Regression and statistical estimation

刘盛

2026 年 1 月 11 日

UESTC

Regression and statistical estimation

Norm approximation

We are given a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $b \in \mathbb{R}^m$. We choose $x \in \mathbb{R}^n$ and compare Ax with b.

Define the residual

$$r(x) = Ax - b \in \mathbb{R}^m,$$

and measure the size of the residual by a norm $\|\cdot\|$ on $\mathbb{R}^m$.

The basic approximation problem is

$$\min_{x \in \mathbb{R}^n} \ \|Ax - b\|. \tag{1}$$

This single template covers: data fitting (regression), parameter estimation from measurements, and many simple design problems. Three common norms :

$$\|r\|_2 = \left(\sum_{i=1}^m r_i^2\right)^{1/2}, \qquad \|r\|_1 = \sum_{i=1}^m |r_i|, \qquad \|r\|_\infty = \max_i |r_i|.$$

Estimation

Think of b as observed outputs and the columns of A as features (regressors).

Given features $a_1, \dots, a_n$, we predict b by a linear model Ax. The coefficients x are chosen by minimizing a norm of the residual.

- ℓ_2 leads to classical least-squares.
- ℓ_1 leads to least absolute deviations (often more robust).
- ℓ_∞ leads to minimax fitting (uniform error control).

A standard measurement model is

$$y = Ax + v,$$

where y is measured, x is unknown, and v is noise.

If we propose an estimate $\hat{x}$, the implied noise is $\hat{v} = y - A\hat{x}$. A common principle is: "choose $\hat{x}$ that makes the implied noise small":

$$\hat{x} \in \arg\min_{x} \|Ax - y\|.$$

So: "regress b on the columns of A" is essentially "solve (1) with a chosen norm".

Weighted norm approximation

Sometimes residual components have different units, different reliability, or different importance. A simple way to reflect this is

$$\min_{x} \; \|W(Ax-b)\|,$$

where $W \in \mathbb{R}^{m\times m}$ is typically diagonal, $W = \mathrm{diag}(w_1,\ldots,w_m)$.

Then the i-th residual effectively becomes $w_i r_i$. Large w_i forces the fit to pay more attention to component i.

This is not a new problem: it is the same as (6.1) with transformed data,

$$\|W(Ax-b)\| = \|\tilde{A}x - \tilde{b}\|, \qquad \tilde{A} = WA, \; \tilde{b} = Wb.$$

Penalty-function approximation: beyond norms

A convenient generalization replaces the norm by a sum of one-dimensional penalties:

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} \phi\big(r_i(x)\big), \qquad r(x) = Ax - b. \tag{6.2}$$

This viewpoint is practical:

- choose ϕ quadratic $\Rightarrow$ least-squares behavior;
- choose ϕ linear in $|u|$ $\Rightarrow$ ℓ_1-type behavior;
- choose ϕ with a deadzone $\Rightarrow$ ignore small errors;
- choose ϕ as a barrier $\Rightarrow$ forbid $|r_i|$ exceeding a limit.

Scaling ϕ by a positive constant does not change the minimizer; the shape matters.

Three illustrative penalties

Let u be a scalar residual.

Quadratic: $\phi(u) = u^2$.

Deadzone-linear (parameter $a > 0$):

$$\phi(u) = \max\{|u| - a,\ 0\}.$$

No cost for $|u| \le a$, linear growth outside.

Log-barrier (limit $a > 0$):

$$\phi(u) = \begin{cases} -a^2 \log\Big(1 - (u/a)^2\Big), & |u| < a, \\ +\infty, & |u| \ge a. \end{cases}$$

This effectively enforces $|u| < a$.

These three already cover: least-squares, “tolerance band”, and hard constraints.

Some common penalty functions

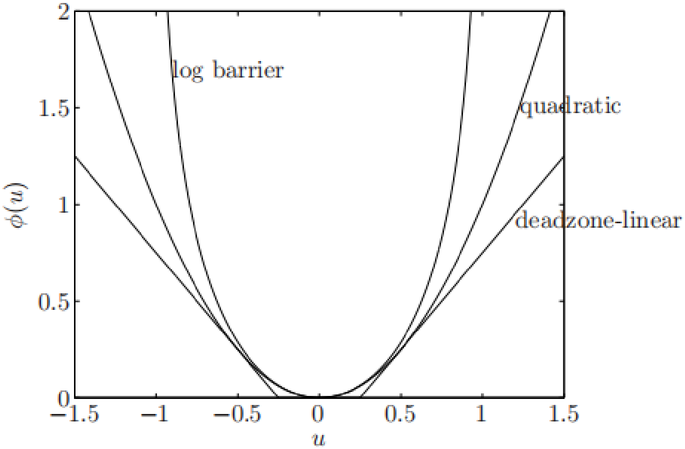

Figure 6.1 Some common penalty functions: the quadratic penalty function $\phi(u) = u^2$, the deadzone-linear penalty function with deadzone width $a = 1/4$, and the log barrier penalty function with limit $a = 1$.

Robust penalties: treating outliers differently

If some measurements are grossly corrupted, squared loss can overreact.

A toy “saturating” idea is

$$\phi(u) = \begin{cases} u^2, & |u| \leq M, \\ M^2, & |u| > M, \end{cases} \tag{6.3}$$

so residuals larger than M no longer become more expensive.

A standard tractable alternative is the Huber penalty:

$$\phi_{\mathrm{hub}}(u) = \begin{cases} u^2, & |u| \leq M, \\ M(2|u| - M), & |u| > M, \end{cases} \tag{6.4}$$

quadratic near 0 and linear in the tails.

In regression, replacing $\sum r_i^2$ by $\sum \phi_{\mathrm{hub}}(r_i)$ typically fits the main cloud instead of chasing a few outliers.

Constrained approximation

In many settings, x is required to satisfy side constraints:

$$\min_{x} \; \|Ax - b\| \quad \text{s.t.} \quad x \in \mathcal{C}.$$

Common choices of $\mathcal{C}$:

- nonnegativity: $x \succeq 0$ (rates, intensities, mixtures);
- simplex: $x \succeq 0, \; \mathbf{1}^T x = 1$ (convex combination / proportions);
- bounds: $\ell \preceq x \preceq u$ (engineering limits);
- norm ball: $\|x\| \leq R$ (controls size/regularization).

Least-norm problems: the model

We chose x to make the residual $Ax - b$ small. Here we flip the emphasis: we enforce $Ax = b$ exactly, and among all feasible x we pick the one with the smallest size.

The basic least-norm problem is

$$\begin{array}{ll} \text{minimize} & \|x\| \\ \text{subject to} & Ax = b, \end{array} \qquad (6.5)$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, and $\| \cdot \|$ is a norm on $\mathbb{R}^n$.

We typically assume the rows of A are independent, so $\operatorname{rank}(A) = m \leq n$. When $m = n$ there is a single feasible point $x = A^{-1}b$. The case that actually has choices is

$$m < n \quad \Longleftrightarrow \quad Ax = b \text{ is underdetermined.}$$

Feasible set and degrees of freedom

Assume $\operatorname{rank}(A) = m < n$. Then the solution set of $Ax = b$ is an affine set:

$$\{x \mid Ax = b\} = x_0 + \mathcal{N}(A),$$

where x_0 is any particular solution and $\mathcal{N}(A) = \{z \mid Az = 0\}$ is the nullspace.

Dimension count:

$$\dim(\mathcal{N}(A)) = n - \operatorname{rank}(A) = n - m.$$

So there are $n - m$ free degrees of freedom. The least-norm problem chooses a particular point on this affine set by minimizing $\|x\|$.

What "small" means depends on the chosen norm:

- $\|x\|_2$: small energy / small Euclidean length;
- $\|x\|_1$: tends to concentrate mass on few coordinates (sparsity);
- $\|x\|_\infty$: keeps every coordinate bounded.

Euclidean case: least-squares solution of $Ax = b$

The most common least-norm choice is the Euclidean norm. Squaring the objective gives an equivalent problem:

$$\begin{array}{ll} \text{minimize} & \|x\|_2^2 \\ \text{subject to} & Ax = b. \end{array}$$

Its unique solution is often called the least-squares solution of the equations $Ax = b$ (or simply the minimum-ℓ_2-norm solution).

Lagrange function： $\min_x \|x\|_2^2 + v^\top (Ax - b)$.

Introduce a dual variable $\nu \in \mathbb{R}^m$ for the constraint. The KKT (first-order optimality) conditions are

$$2x^\star + A^T \nu^\star = 0, \qquad Ax^\star = b.$$

These are linear equations in $(x^\star, \nu^\star)$ and can be solved explicitly when $\text{rank}(A) = m$.

Closed form for the minimum-ℓ_2 solution

From

$$2x^\star + A^T\nu^\star = 0 \quad \Rightarrow \quad x^\star = -\tfrac{1}{2}A^T\nu^\star.$$

Substitute into $Ax^\star = b$:

$$A\big(-\tfrac{1}{2}A^T\nu^\star\big) = b \quad \Rightarrow \quad -\tfrac{1}{2}(AA^T)\nu^\star = b.$$

Hence

$$\nu^\star = -2(AA^T)^{-1}b, \qquad x^\star = A^T(AA^T)^{-1}b.$$

Because $\operatorname{rank}(A) = m < n$, the matrix $AA^T \in \mathbb{R}^{m\times m}$ is invertible. It is also common to write

$$x^\star = A^+b, \qquad A^+ := A^T(AA^T)^{-1},$$

i.e., the Moore–Penrose pseudoinverse formula for full row-rank A.

Least-penalty problems: generalizing "small $\|x\|$"

A useful variant replaces the norm by a separable penalty on components:

$$\begin{array}{ll} \text{minimize} & \phi(x_1) + \cdots + \phi(x_n) \\ \text{subject to} & Ax = b, \end{array} \tag{6.6}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is convex, nonnegative, and satisfies $\phi(0) = 0$.

Read it literally:

- the constraint forces the design/estimate to satisfy $Ax = b$,
- the objective scores how much we "dislike" each component value x_i,
- we pick the feasible x with the smallest total penalty.

This is the analogue of penalty-function approximation in §6.1, with the roles swapped: there we penalized residual components r_i; here we penalize the components of x itself.

Least ℓ_1-norm and sparsity (a typical effect)

A particularly important choice is $\phi(u) = |u|$, i.e., the least ℓ_1-norm solution:

$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax = b. \end{array}$$

Empirical/typical behavior (and a useful rule of thumb): the minimum-ℓ_1 solution often has many components equal to zero, i.e., it tends to produce sparse solutions.

In many instances one observes solutions with roughly m nonzero components (when A is full row rank), meaning the solution uses only as many active variables as the number of constraints.

This is one reason ℓ_1 plays a central role in sparse recovery / compressed sensing, and why least-penalty formulations are not just theoretical generalizations.

Regularized approximation

In many fitting problems we can make the residual small by allowing the coefficient vector x to become large. That is not always a good outcome:

- a large x can mean an unstable design (small perturbations in data lead to big changes in Ax);
- a large x can amplify modeling errors (if $Ax \approx f(x)$ only holds for moderate x);
- in underdetermined settings, there may be infinitely many x with the same fit.

Regularization introduces an explicit tradeoff:

$$\text{make } \|Ax - b\| \text{ small, but also keep } \|x\| \text{ small.}$$

6.3.2 Regularization as scalarization: weighted sum

A common way to select a specific Pareto point is to minimize a weighted sum:

$$\min_{x} \; \|Ax - b\| + \gamma\|x\|, \qquad \gamma > 0. \tag{6.8}$$

What γ does (think “knob”):

- $\gamma \downarrow 0$: you care mostly about fit; x can become large if it helps reduce $\|Ax - b\|$.
- $\gamma \uparrow \infty$: you care mostly about making x small; the solution moves toward $x = 0$.

As γ varies over $(0, \infty)$, the solutions of (6.8) trace the optimal tradeoff curve.

Squared-norm regularization

When Euclidean norms are used, another standard choice is a weighted sum of squares:

$$\min_{x} \ \|Ax - b\|_2^2 + \delta\|x\|_2^2, \qquad \delta > 0. \tag{6.9}$$

Why people like (6.9):

- it is a convex quadratic problem;
- it has a closed-form solution;
- it stays well-posed even when A is rank-deficient.

This is the classical Tikhonov regularization (also called ridge regression in statistics).

Tikhonov regularization: expand and derive the solution

Start from

$$\min_{x} \; \|Ax - b\|_2^2 + \delta\|x\|_2^2. \tag{6.9}$$

Expand the objective:

$$\|Ax - b\|_2^2 + \delta\|x\|_2^2 = x^T(A^TA + \delta I)x - 2b^TAx + b^Tb. \tag{6.10}$$

Differentiate and set the gradient to zero:

$$(A^TA + \delta I)x = A^Tb.$$

Hence the minimizer is

$$x = (A^TA + \delta I)^{-1}A^Tb.$$

Key point: $A^TA + \delta I \succ 0$ for every $\delta > 0$, so the inverse exists with no rank assumptions on A.

Smoothing regularization: penalize variation instead of size

Sometimes “small x” is not the right bias. If x represents samples of a smooth quantity, we may want x to be smooth.

Replace $\|x\|$ by $\|Dx\|$, where D is a (discrete) differentiation operator:

$$\min_{x} \; \|Ax - b\|_2^2 + \delta \|Dx\|_2^2.$$

Example interpretation: $x \in \mathbb{R}^n$ is temperature along $[0, 1]$, and x_i is the temperature at i/n. Then Dx can approximate a first or second derivative, so $\|Dx\|_2^2$ measures roughness.

Second-difference matrix Δ (discrete curvature penalty)

A simple approximation of the second derivative at index i is the second difference

$$n^2(x_{i+1} - 2x_i + x_{i-1}).$$

This can be written as Δx, where $\Delta \in \mathbb{R}^{(n-2)\times n}$ is tridiagonal Toeplitz:

$$\Delta = n^2 \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{bmatrix}.$$

Then $\|\Delta x\|_2^2$ is a discrete measure of mean-square curvature (roughness). A typical smoothing-regularized fit is

$$\min_x \ \|Ax - b\|_2^2 + \delta\|\Delta x\|_2^2.$$

UESTC

Multiple regularizers: control smoothness and size together

You can mix several biases in one objective. A common combination is

$$\min_{x} \ \|Ax - b\|_2^2 + \delta\|\Delta x\|_2^2 + \eta\|x\|_2^2.$$

Here:

- $\delta \geq 0$ sets how strongly we penalize roughness (variation/curvature);
- $\eta \geq 0$ sets how strongly we penalize overall magnitude.

Tuning δ, η moves you along a tradeoff surface: better fit typically requires allowing either larger magnitude, or more oscillation, or both.

ℓ_1-norm regularization: sparsity heuristic

Regularization does not have to be quadratic. Using an ℓ_1 term often promotes sparsity:

$$\min_x \ \|Ax - b\|_2^2 + \gamma \|x\|_1, \qquad \gamma > 0. \tag{6.11}$$

Why this is viewed as a sparsity heuristic:

- The direct combinatorial problem "fit well and use only k nonzeros" can be written as $\min \|Ax - b\|_2^2$ subject to $\mathrm{card}(x) \le k$, but that requires searching over $\binom{n}{k}$ sparsity patterns in general.
- The ℓ_1 penalty is convex and cheap to optimize, yet it often returns solutions with many zeros.

In practice one varies γ until the solution has the desired sparsity level, then (if needed) refits on the selected support.