1 THE CONDITION NUMBER AS A FUNDAMENTAL MATRIX QUANTITY

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

1.1 RICHARDSON

Recall the Richardson iteration for \( Ax = b \):

\[
r^{(k)} = b - Ax^{(k)} \quad x^{(k+1)} = x^{(k)} + \omega r^{(k)}.
\]

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

\[
f(x) = \frac{1}{2}x^T Ax - x^T b \quad \text{with gradient} \quad g(x) = Ax - b = -r(x)
\]

which gives

\[
x^{(k+1)} = x^{(k)} - \omega g^{(k)}.
\]

Now consider the error vector

\[
e^{(k)} = x^{(k)} - x.
\]

The evolution of the error is determined by

\[
e^{(k+1)} = x^{(k+1)} - x = x^{(k)} - \omega r^{(k)} - x = x^{(k)} - \omega b - \omega Ax^{(k)} - x.
\]

But note that \( \omega b = \omega Ax \) for the true solution \( x \). Hence

\[
e^{(k+1)} = x^{(k)} + \omega Ax - \omega Ax^{(k)} - x = (I - \omega A)(x^{(k)} - x)
\]

or simply

\[
e^{(k+1)} = (I - \omega A)e^{(k)} = (I - \omega A)^k e^{(0)}.
\]

This converges quickly if we can make the spectral radius \( \rho(I - \omega A) \) 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 \( \lambda_1 \) and \( \lambda_n \) are the smallest and largest eigenvalues of \( A \) respectively.\(^1\) For this choice we have

\[
\rho(I - \omega A) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}
\]

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

\[
\rho(I - \omega A) = \frac{\frac{1}{\lambda_{\min}} - \frac{1}{\lambda_{\max}}}{\frac{1}{\lambda_{\max}} + \frac{1}{\lambda_{\min}}} = \frac{\frac{1}{\kappa(A)} - 1}{\frac{1}{\kappa(A)} + 1} \leq \frac{\kappa(A) - 1}{\kappa(A)} \leq 1 - \frac{1}{\kappa(A)}.
\]

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

\[
\|e^{(k)}\| \leq (1 - \frac{1}{\kappa})^k \|e^{(0)}\|.
\]

\(^1\) The way to determine this quantity is to look at how \( I - \omega A \) changes the eigenvalues of \( A \). 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 \( \omega \) to minimize \( \max|1 - \omega \lambda_n|, |1 - \omega \lambda_1| \). Note that when \( \omega \) 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 \( \omega \). As \( \omega \) increases, \( 1 - \omega \lambda_n \) goes negative first (assuming \( \lambda_1 < \lambda_n \)) 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 \( \omega = \frac{2}{\lambda_1 + \lambda_n} \) as required. This equivalency point is minimizer as further increasing \( \omega \) just results in a larger spectral radius.
1.2 STEEPEST DESCENT

We will now show that the steepest descent iteration converges faster than Richardson
for a symmetric positive definite system. Recall that steepest descent uses a dynamic
choice of \( \alpha \) called \( \alpha \) or \( \gamma \), that minimizes the function

\[
f(x) = \frac{1}{2} x^T A x - x^T b
\]

at each step. The iteration is

\[
x^{(k+1)} = x^{(k)} - \gamma_k g(x_k) \quad \gamma_k = \frac{g(x_k)^T g(x_k)}{g(x_k)^T A g(x_k)}.
\]

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

\[
f(A^{-1} b) = \frac{1}{2} b^T A^{-T} A A^{-1} b - b^T A^{-T} b = -\frac{1}{2} b^T A^{-1} b.
\]

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

\[
s(x) = \frac{1}{2} x^T A x - x^T b + \frac{1}{2} b^T A^{-1} b
\]

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

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

\[
|x|_{A^{-1}} = \sqrt{x^T A^{-1} x}.
\]

Using this norm, we can write

\[
s(x) = \frac{1}{2} \| Ax - b \|_{A^{-1}}^2 = \frac{1}{2} \| g(x) \|_{A^{-1}}^2.
\]

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

\[
g(x_k) = g_k \quad \gamma_k = \frac{g_k^T g_k}{g_k^T A g_k}
\]

\[
x^{(k+1)} = (I - \gamma_k A) g_k
\]

\[
s(x^{(k+1)}) = \frac{1}{2} \| (I - \gamma_k A) g_k \|_{A^{-1}}^2 = \frac{1}{2} g_k^T A^{-1} g_k - \gamma_k g_k^T g_k + \frac{1}{2} \gamma_k^2 g_k^T A g_k = s(x_k) - \frac{1}{2} \left( \frac{g_k^T g_k}{g_k^T A g_k} \right)^2
\]

\[
s(x^{(k+1)}) = s(x_k) (1 - \frac{1}{2} \left( \frac{g_k^T g_k}{g_k^T A g_k} \right)^2) = s(x_k) (1 - \frac{g_k^T A g_k}{g_k^T g_k} \frac{A^{-1} g_k}{g_k})
\]

because \( s(x_k) = \frac{1}{2} g_k^T A^{-1} g_k \).

The key is that this quantity \( \frac{g_k^T A g_k}{g_k^T g_k} \) is fairly close to a condition number. Let \( \theta \)
be the inverse quantity so

\[
\theta = \frac{g_k^T A g_k}{g_k^T g_k} \frac{A^{-1} g_k}{g_k}
\]

Then we have

\[
\theta = \frac{g_k^T A g_k}{g_k^T g_k} \frac{A^{-1} g_k}{g_k} \leq \max \left[ \frac{g_k^T A g_k}{g_k^T g_k} \right] \max \left[ \frac{g_k^T A^{-1} g_k}{g_k^T g_k} \right] = \lambda_1 \frac{1}{\lambda_n}
\]

So we have

\[
s(x^{(k+1)}) = s(x^{(k)}) \left( 1 - \frac{1}{\kappa(A)} \right)
\]

\[\text{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.}\n
\[\text{Recall that } A \text{ is symmetric and so } A^{-T} = A^{-1}.\]
which is exactly the same rate as Richardson. To improve this, we need a stronger bound on $\theta$.

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

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

This gives us a better bound on $\theta$ and we get

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

This completes the proof.