric spaces are related by isometric embedding. Namely, the Wasserstein-2 distance between two Gaussian distributions $\mu = N(0, \Sigma_\mu)$ and $\nu = N(0, \Sigma_\nu)$ coincides with the Bures-Wasserstein distance between their covariance matrices:

$$W_2(\mu, \nu) = W_2(\Sigma_\mu, \Sigma_\nu).$$

Similarly, the Bures-Wasserstein metric restricted to diagonal PSD matrices is the Hellinger distance:

$$W_2(\operatorname{Diag}(x), \operatorname{Diag}(y)) = W_2(x, y).$$

The following is the main result of this section.

Theorem A.7: (Diagonal reduction)

Consider a symmetric function $f\colon \mathbf{R}^d_+ \to \mathbf{R} \cup \{+\infty\}$ and define the spectral function $F\colon \mathcal{P}_2 \to \mathbf{R} \cup \{+\infty\}$ by setting $F(\mu) = f(\lambda(\Sigma_\mu))$. Then equality holds:

$$F_\rho(\mu) = f_\rho(\lambda(\Sigma_\mu)) \qquad \forall \mu \in \mathcal{P}_2(\mathbf{R}^d).$$

Theorem A.6 follows immediately by applying Theorem A.7 to indicator functions $\delta_{\mathcal{G}}$ and δ_G. The rest of the section is devoted to proving Theorem A.7.

A.2.1 Proof of Theorem A.7

We begin with some notation. The symbol $O(d)$ will denote the set of orthogonal $d \times d$ matrices. The singular values for any matrix $A \in \mathbf{R}^{m \times n}$ (with $m \le n$) in nonincreasing order we written as

$$\sigma_1(A) \ge \sigma_2(A) \ldots \ge \sigma_m(A).$$

We say that two matrices A and B admit a *simultaneous ordered singular-value decomposition (SVD)* if there exist matrices $U \in O(m),\ V \in O(n)$ satisfying $U^\top A V = \operatorname{Diag}(\sigma(A))$ and $U^\top B V =$

$\mathrm{Diag}(\sigma(B))$. The following trace inequality, essentially due to [64, 67], will play a central role in the section. The theorem as stated, along with a proof, may be found in [42, Theorem 4.6].

Theorem A.8 (von Neumann-Theobald)**.** *Any two matrices $A, B \in \mathbf{R}^{m\times n}$ satisfy*

$$\langle \sigma(A), \sigma(B)\rangle \geq \langle A, B\rangle.$$

Moreover, equality holds if and only if A and B admit a simultaneous ordered SVD.

The following lemma shows that the Burer-Wasserstein distance can be written in terms of a Procrustes problem [7, Theorem 1].

Lemma A.9 (Procrustes distance)**.** *For any two matrices $A, B \in \mathcal{S}^d_+$ equality holds:*

$$W_2(A, B) = \min_{U\in O(d)} \|A^{1/2} - B^{1/2}U\|_F$$

We will also need the following variational form of the Burer-Wasserstein distance; see for example [7, Section 3].

Lemma A.10 (Variational form)**.** *For any two matrices $A, B \in \boldsymbol{S}^d_+$ equality holds:*

$$W_2^2(A, B) = \min_{x,y:\ \mathbb{E}[xx^\top]=A,\, \mathbb{E}[yy^\top]=B} \mathbb{E}\,\|x - y\|_2^2$$

With these results in place, we are ready to start proving Theorem A.7. To this end, we will first establish the theorem in the Gaussian setting and then deduce the general case by a reduction. As the first step, we will estimate the W_2-distance of a matrix to an orbit of B under conjugation:

$$\mathcal{O}(B) := \{VBV^\top : V \in O(d)\}.$$

Lemma A.11 (Distance to orbit: Gaussian case)**.** *For any two matrices $A, B \in \boldsymbol{S}^d_+$ it holds:*

$$W_2(A, \mathcal{O}(B)) = \left\|\sqrt{\lambda(A)} - \sqrt{\lambda(B)}\right\|_2. \tag{41}$$

Moreover, the set of nearest points of $\mathcal{O}(B)$ to A is given by

$$\{U\,\mathrm{Diag}(\lambda(B))U^\top : A = U\,\mathrm{Diag}(\lambda(A))U^\top,\ U \in O(d)\}. \tag{42}$$

Proof. To see the inequality $\leq$, consider an eigenvalue decomposition $A = U\,\mathrm{Diag}(\lambda(A))U^\top$ for some $U \in O(d)$. Then we have

$$W_2(A, \mathcal{O}(B)) \leq W_2(U\,\mathrm{Diag}(\lambda(A))U^\top, U\,\mathrm{Diag}(\lambda(B))U^\top) = W_2(\lambda(A), \lambda(B)),$$

as claimed. Next, we show the reverse inequality $\geq$. To this end, set $\bar{A} := A^{1/2}$ and $\bar{B} := B^{1/2}$. Then for any $V \in O(d)$, we successively compute

$$\begin{aligned}
W_2(A, VBV^\top) &= \inf_{U\in O(d)} \|\bar{A} - (V\bar{B}V^\top)U\|_F^2 &(43)\\
&= \|\bar{A}\|^2 + \|\bar{B}\|^2 - 2\sup_{U\in O(d)} \langle \bar{A}, V\bar{B}(V^\top U)\rangle &(44)\\
&= \|\bar{A}\|^2 + \|\bar{B}\|^2 - 2\sup_{Z\in O(d)} \langle \bar{A}, V\bar{B}Z^\top\rangle &(45)\\
&\geq \|\bar{A}\|^2 + \|\bar{B}\|^2 - 2\langle \sigma(\bar{A}), \sigma(\bar{B})\rangle &(46)\\
&= \|\lambda(\bar{A})\|^2 + \|\lambda(\bar{B})\|^2 - 2\langle \lambda(\bar{A}), \lambda(\bar{B})\rangle &(47)\\
&= \|\lambda(\bar{A}) - \lambda(\bar{B})\|_2^2 &\\
&= \left\|\sqrt{\lambda(A)} - \sqrt{\lambda(B)}\right\|_2^2, &(48)
\end{aligned}$$

where (43) follows from Lemma A.9, the estimate (44) follows from expanding the Frobenius norm, (45) uses the variable substitution $Z = V^\top U$, the estimate (46) follows from von Neumann's trace inequality (Lemma A.8), and (47) uses the fact that eigenvalues and singular values coincide for PSD matrices. Taking the infimum over $V \in O(d)$ shows the claimed inequality $\geq$ in (41).

Next, the fact that any matrix in (42) is a nearest point of $\mathcal{O}(B)$ to A follows directly from the expression (41). To see the converse, observe that $VBV \in O(B)$ is the closest point to A if and only if the chain of inequalities (43)-(48) holds as equalities. Since the only inequality appears in (46), applying Lemma A.8 we see that equality holds if and only if there exist matrices $M_1, M_2, Z \in O(d)$ such that

$$\bar{A} = M_1 \operatorname{Diag}(\lambda(\bar{A}))M_2^\top \qquad \text{and} \qquad V\bar{B}Z^\top = M_1 \operatorname{Diag}(\lambda(\bar{B}))M_2^\top.$$

In particular, multiplying each equation by its transpose yields the expressions

$$A = M_1 \operatorname{Diag}(\lambda(A))M_1^\top \qquad \text{and} \qquad VBV^\top = M_1 \operatorname{Diag}(\lambda(B))M_1^\top,$$

thereby concluding the proof. □

We are now ready to complete the proof of Theorem A.7 in the Gaussian setting. In the proof, we will use the basic fact that for any vectors $v, w \in \mathbf{R}^d$, the inequality holds:

$$\|v^\uparrow - w^\uparrow\|_2 \leq \|v - w\|_2, \tag{49}$$

where $v^\uparrow$ and $w^\uparrow$ are the vectors obtained by permuting the coordinates of v and w to be nonincreasing.

Theorem A.12 (Envelope and prox-map in Gaussian space)**.** *Consider a symmetric function $f\colon \mathbf{R}^d_+ \to \mathbf{R} \cup \{+\infty\}$ and define the function on PSD matrices $F\colon \mathcal{S}^d_+ \to \mathbf{R} \cup \{+\infty\}$ by setting $F(A) = f(\lambda(A))$. Then, equality holds:*

$$F_\rho(A) = f_\rho(\lambda(A)) \qquad \forall A \in \boldsymbol{S}^d_+.$$

Moreover, the following expression holds:

$$\operatorname{prox}_{\rho F}(A) = \{U \operatorname{Diag}(v)U^\top : v \in \operatorname{prox}_{\rho f}(\lambda(A)), A = U \operatorname{Diag}(\lambda(A))U^\top,\ U \in O(d)\}.$$

Proof. For any matrix $A \in \boldsymbol{S}^d_+$, we successively compute

$$\begin{aligned}
F_\rho(A) &= \inf_{B \succeq 0}\ F(A) + \tfrac{1}{2\rho}W_2^2(A, B)\\
&= \inf_{v \geq 0} \inf_{B \in \mathcal{O}(\operatorname{Diag}(v))}\ f(\lambda(B)) + \tfrac{1}{2\rho}W_2^2(A, B)\\
&= \inf_{v \geq 0}\ f(v) + \tfrac{1}{2\rho} \inf_{B \in \mathcal{O}(\operatorname{Diag}(v))} W_2^2(A, B)\\
&= \inf_{v \geq 0}\ f(v) + \tfrac{1}{2\rho}W_2^2(A, \mathcal{O}(\operatorname{Diag}(v)))\\
&= \inf_{v \geq 0}\ f(v) + \tfrac{1}{2\rho}\|\sqrt{\lambda(A)} - \sqrt{v^\uparrow}\|_2^2 &(50)\\
&= \inf_{v \geq 0}\ f(v) + \tfrac{1}{2\rho}W_2^2(\lambda(A), v) &(51)\\
&= f_\rho(\lambda(A)),
\end{aligned}$$

where (50) follows from Lemma A.11 and (51) follows from (49). Next, observe from the chain of equalities that $B = U \operatorname{Diag}(v)U^\top$ lies in $\operatorname{prox}_{\rho F}(A)$ for some $U \in O(d)$ if, and only if, v lies in $\operatorname{prox}_{\rho f}(\lambda(A))$ and equality $d_2(A, B) = d_2(A, \mathcal{O}(\operatorname{Diag}(v)))$ holds. Appealing to Lemma A.11, this equality holds if and only if we may write $A = U \operatorname{Diag}(\lambda(A))U^\top$. Thus the proof is complete. □

We now move on to establishing Theorem A.7 in full generality, i.e. outside the Gaussian setting. We begin by extending Lemma A.11.

Lemma A.13 (Distance to orbit: general case)**.** *Fix a matrix $B \in \boldsymbol{S}^d_+$ and define the set of measures*

$$\mathcal{M} := \{\nu \in \mathcal{P}_2 : \ \Sigma_\nu \in \mathcal{O}(B)\}.$$

Then any zero-mean measure $\mu \in \mathcal{P}_2$ satisfies:

$$W_2^2(\mu, \mathcal{M}) = W_2^2(\Sigma_\mu, \mathcal{O}(B)). \tag{52}$$

Proof. We suppose first that Σ_μ is positive definite. Observe now

$$\begin{aligned} W_2(\mu, \mathcal{M}) &= \inf_{\pi \in \Pi(\mu,\nu),\, \nu \in \mathcal{M}} \mathbb{E}_\pi \|x - y\|_2^2 \\ &\geq \inf_{\mathbb{E}[xx^\top] = \Sigma_\mu,\ \mathbb{E}[yy^\top] \in \mathcal{O}(B)} \mathbb{E}\|x - y\|_2^2 = d_2(\Sigma_\mu, \mathcal{O}(B)), \end{aligned}$$

where the inequality follows from Lemma A.10. To see the reverse inequality, consider an eigenvalue decomposition $\Sigma_\mu = V \operatorname{Diag}(\lambda(\Sigma_\mu))V^\top$ for some $V \in O(d)$. Define now the matrix $\bar{B} := V\operatorname{Diag}(\lambda(B))V^\top$ and set $T := \bar{B}^{1/2}\Sigma_\mu^{-1/2}$. Then, clearly the measure $\nu := T_\# \mu$ satisfies $\mathbb{E}_{z\sim\nu}[zz^\top] = T\,\mathbb{E}_\mu[xx^\top]T^\top = T\Sigma_\mu T^\top = \bar{B}$ and therefore $T_\#\mu$ lies in $\mathcal{M}$. Thus, we conclude

$$\begin{aligned} W_2^2(\mu, \mathcal{M}) \leq W_2^2(\mu, \nu) &= \mathbb{E}_\mu \|x - Tx\|_2^2 \\ &= \mathbb{E}_\mu \|x\|^2 - 2\,\mathbb{E}_\mu \operatorname{tr}(Txx^\top) + \mathbb{E}_\mu \operatorname{tr}(T^\top T xx^\top) \\ &= \operatorname{tr}(\Sigma_\mu) - 2\operatorname{tr}(\bar{B}^{1/2}\Sigma_\mu^{1/2}) + \operatorname{tr}(\bar{B}) \\ &= \sum_{i=1}^d \left(\sqrt{\lambda_i(\Sigma_\mu)} - \sqrt{\lambda_i(B)}\right)^2, \end{aligned}$$

Using Lemma A.11 yields the claimed expression (52).

Finally, if Σ_μ is not positive definite, then μ is almost surely supported on some subspace $\mathcal{L}$. We now define the perturbed distribution $\mu_t = \mu \times g_t$ where g_t is a zero-mean Gaussian distribution supported on $\mathcal{L}^\perp$ with covariance $t \cdot I$. Then, by Lemma A.2, $W_2(\mu_t, \mu) = W_2(\mu_t, (\mathsf{P}_{\mathcal{L}})_\#\mu_t) \to 0$, which in turn implies $\mathbb{E}_{\mu_t} xx^\top \to \mathbb{E}_\mu xx^\top$. From (52) we have $W_2^2(\mu_t, \mathcal{M}) = W_2^2(\mathbb{E}_{\mu_t} xx^\top, \mathcal{O}(B))$. Letting t go to zero, we deduce the desired equality $W_2^2(\mu, \mathcal{M}) = W_2^2(\Sigma_\mu, \mathcal{O}(B))$. □

The proof of Theorem A.7 now proceeds in exactly the same way as that of Theorem A.12, with Lemma A.13 being used instead of Lemma A.11.

Proof of Theorem A.7. The argument is essentially the same as in Theorem A.12. We detail it here for completeness. Define the matrix $A := \Sigma_\mu$. We successively compute

$$\begin{aligned} F_\rho(\mu) &= \inf_{\nu \in \mathcal{P}_2} \ F(\nu) + \tfrac{1}{2\rho} W_2^2(\mu, \nu) \\ &= \inf_{u \geq 0} \inf_{\nu:\, \Sigma_\nu \in \mathcal{O}(\operatorname{Diag}(u))} f(u) + \tfrac{1}{2\rho} W_2^2(\mu, \nu) \\ &= \inf_{u \geq 0} \ f(u) + \tfrac{1}{2\rho} \inf_{\nu:\, \Sigma_\nu \in \mathcal{O}(\operatorname{Diag}(u))} W_2^2(\mu, \nu) \\ &= \inf_{u \geq 0} \ f(u) + \tfrac{1}{2\rho} W_2^2(A, \mathcal{O}(\operatorname{Diag}(u))) && (53) \\ &= \inf_{v \geq 0} \ f(v) + \tfrac{1}{2\rho} \|\sqrt{\lambda(A)} - \sqrt{u^\uparrow}\|_2^2 && (54) \\ &= \inf_{u \geq 0} \ f(u) + \tfrac{1}{2\rho} W_2^2(\lambda(A), u) && (55) \\ &= f_\rho(\lambda(A)), \end{aligned}$$

where (53) follows from Lemma A.13, the estimate (54) follows from Lemma A.11, and (55) follows from (49). This completes the proof. □

Connection to existing literature. The results presented in this section have close analogues in the existing literature in matrix analysis and optimization. Namely a function $F\colon \boldsymbol{S}^d \to \mathbf{R}\cup\{+\infty\}$ is called *orthogonally invariant* (or *spectral*) if the equality holds:

$$F(UXU^\top) = F(X) \qquad \forall X \in \boldsymbol{S}^d,\, U \in O(d).$$

Evidently such functions are fully described by their restriction to diagonal matrices. More precisely, a function F is orthogonally invariant if and only if there exists a symmetric function $f\colon \mathbf{R}^d \to \mathbf{R}\cup\{+\infty\}$ satisfying $F(X) = f(\lambda(X))$. A pervasive theme in the study of such functions is that various variational properties of the permutation-invariant function f are inherited by the induced spectral function $F = f\circ\lambda$; see e.g. [22, 23, 25, 40, 41, 42, 60, 72]. For example, f convex if and only if F is convex [25, 40], f is C^p-smooth if and only of F is C^p-smooth [60, 63, 72], and so forth. A useful result in this area is the expression for the Moreau envelope obtained in [22, 33]

$$F_\rho(X) = f_\rho(\lambda(X)) \qquad \forall X \in \boldsymbol{S}^d. \tag{56}$$

For example, as explained in [33], it readily yields expressions relating generalized derivatives of f and F. Crucially, in (56) the Moreau envelope F_ρ is computed with respect to the Frobenius norm on $\boldsymbol{S}^d$ and the Moreau envelope f_ρ is computed with respect to the ℓ_2 norm on $\mathbf{R}^d$. Thus the results presented in the section extend this circle of ideas to functions defined on the Wasserstein space.

B Proofs from Section 2

B.1 Proof of Proposition 2.1

Suppose otherwise that there exists a sequence $\nu_i \in \mathcal{Q}'$ with $W_2(\nu_i,\mu) \le r$ satisfying $\mathrm{REG}(\nu_i) \to \infty$. Then from (7) we deduce $W_2(\nu_i,\mathcal{E}) = \mathrm{RSE}(\nu_i) \to 0$. Subsequently, using the triangle inequality yields $\mathrm{RSE}(\mu) = W_2(\mu,\mathcal{E}) \le W_2(\nu_i,\mu) + W_2(\nu_i,\mathcal{E})$. Letting i tend to infinity, we deduce $W_2(\mu,\mathcal{E}) \le r$, which is a contradiction. Thus no such sequence ν_i exists and (8) holds for some $M > 0$.

Suppose now that for some $q > 0$, the inequality $c \ge \mathrm{REG}(\nu)\cdot\mathrm{RSE}(\nu)^q$ holds. Then, similarly as above, the triangle inequality for any $\nu \in \mathcal{Q}'\setminus\mathcal{E}$ yields

$$\mathrm{RSE}(\mu) = W_2(\mu,\mathcal{E}) \le W_2(\nu,\mu) + W_2(\nu,\mathcal{E}) \le r + \mathrm{RSE}(\nu) \le r + (c/\mathrm{REG}(\nu))^{1/q}.$$

Rearranging yields the estimate $\mathrm{REG}(\nu) \le c(\mathrm{RSE}(\mu) - r)^{-1/q}$, thereby completing the proof.

B.2 Proof of Theorem 2.2

In this section, we prove Theorem 2.2, or rather a stronger version thereof. To this end, we fix a C^1-smooth map $F\colon \mathbf{R}^d \to \boldsymbol{S}^k_+$ satisfying that it is integrable with respect to any measure $\mu \in \mathcal{P}_2$. We define the function $\mathcal{J}\colon \mathcal{P}_2 \to \mathbf{R}$ by setting

$$\mathcal{J}(\mu) = \lambda_{\min}(\mathbb{E}_\mu F(x)).$$

The differential of F will be denoted by $DF(x)\colon \mathbf{R}^d \to \boldsymbol{S}^k$, while the symbol $DF(x)^*\colon \boldsymbol{S}^k \to \mathbf{R}^d$ will denote the adjoint linear map of $DF(x)$. We further assume that there exists a constant $L > 0$ such that $\|DF(x)\|_{\mathrm{op}} \le L(1+\|x\|)$ for all x.

In order to simplify notation, for any matrix $A \in \boldsymbol{S}^k$, we let $E_k(A)$ denote the set of all unit eigenvectors of A corresponding to the minimal eigenvalue $\lambda_{\min}(A)$. We will also use the elementary fact that $\lambda_{\min}$ is a concave function on $\boldsymbol{S}^k$ and its supdifferential at any matrix A is the set

$$\partial\lambda_{\min}(A) = \mathrm{conv}\{uu^\top : u \in E_k(A)\}.$$

See for example [9, Corollary 5.2.3, Corollary 5.2.4 (iii)]. Abusing notation, we will set $E_k(\mu) :=$

$E_k(\mathbb{E}_\mu F(x))$ for any measure $\mu \in \mathcal{P}_2$.

The proof of Theorem 2.2 will be subdivided into two parts, corresponding to the two inequalities in (11). We begin by establishing the first inequality $\mathrm{REG}(\mu)^{q_2-1} \lesssim \mathrm{RSE}(\mu)$. The proof amounts to simply applying the fundamental theorem of calculus to the function $\mathcal{J}$ along a geodesic μ_t joining a measure μ to its nearest point in $\mathcal{E}$. This is the content of the following theorem.

Theorem B.1 (Small distance implies small value)**.** *Fix a measure $\mu \in \mathcal{P}_2$, constants $c, \varepsilon > 0$, and a power $q \in [0,1)$. Suppose that for all measures ν satisfying $W_2(\nu, [\mathcal{J}=0]) \leq W_2(\mu, [\mathcal{J}=0]) + \varepsilon$, the estimate holds:*

$$\min_{u \in E_k(\nu)} \quad \mathbb{E}_\nu \, \|DF(x)^*[uu^\top]\|_2^2 \leq c \cdot \mathcal{J}(\nu)^{2q}.$$

Then, the inequality holds:

$$W_2(\mu, [\mathcal{J}=0]) \geq \tfrac{1}{(1-q)\sqrt{c}} \cdot \mathcal{J}(\mu)^{1-q}.$$

Proof. Fix a measure $\nu \in [\mathcal{J}=0]$ satisfying $W_2(\mu,\nu) \leq W_2(\mu, [\mathcal{J}=0]) + \varepsilon$ and let $\pi \in \Pi(\nu,\mu)$ be an optimal transport plan between ν and μ. Define the functions $\pi_t(x,y) = (1-t)x + ty$. Then the curve $\mu_t = (\pi_t)_\# \mu$ is a constant speed geodesic between μ and ν [59, Theorem 5.27]:

$$W_2(\mu_t, \mu_s) = |t-s| \cdot W_2(\mu,\nu) \qquad \forall t, s \in [0,1].$$

Define the curve of matrices $\gamma(t) = G(\mu_t)$. We would like to now compute $\dot{\gamma}(t)$ by exchanging differentiation and integration in the expression:

$$\dot{\gamma}(t) = \frac{d}{dt} \mathop{\mathbb{E}}_{(x,y)\sim\pi} F((1-t)x + ty). \tag{57}$$

To this end, we bound the derivative of the integrand:

$$\|DF((1-t)x+ty)[y-x]\|_{\mathrm{op}} \leq L(1 + \|x\|_2 + \|y\|_2)\|y-x\|_2.$$

Applying Hölder's inequality, we see that the right side is π-integrable. Therefore, exchanging integration and differentiation in (57) yields the expression $\dot{\gamma}(t) = \mathbb{E}_{(x,y)\sim\pi} DF((1-t)x+ty)[y-x]$.

It is straightforward to see that γ is absolutely continuous and therefore using the subdifferential chain rule for concave functions [11, Lemma 3.3, p. 73], we deduce that for almost every $t \in (0,1)$ we have

$$\frac{d}{dt}\mathcal{J}(\mu_t) = \frac{d}{dt}(\lambda_{\min} \circ \gamma)(t) = \langle U_t, \dot{\gamma}(t)\rangle \qquad \forall U_t \in \partial\lambda_{\min}(\gamma(t)). \tag{58}$$

In particular, for each such t we may choose U_t satisfying the running assumption $\mathbb{E}_{\mu_t} \|DF(x)^*[U_t]\|_2^2 \leq c \cdot \mathcal{J}(\mu_t)^{2q}$. Continuing with (58), we successively compute

$$\begin{aligned}
\frac{d}{dt}\mathcal{J}(\mu_t) &= \mathop{\mathbb{E}}_{(x,y)\sim\pi} \langle DF((1-t)x+ty)^*[U_t], y-x\rangle \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim\pi} \|DF((1-t)x+ty)^*[U_t]\|_2 \cdot \|y-x\|_2 \\
&\leq \sqrt{\mathop{\mathbb{E}}_{(x,y)\sim\pi} \|DF((1-t)x+ty)^*[U_t]\|_2^2} \cdot \sqrt{\mathop{\mathbb{E}}_{(x,y)\sim\pi} \|y-x\|^2} \qquad (59) \\
&\leq \sqrt{c} \cdot W_2(\mu,\nu) \cdot \mathcal{J}(\mu_t)^q,
\end{aligned}$$

where (59) follows from Hölder's inequality. Raising $\mathcal{J}(\mu_t)$ to power $1-q$, we deduce

$$\frac{d}{dt}\mathcal{J}(\mu_t)^{1-q} \leq \sqrt{c}(1-q) \cdot W_2(\mu,\nu) \qquad \text{for a.e. } t \in (0,1).$$

Integrating both sides from $t=0$ to $t=1$, we conclude

$$\mathcal{J}(\mu)^{1-q} - \mathcal{J}(\nu)^{1-q} \leq \sqrt{c}(1-q) \cdot W_2(\mu,\nu).$$

Taking into account the equality $\mathcal{J}(\nu) = 0$ and the estimate $W_2(\mu,\nu) \le W_2(\mu, [\mathcal{J} = 0]) + \varepsilon$, we may now let ε tend to zero 0 thereby completing the proof. $\square$

Next, we pass to the reverse inequality $\mathrm{RSE}(\mu) \lesssim \mathrm{REG}(\mu)^{q_1 - 1}$, which is a more substantive conclusion. The main tool we will use is the characterization of an "error bound property" using the slope. In what follows, for any real number r, the symbol $r_+ = \max\{0, r\}$ denotes its positive part.

Definition B.2 (Slope)**.** Consider a function $f\colon \mathcal{X} \to \mathbf{R} \cup \{+\infty\}$ defined on a metric space $(\mathcal{X}, \mathsf{d})$. The *slope* of f at any point x with $f(x)$ finite is defined by

$$|\nabla f|(x) = \limsup_{x' \to x} \frac{(f(x) - f(x'))_+}{\mathsf{d}(x, x')}.$$

Importantly, if a slope is large on a neighborhood, then the function must decrease significantly. This is the content of the following theorem; see [37, Basic Lemma, Chapter 1]) or [34, Lemma 2.5].

Theorem B.3 (Decrease principle)**.** *Consider a lower semicontinuous function $f\colon \mathcal{X} \to \mathbf{R} \cup \{+\infty\}$ on a complete metric space $(\mathcal{X}, \mathsf{d})$. Fix a point x with $f(x)$ finite, and suppose that there are constants $\alpha < f(x)$ and $r, \kappa > 0$ so that the implication holds:*

$$\alpha < f(u) \le f(x) \quad \text{and} \quad \mathsf{d}(u, x) \le r \qquad \Longrightarrow \qquad |\nabla f|(u) \ge \kappa.$$

If in addition the inequality $f(x) - \alpha < \kappa r$ is valid, then the estimate holds:

$$\mathsf{d}(x, [f \le \alpha]) \le \kappa^{-1}(f(x) - \alpha).$$

We will apply this theorem to the function $\mathcal{J}(\mu)^{1-q}$. The key step therefore is to compute the slope of $\mathcal{J}$. This is the content of the following lemma.

Lemma B.4 (Slope computation)**.** *Suppose that F satisfies $\|DF(x)\|_{\mathrm{op}} \le L(1 + \|x\|_2)$ for all $x \in \mathbf{R}^d$, where L is some constant. Then, for any measure $\mu \in \mathcal{P}_2$, the estimate holds:*

$$|\nabla \mathcal{J}|(\mu) \ge \sup_{u \in E_k(\mu)} \sqrt{\mathbb{E}_\mu \|DF(x)^*[uu^\top]\|_2^2}.$$

Proof. We begin by writing $\mathcal{J}(\mu) = (\lambda_{\min} \circ G)(\mu)$, where we define the map $G(\mu) := \mathbb{E}_\mu F(x)$. Next, fix a measure $\mu \in \mathcal{P}_2$ and a matrix $U \in \partial \lambda_{\min}(G(\mu))$, and define the transport map $T(x) = x - DF(x)^*[U]$. Clearly, we may assume that $DF(x)^*[U]$ is not μ-almost surely zero, since otherwise the theorem holds trivially for $U = uu^\top$. Observe that $I - T$ is square μ-integrable since

$$\mathbb{E}_\mu \|x - T(x)\|_2^2 = \mathbb{E}_\mu \|DF(x)^*[U]\|_2^2 \le L^2 \cdot \|U\|_F^2 \cdot \mathbb{E}_\mu (1 + \|x\|_2)^2 < \infty.$$

Define now the curve $\gamma\colon [0,1) \to \mathcal{P}_2$ by setting $\gamma(t) = (I + t(T - I))_\# \mu$. Note from (34), we have

$$W_2^2(\gamma(t), \gamma(0)) \le t^2\, \mathbb{E}_\mu \|x - T(x)\|_2^2. \tag{60}$$

Next, from concavity of $\lambda_{\min}$ we deduce

$$\mathcal{J}(\gamma(t)) - \mathcal{J}(\gamma(0)) \le \langle U, (G \circ \gamma)(t) - (G \circ \gamma)(0) \rangle. \tag{61}$$

We would like to compute $\frac{d}{dt} \langle U, G \circ \gamma(t) \rangle$ by exchanging integration/differentiation in the expression:

$$\frac{d}{dt} \langle U, G \circ \gamma(t) \rangle = \frac{d}{dt}\, \mathbb{E}_\mu \langle U, F(x + t(T(x) - x)) \rangle. \tag{62}$$

To this end, we bound the derivative of the integrand uniformly in t:

$$\begin{aligned}
|\langle U, DF(x+t(T(x)-x))[T(x)-x]\rangle| &\leq \|DF(x+t(T(x)-x))[T(x)-x]\|_2 \\
&\leq L(1+\|x+t(T(x)-x)\|_2 \cdot \|T(x)-x\|_2) \\
&\leq L(1+\|x\|+t\|T(x)-x\|_2)\|T(x)-x\|_2 \\
&\leq L\|T(x)-x\|_2 + \frac{L}{2}\|x\|_2^2 + \frac{L+2}{2}\|x-T(x)\|_2^2.
\end{aligned}$$

Clearly, the right-side is μ-integrable and therefore by the dominated convergence theorem, we may exchange integration and differentiation in (62) yielding:

$$\frac{d}{dt}\langle U, G\circ\gamma(t)\rangle|_{t=0} = \mathbb{E}_\mu \langle DF(x)^*[U], T(x)-x\rangle = -\mathbb{E}_\mu \|T(x)-x\|_2^2. \tag{63}$$

In particular, we deduce $(\mathcal{J}\circ\gamma)(t) < (\mathcal{J}\circ\gamma)(0)$ for all small $t>0$. Therefore, dividing (61) by $W_2(\gamma(t),\gamma(0))$ and taking the limit as $t\to 0$ yields

$$\begin{aligned}
|\nabla\mathcal{J}|(\mu) &\geq \lim_{t\to 0}\frac{\mathcal{J}(\gamma(0))-\mathcal{J}(\gamma(t))}{W_2(\gamma(0),\gamma(t))} \\
&\geq \lim_{t\to 0}\frac{\mathcal{J}(\gamma(0))-\mathcal{J}(\gamma(t))}{t\sqrt{\mathbb{E}_\mu \|x-T(x)\|_2^2}} && (64)\\
&= \left\langle V, \lim_{t\to 0}\frac{(G\circ\gamma)(0)-(G\circ\gamma)(t)}{t\sqrt{\mathbb{E}_\mu \|x-T(x)\|_2^2}}\right\rangle \\
&= \sqrt{\mathbb{E}_\mu \|T(x)-x\|_2^2} = \sqrt{\mathbb{E}_\mu \|DF(x)^*[U]\|_2^2}, && (65)
\end{aligned}$$

where the estimate (64) follows from (60) and the estimate (65) follows from (63). $\square$

Finally, combining the decrease principle (Theorem B.3) and the estimate on the slope of $\mathcal{J}$ (Lemma B.4) we arrive at the main result.

Theorem B.5 (Small value implies small distance)**.** *Suppose that $F\colon \mathbf{R}^d \to \boldsymbol{S}_+^d$ satisfies $\|DF(x)\|_{\mathrm{op}} \leq L(1+\|x\|_2)$ for all $x\in\mathbf{R}^d$, where L is some constant. Fix a constant $c>0$, a radius $r>0$, and a power $q\in[0,1)$. Suppose that for all measures $\nu\in[0<\mathcal{J}\leq\mathcal{J}(\mu)]\cap\mathbb{B}_2(\mu;r)$, the estimate holds:*

$$\sup_{u\in E_k(\nu)} \mathbb{E}_\nu \|DF(x)^*[uu^\top]\|_2^2 \geq c\cdot\lambda_{\min}(\mathbb{E}_\nu F(x))^{2q}.$$

Then, the inequality holds:

$$W_2(\mu,[\mathcal{J}=0]) \leq \tfrac{1}{(1-q)\sqrt{c}}\cdot\mathcal{J}(\mu)^{1-q},$$

as long as r is large enough so that $\mathcal{J}(\mu)^{1-q} < (1-q)r\sqrt{c}$.

Proof. It follows immediately from [66, Theorem 6.9] and continuity of $\lambda_{\min}(\cdot)$ that the function $\mathcal{J}$ is continuous. Define the function $\mathcal{G}(\nu) := \mathcal{J}(\nu)^{1-q}$ and note that the standard chain rule implies

$$|\nabla\mathcal{G}|(\nu) = (1-q)\cdot\frac{|\nabla\mathcal{J}|(\nu)}{\mathcal{J}(\nu)^q} \geq (1-q)\sqrt{c},$$

whenever $0<\mathcal{G}(\nu)\leq\mathcal{G}(\mu)$ and $W_2(\nu,\mu)\leq r$. Applying Theorem B.3 to $\mathcal{G}$ with $\alpha=0$ and $\kappa=(1-q)\sqrt{c}$ completes the proof. $\square$

Theorem 2.2 follows immediately from Theorems B.1 and B.5.

C Proofs from Section 3

C.1 Proof of Lemma 3.1

We first argue the inclusion $\supset$. Observe that for any measure $\mu \in \mathcal{P}_2^\circ$ with $\lambda_1(\Sigma_\mu) = \lambda_2(\Sigma_\mu)$, the set of maximizers of (12) is the intersection of a sphere and the top eigenspace of Σ_μ. Since none of the maximizers are isolated, they are unstable and therefore μ lies in $\mathcal{E}$.

To see the reverse inclusion $\subset$, fix a measure $\mu \in \mathcal{P}_2^\circ$ and suppose that the top two eigenvalues λ_1 and λ_2 of Σ_μ are distinct. Then the normalized top eigenvector v of Σ_μ is the unique maximizer of (12). It remains to verify that v is a tilt-stable maximizer of (12). To this end, define the Lagrangian function $\mathcal{L}(v, \lambda) = -\frac{1}{2}\mathbb{E}_\mu\langle x, v\rangle^2 + \frac{\lambda}{2}(\|v\|^2 - 1)$. Then equalities hold:

$$\nabla_v \mathcal{L}(v, \lambda) = (\lambda I - \Sigma_\mu)v \qquad \text{and} \qquad \nabla^2_{vv}\mathcal{L}(v, \lambda) = \lambda I - \Sigma_\mu.$$

In particular, the first equation shows that the optimal Lagrange multiplier λ is λ_1. Let $V \in \mathbf{R}^{d\times(d-1)}$ be a matrix with an eigenbasis for $v^\perp$ as its columns. Then an elementary computation yields

$$\min_{y\in\mathbb{S}^{d-2}} \langle \nabla^2_{vv}\mathcal{L}(v, \lambda)(Vy), (Vy)\rangle = \min_{y\in\mathbb{S}^{d-2}} \sum_{i=2}^{d} (\lambda_1 - \lambda_i)y_i^2 = \lambda_1 - \lambda_2 > 0, \tag{66}$$

and therefore v is a tilt-stable maximizer of (12). Moreover, the claimed equality $\mathrm{REG}(\mu)^{-1} = \lambda_1(\Sigma_\mu) - \lambda_2(\Sigma_\mu)$ follows directly from (66), thereby completing the proof.

C.2 Proof of Theorem 3.2

This follows directly by applying Theorem A.6 with $\mathcal{G} = \mathcal{E}$ from Lemma 3.1 and $G = \{v \in \mathbf{R}^d_+ : v^{(1)} = v^{(2)}\}$, where $v^{(i)}$ denotes the i'th largest coordinate value of v.

C.3 Proof of Lemma 3.3

Henceforth, fix a measure $\mu \in \mathcal{P}_2^\circ$ and define the shorthand $\lambda := \lambda(\Sigma_\mu)$. Let's dispense first with the simple direction $\supset$. To this end, suppose that equality $\lambda_q = \lambda_{q+1}$ holds. Then we may form two sets of orthonormal bases $U := \{u_1, \ldots, u_q\}$ and $U' = \{u_1, \ldots, u_{q-1}, u_q'\}$ with $\langle u_q, u_q'\rangle = 0$, and which are contained in the span of the eigenspaces corresponding to the top q eigenvalues. We may further interpolate between the two bases with $U_t = \{u_1, \ldots, tu_q + (1-t)u_q'\}$ for $t \in (0, 1)$. The orthogonal projections onto the span of U_t furnish a path of optimal solutions, which are therefore not tilt-stable. Thus μ lies in $\mathcal{E}$, as claimed.

We now establish the reverse inclusion $\subset$. Suppose therefore that λ_q and λ_{q+1} are distinct. We begin by conveniently parameterizing the Grassmannian manifold $\mathrm{Gr}(q, d)$ as follows. Define the matrix $A := \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix}$. Then using [6, Section 2.1] we may write $\mathrm{Gr}(q, d)$ as the orbit of A under conjugation by orthogonal matrices:

$$\mathrm{Gr}(q, d) = \{UAU^\top : U \in O(d)\}.$$

Fix a skew symmetric matrix $W := \begin{bmatrix} W_1 & W_2 \\ -W_2^\top & W_4 \end{bmatrix}$ and define the curve $\gamma\colon \mathbf{R} \to \mathrm{Gr}(q, d)$ by

$$\gamma(t) = \exp(-tW) \cdot A \cdot \exp(tW).$$

Differentiating the curve yields the expression

$$\dot{\gamma}(0) = AW - WA = \begin{bmatrix} 0 & W_2 \\ W_2^\top & 0 \end{bmatrix}.$$

Moreover, [6, Section 2.3] shows that varying W among all skew-symmetric matrices yields the entire tangent space

$$T_{\mathrm{Gr}(q,d)}(A) = \left\{ \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix} : B \in \mathbf{R}^{q\times(d-q)} \right\}.$$

Now without loss of generality, we may assume that Σ_μ is diagonal, that is $\Sigma_\mu = \mathrm{Diag}(\lambda)$. Then clearly, $R = A$ is the unique maximizer of the problem (14). We now perform the second order expansion

$$(f\circ\gamma)(t) = \langle \gamma(t), \Sigma_\mu\rangle = f(A) + t\underbrace{\langle AW - WA, \Sigma_\mu\rangle}_{=0} + t^2\left\langle \frac{1}{2}(AW^2 + W^2A) - WAW, \Sigma_\mu \right\rangle + O(t^3).$$

In particular, we deduce

$$(f\circ\gamma)''(0) = \left\langle \frac{1}{2}(AW^2 + W^2A) - WAW, \Sigma_\mu \right\rangle.$$

Taking into account the definition of A, a quick computation show

$$\begin{aligned}\frac{1}{2}(AW^2 + W^2A) - WAW &= \begin{bmatrix} W_1^2 - W_2W_2^\top & \frac{1}{2}(W_2W_4 + W_1W_2) \\ -\frac{1}{2}(W_2W_4 + W_1W_2)^\top & 0 \end{bmatrix} - \begin{bmatrix} W_1^2 & W_1W_2 \\ -W_2^\top W_1 & -W_2^\top W_2 \end{bmatrix} \\ &= \begin{bmatrix} -W_2W_2^\top & \frac{1}{2}(W_2W_4 - W_1W_2) \\ -\frac{1}{2}(W_2W_4 - W_1W_2)^\top & W_2^\top W_2 \end{bmatrix}.\end{aligned}$$

Taking the trace product with the diagonal matrix Σ_μ yields

$$\begin{aligned}(f\circ\gamma)''(0) &= -\langle W_2W_2^\top, \mathrm{Diag}(\lambda_{1:q})\rangle + \langle W_2^\top W_2, \mathrm{Diag}(\lambda_{q+1:d})\rangle \\ &\leq -\langle W_2W_2^\top, \lambda_q I_q\rangle + \langle W_2^\top W_2, \lambda_{q+1} I_{d-q}\rangle \\ &= -(\lambda_q - \lambda_{q+1})\cdot \|W_2\|_F^2\end{aligned} \tag{67}$$

In particular, the covariant Hessian $\nabla^2_{\mathcal{M}} f(A)$ is negative definite on the tangent space $T_{\mathrm{Gr}(q,d)}(A)$. Therefore, A is a tilt-stable maximizer of the problem, as we had to show. Moreover, setting $W_2 = \mathrm{Diag}(w)$ with $w_q = w_{q+1} = 1$ and $w_i = 0$ for $i \notin \{q, q+1\}$, yields equality in (67). Therefore we deduce $\mathrm{REG}(\mu)^{-1} = \lambda_q(\Sigma_\mu) - \lambda_{q+1}(\Sigma_\mu)$ as claimed.

C.4 Proof of Theorem 3.4

This follows directly by applying Theorem A.6 with $\mathcal{G} = \mathcal{E}$ from Lemma 3.3 and $G = \{v \in \mathbf{R}^d_+ : v^{(q)} = v^{(q+1)}\}$, where $v^{(i)}$ denotes the i'th largest coordinate value of v.

D Proofs from Section 4

D.1 Proof of Lemma 4.1

For any $\mu \in \mathcal{Q}$, observe

$$\begin{aligned}\nabla f(\beta) = \mathbb{E}\left[(h'(\langle x, \beta\rangle) - y)\, x\right] &= \mathbb{E}\left[\mathbb{E}\left[h'(\langle x, \beta\rangle) - y \mid x\right] x\right] \\ &= \mathbb{E}\left[(h'(\langle x, \beta\rangle) - h'(\langle x, \beta_\star\rangle))x\right]\end{aligned}$$

Therefore, equality $\nabla f(\beta_\star) = 0$ holds for any distribution of x. Hence, $\beta_\star$ is critical for the problem with zero Lagrange multipliers $\lambda = 0$. Differentiating again yields the expression for the Hessian

$$H := \nabla^2 f(\beta_\star) = \sigma^{-2}\cdot \mathbb{E}\left[h''(\langle x, \beta_\star\rangle)xx^\top\right]. \tag{68}$$

Note that H is positive semidefinite since $h'' > 0$. Consequently, the set of ill-conditioned distributions $\mathcal{E}$ corresponds to those distributions on x for which $\ker(H)$ nontrivially intersects $\mathcal{T}$. Clearly, a vector v lies in $\ker(H)$ if and only if $0 = \langle Hv, v\rangle$, or equivalently $0 = \mathbb{E}\big[h''(\langle x, \beta_\star\rangle)\langle x, v\rangle^2\big]$. Taking into account the assumption $h'' > 0$, this occurs precisely when v lies in the nullspace of Σ_μ. Thus $\mathcal{E}$ consists of all measures $\mu \in \mathcal{Q}$ satisfying $\mathcal{T} \cap \ker(\Sigma_\mu) \neq \{0\}$. Finally, it follows directly from (68) that if for some $\alpha, \beta > 0$ the inequality $\alpha \leq h''(\langle x, \beta_\star\rangle) \leq \beta$ holds for μ-almost every x, then $\lambda_{\min}(\Sigma_\mu\,|_{\mathcal{T}}) \in \lambda_{\min}(\Sigma_\mu|_{\mathcal{T}}) \cdot [\alpha, \beta]$, thereby completing the proof.

D.2 Proof of Theorem 4.2

This follows directly from Lemma 4.1 and Theorem A.3.

E Proofs from Section 5.1

E.1 Proof of Lemma 5.1

We begin by verifying the claim for the formulation (21). To this end, a quick computation shows $\nabla f(\beta_\star) = 0$ and $\nabla^2 f(\beta_\star) = \mathbb{E}_\mu\Big[\langle x, \beta_\star\rangle^2 xx^\top\Big]$. Therefore we deduce

$$\lambda_{\min}(\nabla^2 f(\beta_\star)) = \min_{v \in \mathbb{S}^{d-1}} \mathbb{E}_\mu \langle x, \beta_\star\rangle^2 \langle x, v\rangle^2.$$

In particular, $\nabla^2 f(\beta_\star)$ if and only if the support of μ is contained in $\beta_\star^\perp \cup v^\perp$ almost surely.

Next, we verify the claim for the formulation (20). To this end, let $\mathcal{M}$ denote the set of symmetric PSD rank one matrices:

$$\mathcal{M} = \{M \in \boldsymbol{S}^d_+ : \operatorname{rank}(M) = 1\}.$$

A quick computation now yields

$$\nabla f(M) = \mathbb{E}\langle M - M_\star, xx^\top\rangle \qquad \text{and} \qquad \nabla^2 f(M_\star)[\Delta, \Delta] = \mathbb{E}\langle \Delta, xx^\top\rangle^2.$$

In particular, equality $\nabla f(M_\star) = 0$ holds and therefore the optimal Lagrange multipliers $\lambda_\star$ are zero. Hence the Hessian of the Lagrangian $\nabla^2 \mathcal{L}(M_\star, \lambda_\star)$ coincides with $\nabla^2 f(M_\star)$. Classically, if we form the factorization $M_\star = \beta_\star\beta_\star^\top$, then the tangent space to $\mathcal{M}$ at $M_\star$ can be written as

$$\mathcal{T} = \{\beta_\star v^\top + v\beta_\star^\top : v \in \mathbf{R}^d\}.$$

Consequently, for any $\Delta = \beta_\star v^\top + v\beta_\star^\top$ we compute

$$\nabla^2 f(M_\star)[\Delta, \Delta] = \mathbb{E}\langle \Delta, xx^\top\rangle^2 = 4\,\mathbb{E}\langle \beta_\star, x\rangle^2 \langle v, x\rangle^2.$$

Note that $\|\Delta\|_F^2 = 2(\beta_\star^\top v)^2 + 2\|\beta_\star\|^2\|v\|^2$ and therefore $\|\Delta\|_F^2/\|\beta_\star\|^2\|v\|^2 \in [2, 4]$. Therefore Δ is nonzero as long as w is nonzero. We therefore again deduce that $\nabla^2 f(M_\star)[\Delta, \Delta] = 0$ of some nonzero $\Delta \in T_{\mathcal{M}}(M_\star)$ if and only if the support of μ is contained in $\beta_\star^\perp \cup v^\perp$ almost surely. The claimed expression (22) follows immediately.

E.2 Proof of Theorem 5.2

Define the set $\mathcal{K}_v = \{\beta_\star\}^\perp \cup \{v\}^\perp$ for each unit vector $v \in \mathbb{S}^{d-1}$ and define the function $h\colon \mathbb{S}^{d-1} \to [0, \infty)$.

$$h(v) = \mathop{\mathbb{E}}_{x \sim \mu} \operatorname{dist}(x, \mathcal{K}_v)^2 = \mathop{\mathbb{E}}_{x \sim \mu}\left[\left\langle x, \frac{\beta_\star}{\|\beta_\star\|}\right\rangle^2 \wedge \Big\langle x, v\Big\rangle^2\right]$$

Fatou's lemma directly implies that h is lower semicontinuous and therefore admits a minimizer $u_\star \in \operatorname{argmin}_{u \in \mathbb{S}^{d-1}} h(u)$. Let $s\colon \mathbf{R}^d \to \mathbf{R}^d$ be a Borel measurable selection of the metric projection

$\mathsf{P}_{\mathcal{K}_{u_\star}}$. Define the pushforward measure $\bar{\nu} = s_{\#}\mu$. Clearly $\bar{\nu}$ lies in $\mathcal{E}$ and hence

$$\inf_{\nu \in \mathcal{E}} W_2^2(\mu, \nu) \leq W_2^2(\mu, \bar{\nu}) = \mathop{\mathbb{E}}_{x \sim \mu} \operatorname{dist}(x, \mathcal{K}_{u_\star})^2 = h(u_\star) \tag{69}$$

with the first equality holding by (36). On the other hand, for any $\nu \in \mathcal{E}$ there exists $w \in \mathbb{S}^{d-1}$ such that $\operatorname{supp}(\nu) \subset \mathcal{K}_w$ and hence

$$h(u_\star) \leq h(w) = \mathop{\mathbb{E}}_{x \sim \mu} \operatorname{dist}(x, \mathcal{K}_w)^2 \leq W_2^2(\mu, \nu)$$

with the last inequality holding by (35). Taking the infimum over $\nu \in \mathcal{E}$, we deduce that (69) holds with equality, thereby completing the proof.

E.3 Proof of Theorem 5.3

We will need the following two elementary lemmas.

Lemma E.1. *If (y_1, y_2) is a centered Gaussian vector with $\mathbb{E}\, y_1^2 = \sigma_1^2$, $\mathbb{E}\, y_2^2 = \sigma_2^2$, and $\mathbb{E}\, y_1 y_2 = \rho\sigma_1\sigma_2$, then the equations hold:*

$$\mathbb{E}\,|y_1 y_2| = \frac{2}{\pi}\left(\sqrt{1-\rho^2} + \rho \arcsin(\rho)\right)\sigma_1\sigma_2,$$
$$\mathbb{E}(y_1 y_2)^2 = (1 + 2\rho^2)\sigma_1^2\sigma_2^2.$$

Proof. The first equation is proved for example in [43, Corollary 3.1]. To see the second equation, standard results show that the conditional distribution $y_1 \mid y_2$ is Gaussian $N(\rho(\sigma_1/\sigma_2)y_2, \sigma_1^2 - \rho^2\sigma_1^2)$. Thus, the second moment is $\mathbb{E}[y_1^2 \mid y_2] = (1-\rho^2)\sigma_1^2 + \frac{\rho^2\sigma_1^2}{\sigma_2^2}y_2^2$ and therefore iterating expectations gives $\mathbb{E}\, y_1^2 y_2^2 = \mathbb{E}[\mathbb{E}[y_1^2 \mid y_2] y_2^2] = (1+2\rho^2)\sigma_1^2\sigma_2^2$, as claimed. $\square$

Lemma E.2. *Consider the function $\psi \colon [-1,1] \to \mathbf{R}$ defined by $\psi(t) = \sqrt{1-t^2} + t\arcsin(t)$ and the function $\phi \colon [-1,1] \to \mathbf{R}$ given by $\phi(t) = \left(\frac{\pi}{2} - 1\right)t^2 + 1$. Then, for any $t \in [-1,1]$ we have $\psi(t) \leq \phi(t)$.*

Proof. Taking the derivative of $\psi(t)$ and using the fundamental theorem of calculus we get

$$\psi(t) = 1 + \int_0^t \arcsin(s)ds = 1 + \int_0^t \sum_{n=0}^{\infty} \binom{2n}{n} \frac{s^{2n+1}}{4^n(2n+1)}ds = 1 + \sum_{n=0}^{\infty} \binom{2n}{n} \frac{t^{2n+2}}{4^n(2n+2)}, \tag{70}$$

where the second equality follows by taking the Taylor expansion of $\arcsin(s)$. Factorizing a t^2 from the series yields

$$\psi(t) = 1 + t^2 \sum_{n=0}^{\infty} \binom{2n}{n} \frac{|t|^{n+1}}{4^n(2n+2)} \leq 1 + t^2 \sum_{n=0}^{\infty} \binom{2n}{n} \frac{1}{4^n(2n+2)} = 1 + \left(\frac{\pi}{2} - 1\right)t^2 = \phi(t),$$

where the inequality follows since $|t| \leq 1$ and the second to last equality evaluates (70) at $t = 1$. This completes the proof. $\square$

With these two lemmas at hand, we start the proof of Theorem 5.3. We will first verify (24). To this end, notice that we may write $g_\Sigma(u,v) = \mathbb{E}(y_1y_2)^2$, where we define the random variables $y_1 = \langle x, u\rangle$ and $y_2 = \langle x, v\rangle$. We compute $\mathbb{E}\, y_1^2 = \langle \Sigma u, u\rangle$, $\mathbb{E}\, y_2^2 = \langle \Sigma v, v\rangle$, and $\mathbb{E}\, y_1 y_2 = \langle \Sigma u, v\rangle$. Therefore, Lemma E.1 directly implies

$$g_\Sigma(u,v) = \langle \Sigma u, u\rangle\langle \Sigma v, v\rangle + 2\langle \Sigma u, v\rangle^2 = \|\Sigma^{1/2}u\|_2^2 \cdot \|\Sigma^{1/2}v\|_2^2 + 2\langle \Sigma^{1/2}u, \Sigma^{1/2}v\rangle^2$$

Consequently, applying the Cauchy-Schwarz inequality yields the two sided bound

$$\|\Sigma^{1/2}u\|_2^2 \cdot \|\Sigma^{1/2}v\|_2^2 \leq g_\Sigma(u,v) \leq 3\|\Sigma^{1/2}u\|_2^2 \cdot \|\Sigma^{1/2}v\|_2^2.$$

Taking the infimum over $v \in \mathbb{S}^{d-1}$ completes the proof of (24).

Next, we verify (23). Notice that the upper bound follows since

$$\min_{v\in\mathbb{S}^{d-1}} h_\Sigma(u,v) = \min_{v\in\mathbb{S}^{d-1}} \mathbb{E}\left[\min\{\langle x,u\rangle^2, \langle x,v\rangle^2\}\right] \le \min_{v\in\mathbb{S}^{d-1}} \mathbb{E}\left[\langle x,v\rangle^2\right] = \lambda_{\min}(\Sigma).$$

To prove the lower bound we will show the slightly stronger statement that for any $u, v \in \mathbb{S}^{d-1}$,

$$(1-\tfrac{2}{\pi})\lambda_{\min}(\Sigma) \le h_\Sigma(u,v).$$

Recall that $\min\{a,b\} = \frac{a+b}{2} - \frac{|a-b|}{2}$ for any $a, b \in \mathbf{R}$. Therefore, we can write

$$\begin{aligned} h_\Sigma(u,v) &= \mathbb{E}_{x\sim N(0,\Sigma)}\left[\frac{\langle x,u\rangle^2 + \langle x,v\rangle^2}{2} - \frac{|\langle x,u\rangle^2 - \langle x,v\rangle^2|}{2}\right] \\ &= \frac{u^\top\Sigma u + v^\top\Sigma v}{2} - \frac{\mathbb{E}_{x\sim N(0,\Sigma)}|\langle x,u\rangle^2 - \langle x,v\rangle^2|}{2}. \end{aligned}$$

Next, let's compute

$$|\langle x,u\rangle^2 - \langle x,v\rangle^2| = |\underbrace{\langle x,u+v\rangle}_{y_1}\underbrace{\langle x,u-v\rangle}_{y_2}|.$$

Then we get $\sigma_1^2 := \mathbb{E}\, y_1^2 = \|\Sigma^{1/2}(u+v)\|^2$ and $\sigma_2^2 := \mathbb{E}\, y_2^2 = \|\Sigma^{1/2}(u-v)\|^2$, and

$$\mathbb{E}\, y_1y_2 = (u+v)^\top\Sigma(u-v) = \rho\sigma_1\sigma_2,$$

where $\rho := \frac{\langle\Sigma^{1/2}(u+v),\Sigma^{1/2}(u-v)\rangle}{\|\Sigma^{1/2}(u+v)\|\|\Sigma^{1/2}(u-v)\|}$. Thus, by Lemma E.1 we get

$$\mathbb{E}\,|y_1y_2| = \frac{2}{\pi}\left(\underbrace{\sqrt{1-\rho^2} + \rho\arcsin(\rho)}_{=:\psi(\rho)}\right)\sigma_1\sigma_2.$$

Therefore, after the relabeling $\hat{u} = \Sigma^{1/2}u$ and $\hat{v} = \Sigma^{1/2}v$ we deduce

$$h_\Sigma(u,v) = \frac{\|\hat{u}\|^2 + \|\hat{v}\|^2}{2} - \frac{\psi(\rho)}{\pi}\|\hat{u}+\hat{v}\|\|\hat{u}-\hat{v}\|.$$

By Lemma E.2 we have that

$$h_\Sigma(u,v) \ge \frac{\|\hat{u}\|^2 + \|\hat{v}\|^2}{2} - \left(\left(\frac{1}{2} - \frac{1}{\pi}\right)\rho^2 + \frac{1}{\pi}\right)\|\hat{u}+\hat{v}\|\|\hat{u}-\hat{v}\|. \tag{71}$$

In what follows we upper bound the second term on the right-hand-sight of this inequality. Without loss of generality we assume that $\|\hat{v}\| \le \|\hat{u}\|$. By definition ρ is equal to $\cos\alpha$ where α is the angle between $\hat{u}+\hat{v}$ and $\hat{u}-\hat{v}$. Thus, by the cosine law we have that

$$2\rho\|\hat{u}+\hat{v}\|\|\hat{u}-\hat{v}\| = 2\cos(\alpha)\|\hat{u}+\hat{v}\|\|\hat{u}-\hat{v}\| = \|\hat{u}+\hat{v}\|^2 + \|\hat{u}-\hat{v}\|^2 - 4\|\hat{v}\|^2 = 2(\|\hat{u}\|^2 - \|\hat{v}\|^2), \tag{72}$$

where the last equality follows by the parallelogram law. Similarly, using Young's inequality

$$\|\hat{u}+\hat{v}\|\|\hat{u}-\hat{v}\| \le \frac{\|\hat{u}+\hat{v}\|^2 + \|\hat{u}-\hat{v}\|^2}{2} \le \|\hat{u}\|^2 + \|\hat{v}\|^2. \tag{73}$$

Then, applying (72) and (73) yields

$$\begin{aligned}\left(\left(\frac{1}{2}-\frac{1}{\pi}\right)\rho^2+\frac{1}{\pi}\right)\|\hat{u}+\hat{v}\|\|\hat{u}-\hat{v}\| &\le \left(\frac{1}{2}-\frac{1}{\pi}\right)\rho\left(\|\hat{u}\|^2-\|\hat{v}\|^2\right)+\frac{1}{\pi}\left(\|\hat{u}\|^2+\|\hat{v}\|^2\right)\\ &\le \left(\frac{1}{\pi}+\frac{\rho}{2}-\frac{\rho}{\pi}\right)\|\hat{u}\|^2+\left(\frac{1}{\pi}-\frac{\rho}{2}+\frac{\rho}{\pi}\right)\|\hat{v}\|^2\\ &\le \frac{1}{2}\|\hat{u}\|^2+\left(\frac{2}{\pi}-\frac{1}{2}\right)\|\hat{v}\|^2\end{aligned}$$

where the last inequality follows by adding and subtracting $(2/\pi-1/2)$ to the coefficient of $\|\hat{v}\|$ and noting that $\|\hat{v}\|\le\|\hat{u}\|$. Combining this inequality with (71) yields

$$h_\Sigma(u,v)\ge\left(1-\frac{2}{\pi}\right)\|\hat{v}\|^2=\left(1-\frac{2}{\pi}\right)v^\top\Sigma v\ge\left(1-\frac{2}{\pi}\right)\lambda_{\min}(\Sigma),$$

which proves the lower bound.

F Proofs from Section 5.2

F.1 Proof of Lemma 5.5

Set $\Sigma_1=\mathbb{E}_\mu\, xx^\top$ and $\Sigma_2=\mathbb{E}_\nu\, xx^\top$. Define the manifold of rank one $d_1\times d_2$ matrices:

$$\mathcal{M}=\{M\in\mathbf{R}^{d_1\times d_2}:\operatorname{rank}(A)=1\}.$$

A quick computation yields the expression $\nabla f(M)=\mathbb{E}\langle M-M_\star,x_1x_2^\top\rangle$. In particular, equality $\nabla f(M_\star)=0$ holds and therefore the optimal Lagrange multipliers $\lambda^\star$ are zero. Hence the Hessian of the Lagrangian $\nabla^2\mathcal{L}(M_\star,\lambda^\star)$ coincides with $\nabla^2 f(M_\star)$. We now successively compute

$$\begin{aligned}\nabla^2 f(M_\star)[\Delta,\Delta]&=\mathbb{E}\langle\Delta,x_1x_2^\top\rangle^2 &(74)\\ &=\mathbb{E}\,x_1^\top\Delta(x_2x_2^\top)\Delta^\top x_1\\ &=\mathbb{E}_{x_1}\,x_1^\top\Delta\Sigma_2\Delta^\top x_1\\ &=\mathbb{E}_{x_1}\operatorname{tr}(\Delta\Sigma_2\Delta^\top x_1x_1^\top)\\ &=\operatorname{tr}(\Delta\Sigma_2\Delta^\top\Sigma_1)\\ &=\|\Sigma_1^{1/2}\Delta\Sigma_2^{1/2}\|_F^2 &(75)\\ &\ge\lambda_{\min}(\Sigma_1)\lambda_{\min}(\Sigma_2)\|\Delta\|_F^2.\end{aligned}$$

In particular, if both Σ_1 and Σ_2 are nonsingular, then $\mu\times\nu$ does not lie in $\mathcal{E}$. We now leverage the tangent structure to get an upper bound and a better lower bound on the quadratic form $\nabla^2 f(M_\star)[\Delta,\Delta]$. To this end, standard results (see e.g. [10, Section 7.5]) show that the tangent space to $\mathcal{M}$ at $M_\star=\beta_{1\star}\beta_{2\star}^\top$ is given by

$$T_{\mathcal{M}}(M_\star)=\{a\beta_{1\star}\beta_{2\star}^\top+u\beta_{2\star}^\top+\beta_{1\star}v^\top:a\in\mathbf{R},\,u\in\beta_{1\star}^\perp,\,v\in\beta_{2\star}^\perp\}.$$

Let $\Delta=a\beta_{1\star}\beta_{2\star}^\top+u\beta_{2\star}^\top+\beta_{1\star}v^\top\in T_{\mathcal{M}}(M_\star)$. Without loss of generality we might assume $\|\beta_{1\star}\|=\|\beta_{2\star}\|=1$, since otherwise we can make the change of variables $a'/\|\beta_{1\star}\|\|\beta_{2\star}\|\leftarrow a$, $v'/\|\beta_1\|\leftarrow v$, and $u'/\|\beta_2\|\leftarrow u$. Thus, $\|\Delta\|_F^2=a^2+\|v\|^2+\|u\|^2$. Observe that we may rewrite (75) as

$$\begin{aligned}\nabla^2 f(M_\star)[\Delta,\Delta]&=\left\|\Sigma_1^{1/2}\left((a\beta_{1\star}+u)\beta_{2\star}^\top+\beta_{1\star}v^\top\right)\Sigma_2^{1/2}\right\|_F^2\\ &=\left\|\left(\Sigma_2^{1/2}\otimes\Sigma_1^{1/2}\right)\operatorname{vec}\left((a\beta_{1\star}+u)\beta_{2\star}^\top\right)+\left(\Sigma_2^{1/2}\otimes\Sigma_1^{1/2}\right)\operatorname{vec}\left(\beta_{1\star}v^\top\right)\right\|_F^2.\end{aligned}$$

To obtain a bound we leverage the following general claim.

Claim F.1. *Let $A \in S_+^d$ be a positive definite matrix. Let $x, y \in \mathbf{R}^d$ be any pair of orthogonal vectors. Then the estimate holds:*

$$\|Ax + Ay\|^2 \geq \frac{2}{\kappa(A)^2 + 1}\left(\|Ax\|^2 + \|Ay\|^2\right). \tag{76}$$

Proof. The claim will follow from the following inequality

$$|\langle Ax, Ay\rangle| \leq \left(1 - \frac{2}{\kappa(A)^2 + 1}\right)\|Ax\|\|Ay\|. \tag{77}$$

We will come back to the proof of (77) but first we show how it implies the result (76). Expanding the square yields

$$\begin{aligned}
\|Ax + Ay\|^2 &= \|Ax\|^2 + 2\langle Ax, Ay\rangle + \|Ay\|^2 \\
&\geq \|Ax\|^2 - 2|\langle Ax, Ay\rangle| + \|Ay\|^2 \\
&\geq \|Ax\|^2 - 2\left(1 - \frac{2}{\kappa(A)^2 + 1}\right)\|Ax\|\|Ay\| + \|Ay\|^2 &(78)\\
&\geq \|Ax\|^2 - \left(1 - \frac{2}{\kappa(A)^2 + 1}\right)\left(\|Ax\|^2 + \|Ay\|^2\right) + \|Ay\|^2 &(79)\\
&= \frac{2}{\kappa(A)^2 + 1}\left(\|Ax\|^2 + \|Ay\|^2\right),
\end{aligned}$$

where (78) follows by (77) and (79) is an application of Young's inequality. So we now focus on proving (77). This is equivalent to finding an upper bound for the following optimization problem

$$\max_{x,y:\ x\perp y} \frac{\langle Ax, Ay\rangle}{\|Ax\|\|Ay\|} = \max_{x,y;x\perp y} \frac{\|A(x+y)\|^2 - \|A(x-y)\|^2}{4\|Ax\|\|Ay\|}$$

where the equality follows from the parallelogram law. Using the same law yields that

$$\frac{1}{2}\left(\|A(x+y)\|^2 + \|A(x-y)\|^2\right) = \|Ax\|^2 + \|Ay\|^2.$$

We use this constraint to define a relaxation of the problem. We introduce two new variables z and w, which intuitively play the role of $x + y$ and $x - y$, respectively. Without loss of generality we can set $\|x\| = \|y\| = 1$ since we can divide both sides of (77) by $\|x\|\|y\|$ and therefore from orthogonality of x and y we have $\|x - y\|^2 = 1$ and $\|x + y\|^2 = 2$. Define the constraint sets $\mathcal{Z} = \{(z, w) \mid \|z\|^2 = \|w\|^2 = 2\}$ and $\mathcal{X}_{w,z} = \{(x, y) \mid \|Ax\|^2 + \|Ay\|^2 = \left(\|Az\|^2 + \|Aw\|^2\right)/2\}$. We now successively upper bound

$$\begin{aligned}
\max_{x,y;x\perp y} \frac{|\langle Ax, Ay\rangle|}{\|Ax\|\|Ay\|} &\leq \max_{(z,w)\in\mathcal{Z}} \max_{(x,y)\in\mathcal{X}_{w,z}} \frac{\left|\|Az\|^2 - \|Aw\|^2\right|}{4\|Ax\|\|Ay\|} \\
&= \max_{(z,w)\in\mathcal{Z}} \frac{\left|\|Az\|^2 - \|Aw\|^2\right|}{\|Az\|^2 + \|Aw\|^2} &(80)\\
&= \frac{\lambda_{\max}^2(A) - \lambda_{\min}^2(A)}{\lambda_{\max}^2(A) + \lambda_{\min}^2(A)} &(81)
\end{aligned}$$

where (80) follows since the function $(a, b) \mapsto ab$ constrained to $a^2 + b^2 = c$ is minimized at $(a, b) = (\sqrt{c/2}, \sqrt{c/2})$ and (81) follows since $(a, b) \mapsto \frac{|a^2 - b^2|}{a^2 + b^2}$ constrained to the interval $\sqrt{2}[\lambda_{\min}(A), \lambda_{\max}(A)]$ attains a maximum at $(a, b) = \sqrt{2}\left(\lambda_{\max}(A), \lambda_{\min}(A)\right)$. Reordering the terms in (81) proves (77). $\square$

We instantiate the claim with $A = \Sigma_2^{1/2} \otimes \Sigma_1^{1/2}$, $x = \mathrm{vec}\left((a\beta_{1\star} + u)\beta_{2\star}^\top\right)$ and $y = \mathrm{vec}\left(\beta_{1\star}v^\top\right)$. Note that x and y are orthogonal since

$$\langle x, y\rangle = \langle(a\beta_{1\star} + u)\beta_{2\star}^\top, \beta_{1\star}v^\top\rangle = \mathrm{tr}\left(\beta_{2\star}(a\beta_{1\star} + u)^\top\beta_{1\star}v^\top\right) = \mathrm{tr}\left((a\beta_{1\star} + u)^\top\beta_{1\star}v^\top\beta_{2\star}\right) = 0$$

where we used the cyclic invariance of the trace and the fact that $\langle v, \beta_{2\star}\rangle = 0$. Further, $\kappa(A)^2 = \kappa(\Sigma_1)\kappa(\Sigma_2)$ and thus, all together we derive

$$\nabla^2 f(M_\star)[\Delta,\Delta] \geq \frac{2}{\kappa(\Sigma_1)\kappa(\Sigma_2)+1}\left(\langle \beta_{1\star}, \Sigma_1\beta_{1\star}\rangle\langle u, \Sigma_2 u\rangle + \langle \beta_{2\star}, \Sigma_2\beta_{2\star}\rangle\langle a\beta_{1\star}+v, \Sigma_1(a\beta_{1\star}+v)\rangle\right)$$

$$\geq \frac{2}{\kappa(\Sigma_1)\kappa(\Sigma_2)+1}\left(\langle \beta_{1\star}, \Sigma_1\beta_{1\star}\rangle\lambda_{\min}(\Sigma_2)\|u\|^2 + \langle \beta_{2\star}, \Sigma_2\beta_{2\star}\rangle\lambda_{\min}(\Sigma_1)\|a\beta_{1\star}+v\|^2\right)$$

$$\geq \frac{2}{\kappa(\Sigma_1)\kappa(\Sigma_2)+1}\min\left\{\langle \beta_{1\star}, \Sigma_1\beta_{1\star}\rangle\lambda_{\min}(\Sigma_2), \langle \beta_{2\star}, \Sigma_2\beta_{2\star}\rangle\lambda_{\min}(\Sigma_1)\right\}$$

where the last holds since $\|u\|^2 + \|a\beta_{1\star}+v\|^2 = 1$. This establishes the lower bound in (26).

Next we establish the converse. For any $\Delta = a\beta_{1\star}\beta_{2\star}^\top + u\beta_{2\star}^\top + \beta_{1\star}v^\top$, from (74) we have

$$\nabla^2 f(M_\star)[\Delta,\Delta] = \mathbb{E}[a\langle x_1, \beta_{1\star}\rangle\langle x_2, \beta_{2\star}\rangle + \langle x_1, u\rangle\langle x_2, \beta_{2\star}\rangle + \langle x_1, \beta_{1\star}\rangle\langle x_2, v\rangle]^2. \tag{82}$$

Note the equality $\|\Delta\|_F^2 = a^2\|\beta_{1\star}\|_2^2\|\beta_{2\star}\|_2^2 + \|\beta_{1\star}\|^2\|v\|^2 + \|\beta_{2\star}\|^2\|u\|^2$. Now setting $v = 0$, equation (82) becomes

$$\nabla^2 f(M_\star)[\Delta,\Delta] = \mathbb{E}[\langle x_1, a\beta_{1\star}+u\rangle\langle x_2, \beta_{2\star}\rangle]^2 = \langle \Sigma_1(a\beta_{1\star}+u), a\beta_{1\star}+u\rangle\langle \Sigma_2\beta_{2\star}, \beta_{2\star}\rangle.$$

Note the equality $\|a\beta_{1\star}+u\|^2 = a^2\|\beta_{1\star}\|^2 + \|u\|^2 = \frac{\|\Delta\|^2}{\|\beta_{2\star}\|^2}$. Therefore we deduce

$$\min_{\Delta:\|\Delta\|=1}\nabla^2 f(M_\star)[\Delta,\Delta] \leq \min_{w:\|w\|^2=1}\langle \Sigma_1 w, w\rangle\langle \Sigma_2\beta_{2\star}, \beta_{2\star}\rangle/\|\beta_{2\star}\|^2 = \lambda_{\min}(\Sigma_1)\cdot\langle \Sigma_2\beta_{2\star}, \beta_{2\star}\rangle/\|\beta_{2\star}\|^2.$$

A symmetric argument shows

$$\min_{\Delta:\|\Delta\|=1}\nabla^2 f(M_\star)[\Delta,\Delta] \leq \lambda_{\min}(\Sigma_2)\cdot\langle \Sigma_1\beta_{1\star}, \beta_{1\star}\rangle/\|\beta_{1\star}\|^2.$$

In particular, if Σ_1 or Σ_2 are singular then $\mu \times \nu$ lies in $\mathcal{E}$, as claimed. This completes the proof.

F.2 Proof of Theorem 5.6

This follows directly from Lemma 5.5 and Theorem A.3.

G Proofs from Section 5.3

G.1 Proof of Lemma 5.7

Consider any tangent vector $\Delta = \beta_\star v^\top + v\beta_\star^\top$ with $v \in \mathbf{R}^d$. Then we may vectorize Δ as follows:

$$\mathrm{vec}(\Delta) = \mathrm{vec}(\beta_\star v^\top I) + \mathrm{vec}(Iv\beta_\star^\top) = (I \otimes \beta_\star)v + (\beta_\star \otimes I)v = \Phi_\beta v.$$

Therefore, we compute

$$\nabla f(M_\star)[\Delta,\Delta] = \mathrm{vec}(\Delta)^\top\Sigma_\mu\mathrm{vec}(\Delta) = v^\top(\Phi_{\beta_\star}^\top\Sigma_\mu\Phi_{\beta_\star})v, \tag{83}$$

Minimizing the expression (83) in Δ with $\|\Delta\|_F = 1$ yields the guarantee (30).

G.2 Proof of Theorem 5.8

Characterization of well-posedness. To simplify notation, we will relabel $\beta_\star$ to β. Observe for any measure $(p_{ij}) \in \mathcal{Q}$, equation (28) yields

$$\nabla^2 f(M_\star)[\Delta_v,\Delta_v] = \mathbb{E}\langle \Delta_v, X\rangle^2 = \sum_{(i,j)\in E}(v_i\beta_j + v_j\beta_i)^2 p_{ij}, \tag{84}$$

where $\Delta_v := \beta v^\top + v\beta^\top$ is any tangent vector to $M_\star$ at $\mathcal{M}$. Recall moreover that Δ_v is nonzero if and only if v is nonzero.
($\implies$) We prove the contrapositive. Thus, suppose that Assumption 1 does not hold. We will consider two cases separately.

Case 1. Assume that there exists $\bar{i} \in V^0$. Thus for any edge $(\bar{i}, j) \in E$ equality $\beta_j = 0$ holds. Then setting $v = e_{\bar{i}}$, we deduce

$$\nabla^2 f(M_\star)[\Delta_v, \Delta_v] = \sum_{(i,j)\in E} (v_i\beta_j + v_j\beta_i)^2 p_{ij} = \sum_{j:(\bar{i},j)\in E} \beta_j^2 p_{\bar{i}j} = 0.$$

We conclude that μ lies in $\mathcal{E}^{\mathrm{mc}}$, as claimed.

Case 2. Assume that one of the components of G^* is bipartite and the set V^0 is empty. Without loss of generality, we may assume that G^* is connected, since otherwise we can restrict the following argument to any connected component. Thus, there exists a partition $V^* = I \cup J$ with all the edges $(i, j) \in E^*$ satisfying $i \in I$ and $j \in J$. Define the vector

$$v_i = \begin{cases} \beta_i & \text{if } i \in I, \\ -\beta_j & \text{if } j \in J, \\ 0 & \text{otherwise.} \end{cases}$$

Using (84), we obtain

$$\nabla^2 f(M_\star)[\Delta_v, \Delta_v] = \sum_{(i,j)\in E^*} (v_i\beta_j + v_j\beta_i)^2 p_{ij} = \sum_{(i,j)\in E^*} (\beta_i\beta_j - \beta_j\beta_i)^2 p_{ij} = 0.$$

where we used that v is supported on V^* and that V^0 is empty. Thus, μ lies in $\mathcal{E}^{\mathrm{mc}}$, as claimed.
($\impliedby$) Assume that Assumption 1 holds. To this end, let $v \in \mathbf{R}^d$ satisfy $\nabla^2 f(M_\star)[\Delta_v, \Delta_v] = 0$. Our goal is to show that v is the zero vector. To this end, clearly (84) implies

$$v_i\beta_j + v_j\beta_i = 0 \qquad \forall (i, j) \in E. \tag{85}$$

Without loss of generality suppose that G^* is connected, otherwise, we can repeat the argument for each connected component. Since G^* is non-bipartite, it must contain an odd-size cycle $i_1 \to i_2 \to \cdots \to i_k \to i_1$. Consider the expansion

$$2\beta_{i_1} v_{i_k} = (\beta_{i_1} v_{i_k} + \beta_{i_k} v_{i_1}) - \frac{\beta_{i_k}}{\beta_{i_2}} (\beta_{i_2} v_{i_1} + \beta_{i_1} v_{i_2}) + \sum_{j=2}^{k-1} (-1)^j \frac{\beta_{i_1}\beta_{i_k}}{\beta_{i_j}\beta_{i_{j+1}}} \left(\beta_{i_{j+1}} v_{i_j} + \beta_{i_j} v_{i_{j+1}}\right) = 0,$$

where the last equality follows since each term in the parenthesis is zero. Since $\beta_{i_1} > 0$, we deduce $v_{i_k} = 0$. Next, observe from (85) that for any neighbor j of i_k, i.e., satisfying $(i_k, j) \in E^*$, we have that $v_j = \beta_j v_{i_k}/\beta_{i_k} = 0$. Repeating the argument, we deduce $v_j = 0$ for all $v \in V^*$. Next, consider any vertex $i \notin V^*$. Then since V^0 is empty, there exists some $j \in V^*$ with $(i, j) \in E$. Using (85) again, we conclude $v_i\beta_j = 0$. Taking into account $\beta_j > 0$, we deduce $v_i = 0$. Thus v is identically zero.

Distance formula. We have now proved that a measure μ lies in $\mathcal{E}^{\mathrm{mc}}$ if and only if its support $\mathrm{supp}(\mu)$ lies in $\Omega_{\beta_\star}$. Given any set of indices $A \subseteq [d] \times [d]$ define $\mathsf{P}_A \colon \mathbf{R}^{d\times d} \to \mathbf{R}^{d\times d}$ to be the orthogonal projection onto the entries indexed by A. Fix now a set of entries $A \subset \mathrm{supp}(P)$ such that $A \in \Omega_{M_\star}$. Clearly, the pushforward measure $(\mathsf{P}_A)_\#\mu$ lies $\mathcal{E}^{\mathrm{mc}}$, and we compute

$$\min_{\nu\in\mathcal{E}^{\mathrm{mc}}} W_2^2(\mu, \nu) \le W_2^2(\mu, (\mathsf{P}_A)_\#\mu) = \sum_{ij\in \mathrm{supp}(P)\setminus A} p_{ij}.$$

Taking the minimum over A yields the inequality $\le$ in (33).

Conversely, fix a measure $\nu \in \mathcal{E}^{\mathrm{mc}}$ and let A be the indices that are observed with positive probability according to ν. Thus, the support of ν is contained in the subspace range(P_A). Then by Lemma A.1 and (35), we have

$$W_2^2(\mu, \nu) \geq W_2^2(\mu, (\mathsf{P}_A)_{\#}\mu) = \sum_{ij \in \mathrm{supp}(P) \setminus A} p_{ij}.$$

Taking the minimum over all entries $A \subset \mathrm{supp}(P)$ such that $A \in \Omega_{M_\star}$ completes the proof of the reverse inequality $\geq$ in (33).

Hardness. We reduce from the `MAXCUT` problem. Assume that we have a polynomial time algorithm $\texttt{Alg}(M_\star, (p_{ij}))$ to compute (33). Given an instance of `MAXCUT` $G = (V, E)$, we split the graph into its connected components $G_1 = (V_1, E_1), \ldots, G_k = (V_k, E_k)$. For each $\ell \leq k$, we define an instance of matrix completion with $M_\star = \mathbf{1}\mathbf{1}^\top \in \mathbf{R}^{|G_\ell| \times |G_\ell|}$ — with a slight abuse of notation we use $M_\star$ for all k problems — and set the distribution $P^{(\ell)}$ to $p_{ij}^{(\ell)} = \frac{1}{|E_\ell|}$ if $i, j \in E_\ell$, and $p_{ij}^{(\ell)} = 0$, otherwise. Since the entries of $M_\star$ are strictly positive and G_ℓ is connected, the output of $\texttt{Alg}(M_\star, P^{(\ell)})$ times $|E_\ell|$ is equal to the minimum number of edges one needs to remove from G_ℓ to make it bipartite. Thus, $|E_\ell| \left(1 - \texttt{ALG}\left(M_\star, P^{(\ell)}\right)\right)$ is equal to the number of edges of the largest bipartite graph one can construct via edge deletion, which is readily seen to be equal to $\texttt{MAXCUT}(G_\ell)$. Thus summing along connected components yields

$$\texttt{MAXCUT}(G) = |E| - \sum_{\ell \leq k} |E_\ell| \cdot \texttt{ALG}\Big(M_\star, P^{(\ell)}\Big);$$

completing the proof.