1 Condition Numbers & Perturbation (2.2, 2.3)

We are now interested in the *sensitivity* of $A\mathbf{x} = \mathbf{b}$ with respect to perturbations (i.e., error). In other words, does noise in A or $\mathbf{b}$ strongly affect the solution $\mathbf{x}$? Here, we'll deal with two types of perturbations: in $\mathbf{b}$, and in A. Eventually, we'll talk about the case when there's noise in both.

1.1 Motivating Example

To see what we mean, consider the following two examples.

(Example.) Consider the system

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} 2 \\ 0 \end{bmatrix}}_{\mathbf{b}}.$$

1. Solve for $\mathbf{x}$.

   Note that
   $$\mathbf{x} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$
   by backwards substitution.

2. Suppose we introduce a very small error to the entries of $\mathbf{b}$ such that $\hat{\mathbf{b}} = \begin{bmatrix} 2 \\ 0.001 \end{bmatrix}$. Our system now becomes

   $$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \hat{x_1} \\ \hat{x_2} \end{bmatrix}}_{\hat{\mathbf{x}}} = \underbrace{\begin{bmatrix} 2 \\ 0.001 \end{bmatrix}}_{\hat{\mathbf{b}}}.$$

   Solve for $\hat{\mathbf{x}}$. In other words, what happens to $\mathbf{x}$ if we perturb $\mathbf{b}$?

   Here, we have
   $$\hat{\mathbf{x}} = \begin{bmatrix} 1.999 \\ 0.001 \end{bmatrix}.$$
   Here, $\hat{\mathbf{x}}$ is known as a perturbed solution. Notice how the difference between the solution and the perturbed solution is very small, to the point that both $\mathbf{x}$ and $\hat{x}$ are *similar*.

3. Compute the error in $\mathbf{b}$ and in $\mathbf{x}$.

   The error in $\mathbf{b}$ can be found by using the L_2-norm. So, for $\mathbf{b}$, we have
   $$||\mathbf{b} - \hat{\mathbf{b}}||_2 = \left|\left| \begin{bmatrix} 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 \\ 0.001 \end{bmatrix} \right|\right|_2 = \left|\left| \begin{bmatrix} 0 \\ -0.001 \end{bmatrix} \right|\right|_2 = 0.001$$
   Likewise, for $\mathbf{x}$, we have
   $$||\mathbf{x} - \hat{\mathbf{x}}||_2 = \left|\left| \begin{bmatrix} 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 1.999 \\ 0.001 \end{bmatrix} \right|\right|_2 = \left|\left| \begin{bmatrix} 0.001 \\ -0.001 \end{bmatrix} \right|\right|_2 = \sqrt{2} \cdot 0.001 \approx 0.0014.$$

(Example.) Consider a similar system

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 0.001 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} 2 \\ 0 \end{bmatrix}}_{\mathbf{b}}.$$

1. Solve for $\mathbf{x}$.

We have

$$\mathbf{x} = \begin{bmatrix} 2 \\ 0 \end{bmatrix},$$

which we found by backwards substitution.

2. Suppose we introduce a very small error to the entries of $\mathbf{b}$ such that $\hat{\mathbf{b}} = \begin{bmatrix} 2 \\ 0.001 \end{bmatrix}$. Our system now becomes

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 0.001 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \hat{x_1} \\ \hat{x_2} \end{bmatrix}}_{\hat{\mathbf{x}}} = \underbrace{\begin{bmatrix} 2 \\ 0.001 \end{bmatrix}}_{\hat{\mathbf{b}}}.$$

Solve for $\hat{\mathbf{x}}$.

Here, we have

$$\hat{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

One important thing to notice is that the perturbed solution is *quite different* from the actual solution. So, unlike the previous example, $\mathbf{x}$ and $\hat{x}$ are *different*.

3. Compute the error in $\mathbf{b}$ and in $\mathbf{x}$.

The error in $\mathbf{b}$ is the same as in the previous example; therefore,

$$||\mathbf{b} - \hat{\mathbf{b}}||_2 = 0.001$$

But, for $\mathbf{x}$, notice how

$$||\mathbf{x} - \hat{\mathbf{x}}||_2 = \left|\left|\begin{bmatrix} 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right|\right|_2 = \left|\left|\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right|\right|_2 = \sqrt{2} \approx 1.41.$$

In particular, 0.001 is 10^3 larger than 1.41. So, in this linear system, when we perturb $\mathbf{b}$ a little, we can cause a *large* error.

Remark: From this, it follows that the error in $\mathbf{x}$ depends on the matrix A as well.

1.2 Condition Number

How do we measure the dependence on the matrix A? This is related to the **condition number**, known as `cond(A)` in MATLAB. The condition number is a simple but useful measure of the sensitivity of the linear system $A\mathbf{x} = \mathbf{b}$. Although we haven't defined the condition number yet, consider the following examples, which showcase the difference in condition number:

- $\text{cond}\left(\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\right) \approx 2.1618$, which is a small condition number.
- $\text{cond}\left(\begin{bmatrix} 1 & 1 \\ 0 & 0.001 \end{bmatrix}\right) \approx 2 \cdot 10^3$, which is a large condition number, and the error is amplified by this

large condition number.

1.3 Perturbation of b

Now, we want to solve $A\mathbf{x} = \mathbf{b}$, where A is invertible. Instead of $\mathbf{b}$, we only have access to

$$\hat{\mathbf{b}} = \mathbf{b} + \delta\mathbf{b},$$

where $\delta\mathbf{b}$ is the (very small) error, known as the perturbation in $\mathbf{b}$. Then, we can consider the linear system

$$A\hat{\mathbf{x}} = \hat{\mathbf{b}}.$$

If $\hat{\mathbf{b}}$ is close to $\hat{\mathbf{b}}$, is it true that $\hat{\mathbf{x}}$ is close to $\hat{\mathbf{x}}$ as well? **This depends** on the condition number of A. In particular, we'll later see that

$$\boxed{\frac{||\mathbf{x} - \hat{\mathbf{x}}||}{||\mathbf{x}||} \leq \kappa(A) \frac{||\mathbf{b} - \hat{\mathbf{b}}||}{||\mathbf{b}||}}. \tag{1}$$

Here,

- $\frac{||\mathbf{x}-\hat{\mathbf{x}}||}{||\mathbf{x}||}$ is the *relative error* of $\mathbf{x}$,
- $\frac{||\mathbf{b}-\hat{\mathbf{b}}||}{||\mathbf{b}||}$ is the *relative error* of $\mathbf{b}$.
- $\kappa(A)$ is the condition number of the invertible matrix A.

The relative error of $\mathbf{x}$ is bounded by the condition number of matrix A multiplifed by the relative error of $\mathbf{b}$.

What is $\kappa(A)$? We can define it like so:

$$\kappa(A) = ||A|| \cdot ||A^{-1}||,$$

where $||\cdot||$ can be any vector norm. We will use the notation

- κ_p for the p-norm; that is,
$$\kappa_p(A) = ||A||_p \cdot ||A^{-1}||_p.$$
- κ_∞ for the ∞-norm; that is,
$$\kappa_\infty(A) = ||A||_\infty \cdot ||A^{-1}||_\infty.$$

With this in mind, let's prove the inequality in (1).

Proof. We can break this down into two steps.

- Let $\hat{\mathbf{b}} = \mathbf{b} + \delta\mathbf{b}$ (the perturbed $\mathbf{b}$) and $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ (the perturbed $\mathbf{x}$). So,

$$\begin{aligned} A\hat{\mathbf{x}} = \hat{\mathbf{b}} &\implies A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b} \\ &\implies A\mathbf{x} + A\delta\mathbf{x} = \mathbf{b} + \delta\mathbf{b} \\ &\implies A\delta\mathbf{x} = \delta\mathbf{b} && \text{Recall that } A\mathbf{x} = \mathbf{b} \\ &\implies \delta\mathbf{x} = A^{-1}\delta\mathbf{b} && A \text{ is invertible} \\ &\implies ||\delta\mathbf{x}|| = ||A^{-1}\delta\mathbf{b}||. \end{aligned}$$

Recall that $||A|| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{||A\mathbf{x}||}{||\mathbf{x}||}$ is the matrix norm induced by the vector norm. Additionally, note that $||A^{-1}|| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{||A^{-1}\mathbf{x}||}{||\mathbf{x}||}$. Then,

$$||A^{-1}|| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{||A^{-1}\mathbf{x}||}{||\mathbf{x}||} \geq \frac{||A^{-1}\delta\mathbf{b}||}{||\delta\mathbf{b}||} \implies ||A^{-1}|| \cdot ||\delta\mathbf{b}|| \geq ||A^{-1}\delta\mathbf{b}||.$$

So,

$$||\delta\mathbf{x}|| = ||A^{-1}\delta\mathbf{b}|| \leq ||A^{-1}|| \cdot ||\delta\mathbf{b}|| \implies ||\delta\mathbf{x}|| \leq ||A^{-1}|| \cdot ||\delta\mathbf{b}||.$$

- Recall that $\mathbf{b} = A\mathbf{x}$. So,
$$\mathbf{b} = A\mathbf{x} \implies ||\mathbf{b}|| = ||A\mathbf{x}|| \leq ||A|| \cdot ||\mathbf{x}||.$$
Then, we can divide both sides by $||\mathbf{b}|| \cdot ||\mathbf{x}||$ to get
$$\frac{1}{||\mathbf{x}||} \leq ||A|| \frac{1}{||\mathbf{b}||}.$$

With all this, we can combine the inequalities
$$||\delta\mathbf{x}|| \leq ||A^{-1}|| \cdot ||\delta\mathbf{b}||$$
$$\frac{1}{||\mathbf{x}||} \leq ||A|| \frac{1}{||\mathbf{b}||}$$
to get
$$\underbrace{\frac{||\delta\mathbf{x}||}{||\mathbf{x}||}}_{\substack{\text{Relative error}\\ \text{in } \mathbf{x}}} \leq \underbrace{||A^{-1}|| \cdot ||A||}_{\substack{\kappa(A)\\ \text{Condition number.}}} \cdot \underbrace{\frac{||\delta\mathbf{b}||}{||\mathbf{b}||}}_{\substack{\text{Relative error}\\ \text{in } \mathbf{b}}}.$$
This simplifies to $\frac{||\delta\mathbf{x}||}{||\mathbf{x}||} \leq \kappa(A)\frac{||\delta\mathbf{b}||}{||\mathbf{b}||}$, as desired. □

Remarks:

- The matrix norm is the induced matrix norm, e.g., if the vector norm is 2-norm, then the matrix norm is 2-norm. That is,
$$||A||_p = \max_{\mathbf{x}\neq\mathbf{0}} \frac{||A\mathbf{x}||_p}{||\mathbf{x}||_p}$$
$$||A||_\infty = \max_{\mathbf{x}\neq\mathbf{0}} \frac{||A\mathbf{x}||_\infty}{||\mathbf{x}||_\infty}$$
This means that we can use whatever vector norm we want for (1) as long as all vectors use the same norm. Additionally, the induced matrix norm for A should be the same as the one used for the vector norm. **The norms must be consistent in the inequality.**
- When interpreting $\kappa(A)$,
  - If $\kappa(A)$ is small (close to 1), then A is called "well-conditioned."
  - If $\kappa(A)$ is large, then A is called "ill-conditioned."
- A tall matrix does not have a condition matrix because it's not invertible.

1.4 Properties of the Induced Matrix Norm

Proposition. *Let $||\cdot||$ be an induced matrix norm*
$$||A|| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{||A\mathbf{x}||}{||\mathbf{x}||}.$$
Then,

1. $||I|| = 1$. *Here, I is the identity matrix; the condition number of the identity matrix is 1.*
2. $\kappa(A) \geq 1$. *In particular, the condition number of some matrix will be* at least *1.*

Proof. Note that
$$||I|| = \max_{\mathbf{x}\neq\mathbf{0}} \frac{||I\mathbf{x}||}{||\mathbf{x}||} = 1.$$
Also,
$$I = AA^{-1} \implies 1 = ||I|| = ||AA^{-1}|| \leq ||A|| \cdot ||A^{-1}|| = \kappa(A),$$
so we're done. □

Remarks:

- For #2 of the proposition, if we introduce an error in $\mathbf{b}$, the condition number will not make the error smaller.
- $\kappa(I) = 1$. In particular,
$$\kappa(I) = ||I|| \cdot ||I^{-1}|| = 1 \cdot 1 = 1.$$
- The Feobenius norm is not an induced matrix norm. In particular, the above results do not hold for the Feobenius norm $||I||_F$ as $||I||_F \neq 1$. Recall that
$$||A||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij})^2}.$$
However, for $I \in \mathbb{R}^{n\times n}$,
$$||I_n||_F = \sqrt{n} \neq 1.$$

1.5 Perturbation of A

Theorem 1.1

Let A be nonsingular, $\mathbf{b} \neq 0$, and let $\mathbf{x}$ and $\hat{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ be solutions to $A\mathbf{x} = \mathbf{b}$ and $(A + \delta A)\hat{\mathbf{x}} = \mathbf{b}$, respectively. Then,
$$\frac{||\delta\mathbf{x}||}{||\hat{\mathbf{x}}||} \leq \kappa(A)\frac{||\delta A||}{||A||}.$$

Proof. We have

$$\begin{aligned}
(A + \delta A)\hat{\mathbf{x}} = \mathbf{b} &\implies (A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} && \text{We defined } \hat{\mathbf{x}} = \mathbf{x} = \delta\mathbf{x} \\
&\implies A\mathbf{x} + A\delta\mathbf{x} + \delta A\mathbf{x} + \delta A\delta\mathbf{x} = \mathbf{b} \\
&\implies A\mathbf{x} + A\delta\mathbf{x} + \delta A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} \\
&\implies A\mathbf{x} + A\delta\mathbf{x} + \delta A\hat{\mathbf{x}} = \mathbf{b} \\
&\implies A\delta\mathbf{x} + \delta A\hat{\mathbf{x}} = \mathbf{0} && \text{Recall that } A\mathbf{x} = \mathbf{b} \\
&\implies A\delta\mathbf{x} = -\delta A\hat{\mathbf{x}} \\
&\implies \delta\mathbf{x} = -\underbrace{A^{-1}}_{\text{Matrix}}\underbrace{\delta A}_{\text{Matrix}}\underbrace{\hat{\mathbf{x}}}_{\text{Vector}}.
\end{aligned}$$

From here, it follows that
$$\begin{aligned}
||\delta\mathbf{x}|| &= ||-A^{-1}\delta A\hat{\mathbf{x}}|| \\
&\leq ||A^{-1}|| \cdot ||\delta A\hat{\mathbf{x}}|| \\
&\leq ||A^{-1}|| \cdot ||\delta A|| \cdot ||\hat{\mathbf{x}}||
\end{aligned}$$

Dividing through by $||\hat{\mathbf{x}}||$, we have
$$\frac{||\delta\mathbf{x}||}{||\hat{\mathbf{x}}||} \leq ||A^{-1}|| \cdot ||\delta A||.$$

Recall that $\kappa(A) = ||A|| \cdot ||A^{-1}|| \implies ||A^{-1}|| = \frac{\kappa(A)}{||A||}$, so
$$\frac{||\delta\mathbf{x}||}{||\hat{\mathbf{x}}||} \leq \kappa(A)\frac{||\delta A||}{||A||}$$

as desired. □

Remark: δ, in this case, is not a scalar. We can think of $\delta\mathbf{x}$ as $\mathbf{x}$ with a very tiny (and arbitrary) error introduced. Analogously, we can think of δA as A with a very tiny (and arbitrary) error.

1.6 Perturbation of A and $\mathbf{b}$

Theorem 1.2

Instead of $A\mathbf{x} = \mathbf{b}$, we now consider $\hat{A}\hat{\mathbf{x}} = \hat{\mathbf{b}}$ with $\hat{A} = A + \delta A$ and $\hat{\mathbf{b}} = \mathbf{b} + \delta\mathbf{b}$. Then,

$$\frac{||\delta\mathbf{x}||}{||\hat{x}||} \leq \kappa(A)\left(\frac{||\delta A||}{||A||} + \frac{||\delta\mathbf{b}||}{||\hat{\mathbf{b}}||} \cdot \frac{||\delta A||}{||A||} + \frac{||\delta\mathbf{b}||}{||\hat{\mathbf{b}}||}\right)$$

Proof. We can break this down into two steps.

- For $\hat{A}\hat{\mathbf{x}} = \hat{\mathbf{b}}$, we have

$$\begin{aligned} \hat{A}\hat{\mathbf{x}} = \hat{\mathbf{b}} &\implies (A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b} \\ &\implies A\mathbf{x} + A\delta\mathbf{x} + \delta A\mathbf{x} + \delta A\delta\mathbf{x} = \mathbf{b} + \delta\mathbf{b} \\ &\implies A\delta\mathbf{x} + \delta A\mathbf{x} + \delta A\delta\mathbf{x} = \delta\mathbf{b} && \text{Recall that } A\mathbf{x} = \mathbf{b} \\ &\implies A\delta\mathbf{x} + \delta A(\mathbf{x} + \delta\mathbf{x}) = \delta\mathbf{b} \\ &\implies A\delta\mathbf{x} + \delta A\hat{\mathbf{x}} = \delta\mathbf{b} \\ &\implies A\delta\mathbf{x} = \delta\mathbf{b} - \delta A\hat{\mathbf{x}} \\ &\implies \delta\mathbf{x} = A^{-1}(\delta\mathbf{b} - \delta A\hat{\mathbf{x}}). \end{aligned}$$

  Now,

$$\begin{aligned} ||\delta\mathbf{x}|| &= ||A^{-1}(\delta\mathbf{b} - \delta A\hat{\mathbf{x}})|| \\ &\leq ||A^{-1}|| \cdot ||\delta\mathbf{b} - \delta A\hat{\mathbf{x}}|| \\ &\leq ||A^{-1}||(||\delta\mathbf{b}|| + ||\delta A|| \cdot ||\hat{\mathbf{x}}||) && \text{See remark below.} \end{aligned}$$

  Thus,

$$\begin{aligned} ||\delta\mathbf{x}|| &\leq ||A^{-1}||(||\delta\mathbf{b}|| + ||\delta A|| \cdot ||\hat{\mathbf{x}}||) \\ &= \frac{||A^{-1}||}{||\hat{\mathbf{x}}||}(||\delta\mathbf{b}|| + ||\delta A|| \cdot ||\hat{\mathbf{x}}||) \\ &= ||A^{-1}||\left(\frac{||\delta\mathbf{b}||}{||\hat{\mathbf{x}}||} + \frac{||\delta A|| \cdot ||\hat{\mathbf{x}}||}{||\hat{\mathbf{x}}||}\right) \\ &= ||A^{-1}||\left(\frac{||\delta\mathbf{b}||}{||\hat{\mathbf{x}}||} + ||\delta A||\right) \\ &= \frac{\kappa(A)}{||A||}\left(\frac{||\delta\mathbf{b}||}{||\hat{\mathbf{x}}||} + ||\delta A||\right) && \kappa(A) = ||A|| \cdot ||A^{-1}|| \implies ||A^{-1}|| = \frac{\kappa(A)}{||A||} \\ &= \kappa(A)\left(\frac{||\delta\mathbf{b}||}{||A|| \cdot ||\hat{\mathbf{x}}||} + \frac{||\delta A||}{||A||}\right). \end{aligned}$$

- Again, from $\hat{A}\hat{\mathbf{x}} = \hat{\mathbf{b}}$, we have

$$\hat{A}\hat{\mathbf{x}} = \hat{\mathbf{b}} \implies (A + \delta A)\hat{\mathbf{x}} = \hat{\mathbf{b}}.$$

  Then,

$$||\hat{\mathbf{b}}|| = ||(A + \delta A)\hat{\mathbf{x}}|| \leq ||A + \delta A|| \cdot ||\hat{\mathbf{x}}||.$$

  Therefore, dividing by $||\hat{\mathbf{x}}|| \cdot ||\hat{\mathbf{b}}||$ we get

$$\frac{1}{||\hat{\mathbf{x}}||} \leq \frac{||A + \delta A||}{||\mathbf{b}||} \leq \frac{||A|| + ||\delta A||}{||\delta\mathbf{b}||}.$$

Combining the results of the previous two steps yields the desired result. $\square$

Remark: To see why this equality works, consider the following diagram: