Optimization Models

EECS 127 / EECS 227AT

Laurent El Ghaoui

EECS department
UC Berkeley

Fall 2018

LECTURE 8

Least Squares and Variants

If others would but reflect on mathematical truths as deeply and continuously as I have, they would make my discoveries.

C.F. Gauss (1777 – 1855)

Outline

1. Least Squares and Minimum-norm solutions
2. Solving systems of linear equations and LS problems
3. Direct and inverse mapping of a unit ball
4. Variants of the Least-Squares problem
   - ℓ_2-regularized LS
5. Examples

Least Squares

- When $y \notin \mathcal{R}(A)$, the system of linear equations is infeasible: there is no x such that $Ax = y$ (as it happens frequently for overdetermined systems).
- In such cases it may however make sense to determine an "approximate solution" to the system, that is a solution that renders the *residual* vector $r \doteq Ax - y$ as "small" as possible.
- In the most common case, we measure the residual via the Euclidean norm, whence the problem becomes

$$\min_x \quad \|Ax - y\|_2^2.$$

- From this, that is a solution that minimizes the sum of the squares of the equation residuals:

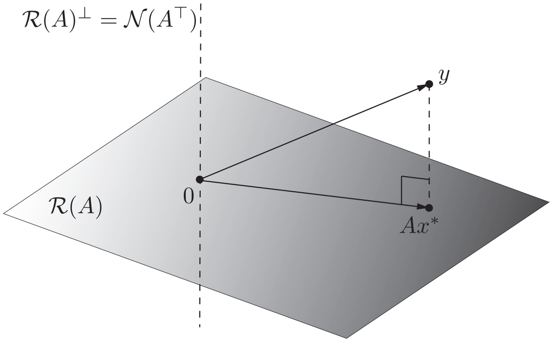
$$\|Ax - y\|_2^2 = \sum_{i=1}^{m} r_i^2, \quad r_i \doteq a_i^\top x - y_i,$$

  where $a_i^\top$ denotes the i-th row of A.

Least Squares

Geometric interpretation

- Since vector Ax lies in $\mathcal{R}(A)$, the problem amounts to determining a point $\tilde{y} = Ax^*$ in $\mathcal{R}(A)$ at *minimum distance* form y.
- The Projection Theorem then tells us that this point is indeed the orthogonal projection of y onto the subspace $\mathcal{R}(A)$.



Least Squares

Solution

- $y - Ax^* \in \mathcal{R}(A)^\perp = \mathcal{N}(A^\top)$, hence

$$A^\top (y - Ax^*) = 0$$

- Solutions x^* to the LS problem must satisfy the Normal Equations:

$$A^\top A x = A^\top y$$

- This system *always* admits a solution.
- If A is full column rank (i.e., $\mathrm{rank}(A) = n$), then the solution is unique, and it is given by

$$x^* = (A^\top A)^{-1} A^\top y.$$

Minimum-norm solutions

- When matrix A has more columns than rows ($m < n$: underdetermined), and $y \in \mathcal{R}(A)$, we have that $\dim \mathcal{N}(A) \geq n - m > 0$, hence the system $y = Ax$ has infinite solutions and that the set of solutions is $\mathcal{S}_{\bar{x}} = \{x : x = \bar{x} + z,\, z \in \mathcal{N}(A)\}$, where $\bar{x}$ is any vector such that $A\bar{x} = y$.
- We single out from $\mathcal{S}_{\bar{x}}$ the one solution x^* with minimal Euclidean norm. That is, we solve

$$\min_{x:Ax=y} \|x\|_2,$$

  which is equivalent to $\min_{x \in \mathcal{S}_{\bar{x}}} \|x\|_2$.
- The solution x^* must be orthogonal to $\mathcal{N}(A)$ or, equivalently, $x^* \in \mathcal{R}(A^\top)$, which means that $x^* = A^\top \xi$, for some suitable ξ.
- Since x^* must solve the system of equations, it must be $Ax^* = y$, i.e., $AA^\top \xi = y$.
- If A is full row rank, $AA^\top$ is invertible and the unique ξ that solves the previous equation is $\xi = (AA^\top)^{-1} y$. This finally gives us the unique minimum-norm solution of the system:

$$x^* = A^\top (AA^\top)^{-1} y.$$

LS solutions and the pseudoinverse

Corollary 1 (Set of solutions of LS problem)

The set of optimal solutions of the LS problem

$$p^* = \min_x \|Ax - y\|_2$$

can be expressed as

$$\mathcal{X}_{\mathrm{opt}} = A^\dagger y + \mathcal{N}(A),$$

where $A^\dagger y$ *is the minimum-norm point in the optimal set. The optimal value* p^* *is the norm of the projection of* y *onto orthogonal complement of* $\mathcal{R}(A)$*: for* $x^* \in \mathcal{X}_{\mathrm{opt}}$*,*

$$p^* = \|y - Ax^*\|_2 = \|(I_m - AA^\dagger)y\|_2 = \|P_{\mathcal{R}(A)^\perp} y\|_2,$$

where matrix $P_{\mathcal{R}(A)^\perp}$ *is the projector onto* $\mathcal{R}(A)^\perp$*. If* A *is full column rank, then the solution is unique, and equal to*

$$x^* = A^\dagger y = (A^\top A)^{-1} A^\top y.$$

Solving systems of linear equations and LS problems

Direct methods

- We discuss techniques for solving a square and nonsingular system of equations of the form
$$Ax = y, \quad A \in \mathbb{R}^{n,n}, \ A \text{ nonsingular.}$$
- If $A \in \mathbb{R}^{n,n}$ has a special structure, such as upper (resp., lower) triangular matrix, then the algorithms of *backward substitution* (resp., *forward substitution*) can be directly applied.
- If A is not triangular, then the method of *Gaussian elimination* applies a sequence of elementary operations that reduce the system in upper triangular form. Then, backward substitution can be applied to this transformed system in triangular form.
- A possible drawback of these methods is that they work simultaneously on the coefficient matrix A and on the right-hand side term y, hence the whole process has to be redone if one needs to solve the system for several different right-hand sides.

Solving systems of linear equations and LS problems

Factorization-based methods

- Another common approach for solving $Ax = y$ is the so-called *factor-solve* method.
- The coefficient matrix A is first factored into the product of matrices having particular structure (such as orthogonal, diagonal, or triangular), and then the solution is found by solving a sequence of simpler systems of equations, where the special structure of the factor matrices can be exploited.
- An advantage of factorization methods is that, once the factorization is computed, it can be used to solve systems for many different values of the right-hand side y.

Factor-solve via SVD

- SVD of $A \in \mathbb{R}^{n,n}$: $A = U\Sigma V^\top$, where $U, V \in \mathbb{R}^{n,n}$ are orthogonal, and Σ is diagonal and nonsingular.
- We write the system $Ax = y$ as a sequence of systems:

$$\xrightarrow{x} \boxed{V^\top} \xrightarrow{z} \boxed{\Sigma} \xrightarrow{w} \boxed{U} \xrightarrow{y}$$

$$Uw = y, \quad \Sigma z = w, \quad V^\top x = z,$$

- These are readily solved sequentially as

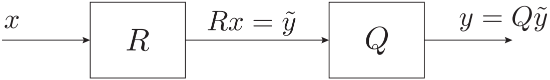
$$w = U^\top y, \quad z = \Sigma^{-1} w, \quad x = Vz.$$

Factor-solve via QR

- Any nonsingular matrix $A \in \mathbb{R}^{n,n}$ can be factored as $A = QR$, where $Q \in \mathbb{R}^{n,n}$ is orthogonal, and R is upper triangular with positive diagonal entries.
- Then, the linear equations $Ax = y$ can be solved by first multiplying both sides on the left by $Q^\top$, obtaining

$$Q^\top Ax = Rx = \tilde{y}, \quad \tilde{y} = Q^\top y,$$

  and then solving the triangular system $Rx = \tilde{y}$ by backward substitution.
- This factor-solve process is represented graphically in the figure below.



SVD method for non-square systems

- Consider the linear equations
$$Ax = y,$$
where $A \in \mathbb{R}^{m,n}$, and $y \in \mathbb{R}^m$, and let $A = U\tilde{\Sigma}V^\top$ be an SVD of A.
- We can completely describe the set of solutions via SVD, as follows. Pre-multiply the linear equation by the inverse of U, $U^\top$; then
$$\tilde{\Sigma}\tilde{x} = \tilde{y}, \quad \tilde{x} = V^\top x,$$
where $\tilde{y} = U^\top y$.
- Due to the diagonal form of $\tilde{\Sigma}$, the above writes
$$\sigma_i \tilde{x}_i = \tilde{y}_i, \ \ i = 1, \ldots, r; \quad 0 = \tilde{y}_i, \ i = r + 1, \ldots, m.$$

SVD method for non-square systems

Two cases can occur:

1. If the last $m - r$ components of $\tilde{y}$ are not zero, then the second set of conditions in the last expression are not satisfied, hence the system is infeasible, and the solution set is empty. This occurs when y is not in the range of A.
2. If y is in the range of A, then the second set of conditions in the last expression hold, and we can solve for $\tilde{x}$ with the first set of conditions, obtaining

$$\tilde{x}_i = \frac{\tilde{y}_i}{\sigma_i}, \quad i = 1, \ldots, r.$$

   The last $n - r$ components of $\tilde{x}$ are free. This corresponds to elements in the nullspace of A.

   If A is full column rank (its nullspace is reduced to $\{0\}$), then there is a unique solution. Once vector $\tilde{x}$ is obtained, the actual unknown x can then be recovered as $x = V\tilde{x}$.

Solving LS problems

- Given $A \in \mathbb{R}^{m,n}$ and $y \in \mathbb{R}^m$, we discuss solution of the LS problem

$$\min_x \|Ax - y\|_2.$$

- All solutions of the LS problem are solutions of the system of normal equations

$$A^\top Ax = A^\top y.$$

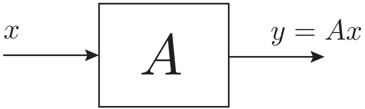
- Therefore, LS solutions can be obtained by either using either direct or factor-solve methods to the normal equations.

Direct and inverse mapping of a unit ball

- We focus on the linear map

$$y = Ax, \quad A \in \mathbb{R}^{m,n},$$

  where $x \in \mathbb{R}^n$ is the input vector, and $y \in \mathbb{R}^m$ is the output.
- We consider two problems that we call the direct and the inverse (or estimation) problem.



Direct and inverse mapping of a unit ball

Direct problem

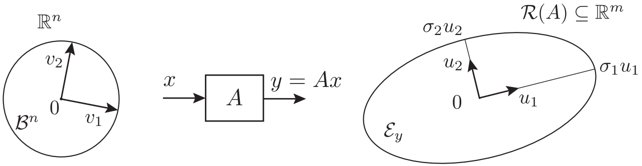
- In the direct problem, we assume that the input x lies in a unit Euclidean ball centered at zero, and we ask where the output y is.
- That is, we let

$$x \in \mathcal{B}^n, \quad \mathcal{B}^n = \{z \in \mathbb{R}^n : \|z\|_2 \le 1\}$$

  and we want to find the output set

$$\mathcal{E}_y = \{y : y = Ax, x \in \mathcal{B}^n\}.$$

- This set is a bounded but possibly degenerate ellipsoid (flat on $\mathcal{R}(A)^\perp$), with the axes directions given by the right singular vectors u_i and with the semi-axes lengths given by σ_i, $i = 1, \ldots, n$.



Direct and inverse mapping of a unit ball

Inverse problem

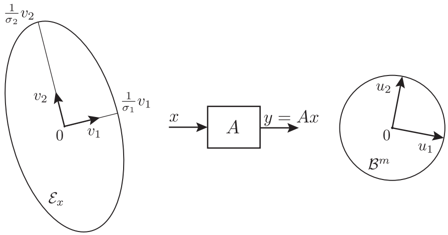
- Suppose $y \in \mathcal{B}^m$, we ask what is the set of input vectors x that would yield such a set as output. Formally, we seek

$$\mathcal{E}_x = \{x \in \mathbb{R}^n : Ax \in \mathcal{B}^m\}.$$

- Since $Ax \in \mathcal{B}^m$ if and only if $x^\top A^\top Ax \leq 1$, we obtain that $\mathcal{E}_x$ is

$$\mathcal{E}_x = \{x \in \mathbb{R}^n : x^\top (A^\top A)x \leq 1\}.$$

  This ellipsoid is unbounded along directions x in the nullspace of A. The axes of $\mathcal{E}_x$ are along the directions of the left singular vectors v_i, and the semi axes lengths are given by σ_i^{-1}, $i = 1, \dots, n$.



Variants of the Least-Squares problem

Linear equality-constrained LS

- A generalization of the basic LS problem allows for the addition of linear equality constraints on the x variable, resulting in the constrained problem

$$\min_x \|Ax - y\|_2^2, \quad \text{s.t. } Cx = d,$$

where $C \in \mathbb{R}^{p,n}$ and $d \in \mathbb{R}^p$.

- This problem can be converted into a standard LS one, by "eliminating" the equality constraints, via a standard procedure. Suppose the problem is feasible, and let $\bar{x}$ be such that $C\bar{x} = d$.
- All feasible points are expressed as $x = \bar{x} + Nz$, where N contains by columns a basis for $\mathcal{N}(C)$, and z is a new variable.
- Problem becomes unconstrained in variable z:

$$\min_z \|\bar{A}z - \bar{y}\|_2^2,$$

where $\bar{A} \doteq AN$, $\bar{y} \doteq y - A\bar{x}$.

Variants of the Least-Squares problem

Weighted LS

- The standard LS objective is a sum of squared equation residuals

$$\|Ax - y\|_2^2 = \sum_{i=1}^{m} r_i^2, \quad r_i = a_i^\top x - y_i.$$

- In some cases, the equation residuals may not be given the same importance, and this relative importance can be modeled by introducing *weights* into the LS objective, that is $f_0(x) = \sum_{i=1}^{m} w_i^2 r_i^2$, where $w_i \geq 0$ are the given weights. This objective is rewritten as

$$f_0(x) = \|W(Ax - y)\|_2^2 = \|A_w x - y_w\|_2^2,$$

where

$$W = \mathrm{diag}\left(w_1, \ldots, w_m\right), \quad A_w \doteq WA, \; y_w = Wy.$$

- The weighted LS problem still has the structure of a standard LS problem, with row-weighted matrix A_w and vector y_w.

Variants of the Least-Squares problem

ℓ_2-regularized LS

- Regularized LS refer to a class of problems of the form

$$\min_x \|Ax - y\|_2^2 + \phi(x),$$

where a "regularization," or *penalty*, term $\phi(x)$ is added to the usual LS objective.

- In the most usual cases, ϕ is proportional either to the ℓ_1 or to the ℓ_2 norm of x. The ℓ_1-regularized case gives rise to the LASSO problem, which is discussed in more detail later. The ℓ_2-regularized case is instead discussed next:

$$\min_x \|Ax - y\|_2^2 + \gamma\|x\|_2^2, \quad \gamma \geq 0$$

Variants of the Least-Squares problem

ℓ_2-regularized LS

$$\min_x \ \|Ax - y\|_2^2 + \gamma\|x\|_2^2, \quad \gamma \geq 0$$

- Recalling that the squared Euclidean norm of a block-partitioned vector is equal to the sum of the squared norms of the blocks, i.e.,

$$\left\| \begin{bmatrix} a \\ b \end{bmatrix} \right\|_2^2 = \|a\|_2^2 + \|b\|_2^2$$

we see that the regularized LS problem can be rewritten in the format of a standard LS problem as follows

$$\|Ax - y\|_2^2 + \gamma\|x\|_2^2 = \|\tilde{A}x - \tilde{y}\|_2^2,$$

where

$$\tilde{A} \doteq \begin{bmatrix} A \\ \sqrt{\gamma} I_n \end{bmatrix}, \quad \tilde{y} \doteq \begin{bmatrix} y \\ 0_n \end{bmatrix}.$$

- $\gamma \geq 0$ is a *tradeoff parameter.* Interpretation in terms of tradeoff between *output tracking accuracy* and *input effort*.

Examples

- Linear regression via least-squares.
- Auto-regressive (AR) models for time-series prediction.