Bounded Approximate Solutions of Linear Systems using SVD

Stephen Brooks

December 18, 2023

1 Definitions

The Singular Value Decomposition (SVD) of a complex matrix is conventionally $A = U\Sigma V^*$, where M^* denotes $\bar{M}^T$. Here, U and V are unitary matrices with $U^{-1} = U^*$ and Σ is diagonal with $\Sigma = \mathrm{diag}[\sigma_n]$. For real matrices this is just $A = U\Sigma V^T$ and unitarity is equivalent to $U^{-1} = U^T$, i.e. orthogonality. In fact, V^T is also orthogonal since $(V^T)^{-1} = (V^{-1})^{-1} = V = (V^T)^T$, which means the simpler definition $A = U\Sigma V$ can be used for the rest of this note.

2 Fundamental Problem

In control systems, one often uses a linear or locally-linear model to determine the required inputs. Suppose an input vector change $\mathbf{x} \in X$ produces an output reponse $A\mathbf{x} \in Y$ that is meant to achieve some desired change $\mathbf{b} \in Y$. The input and output spaces X and Y may have different dimensionalities and therefore A can be a rectangular matrix. This means that an exact solution may not be possible, particularly if $\dim Y > \dim X$. Thus the 'best' solution can be formulated as the minimisation problem of finding $\arg\min |A\mathbf{x} - \mathbf{b}|_Y$.

However, particularly in the case of ill-conditioned matrices, the exact solution may require unacceptably large control inputs. What is required practically is the best approximation that can be achieved while $\mathbf{x}$ is not too large. This suggests casting the fundamental problem as

$$\arg \min_{|\mathbf{x}|_X \le r} |A\mathbf{x} - \mathbf{b}|_Y$$

with $r > 0$ being chosen depending on how large a solution is acceptable. As $r \to \infty$, the value will eventually settle at the exact or optimum solution if one exists.

3 Solution using SVD

The SVD decomposition of A gives

$$\arg \min_{|\mathbf{x}|_X \le r} |A\mathbf{x} - \mathbf{b}|_Y = \arg \min_{|\mathbf{x}|_X \le r} |U\Sigma V\mathbf{x} - \mathbf{b}|_Y.$$

Here, A and Σ are possibly-rectangular matrices mapping from X to Y, V is a square orthogonal matrix mapping X to itself and U is another mapping Y to itself. Note that any orthogonal matrix U preserves the norm as $|U\mathbf{x}|^2 = \mathbf{x}^T U^T U\mathbf{x} = \mathbf{x}^T U^{-1} U\mathbf{x} = \mathbf{x}^T\mathbf{x} = |\mathbf{x}|^2$ so $|U\mathbf{x}| = |\mathbf{x}|$ as norms are non-negative. In particular,

$$|\mathbf{x}|_X = |V\mathbf{x}|_X \qquad \text{and} \qquad |U\Sigma V\mathbf{x} - \mathbf{b}|_Y = |\Sigma V\mathbf{x} - U^{-1}\mathbf{b}|_Y,$$

where the second equality has multiplied by the unitary matrix U^{-1}. This means that

$$\arg\min_{|\mathbf{x}|_X \le r} |A\mathbf{x} - \mathbf{b}|_Y = \arg\min_{|V\mathbf{x}|_X \le r} |\Sigma V\mathbf{x} - U^{-1}\mathbf{b}|_Y.$$

Defining vectors $\mathbf{v} = V\mathbf{x}$ and $\mathbf{u} = U^{-1}\mathbf{b}$ this becomes

$$\arg\min_{|\mathbf{x}|_X \le r} |A\mathbf{x} - \mathbf{b}|_Y = V^{-1} \arg\min_{|\mathbf{v}|_X \le r} |\Sigma\mathbf{v} - \mathbf{u}|_Y,$$

where the right-hand $\arg\min$ is now understood to find the value of $\mathbf{v}$, so the premultiplication for $\mathbf{x} = V^{-1}\mathbf{v}$ is required. The problem has now been simplified into one with a diagonal matrix instead of A.

3.1 Exact Minimum Solution

If the unrestricted $\arg\min$ also satisfies $|\mathbf{x}|_X \le r$ then it is the solution. The unrestricted minimum is a fixed point of the norm expression squared:

$$\begin{aligned} 0 &= \frac{\partial}{\partial v_n} |\Sigma\mathbf{v} - \mathbf{u}|_Y^2 = \frac{\partial}{\partial v_n} \sum_{i=1}^{\dim Y} (\Sigma\mathbf{v} - \mathbf{u})_i^2 = \frac{\partial}{\partial v_n} \sum_{i=1}^{\dim Y} (1_{i \le \dim X} \sigma_i v_i - u_i)^2 \\ &= \frac{\partial}{\partial v_n} (\sigma_n v_n - u_n)^2 = \frac{\partial}{\partial v_n} (\sigma_n^2 v_n^2 - 2\sigma_n v_n u_n + u_n^2) = 2\sigma_n^2 v_n - 2\sigma_n u_n \\ &\Leftrightarrow \quad \sigma_n(\sigma_n v_n - u_n) = 0. \end{aligned}$$

For each n, this is true if either $v_n = u_n/\sigma_n$ or $\sigma_n = 0$. In the latter case, the Σ matrix does not range over the full dimensionality of Y and any value of v_n may be chosen because the minimum is non-unique. It is usually best to choose $v_n = 0$ in all such ambiguous cases, since this corresponds to the minimum with smallest $|\mathbf{v}|_X = |\mathbf{x}|_X$. There is also the case when $\dim Y < \dim X$, where the above equation reduces to $0 = 0$ for $n > \dim Y$, giving no constraint on v_n, which should be set to zero by the same argument. The exact minimum can be written explicitly as

$$\mathbf{x} = V^{-1}[(U^{-1}\mathbf{b})_n /^0 \sigma_n], \qquad \text{where} \qquad x/^0 y = \begin{cases} x/y & \text{if } y \ne 0 \\ 0 & \text{otherwise} \end{cases}.$$

3.2 Constrained Minimum

The function $|\Sigma\mathbf{v} - \mathbf{u}|_Y$ does not have multiple disconnected local minima, so if the exact minimum with smallest norm found in the previous section still has $|\mathbf{x}|_X > r$, the constrained minimum must have $|\mathbf{x}|_X = r$ rather than being an interior point. The local gradient found in the previous section

$$\nabla_{\mathbf{v}} |\Sigma\mathbf{v} - \mathbf{u}|_Y^2 = 2[\sigma_n^2 v_n - \sigma_n u_n]$$

must be a scalar multiple of the position $\mathbf{v}$ because otherwise it has some component parallel to the surface of the radius r hypersphere and the value of the function can be reduced. The gradient is expected to be negative with increasing r, anti-parallel to $\mathbf{v}$, so for some $\lambda > 0$,

$$\begin{aligned} \nabla_{\mathbf{v}}|\Sigma\mathbf{v} - \mathbf{u}|_Y^2 &= -2\lambda^2\mathbf{v} \\ \Leftrightarrow \quad 2(\sigma_n^2 v_n - \sigma_n u_n) &= -2\lambda^2 v_n \\ \Leftrightarrow \quad (\sigma_n^2 + \lambda^2)v_n - \sigma_n u_n &= 0 \\ \Leftrightarrow \quad v_n &= \frac{\sigma_n u_n}{\sigma_n^2 + \lambda^2}. \end{aligned}$$

For the case where $n > \dim Y$, the gradient of that component is zero as before and $0 = -2\lambda^2 v_n$, so $v_n = 0$. The constrained minimum can be written explicitly as

$$\mathbf{x} = V^{-1}\left[\frac{\sigma_n(U^{-1}\mathbf{b})_n}{\sigma_n^2 + \lambda^2}\right], \qquad \text{where we set} \quad (U^{-1}\mathbf{b})_n = 0 \quad \text{if } n > \dim Y.$$

The norm of $\mathbf{x}$ decreases monotonically with λ because $|\mathbf{x}|_X = |\mathbf{v}|_X$ and every element of $\mathbf{v}$ decreases in magnitude with increasing λ. As $\lambda \to 0$ the constrained minimum tends towards the exact minimum. As $\lambda \to \infty$, the constrained minimum tends towards $\mathbf{0}$ but if renormalised, the limit has $v_n = \sigma_n u_n$, which is $-\frac{1}{2}$ times the gradient of $|\Sigma\mathbf{v} - \mathbf{u}|_Y^2$ at $\mathbf{v} = \mathbf{0}$. Thus the large λ limit corresponds to a infinitesimal 'steepest descent' step.

The continuity and monotonicity of $|\mathbf{x}|_X = r(\lambda)$ ensures a value of λ can always be found for any value of r between 0 and the norm of the exact solution point. For example, a bisection search or root-finding algorithm can determine λ for a given r, after first checking the exact solution point does not have norm less than r.

3.3 Implementation Note

Using the orthogonal property of U and V, entries $(U^{-1}\mathbf{b})_n$ should be calculated as the much faster equivalent $(U^T\mathbf{b})_n$ and the premultiplication by V^{-1} should be implemented as V^T. Once the SVD is calculated, nothing slower than matrix-vector multiplication is required.

4 Units

Elements of the vector spaces X and Y can be physical quantities with units $[X]$ and $[Y]$ respectively. By definition, A has units $[Y]/[X]$. In the SVD, the entries of U and V have no units as they map within the same space, leaving Σ and its entries σ_n with units $[Y]/[X]$. The parameter λ in the previous section was defined to also have units $[Y]/[X]$ but r has units $[X]$.

5 Identity with the Levenberg–Marquardt Algorithm

The Levenberg–Marquardt algorithm involves a 'damped' least squares step, which for a Jacobian matrix J involves solving

$$(J^TJ + \lambda_{LM}I)\mathbf{x} = J^T\mathbf{b},$$

where $\lambda_{LM} \geq 0$ is called the damping factor. If the Jacobian is decomposed via SVD as $J = U\Sigma V$, this becomes

$$(V^T\Sigma U^T U\Sigma V + \lambda_{LM}I)\mathbf{x} = V^T\Sigma U^T\mathbf{b}$$

and noting that $U^TU = I$ by orthogonality of U,

$$(V^T\Sigma^2V + \lambda_{LM}I)\mathbf{x} = V^T\Sigma U^T\mathbf{b}.$$

Pre-multipliying both sides by V and using its orthogonality $VV^T = I$ gives

$$\begin{aligned} & (\Sigma^2V + \lambda_{LM}V)\mathbf{x} &=& \Sigma U^T\mathbf{b} \\ \Rightarrow \quad & (\Sigma^2 + \lambda_{LM}I)V\mathbf{x} &=& \Sigma U^T\mathbf{b}. \end{aligned}$$

This is starting to look vaguely familiar. Inverting the left-hand side to give an expression for $\mathbf{x}$ yields

$$\begin{aligned} \mathbf{x} &= V^{-1}(\Sigma^2 + \lambda_{LM}I)^{-1}\Sigma U^T\mathbf{b} \\ &= V^{-1}(\Sigma^2 + \lambda_{LM}I)^{-1}\Sigma U^{-1}\mathbf{b}. \end{aligned}$$

Comparing this to the constrained minimum formula with parameter λ from a previous section:

$$\mathbf{x} = V^{-1}\left[\frac{\sigma_n(U^{-1}\mathbf{b})_n}{\sigma_n^2 + \lambda^2}\right]$$

and noting that $\Sigma = \mathrm{diag}[\sigma_n]$ reveals that these are the same formulae if $\lambda_{LM} = \lambda^2$.

6 Constrained Maximum of a Quadratic

As the $|\Sigma\mathbf{v} - \mathbf{u}|_Y^2$ minimised in the previous sections was a quadratic function of $\mathbf{x}$, it is natural to wonder if an arbitrary (scalar) quadratic function could be maximised using a similar method: that is, find

$$\arg\max_{|\mathbf{x}|\le r} f(\mathbf{x}) = \arg\max_{|\mathbf{x}|\le r}\left(f(\mathbf{0}) + \mathbf{g}\cdot\mathbf{x} + \frac{1}{2}\mathbf{x}^T H\mathbf{x}\right).$$

H is the Hessian matrix of second derivatives, so is symmetric, meaning its SVD decomposition can be written $H = U^T\Sigma U$, with U orthogonal. This permits a change of variable

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{0}) + \mathbf{g}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T U^T\Sigma U\mathbf{x} \\ &= f(\mathbf{0}) + \mathbf{g}^T U^T(U\mathbf{x}) + \frac{1}{2}(U\mathbf{x})^T\Sigma(U\mathbf{x}) \\ \Rightarrow \quad \arg\max_{|\mathbf{x}|\le r} f(\mathbf{x}) &= \arg\max_{|U\mathbf{x}|\le r}\left(f(\mathbf{0}) + (U\mathbf{g})^T(U\mathbf{x}) + \frac{1}{2}(U\mathbf{x})^T\Sigma(U\mathbf{x})\right). \end{aligned}$$

Defining $\mathbf{u} = U\mathbf{x}$ and ignoring the constant term, this becomes

$$\arg\max_{|\mathbf{x}|\le r} f(\mathbf{x}) = U^T \arg\max_{|\mathbf{u}|\le r}\left((U\mathbf{g})^T\mathbf{u} + \frac{1}{2}\mathbf{u}^T\Sigma\mathbf{u}\right).$$

The maximised expression is a single sum as Σ is diagonal, so its gradient vector is

$$\nabla_{\mathbf{u}}\left((U\mathbf{g})^T\mathbf{u} + \frac{1}{2}\mathbf{u}^T\Sigma\mathbf{u}\right) = [(U\mathbf{g})_n + \sigma_n u_n].$$

6.1 Exact Stationary Point

If none of the σ_n are zero, f has a stationary point at $\mathbf{u} = [-(U\mathbf{g})_n/\sigma_n]$, which is only a maximum if all the σ_n are negative.

6.2 Constrained Maximum

A constrained maximum would have, for some $\lambda > 0$,

$$[(U\mathbf{g})_n + \sigma_n u_n] = [\lambda u_n]$$

and thus $\mathbf{u} = [(U\mathbf{g})_n/(\lambda - \sigma_n)]$. The value of λ must satisfy

$$r^2 = |\mathbf{x}|^2 = |\mathbf{u}|^2 = \sum_n \frac{(U\mathbf{g})_n^2}{(\lambda - \sigma_n)^2}.$$

The expression on the right has a $+\infty$ singularity whenever $\lambda = \sigma_n$ for some n. It is also not monotonic, so there could be many solutions. However, note that $\lambda \to \infty$ still corresponds to $r \to 0$, so small r solutions are in the region where $\lambda > \max_n \sigma_n = \sigma_{\max}$.

What does the other end of this region, $\lambda \to \sigma_{\max}^+$ correspond to? First note that if $\sigma_{\max} < 0$ then the other end is actually $\lambda \to 0$, corresponding to the exact maximum (and it really is a maximum because all the σ_n are negative). Otherwise, a vector element u_n with $\sigma_n = \sigma_{\max} \geq 0$ tends to infinity, meaning the solution is asymptotically running up the steepest parabolic ascent direction available to it, as expected of a maximum.

Finally, note that although r^2 is not a monotonic function of λ, it is a (locally) convex one:

$$\frac{\mathrm{d}^2 r^2}{\mathrm{d}\lambda^2} = \sum_n \frac{6(U\mathbf{g})_n^2}{(\lambda - \sigma_n)^4} \geq 0.$$

Taking into account the asymptotic behaviour as $\lambda \to \infty$, this means r^2 in the region $\lambda > \sigma_{\max}$ is monotonically decreasing, so a value of λ can always be found for any value of r between 0 and the norm of the exact solution point (or infinity if $\sigma_{\max} \geq 0$, corresponding to a saddle, ridge or minimum valley).

6.3 Summary

The locus of constrained maxima is

$$\mathbf{x}(\lambda) = U^T \left[\frac{(U\mathbf{g})_n}{\lambda - \sigma_n}\right]$$

for $\lambda > \max\{0, \sigma_{\max}\}$. If $\sigma_{\max} < 0$ then $\mathbf{x}(0)$ is the exact maximum.