ME 7247: Advanced Control Systems **Fall 2022–23**

Lecture 3: Least norm optimization

Friday September 16, 2022

Lecturer: Laurent Lessard *Scribe: Laurent Lessard*

Least norm estimation, optimality conditions, trade-offs and regularization, geometrical intuition,

1 Least norm optimization

Consider the equation $Ax = b$ with $A \in \mathbb{R}^{m\times n}$. This time, imagine we have $m < n$ (A is a wide matrix), and we are in the *control* setup; there are infinitely many x satisfying $Ax = b$, so we want to find the "best" x among all solutions.

In the least norm problem, as the name suggests, we will seek the solution to $Ax = b$ for which $\|x\|$ is as small as possible. In optimization notation, the problem is to

$$\begin{array}{ll} \underset{x\in\mathbb{R}^n}{\text{minimize}} & \|x\|^2 \\ \text{such that} & Ax = b \end{array} \tag{1}$$

Here, $Ax = b$ is a constraint, and we write it beneath the objective function $\|x\|$.

1.1 Geometric intuition

The set of all solutions to $Ax = b$ is the set $X := \{x_p + v \mid v \in \text{null}(A)\}$, where x_p is any point satisfying $Ax_p = b$. We can write this simply as $X = x_p + \text{null}(A)$. The set X is generally *not* a subspace, because it does not include 0. Rather, it is an *affine space*; which is a shifted subspace. Instead of passing through the origin, the set X passes through the point x_p. We can visualize all points in this space as in Fig. 1. Important note: when we drew a picture for least squares, we visualized the output space $\mathbb{R}^m$, in which range(A) is a subspace. Here, we visualize, the input space $\mathbb{R}^n$, in which null(A) is a subspace.

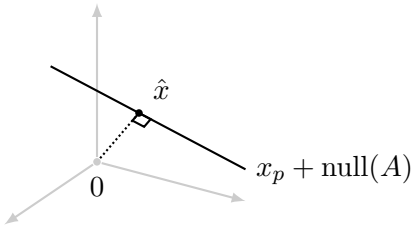

Figure 1: Geometric intuition for minimizing $\|x\|$ subject to $Ax = b$: we are looking for the point in $\hat{x} + \text{null}(A)$ that is closest to 0. This happens when $\hat{x} \in \text{null}(A)^\perp$.

In Fig. 1, we see that among all points in the solution set $x_p + \text{null}(A)$, there is a point $\hat{x}$ that has minimum norm, which means it is closest to the origin. This means that $\hat{x}$ should be orthogonal to all vectors in null(A), so $\hat{x} \in \text{null}(A)^\perp$.

We can also prove the converse; that if we have any vector $\hat{x} \in \text{null}(A)^\perp$, then it must be optimal. To see why, let x be any other feasible point, i.e. a point that satisfies $Ax = b$. Now write:

$$\begin{aligned}\|x\|^2 &= \|x - \hat{x} + \hat{x}\|^2 \\ &= \|\hat{x}\|^2 + \|x - \hat{x}\|^2 + 2\langle \hat{x}, x - \hat{x}\rangle \\ &= \|\hat{x}\|^2 + \|x - \hat{x}\|^2 \\ &\geq \|\hat{x}\|^2\end{aligned}$$

In the third line, we used the fact that $\langle \hat{x}, x - \hat{x}\rangle = 0$. This follows from the fact that x and $\hat{x}$ are both solutions to $Ax = b$. Therefore, we have: $A(x - \hat{x}) = Ax - A\hat{x} = b - b = 0$. So $x - \hat{x} \in \text{null}(A)$. By assumption, $\hat{x} \in \text{null}(A)^\perp$, so $\langle \hat{x}, x - \hat{x}\rangle = 0$. This proves that $\hat{x}$ is an optimal point if and only if $\hat{x} \in \text{null}(A)^\perp$.

1.2 Algebraic solution

To proceed further, we need two useful facts.

Lemma 1.1. *Suppose $A \in \mathbb{R}^{m\times n}$. Then* $\text{range}(A)^\perp = \text{null}(A^\mathsf{T})$.

Proof. Pick any $z \in \text{null}(A^\mathsf{T})$ and $y \in \text{range}(A)$. Then $y = Ax$ for some $x \in \mathbb{R}^n$. Now calculate: $\langle z, y\rangle = \langle z, Ax\rangle = \langle A^\mathsf{T} z, x\rangle = 0$. We just showed that $\langle z, y\rangle$ for all $y \in \text{range}(A)$, which means that $z \in \text{range}(A)^\perp$. Consequently, $\text{null}(A^\mathsf{T}) \subseteq \text{range}(A)^\perp$.

Now pick any $z \in \text{range}(A)^\perp$. Then for any $x \in \mathbb{R}^n$, we have $\langle z, Ax\rangle = 0$, which is equivalent to $\langle A^\mathsf{T} z, x\rangle = 0$. This holds for all x, so we conclude that $A^\mathsf{T} z = 0$, so $z \in \text{null}(A^\mathsf{T})$. Consequently, $\text{range}(A)^\perp \subseteq \text{null}(A^\mathsf{T})$. ■

Lemma 1.2. *Let $S \subseteq \mathbb{R}^n$ be a subspace. Then $S^{\perp\perp} = S$.*

Proof. From the definition: $y \in S^\perp$ means that $\langle y, z\rangle = 0$ for all $z \in S$. Consequently, if $z \in S$, we must have $\langle y, z\rangle = 0$ for all $y \in S^\perp$. But this is precisely the definition of $z \in S^{\perp\perp}$. So we have $S \subseteq S^{\perp\perp}$. To prove the other inclusion, use the fact that we can decompose $\mathbb{R}^n = W \oplus W^\perp$ for any subspace W. Applying this to S and $S^\perp$, we conclude that $n = \dim(S) + \dim(S^\perp) = \dim(S^\perp) + \dim(S^{\perp\perp})$. Therefore $\dim(S) = \dim(S^\perp)$. Together with the fact that $S \subseteq S^{\perp\perp}$, we conclude that $S = S^{\perp\perp}$. ■

Lemma 1.3. *Suppose $A \in \mathbb{R}^{m\times n}$. Then* $\text{range}(A) = \text{range}(AA^\mathsf{T})$.

Proof. We previously proved that $\text{null}(A) = \text{null}(A^\mathsf{T} A)$. Taking the perp of both sides and applying Lemmas 1.1 and 1.2, we conclude that $\text{range}(A^\mathsf{T}) = \text{range}(A^\mathsf{T} A)$. Since A is an arbitrary matrix, we we can replace it by A^T and the result follows. ■

Applying Lemmas 1.1 and 1.2, our condition that $\hat{x} \in \text{null}(A)^\perp$ is equivalent to $\hat{x} \in \text{range}(A^\mathsf{T})$. In other words, we must have $\hat{x} = A^\mathsf{T} w$ for some $w \in \mathbb{R}^m$. But we also know that $A\hat{x} = b$, since $\hat{x}$ must satisfy the linear equations. Substituting, we obtain:

$$AA^\mathsf{T} w = b$$

Therefore, our solution process is clear:

1. Solve the system $AA^\mathsf{T}w = b$.
2. The solution to the minimum norm problem is $\hat{x} = A^\mathsf{T}w$.

Some observations to make:

- What if $AA^\mathsf{T}w = b$ has no solution? In this case $b \notin \text{range}(AA^\mathsf{T})$. From Lemma 1.3, this is equivalent to $b \notin \text{range}(A)$, so there are no solutions to $Ax = b$ at all (the optimization problem is infeasible).
- Can there be infinitely many solutions? For example, suppose we have w_1 and w_2 that both satisfy $AA^\mathsf{T}w = b$. Then, $AA^\mathsf{T}(w_1 - w_2) = 0$, and so $w_1 - w_2 \in \text{null}(AA^\mathsf{T})$. But $\text{null}(AA^\mathsf{T}) = \text{null}(A^\mathsf{T})$ (proved in Lecture 2, Lemma 1.3), so $A^\mathsf{T}(w_1 - w_2) = 0$. Consequently, if we define $\hat{x}_1 = A^\mathsf{T}w_1$ and $\hat{x}_2 = A^\mathsf{T}w_2$, we find that:

  $$\hat{x}_1 - \hat{x}_2 = A^\mathsf{T}(w_1 - w_2) = 0$$

  So although $AA^\mathsf{T}w = b$ may have infinitely many solutions, they all lead to the same solution $\hat{x}$ to the optimization problem.

Remark 1.4. We can use Lemma 1.3 to prove that the normal equations always have a solution. Clearly, we have $A^\mathsf{T}b \in \text{range}(A^\mathsf{T})$, and from Lemma 1.3, we have $\text{range}(A^\mathsf{T}) = \text{range}(A^\mathsf{T}A)$. Therefore, $A^\mathsf{T}b \in \text{range}(A^\mathsf{T}A)$, which means that the equation $A^\mathsf{T}Ax = A^\mathsf{T}b$ has a solution.

1.3 Calculus solution

Given a smooth function f, The vector ∇f points in the direction of greatest increase of f. Meanwhile, vectors orthogonal to ∇f point in directions of no change. This follows from Taylor's theorem in higher dimensions:

$$f(x + \delta x) \approx f(x) + \nabla f(x)^\mathsf{T}\delta x$$

So when $\langle \nabla f(x), \delta x \rangle = 0$ and δx is small, we have no change in f. Likewise, among all vectors δx of equal length, the largest increase is when δx is aligned with $\nabla f(x)$, so $\nabla f(x)$ points in the direction of greatest increase of f at the point x.

Theorem 1.5. *Let $f : \mathbb{R}^n \to 0$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, \ldots, m$ be smooth functions. If $\hat{x}$ minimizes $f(x)$ subject to the constraint $g_i(x) = 0$ for all i, then $\nabla f(\hat{x}) \in \text{span}(\nabla g_i(\hat{x}))$.*

Proof. Suppose instead that $\nabla f(\hat{x}) \notin \text{span}(\nabla g_i(\hat{x}))$. In particular, $\nabla f(\hat{x}) \neq 0$. Therefore, we can pick a nonzero $\delta x \in \text{span}(\nabla g_i(\hat{x}))^\perp$ such that $\langle \delta x, \nabla f(\hat{x}) \rangle < 0$. With this choice, $\langle \delta x, \nabla g_i(\hat{x}) \rangle = 0$ for all i, so by Taylor's theorem, perturbing $\hat{x}$ in the direction of δx will cause all g_i to remain constant but f will decrease, thereby contradicting the optimality of $\hat{x}$. ■

In our case, we want to minimize $f(x) = \|x\|^2$ subject to the constraints (split A into its rows) $g_i(x) = \tilde{a}_i^\mathsf{T}x - b = 0$. The gradient of this constraint is $\tilde{a}_i$. So by Theorem 1.5, we must have:

$$\nabla f(\hat{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\hat{x}) = 0$$

for some choice of constants $\lambda_1, \ldots, \lambda_m$. These constants are called *Lagrange multipliers*. Substituting f and g_i into this equation, we obtain:

$$2\hat{x} + \sum_{i=1}^{m} \lambda_i \tilde{a}_i = 0$$

Assembling the λ_i into a column vector λ, we can write this succinctly as: $2\hat{x} + A^\mathsf{T}\lambda = 0$. This is equivalent to saying that $\hat{x} \in \text{range}(A^\mathsf{T})$; same as we found using geometry.

1.4 Full rank case

When we looked at solutions of $Ax = b$, we saw that when A has full row rank, then there exists a solution for any $b \in \mathbb{R}^m$, so there also exists a solution to the minimum-norm problem. We don't have to worry about uniqueness, since we showed that minimum-norm problems always have a unique solution.

Corollary 1.6. *Suppose $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. If A has full row rank, then there exists a solution to the minimum norm problem: minimize $\|x\|$ subject to $Ax = b$. The solution is also unique, and it is given by $\hat{x} = A^\mathsf{T}(AA^\mathsf{T})^{-1}b$.*

Proof. If A has full row rank, then $\text{range}(A) = \mathbb{R}^m$, so the equation $Ax = b$ has a solution for any b. By Lemma 1.3, $\text{range}(AA^\mathsf{T}) = \text{range}(A) = \mathbb{R}^m$, so AA^T has full row rank as well. This matrix is square and full rank, so it is invertible. Therefore, the equations $AA^\mathsf{T}w = b$ and $\hat{x} = A^\mathsf{T}w$ have a unique solution, and it is given by $A^\mathsf{T}(AA^\mathsf{T})^{-1}b$. ■

When A has full row rank, the matrix $A^\dagger := A^\mathsf{T}(AA^\mathsf{T})^{-1}$ is (also) called the *pseudoinverse* of A. The pseudoinverse is defined for general A as well; we'll see the general definition later. In the full row rank case, we have the following properties:

- If $A \in \mathbb{R}^{m\times n}$, then $A^\dagger \in \mathbb{R}^{n\times m}$. So $A^\dagger$ has the same shape as A^T.
- $AA^\dagger = I_m$. In other words, $A^\dagger$ is a *right*-inverse of A.
- If A is square and full rank (invertible), then both notions of pseudoinverse coincide and we have $A^\dagger = (A^\mathsf{T}A)^{-1}A^\mathsf{T} = A^\mathsf{T}(AA^\mathsf{T})^{-1} = A^{-1}$.

2 Transferring mass a unit distance

In the following example[1], we would like to transfer a mass (initially at rest) a distance of 1 unit in 10 seconds. We can apply a constant force every second. We want to find the least-norm sequence of forces that achieves this. Define the following variables:

- y_t and v_t: position ant velocity at time t, respectively.
- x_t: constant force applied in the time interval $[t, t+1]$.

[1]This example is borrowed from: `http://ee263.stanford.edu/lectures/min-norm.pdf`

We will assume the dynamics are described by the following simple equations:

$$\begin{aligned} v_{t+1} - v_t &= x_t && \text{(force equals change in velocity)} \\ y_{t+1} - y_t &= v_t && \text{(velocity equals change in position)} \end{aligned}$$

We also have the initial conditions $y_0 = 0$, $v_0 = 0$, because the mass is initially at rest. Our goal is to pick $x_0, \ldots, x_9$ so that $y_{10} = 1$ and $v_{10} = 0$, so after 10 seconds, the mass has moved a unit distance and is again at rest. We start by expressing v_{10} and y_{10} in terms of the x_t's:

$$\begin{bmatrix} v_{10} \\ y_{10} \end{bmatrix} = \begin{bmatrix} v_0 \\ y_0 \end{bmatrix} + \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 9 & 8 & 7 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} x_0 \\ \vdots \\ x_9 \end{bmatrix}$$

Substituting the initial and terminal constraints, we have the equation:

$$Ax = b, \qquad \text{where:} \quad b = \begin{bmatrix} v_{10} \\ y_{10} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 9 & 8 & \cdots & 1 & 0 \end{bmatrix}.$$

So finding the minimum-norm input amounts to solving the minimum norm problem (1). Since A has full row rank, the solution is given by the pseudoinverse $\hat{x} = A^\dagger b = A^\mathsf{T}(AA^\mathsf{T})^{-1}b$. This is:

$$\hat{x} = \begin{bmatrix} 1 & 9 \\ 1 & 8 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 10 & 45 \\ 45 & 285 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{165} \begin{bmatrix} 9 \\ 7 \\ 5 \\ \vdots \\ -9 \end{bmatrix}.$$

The solution is plotted in Fig. 2 below.

The optimal input is an *affine* function of time. This is no accident; since our optimal solution belongs to range(A^T), and in this case A^T has columns that are linear (constant rate of change), this means $\hat{x}$ will also have steadily changing components. The optimal input is an affine function of time regardless of the initial and terminal conditions!

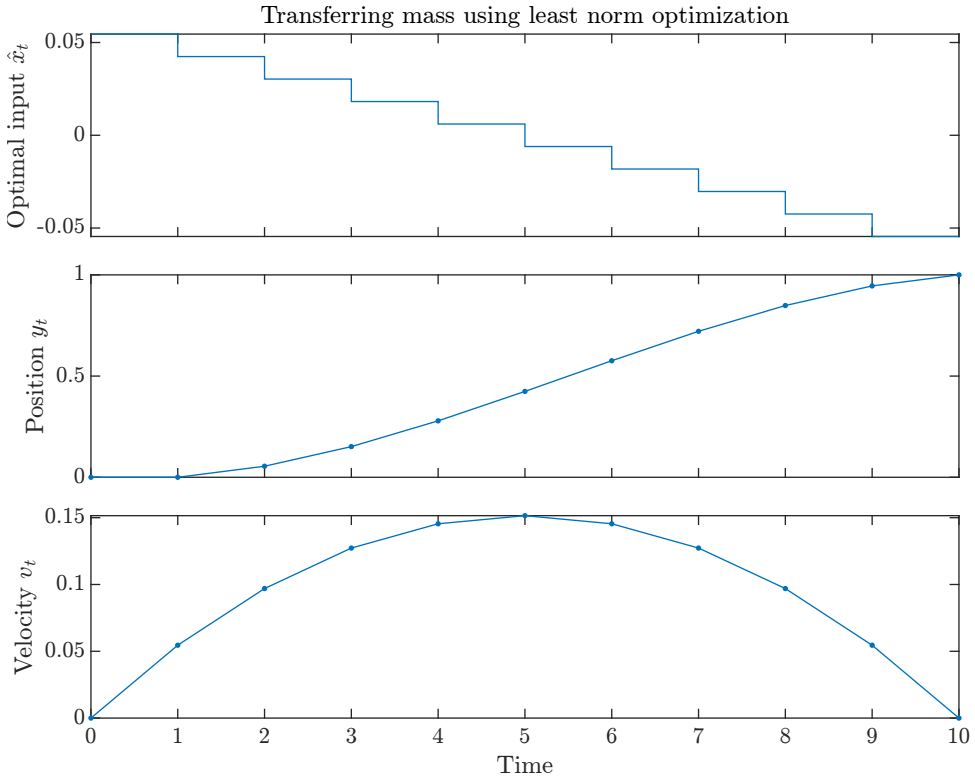

Figure 2: Optimal input, position, and velocity for the minimum-norm mass transfer.