CONDITION NUMBERS AND CONVERGENCE

David F. Gleich

August 21, 2023

1 THE CONDITION NUMBER AS A FUNDAMENTAL MATRIX QUANTITY

Learning objectives

1. See how the matrix condition number arises in terms of converge of Richardson with the optimal parameter.
2. See how the matrix condition number arises in terms of converge of Steepest Descent.
3. Compare the rates of convergence.

Here, we show that the condition number of a matrix determines how quickly various simple iterative methods will converge on symmetric positive $1 - \frac{1}{\kappa(A)}$ definite linear systems. Thus, throughout these notes, we will assume that $\boldsymbol{A}$ is symmetric positive definite.

1.1 RICHARDSON

Recall the Richardson iteration for $\boldsymbol{A}\mathbf{x} = \mathbf{b}$:

$$\mathbf{r}^{(k)} = \mathbf{b} - \boldsymbol{A}\mathbf{x}^{(k)} \qquad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega\mathbf{r}^{(k)}.$$

We can write this in terms of the gradient for the quadratic problem:

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\boldsymbol{A}\mathbf{x} - \mathbf{x}^T\mathbf{b} \quad \text{with gradient} \quad \mathbf{g}(\mathbf{x}) = \boldsymbol{A}\mathbf{x} - \mathbf{b} = -\mathbf{r}(\mathbf{x})$$

which gives

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \omega\mathbf{g}^{(k)}.$$

Now consider the error vector

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}.$$

The evolution of the error is determined by

$$\mathbf{e}^{(k+1)} = \mathbf{x}^{(k+1)} - \mathbf{x} = \mathbf{x}^{(k)} + \omega\mathbf{r}^{(k)} - \mathbf{x} = \mathbf{x}^{(k)} + \omega\mathbf{b} - \omega\boldsymbol{A}\mathbf{x}^{(k)} - \mathbf{x}.$$

But note that $\omega\mathbf{b} = \omega\boldsymbol{A}\mathbf{x}$ for the true solution $\mathbf{x}$. Hence

$$\mathbf{e}^{(k+1)} = \mathbf{x}^{(k)} + \omega\boldsymbol{A}\mathbf{x} - \omega\boldsymbol{A}\mathbf{x}^{(k)} - \mathbf{x} = (\boldsymbol{I} - \omega\boldsymbol{A})(\mathbf{x}^{(k)} - \mathbf{x})$$

or simply

$$\mathbf{e}^{(k+1)} = (\boldsymbol{I} - \omega\boldsymbol{A})\mathbf{e}^{(k)} = (\boldsymbol{I} - \omega\boldsymbol{A})^k\mathbf{e}^{(0)}.$$

Up to this point, for Richardson, this is entirely general and actually has not used the assumption that $\boldsymbol{A}$ is symmetric positive definite.

This converges quickly if we can make the spectral radius $\rho(\boldsymbol{I} - \omega\boldsymbol{A})$ small. For a symmetric positive definite matrix, there is an easy way to do this. The derivation is not particularly interesting. The choice is:

$$\omega = \frac{2}{\lambda_1 + \lambda_n}$$

where λ_1 and λ_n are the smallest and largest eigenvalues of $\boldsymbol{A}$ respectively.[1] For this choice we have

$$\rho(\boldsymbol{I} - \omega\boldsymbol{A}) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}.$$

There is no condition number yet, but it's hiding inside this formula! For a symmetric positive definite system, we have $\kappa(\boldsymbol{A}) = \frac{\lambda_{\max}}{\lambda_{\min}}$, and so we can adjust this expression to include this ratio:

$$\rho(\boldsymbol{I} - \omega\boldsymbol{A}) = \frac{\frac{1}{\lambda_{\min}}}{\frac{1}{\lambda_{\min}}}\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\kappa(\boldsymbol{A}) - 1}{\kappa(\boldsymbol{A}) + 1} \le \frac{\kappa(\boldsymbol{A}) - 1}{\kappa(\boldsymbol{A})} \le 1 - \frac{1}{\kappa(\boldsymbol{A})}.$$

So the asympotic error in Richardson on a symmetric positive definite system goes to 0 at a rate $1 - \frac{1}{\kappa(\boldsymbol{A})}$. Formally,

$$\|\mathbf{e}^{(k)}\| \le (1 - \tfrac{1}{\kappa})^k\|\mathbf{e}^{(0)}\|.$$

[1] The way to determine this quantity is to look at how $\boldsymbol{I} - \omega\boldsymbol{A}$ changes the eigenvalues of $\boldsymbol{A}$. This transform maps the region $[\lambda_1, \lambda_n]$ to the region $[1-\omega\lambda_n, 1-\omega\lambda_1]$. We now want to pick ω to minimize $\max(|1 - \omega\lambda_n|, |1 - \omega\lambda_1|)$. Note that when ω is small enough, then $1 - \omega\lambda_n$ and $1 - \omega\lambda_1$ are both positive and $1 - \omega\lambda_1$ determines the spectral radius, which decreases with ω. As ω increases, $1 - \omega\lambda_n$ goes negative first (assuming $\lambda_1 \neq \lambda_n$) and so at some point we have $|1 - \omega\lambda_n| = -1 + \omega\lambda_n = |1 - \omega\lambda_1| = 1 - \omega\lambda_1$, which gives $\omega = 2/(\lambda_1 + \lambda_n)$ as required. This equivalency point is minimizer as further increasing ω just results in a larger spectral radius.

1.2 STEEPEST DESCENT

We will now show that the steepest descent iteration converges for a symmetric positive definite system that also depends on $\kappa(\boldsymbol{A})$.[2] Recall that steepest descent uses a dynamic choice of ω, called α or γ, that minimizes the function

$$f(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\boldsymbol{A}\mathbf{x} - \mathbf{x}^T\mathbf{b}$$

at each step. The iteration is

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \gamma_k \mathbf{g}(\mathbf{x}_k) \qquad \gamma_k = \frac{\mathbf{g}(\mathbf{x}_k)^T\mathbf{g}(\mathbf{x}_k)}{\mathbf{g}(\mathbf{x}_k)^T\boldsymbol{A}\mathbf{g}(\mathbf{x}_k)}.$$

We are going to tweak this setup slightly. Note that at a solution $\mathbf{x} = \boldsymbol{A}^{-1}\mathbf{b}$ we have

$$f(\boldsymbol{A}^{-1}\mathbf{b}) = \tfrac{1}{2}\mathbf{b}^T\boldsymbol{A}^{-T}\boldsymbol{A}\boldsymbol{A}^{-1}\mathbf{b} - \mathbf{b}^T\boldsymbol{A}^{-T}\mathbf{b} = -\tfrac{1}{2}\mathbf{b}^T\boldsymbol{A}^{-1}\mathbf{b}.$$

The strategy we are going to use is to study the rate $f(\mathbf{x}_k) \to -\frac{1}{2}\mathbf{b}^T\boldsymbol{A}^{-1}\mathbf{b}$. But this is a slightly annoying constant to have around, so we just study the function

$$s(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\boldsymbol{A}\mathbf{x} - \mathbf{x}^T\mathbf{b} + \tfrac{1}{2}\mathbf{b}^T\boldsymbol{A}^{-1}\mathbf{b}$$

instead. This function is just shifted by a constant, and so the gradient is unchanged. Now, we can study the rate at which $s(\mathbf{x}_k) \to 0$ instead, which makes life slightly easier.

This shifted function s is also nice for another reason. Let $n(\mathbf{x}) = \sqrt{\mathbf{x}^T\boldsymbol{A}^{-1}\mathbf{x}}$. Then we can show that $n(\mathbf{x})$ is a vector norm.[3] Typically we write this as

$$\|\mathbf{x}\|_{\boldsymbol{A}^{-1}} = \sqrt{\mathbf{x}^T\boldsymbol{A}^{-1}\mathbf{x}}.$$

Using this norm, we can write

$$s(\mathbf{x}) = \tfrac{1}{2}\|\boldsymbol{A}\mathbf{x} - \mathbf{b}\|^2_{\boldsymbol{A}^{-1}} = \tfrac{1}{2}\|\mathbf{g}(\mathbf{x})\|^2_{\boldsymbol{A}^{-1}}.$$

The goal is to show that $s(\mathbf{x}^{(k+1)}) \leq s(\mathbf{x}^{(k)})$(constant less than 1). We have the following that allow us to do so

$$\mathbf{g}(\mathbf{x}_k) = \mathbf{g}_k \qquad \gamma_k = \frac{\mathbf{g}_k^T\mathbf{g}_k}{\mathbf{g}_k^T\boldsymbol{A}\mathbf{g}_k}$$

$$\mathbf{x}^{(k+1)} = (\boldsymbol{I} - \gamma_k\boldsymbol{A})\mathbf{g}_k$$

$$s(\mathbf{x}^{(k+1)}) = \tfrac{1}{2}\|(\boldsymbol{I} - \gamma_k\boldsymbol{A})\mathbf{g}_k\|_{\boldsymbol{A}^{-1}} = \tfrac{1}{2}\mathbf{g}_k^T\boldsymbol{A}^{-1}\mathbf{g}_k - \gamma_k\mathbf{g}_k^T\mathbf{g}_k + \tfrac{1}{2}\gamma_k^2\mathbf{g}_k^T\boldsymbol{A}\mathbf{g}_k = s(\mathbf{x}_k) - \tfrac{1}{2}\frac{(\mathbf{g}_k^T\mathbf{g}_k)^2}{\mathbf{g}_k^T\boldsymbol{A}\mathbf{g}_k}$$

$$s(\mathbf{x}^{(k+1)}) = s(\mathbf{x}_k)\left(1 - \tfrac{1}{2}\frac{(\mathbf{g}_k^T\mathbf{g}_k)^2}{\mathbf{g}_k^T\boldsymbol{A}\mathbf{g}_k s(\mathbf{x}_k)}\right) = s(\mathbf{x}_k)\left(1 - \frac{(\mathbf{g}_k^T\mathbf{g}_k)^2}{\mathbf{g}_k^T\boldsymbol{A}\mathbf{g}_k\mathbf{g}_k^T\boldsymbol{A}^{-1}\mathbf{g}_k}\right)$$

because $s(\mathbf{x}_k) = \frac{1}{2}\mathbf{g}_k^T\boldsymbol{A}^{-1}\mathbf{g}_k$.

The key is that this quantity $\frac{(\mathbf{g}_k^T\mathbf{g}_k)^2}{\mathbf{g}_k^T\boldsymbol{A}\mathbf{g}_k\mathbf{g}_k^T\boldsymbol{A}^{-1}\mathbf{g}_k}$ is fairly close to a condition number. Let θ be the inverse quantity so

$$\theta = \frac{\mathbf{g}_k^T\boldsymbol{A}\mathbf{g}_k\mathbf{g}_k^T\boldsymbol{A}^{-1}\mathbf{g}_k}{(\mathbf{g}_k^T\mathbf{g}_k)^2}.$$

Then we have

$$\theta = \frac{\mathbf{g}_k^T\boldsymbol{A}\mathbf{g}_k}{\mathbf{g}_k^T\mathbf{g}_k}\frac{\mathbf{g}_k^T\boldsymbol{A}^{-1}\mathbf{g}_k}{\mathbf{g}_k^T\mathbf{g}_k} \leq \max_{\mathbf{g}}\left[\frac{\mathbf{g}_k^T\boldsymbol{A}\mathbf{g}_k}{\mathbf{g}_k^T\mathbf{g}_k}\right]\max_{\mathbf{g}}\left[\frac{\mathbf{g}_k^T\boldsymbol{A}^{-1}\mathbf{g}_k}{\mathbf{g}_k^T\mathbf{g}_k}\right] = \lambda_1\frac{1}{\lambda_n}.$$

So we have

$$s(\mathbf{x}^{(k+1)}) = s(\mathbf{x}^{(k)})\left(1 - \frac{1}{\kappa(\boldsymbol{A})}\right)$$

[2] This is a very slick proof that involves a number of interesting quantities; it's been designed over years to be clever and simple, so it's the sort of proof that would be hard to come up with yourself, so read through it a few times to see what is going on.

Recall that $\boldsymbol{A}$ is symmetric and so $\boldsymbol{A}^{-T} = \boldsymbol{A}^{-1}$.

[3] This is a good exercise. The only challenging step is the triangle inequality. The easy to way to show this is to use the fact that $n(\mathbf{x}) = \|\boldsymbol{F}^{-1}\mathbf{x}\|$ where $\boldsymbol{F}$ is the Choleksy factor of $\boldsymbol{A}^{-1}$ (which exists because we have a symmetric positive definite $\boldsymbol{A}$ and $\boldsymbol{A}^{-1}$.

which is exactly the same rate as Richardson. To improve this, we need a stronger bound on θ.

One such stronger bound is called the Kanterovich inequality. Let $\boldsymbol{A}$ be a symmetric positive definite matrix and let $\mathbf{v}$ be any vector with $\mathbf{v}^T\mathbf{v} = 1$, then

$$(\mathbf{v}^T\boldsymbol{A}\mathbf{v})(\mathbf{v}^T\boldsymbol{A}^{-1}\mathbf{v}) \leq \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n}.$$

This give us a better bound on θ and we get

$$1 - 1/\theta \leq \frac{(\lambda_1 - \lambda_n)^2}{(\lambda_1 + \lambda_n)^2} \leq \left(\frac{\kappa(\boldsymbol{A}) - 1}{\kappa(\boldsymbol{A}) + 1}\right)^2 \leq (1 - 1/\kappa(\boldsymbol{A}))^2.$$

This completes the proof.

1.3 COMPARISON

Which method is faster? For Richardson we have

$$\|\mathbf{e}^{(k)}\| \leq (1 - \tfrac{1}{\kappa})^k \|\mathbf{e}^{(0)}\|$$

whereas for Steepest Descent we have

$$s(\mathbf{x}^{(k+1)}) \leq (1 - \tfrac{1}{\kappa})^(2k) s(\mathbf{x}^{(0)}).$$

To conclude that steepest descent is faster requires a comparison of $s(\mathbf{x}^{(k)})$ and $\|\mathbf{e}^{(0)}\|$. We will continue to study this relationship.