A Big Picture of Variance Analysis*

Karl Stratos

1 Preliminary

In this note, Z denotes a function of n independent variables $X_{1:n} = (X_1 \dots X_n)$. As done in the book, we define the following *random variables* (all assumed to be finite):

$$\begin{aligned}
\mathbf{E}_i[Z] &:= \mathbf{E}\left[Z|X_{1:i}\right] && \text{(random in } X_{1:i}\text{)} \\
\Delta_i &:= \mathbf{E}_i[Z] - \mathbf{E}_{i-1}[Z] && \text{(random in } X_{1:i}\text{)} \\
\mathbf{E}^{(i)}[Z] &:= \mathbf{E}\left[Z|X_{1:i-1}, X_{i+1:n}\right] && \text{(random in } X_{1:i-1}, X_{i+1:n}\text{)} \\
\mathrm{Var}^{(i)}(Z) &:= \mathbf{E}^{(i)}\left[\left(Z - \mathbf{E}^{(i)}[Z]\right)^2\right] && \text{(random in } X_{1:i-1}, X_{i+1:n}\text{)}
\end{aligned}$$

Note that $\mathbf{E}_0[Z] = \mathbf{E}\left[Z\right]$ is a constant and $\mathbf{E}_n[Z] = Z$, allowing for a telescoping sum

$$Z - \mathbf{E}[Z] = \sum_i \Delta_i$$

Let's get some ugly works out of our way upfront. Using the law of iterated expectation $\mathbf{E}[Z] = \mathbf{E}_Y[\mathbf{E}[Z|Y]]$ conditionally (i.e., $\mathbf{E}[Z|X] = \mathbf{E}_{Y|X}[\mathbf{E}[Z|X,Y]|X]$) and using the independence of $X_1 \dots X_n$, we verify that for any $j \geq i$

$$\mathbf{E}_i[\mathbf{E}_j[Z]] = \mathbf{E}\left[\mathbf{E}\left[Z|X_{1:i}, X_{i+1:j}\right]|X_{1:i}\right] = \mathbf{E}\left[Z|X_{1:i}\right] = \mathbf{E}_i[Z]$$

The intuition is that the future value of Z cannot be different from the present estimate *because $X_{1:n}$ are independent*. Therefore $\mathbf{E}_i[\Delta_j] = \mathbf{E}_i[Z] - \mathbf{E}_i[Z] = 0$ for any $j > i$.[1] Also since $\mathbf{E}_i[\Delta_i] = \Delta_i$, we have $\mathbf{E}\left[\Delta_i\Delta_j\right] = \mathbf{E}\left[\mathbf{E}_i[\Delta_i\Delta_j]\right] = \mathbf{E}\left[\Delta_i\mathbf{E}_i[\Delta_j]\right] = 0$ for any $j > i$. Thus the variables (Δ_i, Δ_j) where $i \neq j$ are *orthogonal*.[2] Similarly verify using the independence of $X_{1:n}$,

$$\begin{aligned}
\mathbf{E}_i\left[\mathbf{E}^{(i)}[Z]\right] = \mathbf{E}\left[\underbrace{\mathbf{E}\left[Z|X_{1:i-1}, X_{i+1:n}\right]}_{\text{constant wrt } X_i}\middle| X_{1:i}\right] &= \mathbf{E}\left[\mathbf{E}\left[Z|X_{1:i-1}, X_{i+1:n}\right]|X_{1:i-1}\right] \\
&= \mathbf{E}\left[Z|X_{1:i-1}\right] = \mathbf{E}_{i-1}[Z]
\end{aligned}$$

This allows us to relate the second moment of Δ_i to conditional variance by expressing $\Delta_i = \mathbf{E}_i[Z - \mathbf{E}^{(i)}[Z]]$ and (conditionally) applying Jensen's inequality:

$$\mathbf{E}\left[\Delta_i^2\right] = \mathbf{E}\left[\mathbf{E}_i\left[Z - \mathbf{E}^{(i)}\left[Z\right]\right]^2\right] \leq \mathbf{E}\left[\mathbf{E}_i\left[\left(Z - \mathbf{E}^{(i)}\left[Z\right]\right)^2\right]\right] = \mathbf{E}\left[\mathrm{Var}^{(i)}(Z)\right] \quad (1)$$

*Section 3 of BLM

2 Orthogonal Decomposition

We consider n independent variables $X_i \in \mathcal{X}$ (possibly with different distributions) and some function $f : \mathcal{X}^n \to \mathbb{R}$. The variance of

$$Z := f(X_{1:n})$$

is characterized by the variances of individual variables. This is clear when $f(X_{1:n}) = \sum_i X_i$ since in this case

$$\mathrm{Var}\left(Z\right) = \mathbf{E}\left[\left(\sum_i \overline{X}_i\right)^2\right] = \sum_i \mathbf{E}\left[\overline{X}_i^2\right] = \sum_i \mathrm{Var}\left(X_i\right)$$

where $\overline{X}_i := X_i - \mathbf{E}\left[X_i\right]$. This is an application of the Pythagorean theorem and only involves (a) expressing $Z - \mathbf{E}\left[Z\right]$ as a sum of variables $\overline{X}_i$, and (b) using the pairwise orthogonality of $(\overline{X}_i, \overline{X}_j)$ for $i \neq j$.

We can generalize this result to an arbitrary function f by (a) expressing $Z - \mathbf{E}\left[Z\right]$ as a sum of martingale differences Δ_i and (b) using the pairwise orthogonality of (Δ_i, Δ_j) for $i \neq j$ (which requires the independence of $X_{1:n}$):

$$\mathrm{Var}\left(Z\right) = \mathbf{E}\left[\left(\sum_i \Delta_i\right)^2\right] = \sum_i \mathbf{E}\left[\Delta_i^2\right]$$

Using equation (1) we have

$$\mathrm{Var}\left(Z\right) \leq \sum_i \mathbf{E}\left[\mathrm{Var}^{(i)}(Z)\right] \tag{2}$$

This is known as **Efron-Stein inequality**. Alternative characterizations of (conditional) variance give more usable forms of the inequality.

Re-sampling form. If $X_i' \in \mathcal{X}$ is an independent copy of X_i and Z_i' denotes $f(X_{1:i-1}, X_i', X_{i+1:n})$, then we can write $\mathrm{Var}^{(i)}(Z) = (1/2)\mathbf{E}^{(i)}\left[(Z - Z_i')^2\right]$ and thus

$$\mathrm{Var}\left(Z\right) \leq \frac{1}{2}\sum_i \mathbf{E}\left[(Z - Z_i')^2\right] \tag{3}$$

This is useful when we can show that re-sampling X_i doesn't change Z much.

Variational form. Let $\mathcal{S}_i$ denote the set of square-integrable random variables dependent on $X_{1:i-1}, X_{i+1:n}$. Then we can write $\mathrm{Var}^{(i)}(Z) = \inf_{Z_i \in \mathcal{S}_i} \mathbf{E}^{(i)}\left[\left(Z - Z_i\right)^2\right]$ (since infimum is achieved with $Z_i = \mathbf{E}^{(i)}\left[Z\right]$) and thus

$$\mathrm{Var}\left(Z\right) \leq \sum_i \mathbf{E}\left[\inf_{Z_i \in \mathcal{S}_i} \mathbf{E}^{(i)}\left[(Z - Z_i)^2\right]\right] \leq \sum_i \inf_{Z_i \in \mathcal{S}_i} \mathbf{E}\left[(Z - Z_i)^2\right]$$

where we used the concavity of the infimum. Equivalently,

$$\mathrm{Var}\left(Z\right) \leq \sum_i \mathbf{E}\left[(Z - Z_i)^2\right] \qquad \forall Z_i \in \mathcal{S}_i,\ \forall i = 1 \ldots n \tag{4}$$

This is convenient because we can choose $Z_i \in \mathcal{S}_i$ such that $(Z - Z_i)^2$ is easy to bound.

2.1 Variance of Functions with Bounded Differences

Efron-Stein is particularly usable on functions for which it is guaranteed that a change in a single input value changes the output value by only so much. Given a particular value of the i-th input variable $x_i \in \mathcal{X}$, let $f^{(i)}(x_i)$ denote the random variable $f(X_{1:i-1}, x_i, X_{i+1:n})$. Suppose for some $c_1 \dots c_n \in \mathbb{R}$,

$$\sup_{x_i \in \mathcal{X}} \left| f(X_{1:n}) - f^{(i)}(x_i) \right| \leq c_i \qquad \forall i = 1 \dots n \tag{5}$$

Applying equation (4) with $Z_i = (1/2)(\sup_{x_i \in \mathcal{X}} f^{(i)}(x_i) + \inf_{x_i \in \mathcal{X}} f^{(i)}(x_i))$,

$$Z - Z_i = \frac{1}{2} \left(\underbrace{\sup_{x_i \in \mathcal{X}} \left(f(X_{1:n}) - f^{(i)}(x_i) \right)}_{\in [0, c_i]} + \underbrace{\inf_{x_i \in \mathcal{X}} \left(f(X_{1:n}) - f^{(i)}(x_i) \right)}_{\in [-c_i, 0]} \right) \in \left[-\frac{c_i}{2}, \frac{c_i}{2} \right]$$

Thus

$$\mathrm{Var}\left(Z\right) \leq \frac{1}{4} \sum_i c_i^2 \tag{6}$$

2.2 Variance of Smooth Functions

Efron-Stein can be used to bound the variance of a smooth function $f : \mathbb{R}^n \to \mathbb{R}$ by the squared norm of its gradient $\nabla f(X_{1:n}) \in \mathbb{R}^n$. This additionally requires either Condition 1 or Condition 2 below.

Condition 1. We require that f is "separately convex" and its domain is compact: for simplicity $[0, 1]^n$. In this case, we can set $X_i' = \arg\inf_{x \in \mathbb{R}} f(X_{1:i-1}, x, X_{i+1:n})$ and define $Z_i = f(X_{1:i-1}, X_i', X_{i+1:n})$. Applying equation (4) and using the separate convexity again,

$$\mathrm{Var}\left(f(X_{1:n})\right) \leq \sum_i \mathbf{E}\left[\left([\nabla f(X_{1:n})]_i \underbrace{(X_i - X_i')}_{\in [-1,1]} \right)^2 \right] \leq \mathbf{E}\left[||\nabla f(X_{1:n})||^2 \right] \tag{7}$$

Condition 2. We require only that $X_i \sim \mathcal{N}(0, 1)$ are independently normal. By a nontrivial argument that uses the central limit theorem to decopose a normal variable into an infinite (scaled) sum of Rademacher variables then applies Efron-Stein,[3] we can still show $\mathrm{Var}\left(f(X_{1:n})\right) \leq \mathbf{E}\left[||\nabla f(X_{1:n})||^2 \right]$.

We can apply the above results to Lipschitz continuous functions. This is because of the equivalence between L-Lipschiz and $||\nabla f(x)|| \leq L$ for a smooth function (which is a helpful way to understand Lipschitz continuity).[4] Thus if $f(X_{1:n})$ is a smooth function that meets either Condition 1 or Condition 2, then

$$|f(X_{1:n}) - f(Y_{1:n})| \leq L\, ||X_{1:n} - Y_{1:n}|| \qquad \Longrightarrow \qquad \mathrm{Var}\left(f(X_{1:n})\right) \leq L^2 \tag{8}$$

3 Examples

3.1 Variance of the Largest Singular Value

Let $A \in [0,1]^{n \times m}$ be a random element-wise bounded matrix: that is, it consists of mn bounded random scalars $A_{i,j}$ drawn from some distribution over $[0,1]$. Assume $n \geq m$ WLOG. The largest singular value of A is a random scalar

$$\sigma := \max_{\substack{u' \in \mathbb{R}^n \colon ||u'||=1 \\ v' \in \mathbb{R}^m \colon ||v'||=1}} (u')^\top A v' = u^\top A v$$

where we denote the corresponding left and right singular vectors by $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ (which are also random in A). How does the variance of σ behave wrt. A (e.g., does it increase in the dimensions)? By piggybacking on the stability of the singular value wrt. a random (bounded) matrix followed by Efron-Stein, we can show, in at least two ways, that the variance is in fact constant wrt. the dimensions of A.

By re-sampling. For any $i \in [n]$ and $j \in [m]$, define $A^{(i,j)} \in [0,1]^{n \times m}$ to be a "perturbed" version of A where $A^{(i,j)}_{i,j}$ is an independent copy of $A_{i,j}$ and $A^{(i,j)}_{k,l} = A_{k,l}$ for all $k \neq i$ or $l \neq j$. The largest singular value of $A^{(i,j)}$ is denoted by $\bar{\sigma} = \bar{u}^\top A^{(i,j)} \bar{v}$ where the vectors $\bar{u}$ and $\bar{v}$ are random in $A^{(i,j)}$. We can bound the squared difference between σ and $\bar{\sigma}$ as follows:

$$(\sigma - \bar{\sigma})^2 = \left(u^\top A v - \underbrace{\bar{u}^\top A^{(i,j)} \bar{v}}_{\geq u^\top A^{(i,j)} v} \right)^2 \leq \left(u^\top \left(A - A^{(i,j)} \right) v \right)^2 \leq u_i^2 v_j^2$$

Then by equation (3) and the unit length of u and v:

$$\mathrm{Var}(\sigma) \leq \sum_{i,j} \mathbf{E}\left[\left(A - A^{(i,j)} \right)^2 \right] \leq \mathbf{E}\left[\sum_{i,j} u_i^2 v_j^2 \right] = \mathbf{E}\left[\left(\sum_{i=1}^n u_i^2 \right) \left(\sum_{j=1}^m v_j^2 \right) \right] = 1$$

By convexity and Lipschitz continuity. σ is a smooth function of nm bounded variables $A_{i,j} \in [0,1]$. To see σ is also convex (in particular, separately convex) in A,

$$\begin{aligned} \sigma = \sqrt{\lambda_1(A^\top A)} = \sqrt{\max_{v' \in \mathbb{R}^n \colon ||v'||=1} (v')^\top A^\top A v'} &= \max_{v' \in \mathbb{R}^n \colon ||v'||=1} \sqrt{(v')^\top A^\top A v'} \\ &= \max_{v' \in \mathbb{R}^n \colon ||v'||=1} ||Av'|| \end{aligned}$$

where the norm and max are both convex operations. Thus σ satisfies Condition 1 in Section 2.2. Furthermore, it is also 1-Lipschitz by Mirsky (Appendix A): for any $B \in \mathbb{R}^{n \times m}$ whose largest singular value is denoted by $\tilde{\sigma}$, we have

$$(\sigma - \tilde{\sigma})^2 \leq \sum_{i=1}^m (\sigma_i(A) - \sigma_i(B))^2 \leq ||A - B||_F^2 \quad \Longrightarrow \quad |\sigma - \tilde{\sigma}| \leq ||A - B||_F$$

It follows that $\mathrm{Var}(f(X_{1:n})) \leq 1$ by equation (8).

3.2 Variance of the Longest Common Subsequence Length

Consider the length of the longest common subsequence between random sequences of length n. There are $2n$ random variables each of which can cause at most a difference of 1 in the length. Thus by equation (6), the variance of the length is bounded by $n/2$. Note that it does not matter how complicated the function is: all that matters is that it is a function with bounded differences.

Reference. *Concentration Inequalities* (Boucheron, Lugosi, and Massart)

Notes

[1]This shows that $\Delta_1 \dots \Delta_n$ is a martingale difference sequence with respect to $X_1 \dots X_n$.

[2]Two variables $X, Y \in \mathbb{R}$ are **orthogonal** if $\mathbf{E}[XY] = 0$.

[3]By equation (2), it is sufficient to show the result for $n = 1$: refer the sole input by $X \sim \mathcal{N}(0,1)$. Let ϵ_i denote an independent Rademacher variable. Since $S_m := \sum_{i=1}^m \epsilon_i/\sqrt{m}$ converges in distribution to X as $m \to \infty$, it is sufficient to bound $\mathrm{Var}(f(S_m))$ by $\mathbf{E}[f'(S_m)]$ as $m \to \infty$. Applying equation (2) again, it is sufficient to bound $\mathrm{Var}^{(i)}(f(S_m))$. Writing $\mathrm{Var}^{(i)}(f(S_m)) = (1/2)\mathbf{E}^{(i)}\left[(f(S_m) - f(S'_m))^2\right]$ where S'_m is distributed the same as S_m except for an iid resampling of ϵ_i by ϵ'_i, and using the probability density of Rademacher variables, we have

$$\mathrm{Var}^{(i)}(f(S_m)) \le \frac{1}{2}\left(\frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 0 + \frac{1}{2}\left(f(S_m^{+i}) - f(S_m^{-i})\right)^2\right) = \frac{1}{4}\left(f(S_m^{+i}) - f(S_m^{-i})\right)^2$$

where S_m^{+i} and S_m^{-i} are S'_m with $\epsilon'_i = 1$ and $\epsilon'_i = -1$. To bound the difference between $f(S_m^{+i})$ and $f(S_m^{-i})$ in terms of $f(S_m)$, apply a first-order Taylor's approximation of each around S_m: if $K := \sup_{x\in\mathbb{R}} f''(x)$, by the remainder theorem $f(S_m^{+i}) \le f(S_m) + f'(S_m)(1-\epsilon_i)/\sqrt{m} + K(1-\epsilon_i)^2/m$ and $f(S_m^{-i}) \le f(S_m) + f'(S_m)(-1-\epsilon_i)/\sqrt{m} + K(-1-\epsilon_i)^2/m$. Then $\left|f(S_m^{+i}) - f(S_m^{-i})\right| \le (2/\sqrt{m}\,|f'(S_n)| + 4K/m$, and

$$\begin{aligned}\mathrm{Var}(f(S_m)) \le \sum_{i=1}^m \mathbf{E}\left[\mathrm{Var}^{(i)}(f(S_m))\right] &= \frac{1}{4}\sum_{i=1}^m \mathbf{E}\left[\left(f(S_m^{+i}) - f(S_m^{-i})\right)^2\right] \\ &\le \mathbf{E}\left[f'(S_m)^2 + \frac{4K^2}{m} + \frac{2K}{\sqrt{m}}\left|f'(S_m)\right|\right] \xrightarrow{m\to\infty} \mathbf{E}\left[f'(X)^2\right]\end{aligned}$$

[4]**Claim.** For a smooth function $f : \mathbb{R}^n \to \mathbb{R}$,

$$|f(x) - f(y)| \le L\,||x-y|| \;\;\forall x, y \in \mathbb{R}^n \qquad \iff \qquad ||\nabla f(x)|| \le L \;\; \forall x \in \mathbb{R}^n$$

Proof. Suppose $|f(x) - f(y)| \le L\,||x-y||$. Recall the definition using the directional derivative: the gradient of f at a is a unique vector $\nabla f(a) \in \mathbb{R}^n$ such that

$$\nabla f(a)^\top v := D_v f(a) = \lim_{\epsilon\to 0}\frac{f(a+\epsilon v) - f(a)}{\epsilon} \qquad \forall v \in \mathbb{R}^n :\; ||v|| = 1$$

Plug in $v = \nabla f(a)/\,||\nabla f(a)||$ and use the Lipshitz condition to have

$$||\nabla f(a)|| = \lim_{\epsilon\to 0}\frac{f\left(a + \epsilon\frac{f(a)}{||\nabla f(a)||}\right) - f(a)}{\epsilon} \le \lim_{\epsilon\to 0}\frac{L\left|\left|\epsilon\frac{f(a)}{||\nabla f(a)||}\right|\right|}{\epsilon} \le L$$

Conversely, assume $||\nabla f(a)|| \le L$. By the (multivariate) mean value theorem, given any $x, y \in \mathbb{R}^n$ there is some point $c \in \mathbb{R}^n$ on the line in between x and y such that $f(x) - f(y) = \nabla f(c)^\top (x-y)$. Thus by Cauchy-Schwarz and the assumption, $||f(x) - f(y)|| \le ||\nabla f(c)||\,||x-y|| \le L\,||x-y||$. □

A Proving Mirsky's Singular Value Inequality

Mirsky's inequality states that for any $A, B \in \mathbb{R}^{n\times m}$ where $n \geq m$, if $\sigma_1(A) \geq \cdots \geq \sigma_m(A) \geq 0$ and $\sigma_1(B) \geq \cdots \geq \sigma_m(B) \geq 0$ denote the ordered singular values of A and B, then the sum of the squared difference between the ordered singular values of A and B is at most the sum of the squared singular values of $A - B \in \mathbb{R}^{n\times m}$:

$$\sum_{i=1}^{m} (\sigma_i(A) - \sigma_i(B))^2 \leq \sum_{i=1}^{m} \sigma_i(A-B)^2 \tag{9}$$

This is also written in the Frobenius norm. If $M \in \mathbb{R}^{n\times m}$ is a matrix with SVD $M = U\Sigma V^\top$, then using the cyclic property of trace,

$$\begin{aligned} ||M||_F^2 &:= \sum_{i,j} M_{i,j}^2 =: \operatorname{Tr}\left(M^\top M\right) = \operatorname{Tr}\left(V\Sigma U^\top U\Sigma V^\top\right) = \operatorname{Tr}\left(V\Sigma^2 V^\top\right) \\ &= \operatorname{Tr}\left(\Sigma^2 V^\top V\right) = \operatorname{Tr}\left(\Sigma^2\right) = \sum_i \sigma_i(M)^2 \end{aligned}$$

Thus we can write equation (9) as

$$\sum_{i=1}^{m} (\sigma_i(A) - \sigma_i(B))^2 \leq ||A-B||_F^2 \tag{10}$$

Proving Mirsky is a short and elegant exercise of some fundamental concepts in linear algebra, hence written here. We first prove the eigenvalue version of the inequality, and then extend it to the singular value version by a *linear* relationship between the singular values of M and the eigenvalues of $[0M; M^\top 0]$ (as opposed to the nonlinear relationship between the singular values of M and the eigenvalues of $M^\top M$).

A.1 Eigenvalue Version

Theorem A.1. *For symmetric matrices $A, B \in \mathbb{R}^{n\times n}$ with ordered real eigenvalues $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ and $\lambda_1(B) \geq \cdots \geq \lambda_n(B)$, we have $\sum_{i=1}^{m} (\lambda_i(A) - \lambda_i(B))^2 \leq ||A-B||_F^2$.*

Proof. The proof starts from the RHS of the inequality and establishes an *exact* equality

$$\sum_{i,j} P_{i,j} (\lambda_i(A) - \lambda_j(B))^2 = ||A-B||_F^2 \tag{11}$$

for *some* doubly stochastic matrix $P \in [0,1]^{n\times n}$. It then shows that the LHS is minimized at $P = I_{n\times n}$ over the space of n-dimensional doubly stochastic matrices D_n. Since we know that at least one $P \in D_n$ achieves the exact equality, this proves Theorem A.1.

Showing equation (11). Let $A = V_A\Lambda_A V_A^\top$ and $B = V_B\Lambda_B V_B^\top$ denote eigenvalue decompositions where $V_A, V_B \in \mathbb{R}^{n\times n}$ are orthogonal and $\Lambda_M := \operatorname{diag}(\lambda_1(M)\ldots\lambda_n(M))$ for $M = A, B$. We want to express the RHS in terms of eigenvalues. To this end, note that

$$||A-B||_F^2 = \left|\left|V_A\Lambda_A V_A^\top - V_B\Lambda_B V_B^\top\right|\right|_F^2 = \left|\left|\Lambda_A V_A^\top V_B - V_A^\top V_B\Lambda_B\right|\right|_F^2$$

where we use the orthogonal invarance of the Frobenius norm. Thus defining $Q := V_A^\top V_B$, we have

$$||\Lambda_A Q - Q\Lambda_B||_F^2 = \sum_{i,j} Q_{i,j}^2 \left(\lambda_i(A) - \lambda_j(B)\right)^2 = ||A - B||_F^2$$

where the first equality holds because $\Lambda_A Q - Q\Lambda_B$ is a matrix whose (i,j)-th element is $Q_{i,j}\left(\lambda_i(A) - \lambda_j(B)\right)$. Now, we observe that Q is orthogonal ($Q^\top Q = V_B^\top V_A V_A^\top V_B = I_{n\times n}$) and use the fact that *an element-wise square $Q^{\langle 2\rangle}$ of an orthogonal matrix Q is a doubly stochastic matrix*. This follows because the columns (or the rows) of Q are unit-length, so $\sum_i Q_{i,j}^{\langle 2\rangle} = \sum_i Q_{i,j}^2 = ||Q_{:,j}||_2^2 = 1$ (similarly for other indices). Thus equation (11) holds for $P = Q^{\langle 2\rangle} \in D_n$.

Opitmizing equation (11) over P. We search for

$$P^* \in \operatorname*{arg\,inf}_{P\in D_n} \sum_{i,j} P_{i,j}(\lambda_i(A) - \lambda_j(B))^2$$

For clarity, consider a vectorized form. The vectorization operator $\mathrm{vec}(M)$ concatenates the columns of given matrix M and is an isomorphism between the matrix/vector spaces. Letting $c = \mathrm{vec}([(\lambda_i(A) - \lambda_j(B))^2]_{i,j}) \in \mathbb{R}^{n^2}$, we equivalently search for

$$p^* \in \operatorname*{arg\,inf}_{p\in\{\mathrm{vec}(P):\, P\in D_n\}} c^\top p$$

from which it is clear that we are solving a linear program with a bounded constraint set and feasible solutions. Therefore, the minimum value will be attained at a vertex of the feasible region. By the Birkhoffvon Neumann theorem, the vertices of D_n are the $(n \times n)$ permutation matrices.

Now we argue that the objective (11) has a larger (or equal) value if we set P to be a non-identity permutation matrix rather than $P = I_{n\times n}$. A permutation matrix can be expressed as $I_{n\times n}$ left-multiplied by elementary row-exchanging matrices. Consider any nontrivial row-exchaning matrix $A \in \{0,1\}^{n\times n}$ where for some $i \neq j$ we have $A_{i,j} = A_{j,i} = 1$, $A_{k,k} = 1$ for all $k \notin \{i,j\}$, and zero everywhere else. The contribution to the objective different from using $P = I_{n\times n}$ is $(\lambda_i(A) - \lambda_j(B))^2 + (\lambda_i(A) - \lambda_j(B))^2$. Using the ordering of eigenvalues ($\lambda_i \geq \lambda_j$ if $i < j$), it can be checked that

$$(\lambda_i(A) - \lambda_j(B))^2 + (\lambda_i(A) - \lambda_j(B))^2 \geq (\lambda_i(A) - \lambda_i(B))^2 + (\lambda_j(A) - \lambda_j(B))^2$$

Thus any row exchange can only increase (or not affect) the objective value. □

A.2 Singular Value Version

A valuable link between SVD and eigendecomposition is given by the following lemma.

Lemma A.2 (Theorem 7.3.3, HJ)**.** *Let $n \geq m$ and $M \in \mathbb{R}^{n\times m}$ with a full SVD: $M = [U_1 U_2][\Sigma_m; 0_{(n-m)\times m}]V^\top$ where $\Sigma_m = diag(\sigma_1 \ldots \sigma_m)$ with $\sigma_1 \geq \ldots \geq \sigma_m$. Define an $(n+m)\times(n+m)$ symmetric matrix*

$$\widetilde{M} := \begin{bmatrix} 0_{n\times n} & M \\ M^\top & 0_{m\times m} \end{bmatrix} \tag{12}$$

Then the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{n+m}$ *of* $\widetilde{M}$ *are given by*

$$\sigma_1 \geq \ldots \geq \sigma_m \geq \underbrace{0 \geq \cdots \geq 0}_{n-m} \geq -\sigma_m \geq \ldots \geq -\sigma_1$$

Proof. Define $W, \Lambda \in \mathbb{R}^{(n+m)\times(n+m)}$ by

$$W := \begin{bmatrix} U_1/\sqrt{2} & U_2 & -U_1/\sqrt{2} \\ V/\sqrt{2} & 0_{m\times(n-m)} & V/\sqrt{2} \end{bmatrix} \qquad \Lambda := \operatorname{diag}\left(\Sigma_m, 0_{(n-m)\times(n-m)}, -\Sigma_m,\right)$$

Using the orthonormality between U_1 and U_2 and the scaling by $\sqrt{2}$, we can algebraically verify that $W^\top W = I_{(n+m)\times(n+m)}$ and $\widetilde{M} = W\Lambda W^\top$. Note that the math also works if we move the minus sign from block $(1,3)$ to $(2,3)$ in W. □

Extracting singular vectors. The proof of Lemma A.2 also gives a way to extract singular vectors of A from orthonormal eigenvectors of $\widetilde{M}$. Even though an orthonormal eigendecomposition $\widetilde{M} = W\Lambda W^\top$ is not unique, it finds an orthonormal basis for the eigenspace associated with each eigenvalue σ_i: the choice of the basis does not affect the property of W as the singular vectors of M.

Proof of Mirsky Let $\widetilde{A}, \widetilde{B} \in \mathbb{R}^{(n+m)\times(n+m)}$ denote equation (12) applied to $A, B \in \mathbb{R}^{n\times m}$. By Theorem A.1 and Lemma A.2, we have

$$\begin{aligned} \sum_{i=1}^{n+m} \left(\lambda_i(\widetilde{A}) - \lambda_i(\widetilde{B})\right)^2 &= 2\sum_{i=1}^{m} (\sigma_i(A) - \sigma_i(B))^2 \\ &\leq \left|\left|\widetilde{A} - \widetilde{B}\right|\right|_F^2 = \sum_{i=1}^{n+m} \sigma_i(\widetilde{A} - \widetilde{B})^2 = 2\sum_{i=1}^{m} \sigma_i(A-B)^2 \\ &= 2\,||A-B||_F^2 \end{aligned}$$

Reference. *Matrix Analysis (2nd Edition)* (Horn and Johnson)