V. Beyond Least Squares

Norm approximation

The bulk of this course has been spent studying variations of the least-squares problem: given observations $\boldsymbol{y} \in \mathbb{R}^M$ and a known $M \times N$ matrix $\boldsymbol{A}$, we solve

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \, \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2.$$

When $\boldsymbol{A}$ has full column rank ($M \geq N$ and $\text{rank}(\boldsymbol{A}) = N$), then this problem has a unique solution that can be written in closed form: $\hat{\boldsymbol{x}} = \boldsymbol{A}^\dagger \boldsymbol{y}$, where $\boldsymbol{A}^\dagger = (\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}$ is the pseudo-inverse of $\boldsymbol{A}$.

We did not make this explicit, but we also solved the "closet point" problem using the least-squares program above. Here we ask what is the closest point in an N dimensional subspace $\mathcal{T}$ to a given point $\boldsymbol{y} \in \mathbb{R}^M$; that is, we want to solve

$$\underset{\boldsymbol{w} \in \mathcal{T}}{\text{minimize}} \, \|\boldsymbol{y} - \boldsymbol{w}\|_2^2.$$

If we have a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N$ for $\mathcal{T}$, then we are looking for the closest point in the span of $\{\boldsymbol{v}_n\}$ to $\boldsymbol{y}$. As every point in $\mathcal{T}$ is a linear combination of these vectors, we can express any $\boldsymbol{w} \in \mathcal{T}$ as $\boldsymbol{w} = \boldsymbol{V}\boldsymbol{a}$ for some $\boldsymbol{a} \in \mathbb{R}^N$. So we can transform the problem above into

$$\underset{\boldsymbol{a} \in \mathbb{R}^N}{\text{minimize}} \, \|\boldsymbol{y} - \boldsymbol{V}\boldsymbol{a}\|_2^2,$$

which of course has the solution $\hat{\boldsymbol{a}} = (\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V})^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y}$. The corresponding optimal $\hat{\boldsymbol{w}} \in \mathbb{R}^M$ is then

$$\hat{\boldsymbol{w}} = \boldsymbol{V}\hat{\boldsymbol{a}} = \boldsymbol{V}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V})^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{y}.$$

We also know that this "approximation in a subspace" problem is equally easy if we use any norm which is induced by an inner product.

If $\langle \cdot, \cdot \rangle_H$ is a valid inner product, and $\| \cdot \|_H$ is its induced norm, then we can solve

$$\min_{\boldsymbol{w} \in \mathcal{T}} \|\boldsymbol{y} - \boldsymbol{w}\|_H$$

by forming

$$\boldsymbol{G} = \begin{bmatrix} \langle \boldsymbol{v}_1, \boldsymbol{v}_1 \rangle_H & \langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle_H & \cdots & \langle \boldsymbol{v}_1, \boldsymbol{v}_N \rangle_H \\ \langle \boldsymbol{v}_2, \boldsymbol{v}_1 \rangle_H & \langle \boldsymbol{v}_2, \boldsymbol{v}_2 \rangle_H & & \langle \boldsymbol{v}_2, \boldsymbol{v}_N \rangle_H \\ \vdots & & & \\ \langle \boldsymbol{v}_N, \boldsymbol{v}_1 \rangle_H & \langle \boldsymbol{v}_N, \boldsymbol{v}_2 \rangle_H & \cdots & \langle \boldsymbol{v}_N, \boldsymbol{v}_N \rangle_H \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} \langle \boldsymbol{y}, \boldsymbol{v}_1 \rangle_H \\ \langle \boldsymbol{y}, \boldsymbol{v}_2 \rangle_H \\ \vdots \\ \langle \boldsymbol{y}, \boldsymbol{v}_N \rangle_H \end{bmatrix}$$

and then taking $\hat{\boldsymbol{w}} = \boldsymbol{V}\boldsymbol{G}^{-1}\boldsymbol{b}$.

What if we are interested in finding the closest point in a subspace in a norm that is **not** induced by an inner product? For example[1], we may be interested in solving

$$\underset{\boldsymbol{w} \in \mathcal{T}}{\text{minimize}} \|\boldsymbol{y} - \boldsymbol{w}\|_1$$

or

$$\underset{\boldsymbol{w} \in \mathcal{T}}{\text{minimize}} \|\boldsymbol{y} - \boldsymbol{w}\|_\infty.$$

We can still (partially) characterize the solutions to these types of problems, but in general they do not have closed-form solutions. There is, however, a nice computational framework for solving them — in practice, they are not too much harder to solve than the least-squares problem.

[1] It is a fact that the only ℓ_p norm that is induced by an inner product is the ℓ_2 norm.

Example: Let $\mathcal{T}$ be the subspace of $\mathbb{R}^2$ defined by

$$\mathcal{T} = \{\boldsymbol{w} \in \mathbb{R}^2 \ : \ w[1] + 2w[2] = 0\}.$$

Given

$$\boldsymbol{y} = \begin{bmatrix} 3 \\ 2 \end{bmatrix},$$

find the minimizer $\boldsymbol{w}$ of these three programs:

$$\min_{\boldsymbol{w} \in \mathcal{T}} \ \|\boldsymbol{y} - \boldsymbol{w}\|_2, \qquad \min_{\boldsymbol{w} \in \mathcal{T}} \ \|\boldsymbol{y} - \boldsymbol{w}\|_1, \qquad \min_{\boldsymbol{x} \in \mathcal{T}} \ \|\boldsymbol{y} - \boldsymbol{w}\|_\infty.$$

Example: Filter design

The standard "filter synthesis" problem is to find an finite-impulse response (FIR) filter whose discrete-time Fourier transform (DTFT) is as close to some target $H_*(\omega)$ as possible. When the deviation from the optimal response is measured using a uniform error, this is call "equiripple design", since the error in the solution will tend to have ripples a uniform distance away from the ideal. That is, we would like to solve

$$\min_{H} \sup_{\omega \in [-\pi,\pi]} |H_*(\omega) - H(\omega)|, \quad \text{subject to } H(\omega) \text{ being FIR}$$

If we restrict ourselves to the case where $H_*(\omega)$ has linear phase (so the impulse response is symmetric around some time index)[2] we can recast this as a Chebyshev approximation problem.

A symmetric filter with $2K+1$ taps has a real DTFT that can be written as a superposition of a DC term plus K cosines:

$$h_n = 0 \ |n| > K \quad \Rightarrow \quad H(\omega) = \sum_{k=0}^{K} \tilde{h}_k \cos(k\omega), \quad \tilde{h}_k = \begin{cases} h_0, & k = 0 \\ 2h_k, & 1 \le k \le K. \end{cases}$$

So we are trying to solve

$$\min_{\boldsymbol{x}} \sup_{\omega \in [-\pi,\pi]} \left| H_*(\omega) - \sum_{k=0}^{K} x_k \cos(k\omega) \right|.$$

[2] The case with general phase can also be handled using convex optimization, but it is not naturally stated as a linear program.

We will approximate the supremum on the inside by measuring it at M equally spaced points $\omega_1, \ldots, \omega_M$ between $-\pi$ and π. Then

$$\min_{\boldsymbol{x}} \max_{\omega_m} \left| H_*(\omega_m) - \sum_{k=0}^{K} x_k \cos(k\omega_m) \right| = \min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{F}\boldsymbol{x}\|_\infty,$$

where $\boldsymbol{y} \in \mathbb{R}^M$ and the $M \times (K+1)$ matrix $\boldsymbol{F}$ are defined as

$$\boldsymbol{y} = \begin{bmatrix} H_*(\omega_1) \\ H_*(\omega_2) \\ \vdots \\ H_*(\omega_M), \end{bmatrix} \quad \boldsymbol{F} = \begin{bmatrix} 1 & \cos(\omega_1) & \cos(2\omega_1) & \cdots & \cos(K\omega_1) \\ 1 & \cos(\omega_2) & \cos(2\omega_2) & \cdots & \cos(K\omega_2) \\ \vdots & & \ddots & & \\ 1 & \cos(\omega_M) & \cos(2\omega_M) & \cdots & \cos(K\omega_M) \end{bmatrix}$$

It should be noted that since the ω_m are equally spaced, the matrix $\boldsymbol{F}$ (and its adjoint) can be applied efficiently using a fast discrete cosine transform. Just as with least-squares (recall steepest descent and conjugate gradients), this has a direct impact on the number of computations we need to solve the Chebyshev approximation problem above.

ℓ_1 and ℓ_∞ minimization

As we can see from the example above, minimizing $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1$ or $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_\infty$ is doing something quite different than least squares. We can say the following qualitative things about these programs:

- Minimizing $\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_1$ encourages the the residual $\boldsymbol{r} = \boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}$ to be *sparse*. That is, many of the components of the optimal $\boldsymbol{r}$ will be exactly equal to zero.
- Minimizing $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_\infty$ encourages the residual $\boldsymbol{r}$ to be *uniform*. That is, many of the components of the optimal $\boldsymbol{r}$ will have exactly the same magnitude.

In both of these cases, it is also easy to see that the solution need not be unique (although for "generic" problem instances, it will be — you need the column space of $\boldsymbol{A}$ and $\boldsymbol{y}$ to line up just right for the solution to be non-unique).

We can actually characterize these properties a little more carefully. Here is concrete result for ℓ_1:

Theorem: Given $\boldsymbol{y} \in \mathbb{R}^M$ and an $M \times N$ matrix $\boldsymbol{A}$ with full column rank, there is a solution $\hat{\boldsymbol{x}}_1$ to

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1 \tag{1}$$

such that $\hat{\boldsymbol{r}}_1 = \boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}_1$ has at most $M - N$ non-zero terms (and so at least N terms which are exactly equal to zero).

Proof: Suppose $\boldsymbol{x}$ is a vector such that $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$ has more than $M - N$ non-zero terms. We will show that there is another vector $\boldsymbol{x}' \neq \boldsymbol{x}$ such that one of two things is true:

1. $\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}'\|_1 < \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_1$, and so $\boldsymbol{x}$ can not be the minimizer;

2. $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}'\|_1 = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1$, and $\boldsymbol{r}' = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}'$ has at least one fewer non-zero term than $\boldsymbol{r}$.

Let $\Gamma \subset \{1, \ldots, M\}$ be the subset of locations where $\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}$ is equal to zero. For now we will assume that Γ is not empty, we will remove this restriction below. By assumption, Γ has fewer than N members. We will use $|\Gamma|$ the number of terms in Γ, so $|\Gamma| < N$.

Let $\boldsymbol{A}_\Gamma$ be the $|\Gamma| \times N$ matrix formed by extracting the rows of $\boldsymbol{A}$ corresponding to Γ. Since $|\Gamma| < N$, this matrix has fewer rows than columns, and so it has a non-trivial null space. Let $\boldsymbol{d} \in \mathbb{R}^N$ be any non-zero vector in this null space. By construction, $\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x} + \delta\boldsymbol{d})$ will also be 0 on the set Γ for all δ.

Let Γ^c be the complement of Γ (the set of all indices that are not in Γ). Define $\boldsymbol{a}_m \in \mathbb{R}^N$ so that $\boldsymbol{a}_m^{\mathrm{T}}$ is the mth row of $\boldsymbol{A}$. Then for any δ, we can write

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x} + \delta\boldsymbol{d})\|_1 &= \sum_{m \in \Gamma^c} |y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - \delta\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{d}| \\ &= \sum_{m \in \Gamma^c} \mathrm{sign}(y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - \delta\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{d})(y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - \delta\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{d}).\end{aligned}$$

Since δ is arbitrary, we can make its magnitude small enough so that

$$\mathrm{sign}(y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - \delta\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{d}) = \mathrm{sign}(y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x}) \quad \text{for all} \quad m \in \Gamma^c, \tag{2}$$

and so for appropriately small δ,

$$\begin{aligned}\|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x} + \delta\boldsymbol{d})\|_1 &= \sum_{m \in \Gamma^c} \mathrm{sign}(y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x})(y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - \delta\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{d}) \\ &= \sum_{m=1}^{M} \mathrm{sign}(y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x})(y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - \delta\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{d}),\end{aligned}$$

where we have added back in the terms on Γ since by construction $y[m] - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x} - \delta\boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{d} = 0$ for $m \in \Gamma$. We can now break that sum

into two parts, yielding

$$\|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x} + \delta \boldsymbol{d})\|_1 = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1 - \delta \sum_{m=1}^{M} \text{sign}(y[m] - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{x}) \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{d}. \tag{3}$$

If the sum on the right above is exactly equal to zero, we can choose any δ that maintains (2). This equality is maintained exactly until an additional term in $y[m] - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{x} - \delta \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{d}$ is driven to zero (and this is guaranteed to happen for some $\delta = \delta_0$), meaning that condition (2) above holds for $\boldsymbol{x}' = \boldsymbol{x} + \delta_0 \boldsymbol{d}$: $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}'\|_1 = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1$ and there is at least one fewer nonzero term in $\boldsymbol{r}' = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}'$ than in $\boldsymbol{r}$.

Now consider the case where the sum on the right in (3) is nonzero. Since the sign of δ is arbitrary, so we can choose so that $\delta \sum_{m=1}^{M} \text{sign}(y[m] - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{x}) \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{d} > 0$, and so

$$\|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{x} + \delta \boldsymbol{d})\|_1 < \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1.$$

Hence, $\boldsymbol{x}$ cannot be a solution to (1) since there is another vector $\boldsymbol{x}' = \boldsymbol{x} + \delta \boldsymbol{d}$ that results in a smaller value of the functional. This establishes the theorem.

The key part of finding $\boldsymbol{x}'$ was the fact that $\boldsymbol{A}_\Gamma$ was guaranteed to have a non-trivial null space. This was true simply because it had more columns than rows — if $\Gamma \geq N$ (meaning the solution has at least N terms exactly equal to zero), no such guarantee exists.

A similar statement can be made for ℓ_∞ norm approximation — we may explore this more on the homework.

ℓ_1 and ℓ_∞ as linear programs

Unlike the ℓ_2 case, the solutions to the problems

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_1 \quad \text{and} \quad \min_{\boldsymbol{x}\in\mathbb{R}^N} \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_\infty,$$

do not have closed form solutions, and the mapping from the data $\boldsymbol{y}$ to the solution $\hat{\boldsymbol{x}}$ is not linear at all. In addition, these functionals are not differentiable, and so the gradient is not defined at every point, meaning that we cannot use a standard gradient descent algorithms.

There is, however, a very well-defined computational framework for solving these two problems. They can be recast as **linear programs**. This is a type of (convex) optimization program that appears all over applied mathematics (operations research, statistics, machine learning, etc), and so there has been a lot of work in the last 30 on efficient methods for solving them.

A **linear program** (LP) is an optimization program of the form

$$\min_{\boldsymbol{z}\in\mathbb{R}^Q} \boldsymbol{c}^{\mathrm{T}}\boldsymbol{z} \quad \text{subject to} \quad \boldsymbol{M}\boldsymbol{z} \leq \boldsymbol{b}. \tag{4}$$

Above, $\boldsymbol{c} \in \mathbb{R}^Q$, $\boldsymbol{M}$ is a $P \times Q$ matrix, and $\boldsymbol{b} \in \mathbb{R}^P$. The nomenclature comes from the fact that both the functional $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{z}$ and the constraints $\boldsymbol{M}\boldsymbol{z}$ are linear in $\boldsymbol{z}$. Notice that the presence of the constraints are necessary for this even to make sense, as $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{z}$ is unbounded below if left unconstrained.

There are several ways in which the ℓ_1 (and ℓ_∞) norm approximation problem can be recast as a linear program; we will illustrate one of them here. The trick we will use is to introduce a dummy variable

$\boldsymbol{u} \in \mathbb{R}^M$ whose entries bound the magnitudes of the residuals:

$$|y[m] - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{x}| \;\leq\; u[m], \quad \text{for all } m = 1, \ldots, M.$$

We can write this as

$$-u[m] \;\leq\; y[m] - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{x} \;\leq\; u[m],$$

or combining all the constraints together,

$$-\boldsymbol{u} \;\leq\; \boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} \;\leq\; \boldsymbol{u}$$

which indicates that for a pair $\boldsymbol{x}, \boldsymbol{u}$ to be feasible, we will need $\boldsymbol{u} \geq 0$. Then

$$\min_{\substack{\boldsymbol{x} \in \mathbb{R}^N \\ \boldsymbol{u} \in \mathbb{R}^M}} \; \sum_{m=1}^{M} u[m] \quad \text{subject to} \quad \begin{array}{r} -\boldsymbol{A}\boldsymbol{x} - \boldsymbol{u} \leq -\boldsymbol{y} \\ \boldsymbol{A}\boldsymbol{x} - \boldsymbol{u} \leq \boldsymbol{y} \end{array} \tag{5}$$

will have $|y[m] - \boldsymbol{a}_m^{\mathrm{T}} \hat{\boldsymbol{x}}| = \hat{u}[m]$ at its minimum, which means the $\hat{\boldsymbol{x}}$ in the solution above will be the same as the $\hat{\boldsymbol{x}}$ which solves (1). (The constraint $\boldsymbol{u} \geq 0$ is implicit above, and so does not need to be incorporated explicitly.) We can also see that (5) is a linear program of the form (4) with $Q = M + N$ variables and $P = 2M$ constraints, and

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{u} \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{1} \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} -\boldsymbol{A} & -\mathbf{I} \\ \boldsymbol{A} & -\mathbf{I} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} -\boldsymbol{y} \\ \boldsymbol{y} \end{bmatrix}.$$

Recasting $\min \; \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_\infty$ as a linear program is even easier. Now we just need a single dummy variable $t \in \mathbb{R}$, which constrains

$$|y[m] - \boldsymbol{a}_m^{\mathrm{T}} \boldsymbol{x}| \;\leq\; t \quad \text{for all } m = 1, \ldots, M.$$

Then

$$\min_{\substack{\boldsymbol{x}\in\mathbb{R}^N \\ t\in\mathbb{R}}} \; t \quad \text{subject to} \quad \begin{aligned} -\boldsymbol{A}\boldsymbol{x} - t\mathbf{1} &\leq -\boldsymbol{y} \\ \boldsymbol{A}\boldsymbol{x} - t\mathbf{1} &\leq \boldsymbol{y}, \end{aligned}$$

will have $\hat{t} = \|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{x}}\|_\infty$ at its minimum, and so the $\hat{\boldsymbol{x}}$ in the solution above will match the solution to $\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_\infty$. This is a linear program with $Q = N + 1$ variables and $P = 2M$ constraints — we can plug into (4) with

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{x} \\ t \end{bmatrix}, \quad \boldsymbol{c} = \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} -\boldsymbol{A} & -\mathbf{1} \\ \boldsymbol{A} & -\mathbf{1} \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} -\boldsymbol{y} \\ \boldsymbol{y} \end{bmatrix}.$$

Convexity and uniqueness

From our geometric reasoning above, we can see that the ℓ_1 and ℓ_∞ norm approximation problems do not always have unique solutions. For example:

$$\min_{\boldsymbol{w}\in\mathcal{T}} \|\boldsymbol{y}-\boldsymbol{w}\|_\infty, \quad \mathcal{T} = \text{span}\left(\begin{bmatrix}0\\1\end{bmatrix}\right), \quad \boldsymbol{y} = \begin{bmatrix}3\\2\end{bmatrix}$$

has solutions

$$\hat{\boldsymbol{w}} = \begin{bmatrix}0\\\beta\end{bmatrix} \quad \text{for all} \;\; \beta \in [-1,5].$$

But it is true that all of the solutions for both of these problems will be clustered together in a connected set.

In general, for any valid norm $\|\cdot\|$, the approximation problem

$$\min_{\boldsymbol{w}\in\mathcal{T}} \|\boldsymbol{y}-\boldsymbol{w}\|, \quad \mathcal{T} = \text{ subspace}$$

is **convex**. Here are the require formal definitions:

Convexity: A subset $\mathcal{U}$ of a vector space is called convex if

$$\boldsymbol{x},\boldsymbol{y}\in\mathcal{U} \quad \Rightarrow \quad (1-\lambda)\boldsymbol{x}+\lambda\boldsymbol{y}\in\mathcal{U} \;\; \text{for all } 0\le\lambda\le 1.$$

A real-valued functional on a vector space $f(\cdot) : \mathcal{S} \to \mathbb{R}$ is called convex if

$$f((1-\lambda)\boldsymbol{x}+\lambda\boldsymbol{y}) \;\le\; (1-\lambda)f(\boldsymbol{x})+\lambda f(\boldsymbol{y}) \quad \text{for all} \;\; 0\le\lambda\le 1.$$

If we replace the $\le$ above with a strict inequality,

$$f((1-\lambda)\boldsymbol{x}+\lambda\boldsymbol{y}) \;<\; (1-\lambda)f(\boldsymbol{x})+\lambda f(\boldsymbol{y}) \quad \text{for all} \;\; 0<\lambda<1,$$

for all $\boldsymbol{x} \neq \boldsymbol{y}$, then $f(\cdot)$ is called strictly convex. There is a similar definition[3] for sets: $\mathcal{U} \subset \mathcal{S}$ is called strictly convex if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{U}$

$$(1-\lambda)\boldsymbol{x} + \lambda \boldsymbol{y} \in \text{interior}(\mathcal{U}), \quad 0 < \lambda < 1.$$

An optimization program in a vector space $\mathcal{S}$ is called convex (or a "convex program") if it has the form

$$\min \ f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{U},$$

where $\mathcal{U} \subset \mathcal{S}$ is a convex set, and $f(\cdot)$ is a convex function.

The main consequence of an optimization program being convex is that all local minima are also global minima. This fact is extremely useful computational, since we know a lot of simple ways that are guaranteed to find local minima (e.g. steepest descent).

Fact: Any valid norm $\|\cdot\|$ on a vector space $\mathcal{S}$ is a convex function. You can prove this at home; it is a direct consequence of the triangle inequality.

Fact: For any valid norm $\|\cdot\|$ on a vector space $\mathcal{S}$, the unit ball

$$\mathcal{B} = \{\boldsymbol{x} \in \mathcal{S} \ : \ \|\boldsymbol{x}\| \leq 1\}$$

is convex. This is also easy enough to prove that you can do it at home in your spare time.

[3] We will need to recall the definition of the interior of a set. We say that $\boldsymbol{z} \in \text{interior}(\mathcal{U})$ if there exists any $\epsilon > 0$ such that the neighborhood around $\boldsymbol{z}$ given by $\mathcal{B}_\epsilon(\boldsymbol{z}) = \{\boldsymbol{z}' \in \mathcal{S} \ : \ \|\boldsymbol{z} - \boldsymbol{z}'\| \leq \epsilon\}$ is also included in $\mathcal{U}$, $\mathcal{B}_\epsilon(\boldsymbol{z}) \subset \mathcal{U}$.

Examples:

- The ℓ_1 norm (and unit ball) is convex, but not strictly convex.
- The ℓ_2 norm (and unit ball) is strictly convex.
- The ℓ_∞ norm (and unit ball) is convex, but not strictly convex.

In fact, the ℓ_p norm (and unit ball) is strictly convex for all $1 < p < \infty$. This fact follows directly from the **Minkowski Inequality**[4]. If $\{x_n\}$ and $\{y_n\}$ are sequences of real numbers, then for any $1 < p < \infty$,

$$\left(\sum_{n=1}^{N} |x_n + y_n|^p\right)^{1/p} \leq \left(\sum_{n=1}^{N} |x_n|^p\right)^{1/p} + \left(\sum_{n=1}^{N} |y_n|^p\right)^{1/p},$$

with equality holding only when there exists an $\alpha \in \mathbb{R}$ such that

$$x_n = \alpha y_n \quad \text{for all } n = 1, \ldots, N.$$

Uniqueness of ℓ_p norm minimization

In a vector space $\mathcal{S}$ with a strictly convex norm $\|\cdot\|$, the program

$$\min_{\boldsymbol{w} \in \mathcal{T}} \|\boldsymbol{y} - \boldsymbol{w}\|, \quad \mathcal{T} = \text{ subspace},$$

has a unique solution for all $\boldsymbol{y} \in \mathcal{S}$.

Proof: Consider two vectors $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathcal{T}$ that have equal distance to $\boldsymbol{y}$:

$$\|\boldsymbol{y} - \boldsymbol{w}_1\| = \|\boldsymbol{y} - \boldsymbol{w}_2\| = \rho.$$

[4] Proving this would make another good homework problem, but if you are impatient you can look at http://en.wikipedia.org/wiki/Minkowski_inequality.

We first show that if $\boldsymbol{w}_1 \neq \boldsymbol{w}_2$, we can always find another point in $\mathcal{T}$ that is even closer to $\boldsymbol{y}$. By strict convexity,

$$\|0.5(\boldsymbol{y} - \boldsymbol{w}_1) + 0.5(\boldsymbol{y} - \boldsymbol{w}_2)\| < \rho,$$

and so re-arranging the terms inside the norm above,

$$\left\|\boldsymbol{y} - \frac{\boldsymbol{w}_1 + \boldsymbol{w}_2}{2}\right\| < \rho,$$

and so $\boldsymbol{w}_3 = (\boldsymbol{w}_1 + \boldsymbol{w}_2)/2$, which of course is also a member of $\mathcal{T}$, is closer to $\boldsymbol{y}$ that both $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. Thus the closest point to $\boldsymbol{y}$ in $\mathcal{T}$ must be unique.

Closest point in a convex set

Let $\boldsymbol{x}_0 \in \mathbb{R}^N$ be given, and let $\mathcal{C}$ be a non-empty, closed, convex set. The **projection** of $\boldsymbol{x}_0$ onto $\mathcal{C}$ is the closest point (in the standard Euclidean distance, for now) in $\mathcal{C}$ to $\boldsymbol{x}_0$:

$$P_{\mathcal{C}}(\boldsymbol{x}_0) = \arg\min_{\boldsymbol{y} \in \mathcal{C}} \|\boldsymbol{x}_0 - \boldsymbol{y}\|_2$$

We will see below that there is a unique minimizer to this problem, and that the solution has geometric properties that are analogous to the case where $\mathcal{C}$ is a subspace.

Projection onto a subspace

Let's recall how we solve this problem in the special case where $\mathcal{C} := \mathcal{T}$ is a K-dimensional *subspace*. In this case, the solution $\hat{\boldsymbol{x}} = P_{\mathcal{T}}(\boldsymbol{x}_0)$ is unique, and is characterized by the **orthogonality principle**:

$$\boldsymbol{x}_0 - \hat{\boldsymbol{x}} \perp \mathcal{T}$$

meaning that $\langle \boldsymbol{y}, \boldsymbol{x}_0 - \hat{\boldsymbol{x}} \rangle = 0$ for all $\boldsymbol{y} \in \mathcal{T}$.

Affine sets are not fundamentally different than subspaces. Any affine set $\mathcal{C}$ can be written as a subspace $\mathcal{T}$ plus an offset $\boldsymbol{v}_0$:

$$\mathcal{C} = \mathcal{T} + \boldsymbol{v}_0 = \{\boldsymbol{x} \ : \ \boldsymbol{x} = \boldsymbol{y} + \boldsymbol{v}_0, \ \boldsymbol{y} \in \mathcal{C}\}.$$

It is easy to translate the results for subspaces above to say that the projection onto an affine set is unique, and obeys the orthogonality principle

$$\langle \boldsymbol{y} - \hat{\boldsymbol{x}}, \boldsymbol{x}_0 - \hat{\boldsymbol{x}} \rangle = 0, \quad \text{for all} \ \ \boldsymbol{y} \in \mathcal{C}. \tag{6}$$

You can solve this problem by shifting $\boldsymbol{x}_0$ and $\mathcal{C}$ by negative $\boldsymbol{v}_0$, projecting $\boldsymbol{x}_0 - \boldsymbol{v}_0$ onto the subspace $\mathcal{C} - \boldsymbol{v}_0$, and then shifting the answer back.

Projection onto a general convex set

In general, there is no closed-form expression for the projector onto a given convex set. However, the concepts above (orthogonality, projection onto a subspace) can help us understand the solution for an arbitrary convex set.

Uniqueness If $\mathcal{C}$ is closed and convex, then for any $\boldsymbol{x}_0$, the program

$$\min_{\boldsymbol{y} \in \mathcal{C}} \|\boldsymbol{x}_0 - \boldsymbol{y}\|_2 \tag{7}$$

has a **unique** solution.

First of all, that there is at least one point in $\mathcal{C}$ such that the minimum is obtained is a consequence of $\mathcal{C}$ being closed and of $\|\boldsymbol{x}_0 - \boldsymbol{y}\|_2$ being a continuous function of $\boldsymbol{y}$. Let $\hat{\boldsymbol{x}}$ be one such minimizer; we will show that $\|\boldsymbol{x}_0 - \boldsymbol{y}\|_2 > \|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|_2$ for all $\boldsymbol{y} \in \mathcal{C}$, $y \neq \hat{\boldsymbol{x}}$.

Consider first all the points in $\mathcal{C}$ which are co-aligned with $\hat{\boldsymbol{x}}$. Let

$$\mathcal{I} = \{\alpha \in \mathbb{R} \ : \ \alpha \hat{\boldsymbol{x}} \in \mathcal{C}\}.$$

Since $\mathcal{C}$ is convex and closed, this is a closed interval of the real line (that contains at least the point $\alpha = 1$). The function

$$g(\alpha) = \|\boldsymbol{x}_0 - \alpha \hat{\boldsymbol{x}}\|_2^2 = \alpha^2 \|\hat{\boldsymbol{x}}\|_2^2 - 2\alpha \langle \hat{\boldsymbol{x}}, \boldsymbol{x}_0 \rangle + \|\boldsymbol{x}_0\|_2^2,$$

is minimized at $\alpha = 1$ by construction, and since its second derivative is strictly positive, we know that this is the unique minima. So if there is another minimizer of (7), then it is not co-aligned with $\hat{\boldsymbol{x}}$.

Now let $\boldsymbol{y}$ be any point in $\mathcal{C}$ that is not co-aligned with $\hat{\boldsymbol{x}}$. We will show that $\boldsymbol{y}$ cannot minimize (7) because the point $\hat{\boldsymbol{x}}/2 + \boldsymbol{y}/2 \in \mathcal{C}$

is definitively closer to $\boldsymbol{x}_0$. We have

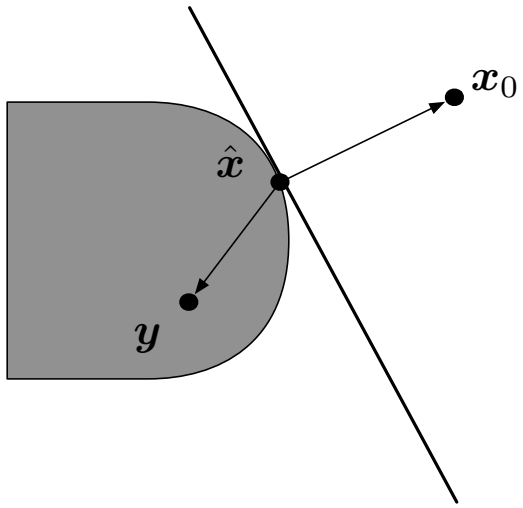
$$\begin{aligned}
\left\| \boldsymbol{x}_0 - \frac{\hat{\boldsymbol{x}}}{2} - \frac{\boldsymbol{y}}{2} \right\|_2^2 &= \left\| \frac{\boldsymbol{x}_0 - \hat{\boldsymbol{x}}}{2} + \frac{\boldsymbol{x}_0 - \boldsymbol{y}}{2} \right\|_2^2 \\
&= \frac{\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|_2^2}{4} + \frac{\|\boldsymbol{x}_0 - \hat{\boldsymbol{y}}\|_2^2}{4} + \frac{\langle \boldsymbol{x}_0 - \hat{\boldsymbol{x}}, \boldsymbol{x}_0 - \boldsymbol{y} \rangle}{2} \\
&< \frac{\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|_2^2}{4} + \frac{\|\boldsymbol{x}_0 - \hat{\boldsymbol{y}}\|_2^2}{4} + \frac{\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}|_2 \, \|\boldsymbol{x}_0 - \boldsymbol{y}\|_2}{2} \\
&= \left(\frac{\|\boldsymbol{x}_0 - \hat{\boldsymbol{x}}\|_2}{2} + \frac{\|\boldsymbol{x}_0 - \boldsymbol{y}\|_2}{2} \right)^2 \\
&\leq \|\boldsymbol{x}_0 - \boldsymbol{y}\|_2^2.
\end{aligned}$$

The strict inequality above follows from Cauchy-Schwarz, while the last inequality follows from the fact that $\hat{\boldsymbol{x}}$ is a minimizer. This shows that no $\boldsymbol{y} \neq \hat{\boldsymbol{x}}$ can also minimize (7), and so $\hat{\boldsymbol{x}}$ is unique.

Obtuseness. Relative to the solution $\hat{\boldsymbol{x}}$, every vector in $\mathcal{C}$ is at an obtuse angle to the error $\boldsymbol{x}_0 - \hat{\boldsymbol{x}}$. More precisely, $P_{\mathcal{C}}(\boldsymbol{x}_0) = \hat{\boldsymbol{x}}$ if and only if

$$\langle \boldsymbol{y} - \hat{\boldsymbol{x}}, \boldsymbol{x}_0 - \hat{\boldsymbol{x}} \rangle \leq 0. \tag{8}$$

Compare with (6) above. Here is a picture:



We first prove that (8)$\Rightarrow P_{\mathcal{C}}(\boldsymbol{x}_0) = \hat{\boldsymbol{x}}$. For any $\boldsymbol{y} \in \mathcal{C}$, we have

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{x}_0\|_2^2 &= \|\boldsymbol{y} - \hat{\boldsymbol{x}} + \hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2 \\ &= \|\boldsymbol{y} - \hat{\boldsymbol{x}}\|_2^2 + \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2 + 2\langle \boldsymbol{y} - \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}} - \boldsymbol{x}_0 \rangle \\ &\geq \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2 + 2\langle \boldsymbol{y} - \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}} - \boldsymbol{x}_0 \rangle. \end{aligned}$$

Note that the inner product term above is the same as (8), but with the sign of the second argument flipped, so we know this term must be non-negative. Thus

$$\|\boldsymbol{y} - \boldsymbol{x}_0\|_2^2 \;\geq\; \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2^2.$$

Since this holds uniformly over all $\boldsymbol{y} \in \mathcal{C}$, $\hat{\boldsymbol{x}}$ must be the closest point in $\mathcal{C}$ to $\boldsymbol{x}_0$.

We now show that $P_{\mathcal{C}}(\boldsymbol{x}_0) = \hat{\boldsymbol{x}} \;\Rightarrow\; (8)$. Let $\boldsymbol{y}$ be an arbitrary point in $\mathcal{C}$. Since $\mathcal{C}$ is convex, the point $\hat{\boldsymbol{x}} + \theta(\boldsymbol{y} - \hat{\boldsymbol{x}})$ must also be in $\mathcal{C}$ for all $\theta \in [0, 1]$. Since $\hat{\boldsymbol{x}} = P_{\mathcal{C}}(\boldsymbol{x}_0)$,

$$\|\hat{\boldsymbol{x}} + \theta(\boldsymbol{y} - \hat{\boldsymbol{x}}) - \boldsymbol{x}_0\|_2 \;\geq\; \|\hat{\boldsymbol{x}} - \boldsymbol{x}_0\|_2 \quad \text{for all} \quad \theta \in [0, 1]. \tag{9}$$

It is a basic geometric fact that that for any two vectors $\boldsymbol{x}, \boldsymbol{y}$,

$$\text{if} \;\; \|\boldsymbol{x} + \theta\boldsymbol{y}\|_2 \geq \|\boldsymbol{x}\|_2 \;\; \text{for all} \;\; \theta \in [0, 1] \quad \text{then} \quad \langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq 0. \tag{10}$$

Applying this to (9) means

$$\langle \boldsymbol{y} - \hat{\boldsymbol{x}}, \hat{\boldsymbol{x}} - \boldsymbol{x}_0 \rangle \;\geq\; 0 \quad \Rightarrow \quad \langle \boldsymbol{y} - \hat{\boldsymbol{x}}, \boldsymbol{x}_0 - \hat{\boldsymbol{x}} \rangle \;\leq\; 0.$$

Finally, to prove (10), we expand the norm as

$$\|\boldsymbol{x} + \theta\boldsymbol{y}\|_2^2 = \|\boldsymbol{x}\|_2^2 + \theta^2\|\boldsymbol{y}\|_2^2 + 2\theta\langle \boldsymbol{x}, \boldsymbol{y} \rangle,$$

from which we can immediately deduce that

$$\begin{aligned} &\frac{\theta}{2}\|\boldsymbol{y}\|_2^2 + \langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq 0 \quad \text{for all} \;\; \theta \in [0, 1] \\ &\qquad \Rightarrow \langle \boldsymbol{x}, \boldsymbol{y} \rangle \geq 0. \end{aligned}$$