The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381K: Convex Optimization — Fall 2019

Lecture 21

Aryan Mokhtari Wednesday, November 13, 2019.

Goal: In this lecture, we talk about approximation and fitting. In particular, we study norm approximation, least-norm problems, regularized approximation, and robust approximation.

1 Norm approximation

Consider solving the following optimization problem:

$$\min : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|,$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$.

Let's consider the definition $\mathbf{x}^* = \text{argmin}_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|$. Then, we can come up with different interoperation for the norm approximation problem and optimal solution $\mathbf{x}^*$.

1. **Geometric interpretation:** $\mathbf{x}^*$ is such that $\mathbf{A}\mathbf{x}^*$ is the closest (with respect to the norm) point to $\mathbf{b}$ among the points in $\mathcal{R}(\mathbf{A})$.

   Also, we can rewrite this problem as projection of $\mathbf{b}$ into the Range of $\mathbf{A}$ in the considered norm

   $$\begin{aligned} \min : \quad & \|\mathbf{u} - \mathbf{b}\| \\ \text{s.t.} : \quad & \mathbf{u} \in \mathcal{R}(\mathbf{A}) \end{aligned}$$

   (to see the connection set $\mathbf{u} = \mathbf{A}\mathbf{x}$.)

2. **Estimation interpretation:** Consider a linear measurement model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{v}$ where

   (a) $\mathbf{y} \in \mathbb{R}^m$ is the vector of measurements

   (b) $\mathbf{x} \in \mathbb{R}^n$ is the vector of parameters to be estimated

   (c) $\boldsymbol{\nu} \in \mathbb{R}^m$ is some measurement error that is unknown, but presumed to be small.

   Here the goal is to estimate $\mathbf{x}$ given $\mathbf{y}$, and $\mathbf{x}^*$ is the best estimate of $\mathbf{x}$ when $\mathbf{y} = \mathbf{b}$.

3. **Optimal design interpretation:** In the optimal design problem we think of $\mathbf{x}$ as our design variables (input) and $\mathbf{A}\mathbf{x}$ as result (output). The goal is to find the input which leads to the desired output $\mathbf{b}$. We can also think of $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{b}$ as the deviation between the actual output and the desired output.

1.1 Least-squares approximation (ℓ_2 norm)

Consider the case that we solve the norm approximation problem using ℓ_2 norm. Then,

$$\begin{aligned} &\min : \|\mathbf{Ax}-\mathbf{b}\|_2^2 && \min : r_1^2+\cdots+r_m^2 \\ & && \text{s.t.} : \mathbf{r}=\mathbf{Ax}-\mathbf{b}, \end{aligned}$$

By optimality condition, we know that

$$\mathbf{A}^\top\mathbf{Ax}^* = \mathbf{A}^\top\mathbf{b}$$

and if $rank(\mathbf{A}) = n$, then we have $\mathbf{x}^* = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b}$.

1.2 Chebyshev approximation (ℓ_∞ norm)

Consider the case that we solve the norm approximation problem using ℓ_∞ norm. Then,

$$\begin{aligned} &\min : \|\mathbf{Ax}-\mathbf{b}\|_\infty && \min : \max_i\{|r_1|,\ldots,|r_m|\} \\ & && \text{s.t.} : \mathbf{r}=\mathbf{Ax}-\mathbf{b}, \end{aligned}$$

This problem can be written as an LP:

$$\begin{aligned} &\min : \ t \\ &\text{s.t.} : -t\mathbf{1} \le \mathbf{Ax}-\mathbf{b} \le t\mathbf{1}. \end{aligned}$$

1.3 Sum of absolute residuals approximation (ℓ_1 norm)

Consider the case that we solve the norm approximation problem using ℓ_1 norm. Then,

$$\begin{aligned} &\min : \|\mathbf{Ax}-\mathbf{b}\|_1 && \min : |r_1|+\cdots+|r_m| \\ & && \text{s.t.} : \mathbf{r}=\mathbf{Ax}-\mathbf{b}, \end{aligned}$$

This problem can be written as an LP:

$$\begin{aligned} &\min : \ \mathbf{1}^\top\mathbf{t} \\ &\text{s.t.} : -\mathbf{t} \le \mathbf{Ax}-\mathbf{b} \le \mathbf{t}. \end{aligned}$$

1.4 General norm form (ℓ_p norm)

Consider the case that we solve the norm approximation problem using ℓ_p norm. Then,

$$\begin{aligned} &\min : \|\mathbf{Ax}-\mathbf{b}\|_p && \min : \left(|r_1|^p+\cdots+|r_m|^p\right)^{1/p}, \\ & && \text{s.t.} : \mathbf{r}=\mathbf{Ax}-\mathbf{b}, \end{aligned}$$

1.5 Penalty function approximation

Consider a convex function $\phi : \mathbb{R} \to \mathbb{R}$.

$$\begin{aligned} &\min : \phi(r_1)+\cdots+\phi(r_m), \\ &\text{s.t.} : \mathbf{r}=\mathbf{Ax}-\mathbf{b}, \end{aligned}$$

In this formulation ϕ captures the price that we pay by having a nonzero residual.

1. $\phi(u) = |u|^p$: leads to ℓ_p norm approximation. Note that by choosing a larger p we care less about small residuals and we are more sensitive to big residuals.

2. Deadzone-linear penalty with deadzone $a > 0$:

$$\phi(u) = \begin{cases} 0 & \text{if } |u| \leq a \\ |u| - a & \text{if } |u| > a \end{cases}$$

   Not sensitive at all to residuals smaller than a, and we have to pay a linear penalty for residuals larger than a.

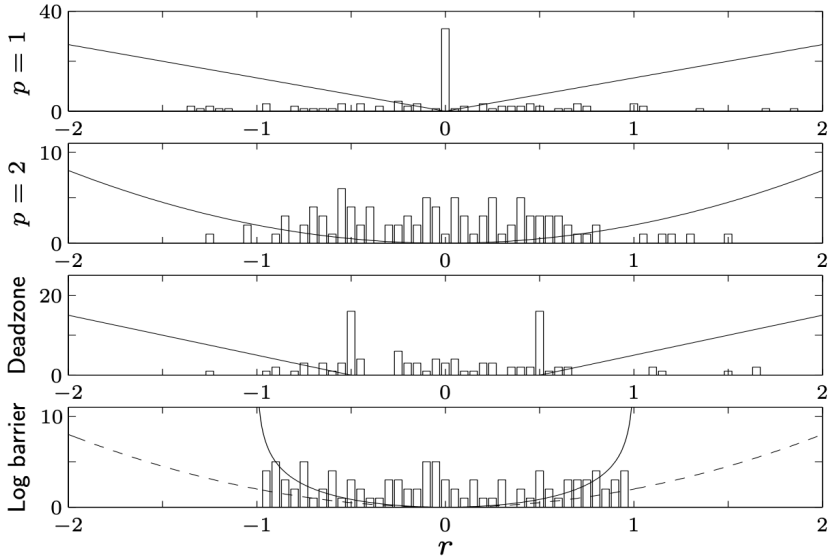
3. Log-barrier penalty function with limit $a > 0$:

$$\phi(u) = \begin{cases} -a^2 \log\left(1 - \left(\frac{u}{a}\right)^2\right) & \text{if } |u| < a \\ \infty & \text{if } |u| \geq a \end{cases}$$

   very sensitive to large residuals, and we pay a quadratic penalty for residuals that have norm smaller than a.

example ($m = 100$, $n = 30$): histogram of residuals for penalties

$$\phi(u) = |u|, \quad \phi(u) = u^2, \quad \phi(u) = \max\{0, |u| - a\}, \quad \phi(u) = -\log(1 - u^2)$$



shape of penalty function has large effect on distribution of residuals

Figure 1: Shape of penalty function has large effect on distribution of residuals

4. Huber penalty function (with parameter M)

$$\phi(u) = \begin{cases} u^2 & \text{if } |u| \leq M \\ M(2|u| - M) & \text{if } |u| > M \end{cases}$$

It is a combination for ℓ_2 and ℓ_1 approximations. It is not very sensitive to very small residuals, and also linearly sensitive to large residuals which makes approximation less sensitive to outliers!

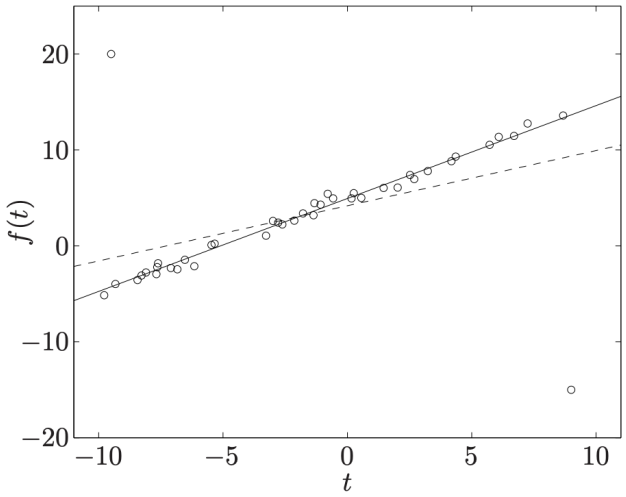


Figure 6.5 The 42 circles show points that can be well approximated by an affine function, except for the two outliers at upper left and lower right. The dashed line is the least-squares fit of a straight line $f(t) = \alpha + \beta t$ to the points, and is rotated away from the main locus of points, toward the outliers. The solid line shows the robust least-squares fit, obtained by minimizing Huber's penalty function with $M = 1$. This gives a far better fit to the non-outlier data.

Figure 2: Shape of penalty function has large effect on distribution of residuals

2 Least-norm problems

Now consider the following optimization problem

$$\begin{aligned} &\text{min}: \ \|\mathbf{x}\| \\ &\text{s.t.}: \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $m \leq n$. WLOG, we assume that the rows of $\mathbf{A}$ are linearly independent. When $m = n$, then the only feasible solution is $\mathbf{A}^{-1}\mathbf{b}$ and when $m < n$ the problem is underdetermined.

We can come up with different interoperation for the least-norm problems problem:

1. **Geometric interpretation:** $\mathbf{x}^*$ is a point in the affine set $\{\mathbf{x}|\mathbf{A}\mathbf{x} = \mathbf{b}\}$ that has the smallest norm (or is the closest point to the origin).

2. **Estimation interpretation:** we do not have enough measurements and we need to come up with a model $\mathbf{x}$ that has the smallest norm (simplest model) and is consistent with our measurements $\mathbf{A}\mathbf{x} = \mathbf{b}$.

3. **Optimal design interpretation:** We plan to find the simplest input $\mathbf{x}$ (smallest norm) that leads to our desired output $\mathbf{b}$.

2.1 Least-squares approximation (ℓ_2 norm)

Consider the case that we solve the norm approximation problem using ℓ_2 norm. Then,

$$\begin{aligned} &\min: \ \|\mathbf{x}\|_2^2 \\ &\text{s.t.}: \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned}$$

By optimality conditions, we know that

$$2\mathbf{x}^* + \mathbf{A}^\top \mathbf{v}^* = \mathbf{0}, \qquad \mathbf{A}\mathbf{x}^* = \mathbf{b}$$

Hence,

$$\boldsymbol{\nu}^* = -2(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{b}, \qquad \mathbf{x}^* = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{b}$$

2.2 Minimum sum of absolute values (ℓ_1 norm)

Consider the case that we solve the norm approximation problem using ℓ_1 norm. Then,

$$\begin{aligned} &\min: \ \|\mathbf{x}\|_1 \\ &\text{s.t.}: \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned}$$

This problem can be written as an LP:

$$\begin{aligned} &\min: \ \mathbf{1}^\top \mathbf{y} \\ &\text{s.t.}: -\mathbf{y} \leq \mathbf{x} \leq \mathbf{y}, \qquad \mathbf{A}\mathbf{x} = \mathbf{b}. \end{aligned}$$

2.3 Least-penalty problem

Consider a convex function $\phi : \mathbb{R} \to \mathbb{R}$.

$$\begin{aligned} &\min: \ \phi(x_1) + \cdots + \phi(x_n) \\ &\text{s.t.}: \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned}$$

In this case we put penalty on the norm of the variables.

3 Regularized approximation

In the most general case, we want to find a solution $\mathbf{x}$ that minimizes both $\|\mathbf{x}\|$ and $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|$, which is a bi-criterion optimization problem. Instead, we minimize their weighted sum with some parameter $\gamma > 0$, i.e.,

$$\min: \ \|\mathbf{A}\mathbf{x} - \mathbf{b}\| + \gamma\|\mathbf{x}\|$$

By choosing a larger γ we care more about the simplicity of the model and its norm and we care less about fitting the measurements. Conversely, by choosing a small γ, we emphasize more on fitting and less on simplicity of the model.

3.1 Tikhonov regularization

When we use Euclidean norms then we have

$$\min: \ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \delta\|\mathbf{x}\|_2^2$$

which has the following closed-form solution: (requires no rank assumptions on the matrix $\mathbf{A}$)

$$\mathbf{x}^* = (\mathbf{A}^\top\mathbf{A} + \delta\mathbf{I})^{-1}\mathbf{A}^\top\mathbf{b}$$

3.2 Signal reconstruction

Consider the following problem:

1. $\mathbf{x} \in \mathbb{R}^n$ is an unknown signal
2. $\mathbf{x}_{cor} = \mathbf{x} + \mathbf{v}$ is the corrupted signal with noise $\mathbf{v}$
3. $\hat{\mathbf{x}}$ is an estimate of $\mathbf{x}$
4. We measure smoothness of a signal by $\phi_{quad}(\hat{\mathbf{x}}) = \sum_{i=1}^{n-1}(\hat{x}_i - \hat{x}_{i+1})^2$
5. Goal: Find $\hat{\mathbf{x}}$ that is smooth (does not have a lot of variations) and is close to $\mathbf{x}_{cor}$

The regularized version of this problem can be written as

$$\text{min}: \ \|\mathbf{x}_{cor} - \hat{\mathbf{x}}\|_2^2 + \delta \sum_{i=1}^{n-1}(\hat{x}_{i+1} - \hat{x}_i)^2$$

which can be written as

$$\text{min}: \ \|\mathbf{x}_{cor} - \hat{\mathbf{x}}\|_2^2 + \delta \|\mathbf{D}\hat{\mathbf{x}}\|_2^2$$

where $\mathbf{D} \in \mathbb{R}^{(n-1)\times n}$ is given by $\begin{bmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}$ and hence $\hat{\mathbf{x}}^* = (\mathbf{I} + \delta \mathbf{D}^\top \mathbf{D})^{-1}\mathbf{x}_{cor}$

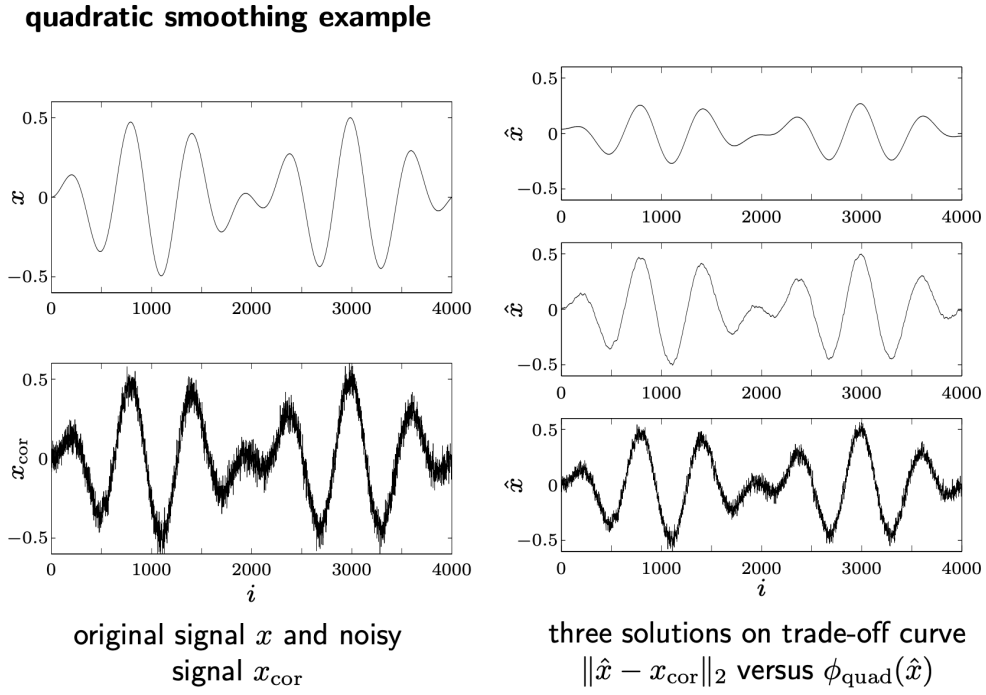


Figure 3: Shape of penalty function has large effect on distribution of residuals

4 Robust approximation

We aim to minimize $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|$ but we are not certain about $\mathbf{A}$.

4.1 Stochastic approach

In the stochastic approach, we assume that $\mathbf{A}$ has a probability distribution and we minimize

$$\min: \ \mathbf{E}[\|\mathbf{A}\mathbf{x}-\mathbf{b}\|]$$

E.g., consider the case that we use ℓ_2 norm and $\mathbf{A}=\bar{\mathbf{A}}-\mathbf{U}$ where $\mathbf{E}[\mathbf{U}]=\mathbf{0}$ and $\mathbf{E}[\mathbf{U}^\top\mathbf{U}]=\mathbf{P}$:

$$\begin{aligned}\mathbf{E}[\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2] &= \mathbf{E}(\mathbf{A}\mathbf{x}-\mathbf{b})^\top(\mathbf{A}\mathbf{x}-\mathbf{b})\\ &= (\bar{\mathbf{A}}\mathbf{x}-\mathbf{b})^\top(\bar{\mathbf{A}}\mathbf{x}-\mathbf{b})+\mathbf{E}[\mathbf{x}^\top\mathbf{U}^\top\mathbf{U}\mathbf{x}]\\ &= \|\bar{\mathbf{A}}\mathbf{x}-\mathbf{b}\|_2^2+\|\mathbf{P}^{1/2}\mathbf{x}\|_2^2\end{aligned}$$

This can be considered as Tikhonov regularization when we set $\mathbf{P}=\delta\mathbf{I}$. This means that in Tikhonov regularization we are not certain about our measurements and to robustify our solution with assume a prior model for the noise $\mathbf{U}$ that $\mathbf{E}[\mathbf{U}]=\mathbf{0}$ and $\mathbf{E}[\mathbf{U}^\top\mathbf{U}]=\delta\mathbf{I}$.

Another example is when $\mathbf{A}$ has only a finite number of values, and therefore we can rewrite the problem $\min: \ \mathbf{E}[\|\mathbf{A}\mathbf{x}-\mathbf{b}\|]$ as

$$\min: \ \sum_{i=1}^{k} p_i\|\mathbf{A}_i\mathbf{x}-\mathbf{b}\|$$

which can be written as

$$\begin{aligned}&\min: \ \mathbf{p}^\top\mathbf{t}\\ &\text{s.t.}: \|\mathbf{A}_i\mathbf{x}-\mathbf{b}\|\le t_i, \quad \text{for } i=1,\dots,k.\end{aligned}$$

For ℓ_2 norm this is an SOCP and for ℓ_1 and ℓ_∞ norms this is an LP.

4.2 Deterministic approach (worst-case)

In the deterministic approach we assume that $\mathbf{A}$ belongs to a set $\mathcal{A}$ and we aim to minimize the worst case norm approximation error by solving

$$\min: \sup_{\mathbf{A}\in\mathcal{A}} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|$$

A special case is when $\mathcal{A}$ has a finite number of elements and the problem can be written as

$$\begin{aligned}&\min: \ t\\ &\text{s.t.}: \|\mathbf{A}_i\mathbf{x}-\mathbf{b}\|\le t, \quad \text{for } i=1,\dots,k.\end{aligned}$$