Derivative of the first order SVD

Randolf Scholz

May 16, 2023

Consider computing the first order SVD expansion. By the Eckart–Young–Mirsky theorem, this is equivalent to solving

$$\underset{\sigma,\,u,\,v}{\text{minimize}}\ \tfrac{1}{2}\|A-\sigma uv^\top\|_F^2 \quad \text{s.t.} \quad \|u\|=1 \quad \text{and} \quad \|v\|=1 \quad \text{and} \quad \sigma \geq 0$$

Equivalently this may be formalized as

$$\sigma = \max_{u,\,v} u^\top Av \quad \text{s.t.} \quad \|u\|=1 \quad \text{and} \quad \|v\|=1$$

Which is a non-convex quadratically constrained quadratic program (QCQP)

$$\sigma = \max_{u,v} \frac{1}{2}\begin{bmatrix}u\\v\end{bmatrix}^\top \cdot \begin{bmatrix}\mathbf{0}_{m\times m} & A\\ A^\top & \mathbf{0}_{n\times n}\end{bmatrix} \cdot \begin{bmatrix}u\\v\end{bmatrix} \text{ s.t. } \begin{matrix}\begin{bmatrix}u\\v\end{bmatrix}^\top \cdot \begin{bmatrix}\mathbb{I}_m & \mathbf{0}_{m\times n}\\ \mathbf{0}_{n\times m} & \mathbf{0}_{n\times n}\end{bmatrix} \cdot \begin{bmatrix}u\\v\end{bmatrix} = 1\\ \begin{bmatrix}u\\v\end{bmatrix}^\top \cdot \begin{bmatrix}\mathbf{0}_{m\times m} & \mathbf{0}_{m\times n}\\ \mathbf{0}_{n\times m} & \mathbb{I}_n\end{bmatrix} \cdot \begin{bmatrix}u\\v\end{bmatrix} = 1\end{matrix}$$

The Jacobian and Lagrangian The derivative of the objective function is

$$\mathbf{J}_f(A,\, \left[\begin{smallmatrix}\sigma\\u\\v\end{smallmatrix}\right]) = \begin{bmatrix} & \sigma - u^\top Av\\ A-\sigma uv^\top & \sigma^2 u - \sigma Av\\ & \sigma^2 v - \sigma A^\top u\end{bmatrix} \implies \mathbf{H}_f(\left[\begin{smallmatrix}\sigma\\u\\v\end{smallmatrix}\right]) = \begin{bmatrix}1 & 2\sigma u - Av & 2\sigma v - A^\top u\\ -Av & \sigma^2\mathbb{I}_m & -\sigma A\\ -A^\top u & -\sigma A^\top & \sigma^2 \mathbb{I}_n\end{bmatrix}$$

Consider the function

$$f(A,\, \left[\begin{smallmatrix}\sigma\\u\\v\end{smallmatrix}\right]) = \begin{pmatrix}\sigma - u^\top Av\\ \sigma^2 u - \sigma Av\\ \sigma^2 v - \sigma A^\top u\end{pmatrix} \equiv \mathbf{0} \implies \mathbf{J}_f(A,\, \left[\begin{smallmatrix}\sigma\\u\\v\end{smallmatrix}\right]) = \left[\begin{array}{c|ccc} -\xi vu^\top & 1 & 2\sigma u - Av & 2\sigma v - A^\top u\\ -\sigma v\phi^\top & -Av & \sigma^2\mathbb{I}_m & -\sigma A\\ -\sigma u\psi^\top & -A^\top u & -\sigma A^\top & \sigma^2\mathbb{I}_n\end{array}\right]$$

Thus, gradient descent schema is

$$\begin{aligned}\sigma' &= \sigma - \eta_\sigma(\sigma - u^\top Av)\\ u' &= u - \eta_u(\sigma^2 u - \sigma Av)\\ v' &= v - \eta_v(\sigma^2 v - \sigma A^\top u)\end{aligned}$$

And the newton step with diagonal approximation of the hessian:

$$\begin{aligned}
\sigma' &= \sigma - 1(\sigma - u^\top Av) &&= u^\top Av \\
u' &= u - \tfrac{1}{\sigma^2}(\sigma^2 u - \sigma Av) &&= \tfrac{1}{\sigma}Av \\
v' &= v - \tfrac{1}{\sigma^2}(\sigma^2 v - \sigma A^\top u) &&= \tfrac{1}{\sigma}A^\top u
\end{aligned}$$

1 Analysis of the backward

At the equilibrium point, we have:

$$\sigma = u^\top Av \qquad Av = \sigma u \qquad A^\top u = \sigma v \qquad u^\top u = 1 \qquad v^\top v = 1$$

Note that this states that σ is an eigenvalue:

$$\begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} = \sigma \begin{bmatrix} u \\ v \end{bmatrix}$$

In particular, Rayleigh iteration could be useful. from this we can derive

$$\Delta\sigma = \Delta u^\top Av + u^\top \Delta A v + u^\top A \Delta v = \Delta u^\top u + u^\top \Delta A v + v^\top \Delta v = u^\top \Delta A v$$

Where in the last step we used $\Delta u \perp u$ and $\Delta v \perp v$, which follows from the side condition. Further we have:

$$\begin{aligned}
\Delta\sigma u + \sigma \Delta u &= \Delta A v + A\Delta v \\
\Delta\sigma v + \sigma \Delta v &= \Delta A^\top u + A^\top \Delta u
\end{aligned}
\iff
\underbrace{\begin{bmatrix} \sigma \mathbb{I}_m & -A \\ -A^\top & \sigma \mathbb{I}_n \end{bmatrix}}_{=:K} \cdot \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = \begin{bmatrix} \Delta A v - \Delta\sigma u \\ \Delta A^\top u - \Delta \sigma v \end{bmatrix}$$

which allows us to express Δu and Δv in terms of ΔA. The constraints yield

$$\begin{aligned}
u^\top \Delta u + \Delta u^\top u = 0 &\iff u \perp \Delta u \\
v^\top \Delta v + \Delta v^\top v = 0 &\iff v \perp \Delta v
\end{aligned}$$

We can augment the original system with these:

$$\underbrace{\begin{bmatrix} \sigma \mathbb{I}_m & -A \\ -A^\top & \sigma \mathbb{I}_n \\ u^\top & \mathbf{0}_n^\top \\ \mathbf{0}_m^\top & v^\top \end{bmatrix}}_{=:\widetilde{K}} \cdot \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = \underbrace{\begin{bmatrix} \Delta A v - \Delta\sigma u \\ \Delta A^\top u - \Delta \sigma v \\ 0 \\ 0 \end{bmatrix}}_{=:\widetilde{c}}$$

2 VJP with modified K matrix

$$\begin{aligned}
\left\langle \begin{bmatrix}\phi\\ \psi\end{bmatrix} \;\middle|\; \begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right\rangle &= \left\langle \begin{bmatrix}\phi\\ \psi\end{bmatrix} \;\middle|\; \widetilde{K}^{-1}\widetilde{c}\right\rangle \\
&= \left\langle \widetilde{K}^{-\top}\begin{bmatrix}\phi\\ \psi\end{bmatrix} \;\middle|\; \widetilde{c}\right\rangle \\
&= \left\langle \begin{bmatrix}\sigma\mathbb{I}_m & -A & u & \mathbf{0}_m\\ -A^\top & \sigma\mathbb{I}_n & \mathbf{0}_n & v\end{bmatrix}^{-1}\begin{bmatrix}\phi\\ \psi\end{bmatrix} \;\middle|\; \begin{bmatrix}\Delta Av-\Delta\sigma u\\ \Delta A^\top u-\Delta\sigma v\\ 0\\ 0\end{bmatrix}\right\rangle \\
&= \left\langle \begin{bmatrix}\sigma\mathbb{I}_m & -A & u & \mathbf{0}_m\\ -A^\top & \sigma\mathbb{I}_n & \mathbf{0}_n & v\end{bmatrix}\begin{bmatrix}p\\ q\\ \lambda\\ \mu\end{bmatrix} = \begin{bmatrix}\phi\\ \psi\end{bmatrix} \;\middle|\; \begin{bmatrix}\Delta Av-\Delta\sigma u\\ \Delta A^\top u-\Delta\sigma v\\ 0\\ 0\end{bmatrix}\right\rangle
\end{aligned}$$

2.1 augmented part multiplied with inverse K

$$\begin{aligned}
K^{-1}\begin{bmatrix}u & 0\\ 0 & v\end{bmatrix} &= \begin{bmatrix}\sigma(\sigma^2\mathbb{I}_m - AA^\top)^{-1} & (\sigma^2\mathbb{I}_m - AA^\top)^{-1}A\\ (\sigma^2\mathbb{I}_n - A^\top A)^{-1}A^\top & \sigma(\sigma^2\mathbb{I}_n - A^\top A)^{-1}\end{bmatrix}\begin{bmatrix}u & 0\\ 0 & v\end{bmatrix} \\
&= \begin{bmatrix}(\sigma^2\mathbb{I}_m - AA^\top)^{-1}\sigma u & (\sigma^2\mathbb{I}_m - AA^\top)^{-1}Av\\ (\sigma^2\mathbb{I}_n - A^\top A)^{-1}A^\top u & (\sigma^2\mathbb{I}_n - A^\top A)^{-1}\sigma v\end{bmatrix} \\
&= \begin{bmatrix}(\sigma^2\mathbb{I}_m - AA^\top)^{-1}\sigma u & (\sigma^2\mathbb{I}_m - AA^\top)^{-1}Av\\ (\sigma^2\mathbb{I}_n - A^\top A)^{-1}A^\top u & \sigma(\sigma^2\mathbb{I}_n - A^\top A)^{-1}v\end{bmatrix}
\end{aligned}$$

3 The VJP

The last equation allows us to compute the VJP at ease:

$$\begin{aligned}
\left\langle \begin{bmatrix}\phi\\ \psi\end{bmatrix} \;\middle|\; \begin{bmatrix}\Delta u\\ \Delta v\end{bmatrix}\right\rangle &= \left\langle \begin{bmatrix}\phi\\ \psi\end{bmatrix} \;\middle|\; K^{-1}\begin{bmatrix}\Delta Av-\Delta\sigma u\\ \Delta A^\top u-\Delta\sigma v\end{bmatrix}\right\rangle \\
&= \left\langle K^{-\top}\begin{bmatrix}\phi\\ \psi\end{bmatrix} \;\middle|\; \begin{bmatrix}\Delta Av-\Delta\sigma u\\ \Delta A^\top u-\Delta\sigma v\end{bmatrix}\right\rangle \\
&= \left\langle \begin{bmatrix}\tilde{\phi}\\ \tilde{\psi}\end{bmatrix} \;\middle|\; \begin{bmatrix}\Delta Av-\Delta\sigma u\\ \Delta A^\top u-\Delta\sigma v\end{bmatrix}\right\rangle
\end{aligned}$$

Now, we compute the terms individually:

$$\begin{aligned}
\langle\tilde{\phi} \mid \Delta Av - \Delta\sigma u\rangle &= \langle\tilde{\phi}v^\top \mid \Delta A\rangle - \langle u^\top\tilde{\phi} \mid \Delta\sigma\rangle \\
&= \langle\tilde{\phi}v^\top \mid \Delta A\rangle - \langle u^\top\tilde{\phi} \mid u^\top\Delta Av\rangle \\
&= \langle(\mathbb{I}_m - uu^\top)\tilde{\phi}v^\top \mid \Delta A\rangle
\end{aligned}$$

And for the second term we get

$$\begin{aligned}
\langle\tilde{\psi} \mid \Delta A^\top u - \Delta\sigma v\rangle &= \langle\tilde{\psi}u^\top \mid \Delta A^\top\rangle - \langle v^\top\tilde{\psi} \mid \Delta\sigma\rangle \\
&= \langle u\tilde{\psi}^\top \mid \Delta A\rangle - \langle\tilde{\psi}^\top v \mid u^\top\Delta Av\rangle \\
&= \langle u\tilde{\psi}(\mathbb{I}_n - vv^\top) \mid \Delta A\rangle
\end{aligned}$$

Using the formula for inverting a 2×2 block-matrix, we can give an explicit solution to $K^{-\top}\left[\begin{smallmatrix}\phi\\ \psi\end{smallmatrix}\right]$:

$$\begin{aligned}
K^{-1} = \begin{bmatrix}\sigma\mathbb{I}_m & -A\\ -A^\top & \sigma\mathbb{I}_n\end{bmatrix}^{-1} &= \begin{bmatrix}(\sigma\mathbb{I}_m - \frac{1}{\sigma}AA^\top)^{-1} & \mathbf{0}_{m\times n}\\ \mathbf{0}_{n\times m} & (\sigma\mathbb{I}_n - \frac{1}{\sigma}A^\top A)^{-1}\end{bmatrix} \cdot \begin{bmatrix}\mathbb{I}_m & \frac{1}{\sigma}A\\ \frac{1}{\sigma}A^\top & \mathbb{I}_n\end{bmatrix}\\
&= \begin{bmatrix}\sigma(\sigma^2\mathbb{I}_m - AA^\top)^{-1} & (\sigma^2\mathbb{I}_m - AA^\top)^{-1}A\\ (\sigma^2\mathbb{I}_n - A^\top A)^{-1}A^\top & \sigma(\sigma^2\mathbb{I}_n - A^\top A)^{-1}\end{bmatrix}
\end{aligned}$$

And we see it's basically projection operators with respect to the image/kernel of $\tilde{A} = \frac{1}{\sigma}A$. In summary, we obtain the following formula for the VJP:

$$K\begin{bmatrix}p\\ q\end{bmatrix} = \begin{bmatrix}\phi\\ \psi\end{bmatrix} \iff \begin{bmatrix}p\\ q\end{bmatrix} = \begin{bmatrix}\sigma(\sigma^2\mathbb{I}_m - AA^\top)^{-1} & (\sigma^2\mathbb{I}_m - AA^\top)^{-1}A\\ (\sigma^2\mathbb{I}_n - A^\top A)^{-1}A^\top & \sigma(\sigma^2\mathbb{I}_n - A^\top A)^{-1}\end{bmatrix}\begin{bmatrix}\phi\\ \psi\end{bmatrix}$$

In particular, we can find the solution by solving 4 smaller linear systems:

$$\begin{aligned}
\sigma(\sigma^2\mathbb{I}_m - AA^\top)^{-1}\phi &= x & \qquad (\sigma^2\mathbb{I}_m - AA^\top)^{-1}A\psi &= y\\
(\sigma^2\mathbb{I}_n - A^\top A)^{-1}A^\top\phi &= w & \qquad \sigma(\sigma^2\mathbb{I}_n - A^\top A)^{-1}\psi &= z
\end{aligned}$$

Or, equivalently:

$$\begin{aligned}
(\sigma^2\mathbb{I}_m - AA^\top)x &= \sigma\phi & \qquad (\sigma^2\mathbb{I}_m - AA^\top)y &= A\psi\\
(\sigma^2\mathbb{I}_n - A^\top A)w &= A^\top\phi & \qquad (\sigma^2\mathbb{I}_n - A^\top A)z &= \sigma\psi
\end{aligned}$$

Note how this shows that the off-diagonal entries are solutions to regularized least squares problems! However, we really do not want to compute the matrices $AA^\top$ and $A^\top A$ since this leads to numerical stability (squared condition number!) To circumvent this issue, we do a reformulation

$$\begin{aligned}
(\sigma^2\mathbb{I}_m - AA^\top)y = A\psi \quad &\iff y = \operatorname*{argmin}_y \| - A^\top y - \psi\|_2^2 - \sigma^2\|y\|_2^2\\
&\iff y = \operatorname*{argmin}_y \left\|\begin{bmatrix}A^\top\\ \sigma^2\mathbb{I}_m\end{bmatrix} y - \begin{bmatrix}-\psi\\ \mathbf{0}_m\end{bmatrix}\right\|_2^2\\
(\sigma^2\mathbb{I}_n - A^\top A)w = A^\top\phi &\iff w = \operatorname*{argmin}_w \|Aw + \phi\|_2^2 - \sigma^2\|w\|_2^2\\
&\iff w = \operatorname*{argmin}_w \left\|\begin{bmatrix}A\\ \sigma^2\mathbb{I}_n\end{bmatrix} w - \begin{bmatrix}-\phi\\ \mathbf{0}_n\end{bmatrix}\right\|_2^2
\end{aligned}$$

Remark 1 (When is Ridge Regression unconstrained?). Consider the problem

$$\beta^* = \operatorname*{argmin}_\beta \|X\beta - y\|^2 + \lambda\|\beta\|^2$$

Question: When is there an unconstrained solution? The solution satisfies the normal equation

$$(X^T X + \lambda\mathbb{I})\beta = X^\top y$$

If $\lambda > 0$, then $(X^T X + \lambda\mathbb{I})$ is positive definite and hence invertible. If $\lambda < 0$, then $(X^T X + \lambda\mathbb{I})$ is singular whenever λ is an eigenvalue of $X^T X$. In particular, the 4 systems listed before are all ill-conditioned! The central issue is that the constraint is missing! $\|u\|^2 = 1$ and $\|v\|^2 = 1$ translate to $u \perp \Delta u$ and $v \perp \Delta v$. Since u, v are singular vectors, this means we avoid the singular subspace when solving these equations!

What we should do is use **Riemannian Optimization**.

3.1 What happens if ϕ or ψ are zero?

In this case we want to fast track the calculation, meaning skip half of the necessary inversions. Looking at the equations we find that if $\phi = 0$ then $x = 0$ and $w = 0$, and if $\psi = 0$ then $y = 0$ and $z = 0$. This suggests that backward substitution is better than forward substitution, since it allows decoupling of the two gradient contributions.

3.2 Via Forward Substitution

Now, the diagonal entries we have a problem: the RHS lacks the A matrix. Thus, we solve in two steps instead:

$$\begin{aligned} A\mu = \sigma\phi &\implies x = \operatorname*{argmin}_x \left\| \begin{bmatrix} A^\top \\ \sigma^2\mathbb{I}_m \end{bmatrix} x - \begin{bmatrix} -\mu \\ \mathbf{0}_m \end{bmatrix} \right\|_2^2 \\ A^\top\nu = \sigma\psi &\implies z = \operatorname*{argmin}_z \left\| \begin{bmatrix} A \\ \sigma^2\mathbb{I}_n \end{bmatrix} z - \begin{bmatrix} -\nu \\ \mathbf{0}_n \end{bmatrix} \right\|_2^2 \end{aligned}$$

We can optimize further by performing a simultaneous solve:

$$\begin{aligned} \big[x,\, y\big] &= \operatorname*{argmin}_{x,y} \left\| \begin{bmatrix} A^\top \\ \sigma^2\mathbb{I}_m \end{bmatrix} \big[x,\, y\big] - \begin{bmatrix} -\mu & -\psi \\ \mathbf{0}_m & \mathbf{0}_m \end{bmatrix} \right\|_2^2 & \mu &= \operatorname*{argmin}_\mu \|A\mu - \sigma\phi\|_2^2 \\ \big[w,\, z\big] &= \operatorname*{argmin}_{w,z} \left\| \begin{bmatrix} A \\ \sigma^2\mathbb{I}_n \end{bmatrix} \big[w,\, z\big] - \begin{bmatrix} -\phi & -\nu \\ \mathbf{0}_n & \mathbf{0}_n \end{bmatrix} \right\|_2^2 & \nu &= \operatorname*{argmin}_\nu \|A^\top\nu - \sigma\psi\|_2^2 \end{aligned}$$

3.3 Via Backward Substitution

We need to introduce an additional modification:

If $A\mu = \sigma\phi$ not solveable, we instead can multiply the equation by $A^\top$ to obtain:

$$\begin{aligned} (\sigma^2\mathbb{I}_m - AA^\top)x = \sigma\phi \quad &\implies \quad (\sigma^2\mathbb{I}_n - A^\top A)\mu = \sigma A^\top\phi \quad & A^\top x = \mu \\ (\sigma^2\mathbb{I}_n - A^\top A)z = \sigma\psi \quad &\implies \quad (\sigma^2\mathbb{I}_n - A^\top A)A\nu = \sigma A\psi \quad & Az = \nu \end{aligned}$$

So:

$$\begin{aligned} \mu &= \operatorname*{argmin}_\mu \left\| \begin{bmatrix} A \\ \sigma^2\mathbb{I}_n \end{bmatrix} \mu - \begin{bmatrix} -\sigma\phi \\ \mathbf{0}_n \end{bmatrix} \right\|_2^2 & A^\top x &= \mu \\ \nu &= \operatorname*{argmin}_\mu \left\| \begin{bmatrix} A^\top \\ \sigma^2\mathbb{I}_m \end{bmatrix} \nu - \begin{bmatrix} -\sigma\psi \\ \mathbf{0}_m \end{bmatrix} \right\|_2^2 & Az &= \nu \end{aligned}$$

So

$$\begin{bmatrix}\mu & w\end{bmatrix} = \begin{bmatrix}A \\ \sigma^2\mathbb{I}_n\end{bmatrix}\begin{bmatrix}-\sigma\phi & -\phi \\ \mathbf{0}_n & \mathbf{0}_n\end{bmatrix}$$
$$\begin{bmatrix}y & \nu\end{bmatrix} = \begin{bmatrix}A^\top \\ \sigma^2\mathbb{I}_m\end{bmatrix}\begin{bmatrix}-\psi & -\sigma\psi \\ \mathbf{0}_m & \mathbf{0}_m\end{bmatrix}$$

In principle, one could try to rephrase these as smaller problems, but for now, it's better to just stick to the bigger system. We can use the **push-through identity** to convert these into 4 linear systems:

$$Px = \phi \qquad Py = \tilde{A}\psi$$
$$Qz = \tilde{A}^\top\phi \qquad Qw = \psi$$

Then $\tilde{\phi} = x + y$ and $\tilde{\psi} = z + w$, and the VJP are given by the previous equations:

$$\xi^\top \frac{\partial \sigma}{\partial A} = \xi u v^\top$$
$$\phi^\top \frac{\partial u}{\partial A} = (\mathbb{I}_m - uu^\top)\tilde{\phi}v^\top = (\tilde{\phi} - (u^\top\tilde{\phi})u)v^\top$$
$$\psi^\top \frac{\partial v}{\partial A} = u\tilde{\psi}^\top(\mathbb{I}_n - vv^\top) = u(\tilde{\psi} - (v^\top\tilde{\psi})v)^\top$$

4 Spectral Normalization

The VJP of spectral normalization can be computed as follows: let $g(A) = \|A\|_2$ and V be the vector in the VJP. then

$$\begin{aligned}\nabla_A\langle V \mid \frac{A}{\|A\|_2}\rangle &= \langle V|\frac{A+\Delta A}{g(A+\Delta A)} - \frac{A}{g(A)}\rangle \\ &= \langle V|\frac{A+\Delta A}{g(A)+\nabla g(A)\Delta A} - \frac{A}{g(A)}\rangle \\ &= \langle V|\frac{(A+\Delta A)(g(A)-\nabla g(A)\Delta A)}{(g(A)+\nabla g(A)\Delta A)(g(A)-\nabla g(A)\Delta A)} - \frac{A}{g(A)}\rangle \\ &= \langle V|\frac{\Delta A g(A) - A\nabla g(A)\Delta A}{g(A)^2}\rangle \\ &= \langle \tfrac{1}{g(A)}V - \tfrac{\langle V|A\rangle}{g(A)}\nabla g(A)|\Delta A\rangle\end{aligned}$$

$$g(A) = 1 \implies \nabla_A\langle V \mid \frac{A}{\|A\|_2}\rangle = \langle V - \langle V \mid A\rangle\nabla g(A) \mid \Delta A\rangle$$

5 Projected gradient

When using spectral normalization we want to do the following:

$$\text{update:} \quad A' = A - \nabla_A \mathcal{L}(\frac{A}{\|A\|_2})$$
$$\text{project:} \quad A = \frac{A'}{\|A'\|_2}$$

Moreover, we want:

- During forward, compute $\frac{A}{\|A\|_2}$ only once and then reuse this node.
- Compute $\|A\|_2$ effectively between gradient updates.
  - Avoid built-in torch algos, as they make use of full SVD algos.
- After gradient update, perform projection step. (maybe unnecessary)

NOTE: gradients are different if we include normalization!