Condition Numbers

Brent Koogler

March 3, 2025

1 Introduction

We begin with a geometric definition of condition numbers due to the classical paper [Ric66]. This requires a brief review of Riemannian manifolds. These condition numbers sometimes have a geometric interpretation. We have already seen the condition number for matrix inversion. This can be interpreted as a distance to a bad discriminant set. We provide several proofs: two computational [Wikb; Wika] and one geometric [Bou23; BKS24].

2 A geometric definition of conditioning

We limit our discussion to regular embedded Riemannian submanifolds of the Euclidean space $\mathbb{R}^n$, endowed with the standard inner product. This allows us to prove the fundamental result from [Ric66] with nice Euclidean intuition. This is all that is needed for the condition number theorem results in [BKS24].

2.1 Riemannian submanifold background

Definition 2.1 (Manifold)**.** We call a subset $M \subseteq \mathbb{R}^n$ an m-dimensional regular Riemannian submanifold if for each $p \in M$, there is an open neighborhood $p \in U_p \subseteq \mathbb{R}^n$ and a smooth function $F_p : U_p \to \mathbb{R}^{n-m}$ such that $F_p^{-1}(0) \subseteq M$ and such that the differential (Jacobean matrix) has full rank. We endow M with the induced inner product from $\mathbb{R}^n$:

$$\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_M = \langle \cdot, \cdot \rangle_{\mathbb{R}^n}\big|_{M \times M}.$$

In addition, we call M an n-dimensional regular Riemannian submanifold if $M \subseteq \mathbb{R}^n$ is open.

For example, we can consider $S^2 \subseteq \mathbb{R}^3$. Define $F : \mathbb{R}^3 \setminus \{0\} \to \mathbb{R}$ by $F(x) = \|x\|^2 - 1$. Certainly $F^{-1}(0) = S^2$ and we have the derivative (differential)

$$\mathrm{D}F(x)[v] = \langle x, v \rangle, \qquad x, v \in \mathbb{R}^3,$$

which has full rank for $x \neq 0$. In contrast, the set of matrices of fixed rank $\mathbb{R}^{n \times m}_{=r}$ usually needs several such charts depending on the location of the nonzero $r \times r$ minor. More on this later.

Notation 2.2. Let $(M_k, \langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_k)$ denote an m_k-dimensional regular Riemannian submanifold of $\mathbb{R}^{n_k}$. For $k = 0$, we omit the subscript k.

Definition 2.3 (Tangent and normal spaces)**.** Let $p \in M$, and let $F : U \to \mathbb{R}^{n-m}$ denote a smooth zero function at p. The tangent space at p is

$$\mathrm{T}_p M = \operatorname{null} \mathrm{D}F(p),$$

and it has dimension $\dim \mathrm{T}_p M = n - (n - m) = m$. The normal space is the orthogonal complement in the embedding space $\mathbb{R}^n$:

$$\mathrm{N}_p M = (\mathrm{T}_p M)^\perp \qquad \Longrightarrow \qquad \mathrm{T}_p M \oplus \mathrm{N}_p M = \mathbb{R}^n.$$

Proposition 2.4 (Tangents as curves)**.** *Let $p \in M$ be a point. For each vector $v \in \mathrm{T}_p M$, there is a smooth curve $c : (-\epsilon, \epsilon) \to M \subseteq \mathbb{R}^n$ with $c(0) = p$ such that $\dot{c}(0) = v$. Conversely, $\dot{c}(0) \in \mathrm{T}_p M$ for each such curve.*

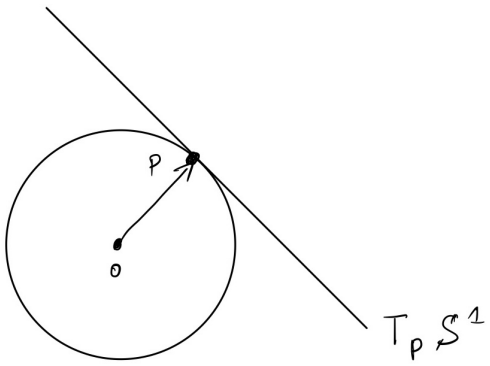
Proof. By picture. Rigorously, this is an implicit function theorem result. □

On S^1, let $p = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \in S^1 \subseteq \mathbb{R}^2$, and let $F(x) = \|x\|^2 - 1$. Then $\mathrm{D}F(x)[v] = \langle x, v \rangle$. For $\alpha > 0$, define

$$c(t) = \big(\cos(\alpha\, t + \tfrac{\pi}{4}),\, \sin(\alpha\, t + \tfrac{\pi}{4})\big).$$

Then we have the 1-dimensional tangent space representation

$$\dot{c}(t) = \alpha\, \big(-\sin(\alpha\, t + \tfrac{\pi}{4}),\, \cos(\alpha\, t + \tfrac{\pi}{4})\big) \quad \text{and} \quad \dot{c}(0) = \alpha\, \big(-\tfrac{1}{\sqrt{2}},\, \tfrac{1}{\sqrt{2}}\big).$$



Definition 2.5 (Smooth function and its differential)**.** We call a function $f : M \to \mathbb{R}$ smooth if there exists an open neighborhood $U \supseteq M$ and a smooth extension $\overline{f} : U \to \mathbb{R}$ with $\overline{f}|_M = f$. Let $p \in M$, let $v \in \mathrm{T}_p M$, and let $c(0) = p$ and $\dot{c}(0) = v$ with $c : (-\epsilon, \epsilon) \to M$ smooth. The differential of f at p in the direction v is defined as

$$\mathrm{D}f(p)[v] = \mathrm{D}\overline{f}(p)[v] = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} (f \circ c)(t).$$

The differential $\mathrm{D}f(p) : \mathrm{T}_p M \to \mathbb{R}$ is linear (clearly seen by transferring to $\overline{f}$).

Definition 2.6 (Geodesic)**.** We call a curve $\gamma : (-\epsilon, \epsilon) \to M$ a geodesic if γ is smooth, has unit velocity, and has zero accelearation. (The unit velocity is nonstandard.) Unit velocity:

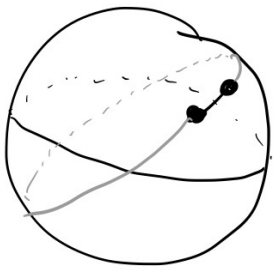
$$\|\dot{\gamma}(0)\| = 1.$$

Unit acceleration:

$$\mathrm{proj}_{\mathrm{T}_{\gamma(t)} M}\, \ddot{\gamma}(t) = 0.$$

Theorem 2.7. *Locally, the shortest arc length between any two points is given by a geodesic.*

For geodesics, think "great arcs" on S^2. Locally, there are unique distance minimizing arcs.



Definition 2.8 (Local distance)**.** Let $p, q \in M$. The distance between p and q is

$$d(p,q) = \inf_c \int_a^b \|\dot{c}(t)\| \, \mathrm{d}t,$$

where $c : (a,b)\mathbb{R} \to M$ is smooth with $c(a) = p$ and $c(b) = q$. If no such path c exists, the distance function is not defined.

Theorem 2.9. *For all $p \in M$, there is an open neighborhood $U \subseteq M$ containing p such that $d(q,p)$ exists for all $q \in U$. In fact, the infinum is achieved by a geodesic parameterized by arc length (unit speed).*

Definition 2.10. Let $B(p,\delta)$ denote the ball centered at $p \in M$ of radius $\delta > 0$, using our local distance d. (This is only defined for all δ sufficiently small, and this notion of sufficiently small need not be uniform over M.)

Proposition 2.11. *Let $p \in M$. There is a one-to-one correspondence between $\partial B(p,\delta)$ and the unit sphere in $\mathrm{T}_p M$ given by the geodesics of constant unit velocity. The correspondence is given by arc-length parameterized geodesics.*

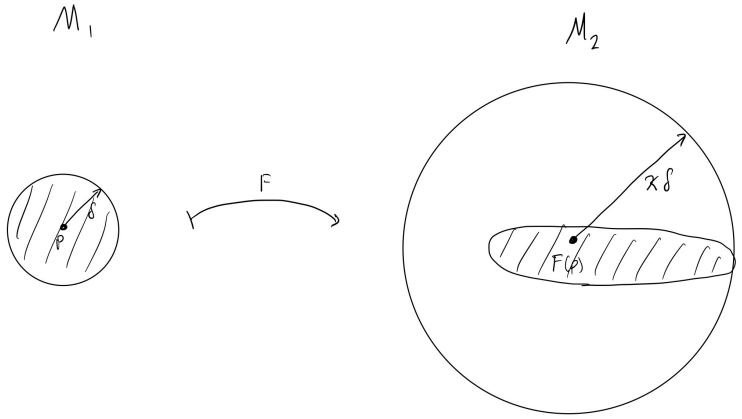
2.2 Conditioning

Definition 2.12 (Condition number)**.** Let $F : M_1 \to M_2$ be a function. The condition number of F at $p \in M$ is

$$\kappa(F)(p) = \lim_{\delta \searrow 0} \inf \big\{ \kappa > 0 : F\big(B(p,\delta)\big) \subseteq B\big(F(p), \kappa\,\delta\big) \big\},$$

if it exists. (We did not require F to be smooth here.)

So, $\kappa(F)(p)$ quantifies infinitesimal perturbations, which we can visualize.



Theorem 2.13 (Smooth condition number)**.** *Let $F : M_1 \to M_2$ be smooth, and let $p \in M_1$. Then*

$$\kappa(F)(p) = \| \mathrm{D}F(p) \|_{\mathrm{op}} = \sup_{\substack{\|v\|_1 = 1 \\ v \in \mathrm{T}_p M}} \| \mathrm{D}F(p)[v] \|_2.$$

Lemma 2.14. *The (local) map $q \mapsto d(q,p)$ has (one-sided) differential*

$$\mathrm{D}\big(d(\cdot,p)\big)(p)[v] = \|v\|_{\mathrm{T}_p M}, \quad \forall v \in \mathrm{T}_p M.$$

Proof. Let $v \in \mathrm{T}_p M$ have unit length $\|v\| = 1$. Let $\gamma : [0,\epsilon) \to M$ be an arc-length parameterized geodesic with $\gamma(0) = p$ and $\dot{\gamma}(0) = v$. Then $d(\gamma(t),p) = t$ and

$$\mathrm{D}\big(d(\cdot,p)\big)(p)[v] = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} d\big(\gamma(t),p\big) = 1.$$

So, for $v \in \mathrm{T}_p M$ arbitrary, we have linearity: for nonzero v,

$$\mathrm{D}\big(d(\cdot,p)\big)(p)[v] = \cancel{\mathrm{D}\big(d(\cdot,p)\big)\Big[\frac{v}{\|v\|}\Big]}^{1}\|v\| = \|v\|. \qquad \square$$

Proof of theorem. Let $M_1 = M_2 = M$ for notational convenience. We rewrite

$$\kappa(F)(p) = \lim_{\delta \searrow 0} \sup_{d(q,p)\le\delta} \frac{d\big(F(q),F(p)\big)}{\delta}.$$

Let $0 < \delta \ll 1$ be small, let $\gamma : [0, d_{p,q}] \to M$ be a geodesic with $\gamma(0) = p$ and $\gamma(d_{p,q}) = q$ (arc-length parameterized). Because $d(F(\gamma(t)), F(p))$ is a (local) smooth real valued function (somewhat subtle), we expand

$$\begin{aligned}
&d\big(F(\gamma(t)),F(p)\big)\\
&= d(\cdot,F(p)) \circ F \circ \gamma(t)\\
&= \cancel{d\big(F(\gamma(0)),F(p)\big)}^{0} + \Big(\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} d(\cdot,F(p)) \circ F \circ \gamma(t)\Big)\, t + o(t)\\
&= \mathrm{D}\big(d(\cdot,F(p))\big)\big(F(p)\big)\Big[\mathrm{D}F(p)[\dot\gamma(0)]\Big]\, t + o(t) \qquad \text{(chain rule)}\\
&= \|\,\mathrm{D}F(p)[\dot\gamma(0)]\|\, t + o(t).
\end{aligned}$$

So, for $v = \gamma'(0) \in \mathrm{T}_p M$ arbitrary of unit length,

$$\frac{d\big(F(q),F(p)\big)}{\delta} = \|\,\mathrm{D}F(p)[v]\|\,\frac{d_{p,q}}{\delta} + o\Big(\frac{d_{p,q}}{\delta}\Big),$$

and consequently

$$\lim_{\delta\searrow 0}\sup_{d(q,p)\le\delta}\frac{d\big(F(q),F(p)\big)}{\delta} = \lim_{\delta\searrow 0}\sup_{d(q,p)=\delta}\|\,\mathrm{D}F(p)[v]\| = \sup_{\substack{\|v\|=1\\ v\in\mathrm{T}_pM}}\|\,\mathrm{D}F(p)[v]\|,$$

as desired. $\square$

3 Eckart-Young theorem: a condition number theorem

We recall our condition number result from class. Let

$$\begin{aligned}\mathrm{inv} : \mathrm{GL}(n,\mathbb{R}) &\to \mathrm{GL}(n,\mathbb{R})\\ A &\mapsto A^{-1}\end{aligned}$$

denote the matrix inversion map. We may view $\mathrm{GL}(n,\mathbb{R})$ as an (open) Riemannian submanifold of $\mathbb{R}^{n\times n}$ when equipped with the Frobenious inner product: for $A = [a_{k,\ell}]$ and $B = [b_{k,\ell}]$ in $\mathbb{R}^{n\times m}$,

$$\langle A, B\rangle = \mathrm{tr}(A^T B) = \sum_{k=1}^{n}\sum_{\ell=1}^{m} a_{k,\ell}\, b_{k,\ell} = \langle \mathrm{vec}\, A, \mathrm{vec}\, B\rangle.$$

For notation, let $\|A\| = \sqrt{\langle A, A\rangle}$ denote the induced Frobenious norm, and let $\|A\|_{\mathrm{op}} = \sup_{\|v\|=1}\|A\,v\|$ denote the induced ℓ^2 operator norm.

Definition 3.1 (Block diagonal)**.** In $\mathbb{R}^{n\times m}$, for numbers $\sigma_1,\dots,\sigma_r$ with $r \le \min n, m$, define the block diagonal matrix

$$\mathrm{block\,diag}(\sigma_1,\dots,\sigma_r) = \begin{bmatrix}\mathrm{diag}(\sigma_1,\dots,\sigma_r) & 0\\ 0 & 0\end{bmatrix}.$$

Theorem 3.2 (Full Singular Value Decomposition (SVD))**.** *Let $A \in \mathbb{R}^{n\times m}$. Then there exists unitary $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ (i.e., $U\,U^T = I_n$ and $V\,V^T = I_m$) and a block diagonal $\Sigma \in \mathbb{R}^{n\times m}$ such that*

$$A = U\,\Sigma\,V^T.$$

In particular, $\Sigma = \operatorname{block\,diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \geq \ldots \geq \sigma_r > 0$ with $r = \operatorname{rk} A$. If we enumerate the columns $U = [u_1 \ldots u_n]$ and $V = [v_1 \ldots v_m]$, then we have the rank 1 expansion

$$A = \sum_{k=1}^{r} \sigma_k\, u_k\, v_k^T.$$

Observation 3.3. Because the columns of V are orthonormal, note that

$$A\,v_k = \sigma_k u_k.$$

Theorem 3.4. *The singular values are continuous with respect to the matrix entries.*

Now, we have the result from class.

Theorem 3.5 (Inverse condition number)**.** *For $A \in \mathrm{GL}(n, \mathbb{R})$,*

$$\kappa(\mathrm{inv})(A) = \frac{1}{\sigma_n(A)^2}$$

with $\sigma_n(A) > 0$ the smallest singular value of A, with $\operatorname{rk} A = n$.

A condition number theorem is a theorem that relates the condition number of a problem to the "distance" to a "bad" set. We give one such notion of "distance" and "bad".

Definition 3.6 (Rank based spaces)**.** For $0 \leq r, t \leq \min\{n, m\}$, define

$$\mathbb{R}^{n\times m}_{\leq r} = \{A \in \mathbb{R}^{n\times m} : \operatorname{rk} A \leq r\}$$

and

$$\mathbb{R}^{n\times m}_{=t} = \{A \in \mathbb{R}^{n\times m} : \operatorname{rk} A = t\}.$$

Definition 3.7 (SVD truncation)**.** Let $A \in \mathbb{R}^{n\times m}$ have SVD $A = U\,\Sigma\,V^T$ with $\Sigma = \operatorname{block\,diag}(\sigma_1, \ldots, \sigma_{\min\{n,m\}})$ and $\sigma_1 \geq \ldots \geq \sigma_{\min\{n,m\}} \geq 0$. For $0 \leq r \leq \min\{n, m\}$, we define the rank r truncation of A to be

$$A_r = U\,\Sigma_r\,V^T \quad \text{where} \quad \Sigma_r = \operatorname{block\,diag}(\sigma_1, \ldots, \sigma_r).$$

Theorem 3.8 (Eckart-Young)**.** *Let $A \in \mathbb{R}^{n\times m} \setminus \mathbb{R}^{n\times m}_{\leq r} =: \mathbb{R}^{n\times m}_{>r}$. Consider the Euclidean Distance (ED) problem (in the Frobenius norm)*

$$\min_{B \in \mathbb{R}^{n\times m}_{\leq r}} \|A - B\|^2 = \sum_{k=1}^{n}\sum_{\ell=1}^{m} |a_{k,\ell} - b_{k,\ell}|^2.$$

Then a minimum exists. Moreover, the minimizer B_A is given by $B_A = A_r$, the rank r truncation of A.

This relates to our matrix-inverse condition number. Note that inv is defined on $\mathrm{GL}(n, \mathbb{R})$ only, and it is not meaningful on the set

$$\mathbb{R}^{n\times n} \setminus \mathrm{GL}(n, \mathbb{R}) = \mathbb{R}^{n\times n}_{\leq n-1}.$$

Moreover, for $A \in \mathrm{GL}(n, \mathbb{R})$ and SVD $A = U\,\Sigma\,V^T$ with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}, \sigma_n)$, we can compute

$$\begin{aligned}
\min_{B \in \mathbb{R}^{n\times n}_{\leq n-1}} \|A - B\|^2 &= \|A - A_{n-1}\|^2 \\
&= \|U\,\Sigma\,V^T - U\,\Sigma_{n-1}\,V^T\|^2 \\
&= \|\Sigma - \Sigma_{n-1}\|^2 && \text{(orthogonal invariance)} \\
&= \|\operatorname{diag}(0, \ldots, 0, \sigma_n)\|^2 \\
&= \sigma_n^2.
\end{aligned}$$

So,

$$\kappa(\mathrm{inv})(A) = \frac{1}{\sigma_n^2} = \frac{1}{\operatorname{dist}(A, \mathbb{R}^{n\times n}_{\leq n-1})},$$

i.e., the condition number is inversely proportional to the "distance" from the "bad" set.

We now give several proofs of Eckart-Young. We begin with a computational approach found in [Wikb; Wika]. Then we describe a "geometric" approach using normal spaces following [BKS24], which requires some more background from [Bou23].

3.1 Computational Eckart-Young

We provide two computational proofs for Eckart-Young, but one uses a different norm. In the minimization $\min_{B\in\mathbb{R}^{n\times m}_{\leq r}} \|A - B\|^2$, we could have used a norm other than the Frobenius norm (and then lose the ED classification). Notably, many norms on $\mathbb{R}^{n\times m}$ induce the same minimizer A_r, the rank r truncation. Most of these norms can be defined in terms of the singular values of the input matrix. So, we first prove the result for the ℓ^2 operator norm

$$\|A\|_{\mathrm{op}} := \max_{\|v\|=1} \|A\,v\|_2 \equiv \sigma_1(A),$$

and then we show the Frobenius case.

Proof of the ℓ^2 case. Recall that the operator norm $\|\cdot\|_{\mathrm{op}}$ orthogonally invariant and is equal to the largest singular value of a matrix. Write the SVD $A = U\,\Sigma\,V^T$ with $\Sigma = \operatorname{block\,diag}(\sigma_1, \ldots, \sigma_{\mathrm{rk}\,A})$, where $\mathrm{rk}\,A > r$ by hypothesis. Recall that we implicitly order $\sigma_1 \geq \ldots \geq \sigma_{\mathrm{rk}\,A} \geq 0$. Precompute

$$\begin{aligned}
\|A - A_r\|_{\mathrm{op}}^2 &= \|U\,\Sigma\,V^T - U\,\Sigma_r\,V^T\|_{\mathrm{op}}^2 \\
&= \|\Sigma - \Sigma_r\|_{\mathrm{op}}^2 && \text{(orthogonal invariance)} \\
&= \|\operatorname{block\,diag}(0, \ldots, 0, \sigma_{r+1}, \ldots, \sigma_{\mathrm{rk}\,A})\|_{\mathrm{op}}^2 \\
&= \sigma_{r+1}^2.
\end{aligned}$$

Now, let $B \in \mathbb{R}^{n\times m}_{\leq r}$ be arbitrary, say $B \in \mathbb{R}^{n\times m}_{=t}$ with $0 \leq t \leq r$. Decompose $B = B_1\,B_2$ (not rank truncations) with $B_1 \in \mathbb{R}^{n\times t}$, $B \in \mathbb{R}^{t\times m}$, and $\mathrm{rk}\,B_1 = \mathrm{rk}\,B_2 = t$. Because $V = [v_1\ \ldots\ v_m]$ has $m > t$ linearly independent and orthogonal columns, then there is a nontrivial vector w in

$$\operatorname{span}\{v_1, \ldots, v_{t+1}\} \cap \operatorname{null} B_2 \neq \varnothing, \qquad \text{(somewhat subtle)}$$

say

$$w = \gamma_1\,v_1 + \cdots + \gamma_{t+1}\,v_{t+1}$$

with $\|w\|_2 = 1$, i.e., with $\gamma_1^2 + \cdots + \gamma_{t+1}^2 = 1$. Enumerate the orthonormal columns $U = [u_1\ \cdots\ u_n]$. Compute

$$(A - B)\,w = A\,w - B_1\,\cancelto{0}{(B_2\,w)} = \sigma_1\,\gamma_1\,u_1 + \cdots + \sigma_{t+1}\,\gamma_{t+1}\,u_{t+1}.$$

Then using the supremum definition of the operator norm, we can lower bound

$$\|A - B\|_{\mathrm{op}}^2 \geq \|(A - B)\,w\|_2^2 = \sigma_1\,\gamma_1^2 + \cdots + \sigma_{t+1}^2\,\gamma_{t+1}^2 \geq \sigma_{t+1}^2 = \|A - A_r\|_{\mathrm{op}}^2. \qquad \square$$

Proof of the Frobenius case. For SVD $A = U\,\Sigma\,V^T$, we have the proposed minimum (in the Frobenius norm)

$$\|A - A_r\|^2 = \|\Sigma - \Sigma_r\|^2 = \sum_{k\geq r+1} \big(\sigma_k(A)\big)^2.$$

Let $B \in \mathbb{R}^{n\times m}_{\leq r}$, where we aim to estimate $\|A - B\|^2$. By the triangle inequality of the operator norm (i.e., the largest singular value), let $k \in \mathbb{Z}_{\geq 1}$ and estimate

$$\begin{aligned}
\sigma_{k+r}(A) &= \sigma_1(A - A_{r+k-1}) \\
&\leq \sigma_1\big(A - [(A - B)_{k-1} + B_r]\big) && (\mathrm{rk}((A - B)_{k-1} + B_r) \leq r + k + 1) \\
&= \sigma_1\big((A - B) - (A - B)_{k-1}\big) && (B \in \mathbb{R}^{n\times m}_{\leq r}) \\
&= \sigma_k(A - B).
\end{aligned}$$

Thus,

$$\|A - A_r\|^2 = \sum_{k \geq r+1} \sigma_k(A)^2 = \sum_{k \geq 1} \sigma_{k+r}(A)^2 \leq \sum_{k \geq 1} \sigma_k(A-B)^2 = \|A - B\|^2,$$

as desired. □

3.2 Geometric Eckart-Young: outline

We now give a geometric proof, which is technically more subtle than the presentation given by [BKS24]. We want to optimize over $\mathbb{R}^{n \times m}_{\leq r}$, which is an algebraic variety and a stratified smooth manifold. But $\mathbb{R}^{n \times m}_{\leq r}$ is itself *not* a smooth manifold. (The tangent spaces — as defined through the derivatives of smooth curves — do not all have the same dimension as linear spaces.) However, by being stratified, we mean that

$$\mathbb{R}^{n \times m}_{\leq r} = \bigsqcup_{t=0}^{r} \mathbb{R}^{n \times m}_{=t}$$

with each $\mathbb{R}^{n \times m}_{=t}$ a smooth embedded Riemannian submanifold of $\mathbb{R}^{n \times m}$, endowed with the Frobenius inner product. The advantage of working in a Riemannian submanifold is the added Euclidean geometry.

Namely, we can talk about tangent and normal planes (instead of the more general notion of tangent and normal cones). Essentially, we may optimize on each $\mathbb{R}^{n \times m}_{=t}$, $t \in \{0, \ldots, r\}$, and then we choose the best solution, which apparently corresponds to $t = r$. To discuss this geometry, we discuss some more background from [Bou23].

3.3 Optimization on Riemannian submanifolds $\rightsquigarrow$ Lagrange multipliers

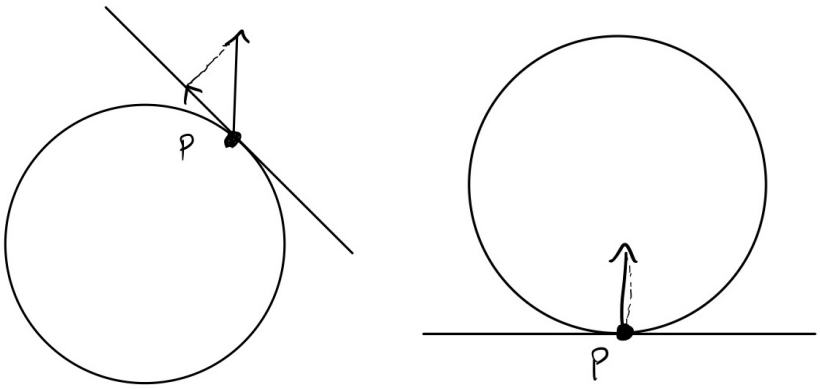
We consider a motivational example. Consider $M = S^1 \subseteq \mathbb{R}^2$ endowed with the usual Euclidean inner product, and define

$$\begin{aligned} f : S^1 &\to \mathbb{R} & \overline{f} : \mathbb{R}^2 &\to \mathbb{R} \\ (x, y) &\mapsto y & (x, y) &\mapsto y \\ \nabla f &= ? & \nabla \overline{f} &= (0, 1). \end{aligned}$$

We want to optimize

$$\max_{p \in S^1} f(p).$$

Our key idea is to look at the gradient. We visualize a couple of cases. The main upshot is that the projected gradient gives us information about how to locally optimize f.



(a) Nonzero projected gradient. (b) Zero projected gradient.

Notation 3.9. Let $(M, \langle \cdot, \cdot \rangle)$ be a smooth regular embedded Riemannian submanifold of $\mathbb{R}^n$. Let $f : M \to \mathbb{R}$ be smooth, i.e., we have an open $U \subseteq M$ and a smooth extension $\overline{f} : U \to \mathbb{R}$ with $\overline{f}|_M = f$. Let $p \in M$. Let $c :\subseteq \mathbb{R} \to M$ be a smooth curve with $c(0) = p$.

Definition 3.10 (First order critical)**.** We say that $p \in M$ is first order critical for f if for all smooth curves c with $c(0) = p$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} f\big(c(t)\big) = 0. \qquad (f \circ c :\subseteq \mathbb{R} \to \mathbb{R})$$

Definition 3.11 (Riemannian gradient)**.** Let $p \in M$. We call $\nabla f(p) \in \mathrm{T}_p M$ the Riemannian gradient of f at p if we have the Riesz representation

$$\mathrm{D}f(p)[v] = \langle \nabla f(p), v \rangle, \qquad \forall v \in \mathrm{T}_p M,$$

i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} f\big(c(t)\big) = \langle \nabla f(p), \dot{c}(0) \rangle, \qquad \forall c \text{ smooth with } c(0) = p.$$

Note that if c is smooth with $c(0) = p$, we may form a first order Taylor expansion:

$$\begin{aligned} f\big(c(t)\big) &= f\big(c(0)\big) + (f \circ c)'(0)\, t + o(t) \\ &= f(p) + \langle \nabla f(p), \dot{c}(0) \rangle\, t + o(t). \end{aligned}$$

So, if $\nabla f(p) = 0$ (in $\mathrm{T}_p M$), then it is not clear if moving forwards or backwards along (arbitrary) c will increase or decrease f. For computation, we have the following geometric intuition.

Theorem 3.12. *With* proj *the projection operator onto a linear space, the Riemannian gradient is given by*

$$\nabla f(p) = \mathrm{proj}_{\mathrm{T}_p M} \nabla \overline{f}(p).$$

Proof. By the usual Euclidean chain rule

$$(f \circ c)'(0) = (\overline{f} \circ c)'(0) = \langle \nabla \overline{f}(p), \dot{c}(0) \rangle.$$

The problem is that the Euclidean gradient $\nabla \overline{f}(p)$ is mostly likely not in the tangent space $\mathrm{T}_p M$. Because $\mathrm{T}_p M$ is a linear subspace of $\mathbb{R}^n$, we may decompose $\nabla \overline{f}(p)$ into tangential and orthogonal (normal) components, say $\nabla f(p)$ and $\mathrm{N}f(p)$, respectively. Then

$$\begin{aligned} (f \circ c)'(0) &= \langle \nabla f(p) + \mathrm{N}f(p), \dot{c}(0) \rangle \\ &= \langle \nabla f(p), \dot{c}(0) \rangle + \cancelto{0}{\langle \mathrm{N}f(p), \dot{c}(0) \rangle} \\ &= \langle \nabla f(p), \dot{c}(0) \rangle. \end{aligned}$$

Because c is arbitrary, we have shown the claim. □

Theorem 3.13. *If p is a minimizer of f, then $\nabla f(p) = 0 \in \mathrm{T}_p M$, i.e., $\nabla \overline{f}(p) \in \mathrm{N}_p M$.*

Proof. Duh. □

Proposition 3.14. *Define $f : \mathbb{R}^{n \times m}_{=t} \to \mathbb{R}$ by $f(B) = \|A - B\|^2$. Then the critical points B are characterized by their normal spaces: B is a critical point if and only if $A \in B + \mathrm{N}_B \mathbb{R}^{n \times m}_{=t}$.*

Proof. Let $B(t)$ be a smooth path in $\mathbb{R}^{n \times m}_{=t}$. Compute the differential

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \|A - B(t)\|^2 &= \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \sum_{k=1}^{n} \sum_{\ell=1}^{m} \big(a_{k,\ell} - b_{k,\ell}(t)\big)^2 \\ &= \sum_{k=1}^{n} \sum_{\ell=1}^{m} \big(b_{k,\ell}(0) - a_{k,\ell}\big)\, \dot{b}_{k,\ell}(0) \\ &= \langle B - A, \dot{B} \rangle. \end{aligned}$$

So, our first order critical condition requires $B - A \in \mathrm{N}_B \mathbb{R}^{n \times m}_{=t}$. □

3.4 Geometric Eckart-Young: refined outline

We take a stratified inverse approach.

(1) Let $B \in \mathbb{R}^{n\times m}_{=t}$, $t \in \{0,\dots,r\}$, and compute the normal space $\mathrm{N}_B\mathbb{R}^{n\times m}_{=t}$.

(2) Let $A \in \mathbb{R}^{n\times m}_{>r}$, for the nontrivial case.

(3) Minimize over $B \in \mathbb{R}^{n\times m}_{=t}$ such that $A \in B + \mathrm{N}_B\mathbb{R}^{n\times m}_{=t}$, say B_t.

(4) Choose $\min_{t\in\{0,\dots,r\}} \|A - B_t\|^2$.

This construction obviously produces a minimizer, if a minimizer exists. We outline existence. In the Frobenious norm (sum of squares of matrix entries), we have $\|A-B\|^2 \to \infty$ as $B \to \infty$. So, we may choose $R > 0$ large enough such that all $\|B\| > R$ with $B \in \mathbb{R}^{n\times m}_{\le r}$ have $\|A-0\|^2 < \|A-B\|^2$ (where $0 \in \mathbb{R}^{n\times m}_{\le r}$). Next, $\mathbb{R}^{n\times m}_{\le r}$ is closed in $(\mathbb{R}^{n\times m}, \|\cdot\|)$. If we have a sequence $B_k \in \mathbb{R}^{n\times m}_{\le r}$ and if $B_k \to B$, then we are not going to gain any extra nonzero singular values, by continuity. (We can lose singular values, but not gain one.) Then $\mathbb{R}^{n\times m}_{\le r} \cap \overline{B(0,R)}$ is closed and bounded, and consequently $B \mapsto \|A-B\|^2$ achieves its minimum, namely in the interior.

Because we have the partition $\mathbb{R}^{n\times m}_{\le r} = \bigsqcup_{t=0}^{r} \mathbb{R}^{n\times m}_{=t}$, our minimizer must lie on one of the manifolds $\mathbb{R}^{n\times m}_{=t}$. So, it suffices to minimize over the normal spaces in all the $\mathbb{R}^{n\times m}_{=t}$, which is our first order optimality condition.

3.5 Normal spaces of $\mathbb{R}^{n\times m}_{=t}$

To compute the normal spaces of $\mathbb{R}^{n\times m}_{=t}$, it is productive to understand its dimension. So, we compute some "charts". For a Riemannian manifold, we look for smooth $F : \mathbb{R}^{n\times m} \to \mathbb{R}^d$ such that $F^{-1}(0) = \mathbb{R}^{n\times m}_{=t}$. Certainly $F(B) = \mathrm{rk}(B) - t$ is a natural choice. But the rank function is integer valued, and hence not continuous, and hence not smooth. (So, a differential doesn't exist to give us our tangent spaces.) Alternatively, we can define a collection of these F about each entry in $\mathbb{R}^{n\times m}_{=t}$.

Let $B \in \mathbb{R}^{n\times m}_{=t}$. Then B has an invertible $t \times t$ submatrix, say that it is in the upper left:

$$B = \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}$$

with $B_{1,1} \in \mathrm{GL}(t,\mathbb{R})$, $B_{1,2} \in \mathbb{R}^{t\times(m-t)}$, $B_{2,1} \in \mathbb{R}^{(n-t)\times t}$, and $B_{2,2} \in \mathbb{R}^{(n-t)\times(m-t)}$. Because B has rank t, then the columns of $\begin{bmatrix} B_{1,2} \\ B_{2,2} \end{bmatrix}$ are a linear combination of the full rank columns $\begin{bmatrix} B_{1,1} \\ B_{2,1} \end{bmatrix}$, say

$$\begin{bmatrix} B_{1,2} \\ B_{2,2} \end{bmatrix} = \begin{bmatrix} B_{1,1} \\ B_{2,1} \end{bmatrix} W, \qquad W \in \mathbb{R}^{t\times(m-t)}.$$

Then

$$W = B_{1,1}^{-1} B_{1,2} \qquad \Longrightarrow \qquad B_{2,2} = B_{2,1}\, B_{1,1}^{-1}\, B_{1,2}.$$

Then we propose $F : U \to \mathbb{R}^{(n-t)\times(m-t)}$, with U open, defined by

$$F(B) = B_{2,1}\, B_{1,1}^{-1}\, B_{1,2} - B_{2,2}, \qquad U = \left\{ B \in \mathbb{R}^{n\times m} : B = \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix},\ \det B_{1,1} \neq 0 \right\}.$$

Certainly F is nice and smooth, and its differential has full rank: compute

$$\mathrm{D}F(B)[V] = V_{2,1}\, B_{1,1}^{-1} B_{1,2} + B_{2,1}\, B_{1,1}^{-1}\, V_{1,2} + B_{2,1}\, B_{1,1}^{-1}\, V_{1,1}\, B_{1,1}^{-1}\, B_{1,2} - V_{2,2},$$

where we set $V_{1,1} = 0$, $V_{1,2} = 0$, and $V_{2,1} = 0$ to demonstrate the full rank $\mathrm{D}F(B)[V] = V_{2,2}$ with $V \in \mathbb{R}^{n\times m}$ arbitrary. Thus, F is a smooth local defining zero function for $B \in \mathbb{R}^{n\times m}_{=t}$. Note that we need a different F depending on the entry positions of the invertible $t \times t$ submatrix.

This also shows the dimension of $\mathbb{R}^{n\times m}_{=t}$:

$$\begin{aligned}\dim \mathbb{R}^{n\times m}_{=t} &= \dim \operatorname{null} F \\ &= n\,m - (n-t)\,(m-t) \\ &= t\,(n+m-t).\end{aligned}$$

For $B \in \mathbb{R}^{n\times m}_{=t}$, we have $\dim \mathrm{T}_B \mathbb{R}^{n\times m}_{=t} = t\,(n+m-t)$, and consequently

$$\dim \mathrm{N}_B \mathbb{R}^{n\times m}_{=t} = (n-t)\,(m-t).$$

So, we are looking for $(n-t)\,(m-t)$ linearly independent (e.g., orthogonal) matrices to span the normal space $\mathrm{N}_B \mathbb{R}^{n\times m}_{=t}$.

Let B be a smooth path through $\mathbb{R}^{n\times m}_{=t}$. Then we may factor $B = R\,S$ with $R \in \mathbb{R}^{n\times t}_{=t}$ and $S \in \mathbb{R}^{t\times m}_{=t}$ both of full rank. Compute

$$\dot{B} = \dot{R}\,S + R\,\dot{S}$$

with $\dot{R} \in \mathbb{R}^{n\times t}$ ad $S \in \mathbb{R}^{t\times m}$ arbitrary. We add rows $[S^T\, s_1^T\, s_2^T\, \ldots\, s_{m-t}^T]^T \in \mathbb{R}^{m\times m}$ with the $s_k \in \mathbb{R}^m$ orthonormal to S and the other s_ℓ. Similarly, orthonormally extend the columns $[R\, r_1\, r_2\, \ldots\, r_{n-t}] \in \mathbb{R}^{n\times n}$, each $r_k \in \mathbb{R}^n$. Then for the $(n-t)\,(m-t)$ vectors $r_k\, s_\ell^T$, we have the following Frobenious orthogonality:

$$\begin{aligned}\langle \dot{B}, r_k\, s_\ell^T \rangle &= \operatorname{tr}\left(\dot{B}\,(r_k\, s_\ell)^T\right) \\ &= \operatorname{tr}\left((\dot{R}\,S + R\,\dot{S})\, s_\ell\, r_k^T\right) \\ &= \operatorname{tr}\left(r_k^T\,(\dot{R}\,S + R\,\dot{S})\, s_\ell\right) \\ &= \operatorname{tr}\left(r_k^T\,\dot{R}\,\cancelto{0}{(S\, s_\ell)} + \cancelto{0}{(r_k^T R)}\,\dot{S}\, s_\ell\right) \\ &= 0.\end{aligned}$$

Note that the $r_k\, s_\ell^T$ are orthogonal, via a similar argument. Thus,

Observation 3.15.

$$\mathrm{N}_B \mathbb{R}^{n\times m}_{=t} = \operatorname{span}\left\{r_k\, s_\ell^T : k \in \{1,\ldots,m-t\},\, \ell \in \{1,\ldots,n-t\}\right\}.$$

3.6 Geometric Eckart-Young: the main argument

Now, we begin the minimization procedure over the first order optimality condition. Let $A \in \mathbb{R}^{n\times m}_{>r}$, and let $0 \le t \le r$. On $\mathbb{R}^{n\times m}_{=r}$, let B be a critical point to $\|A-B\|^2$, i.e., let $A \in B + \mathrm{N}_B \mathbb{R}^{n\times m}_{=t}$. Due to the orthogonal invariance of the Frobenious inner product, we may take the SVD of B to write $B = \operatorname{block\,diag}(\sigma_1, \ldots, \sigma_t)$ with $\sigma_1 \ge \ldots \ge \sigma_t > 0$. Moreover, we can take our orthogonal spanning set of $\mathrm{N}_B \mathbb{R}^{n\times m}_{=t}$ to be the standard basis:

$$\mathrm{N}_B \mathbb{R}^{n\times m}_{=t} = \operatorname{span}\left\{e_k^{(n)}\,(e_\ell^{(m)})^T : k \in \{t+1,\ldots,m\},\, \ell \in \{t+1,\ldots,n\}\right\}.$$

(This is only slightly subtle. In essence, we have the direct sum decomposition $\mathrm{T}_B \mathbb{R}^{n\times m}_{=t} \oplus \mathrm{N}_B \mathbb{R}^{n\times m}_{=t} = \mathbb{R}^{n\times m}$, and we are using an orthogonal transformation to get disjoint standard bases for these two blocks.) So, for some basis coefficients $a_{k,\ell}$, we have the form

$$A = B + \sum_{k=t+1}^{n} \sum_{\ell=t+1}^{m} a_{k,\ell} e_k\, e_\ell^T =: \begin{bmatrix}\Sigma & 0 \\ 0 & A_{\mathrm{N}}\end{bmatrix}.$$

Write the SVD $A_{\mathrm{N}} = U_{\mathrm{N}}\,\Sigma_{\mathrm{N}}\,V_{\mathrm{N}}^T$, with Σ_{N} diagonal. Consequently,

$$\begin{aligned}\|A-B\|^2 = \| \begin{bmatrix}\Sigma & 0 \\ 0 & A_{\mathrm{N}}\end{bmatrix} - \begin{bmatrix}\Sigma & 0 \\ 0 & 0\end{bmatrix} \|^2 &= \| \begin{bmatrix}0 & 0 \\ 0 & A_{\mathrm{N}}\end{bmatrix} \|^2 \\ &= \|A_{\mathrm{N}}\|^2 = \|\Sigma_{\mathrm{N}}\|^2 \qquad \text{(orthogonal invariance)} \\ &= \text{ sum of squares of diagonal entries.}\end{aligned}$$

Interpretation: if $B \in \mathbb{R}^{n\times m}_{=t}$ is a first order critical point of $\|A - \cdot\|^2$, then B shares t singular values of A, and the loss function evaluation $\|A - B\|^2$ is equal to the sum of the squares of the remaining singular values of A. Therefore, the rank t truncation A_t provides a lower bound on the first order critical points, as desired.

References

[BKS24] P. Breiding, K. Kohn, and B. Sturmfels. *Metric Algebraic Geometry*. Oberwolfach Seminars Series. Springer Nature, 2024. ISBN: 9783031514623. DOI: 10.1007/978-3-031-51462-3. URL: https://link.springer.com/book/10.1007/978-3-031-51462-3.

[Bou23] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023. DOI: 10.1017/9781009166164. URL: https://www.nicolasboumal.net/book.

[Ric66] John R. Rice. "A Theory of Condition". In: *SIAM Journal on Numerical Analysis* 3.2 (1966), pp. 287–310. ISSN: 00361429. URL: http://www.jstor.org/stable/2949623 (visited on 03/01/2025).

[Wika] Wikipedia. *Low-rank approximation*. URL: https://en.wikipedia.org/wiki/Low-rank_approximation (visited on 03/02/2025).

[Wikb] Wikipedia. *Wikipedia:Wikipe-tan*. URL: https://en.wikipedia.org/wiki/Wikipedia:Wikipe-tan (visited on 03/02/2025).