AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

April 23, 2026

Outline

Sparsity is Everywhere in ML

Feature selection: in modern genomics, a study measures gene expression for $p \approx 20000$ genes and $n = 100$ patients.

Goal: predict disease status $y_i \in \mathbb{R}$ from expression $x_i \in \mathbb{R}^p$.

Ordinary Least Squares (OLS):

$$y_i = x_i^\top \beta + \varepsilon_i, \qquad y = X\beta + \varepsilon, \qquad X \in \mathbb{R}^{n \times p}.$$

OLS problem $\min \|y - X\beta\|^2$ has infinitely many solutions when $n < p$.

Observation: only a handful of genes actually matter. We want $\|\beta\|_0 \leq s$ for some small s. But this is NP-hard.

LASSO:

$$\min_{\beta \in \mathbb{R}^p} \ \frac{1}{2} \|y - X\beta\|^2 + \lambda \|\beta\|_1 .$$

Sparsity is Everywhere in ML

Compressed sensing: classical theory says to reconstruct a bandlimited signal, we must sample at twice the bandwidth.

But natural signals, such as images or audio, are sparse in a suitable basis. Can we sample proportional to the sparsity, not the bandwidth?

Formulation:

- Unknown signal $x \in \mathbb{R}^n$ that is s-sparse: $\|x\|_0 \leq s$, with $s \ll n$.
- Take m linear measurements: $y = Ax \in \mathbb{R}^m$, with $m \ll n$.
- Recovery is ill-posed: infinitely many x satisfy $Ax = y$.

Basis pursuit:

$$\min_{x \in \mathbb{R}^n} \|x\|_1 \quad \text{s.t.} \quad Ax = y.$$

[Candes, 2008] shows that if A satisfies some RIP condition, basis pursuit recovers x exactly.

Sparsity is Everywhere in ML

Robust PCA: A surveillance camera records T frames of a scene, and we vectorize each frame and stack them into a matrix $M \in \mathbb{R}^{p \times T}$.

Goal: separate background and moving objects

Question: why not do standard PCA, i.e. truncated SVD?

Observation: background is almost unchanged across frame and moving object only change a few pixels each frame.

Ideal Formulation:

$$\min_{L,S} \ \operatorname{rank}(L) + \lambda \|S\|_0 \quad \text{s.t.} \quad L + S = M.$$

Double Relaxation: [Candès et al., 2011]

$$\min_{L,S} \ \|L\|_* + \lambda \|S\|_1 \quad \text{s.t.} \quad L + S = M.$$

Problem Formulation

Generic sparse optimization:

$$\min_{x \in \mathbb{R}^d} \ f(x) \quad \text{s.t.} \quad \|x\|_0 \le s,$$

where sparsity is measured by $\|x\|_0 := \#\{i : x_i \neq 0\}$.

Bad news:

- this problem is NP-hard in general.
- the ℓ_0-norm function is discontinuous and nonconvex.

Easy case: if $f(x) = \frac{1}{2}\|x - a\|_2^2$, then the optimal x is hard-thresholding, i.e., keep the top-s entries of a in magnitude and zero out others.

Two general routes:

- Convex relaxation: replace $\|x\|_0$ by the ℓ_1-norm (not a norm) $\|x\|_1$.
- Nonconvex direct: hard-thresholding, iterative hard-thresholding, or orthogonal matching pursuit.

ℓ_1-Norm as Convex Envelope of ℓ_0-Norm

The ℓ_0-norm is nonconvex and discontinuous. We want a convex surrogate that promotes sparsity. Candidate: the ℓ_1-norm.

Theorem

On the ℓ_∞ unit ball $\{x : \|x\|_\infty \leq 1\}$, the ℓ_1 norm $\|x\|_1$ is the convex envelope of $\|x\|_0$, i.e., the largest convex function that lower bounds $\|\cdot\|_0$.

Sanity check: on the unit ball, every entry satisfies $|x_i| \leq 1$, so

$$\|x\|_1 = \sum_i |x_i| \;\leq\; \sum_{i:x_i \neq 0} 1 = \|x\|_0 \,.$$

The ℓ_1 norm is a lower bound on ℓ_0. The theorem says it is the tightest such convex lower bound.

Consequence: replacing $\|\cdot\|_0$ by $\|\cdot\|_1$ gives the tightest convex relaxation. This is the foundation of LASSO, basis pursuit, etc.

A Closer Look at LASSO

LASSO: given $A \in \mathbb{R}^{n \times d}$ with columns a_i and $b \in \mathbb{R}^n$,

$$\min_{\beta \in \mathbb{R}^d} \; F(\beta) := \frac{1}{2} \|A\beta - b\|^2 + \lambda \|\beta\|_1 .$$

Subdifferential:

$$\partial \|\beta\|_1 = \{g \in \mathbb{R}^d \; : \; g_i = \operatorname{sign}(\beta_i) \text{ if } \beta_i \neq 0, \;\; g_i \in [-1, 1] \text{ if } \beta_i = 0\}.$$

Optimality condition: $\beta^\star$ is optimal iff $0 \in \partial F(\beta^\star)$, i.e., for each i:

$$a_i^\top (A\beta^\star - b) \in -\lambda \, \partial |\beta_i^\star| \iff \begin{cases} a_i^\top (A\beta^\star - b) = -\lambda \operatorname{sign}(\beta_i^\star) & \text{if } \beta_i^\star \neq 0, \\ |a_i^\top (A\beta^\star - b)| \leq \lambda & \text{if } \beta_i^\star = 0. \end{cases}$$

Observation: a coordinate is zero at optimum iff the residual correlation $|a_i^\top r^\star|$ is below the threshold λ. This leads to Coordinate Descent.

Outline

A New Algorithm

You already know three first-order methods that solve LASSO.

- ISTA/FISTA: proximal gradient with soft-thresholding;
- ADMM: variable splitting, ridge step with soft-thresholding;
- Subgradient Method: works theoretically but slow in practice.

Which is the best algorithm for solving LASSO?

- In practice, the fastest solver for LASSO is none of these. It is coordinate descent!

The R package `glmnet` [Friedman et al., 2010] is built on coordinate descent and solves a LASSO problem with $d \approx 10^5$ features in seconds.

Why is coordinate descent so good at sparse problems?

Coordinate Descent

Problem: $\min_{x \in \mathbb{R}^d} \; f(x)$.

Idea: at each iteration, minimize f along one coordinate direction.

Coordinate Descent (CD): Initialize $x^{(1)} \in \mathbb{R}^d$. For $k = 1, 2, \ldots, K$,

- Choose a coordinate $i_k \in \{1, \ldots, d\}$,
- Update

$$x_j^{(k+1)} = \begin{cases} \arg\min_{\xi \in \mathbb{R}} \; f(x_1^{(k)}, \ldots, \xi, \ldots, x_d^{(k)}) & \text{if } j = i_k, \\ x_j^{(k)} & \text{if } j \neq i_k. \end{cases}$$

Coordinate Selection:

- Cyclic: sweep through $1, 2, \ldots, d, 1, 2, \ldots$ in order.
- Randomized: sample $i_k \sim P$ for some distribution P.
- Greedy: pick the coordinate with the largest gradient component.

Coordinate Gradient Descent

Issue: $\min_\xi f(x_1, \ldots, \xi, \ldots, x_d)$ may still be too hard to solve efficiently.

Fix: inexactly solve by a gradient step along the chosen coordinate.

Coordinate Gradient Descent (CGD): Initialize $x^{(1)} \in \mathbb{R}^d$, step size $\eta > 0$. For $k = 1, 2, \ldots, K$,

- Choose a coordinate $i_k \in \{1, \ldots, d\}$,
- Update

$$x_j^{(k+1)} = \begin{cases} x_j^{(k)} - \eta \nabla_j f(x^{(k)}) & \text{if } j = i_k, \\ x_j^{(k)} & \text{if } j \neq i_k. \end{cases}$$

CD vs CGD:

- Use CD when the univariate min has a closed form (e.g., LASSO).
- Use CGD when it does not (e.g., ℓ_1-regularized logistic regression).

Why CD is Made for LASSO?

Separability. The regularizer $\|\beta\|_1 = \sum_i |\beta_i|$ decomposes across coordinates. Each subproblem becomes a clean univariate ℓ_1 problem.

Closed-form. The univariate subproblem is cheap. Just need one dot product and one soft-thresholding.

Sparsity. Soft-thresholding returns 0 whenever the residual below λ. Cycle only on nonzero coordinates, periodically check optimality on the rest. Per-iteration cost scales with the number of nonzeros, not d.

Adaptivity. If the Lipschitz constants ℓ_i of $\nabla_j f$ are much smaller than the Lipschitz constant L of ∇f, CD can take much larger step size than GD.

CD Update for LASSO

Fix all coordinates except β_i. The univariate Lasso subproblem is

$$\min_{\beta_i \in \mathbb{R}} \ \frac{1}{2} \left\| a_i \beta_i - r_{-i} \right\|^2 + \lambda |\beta_i|, \qquad r_{-i} := b - \sum_{j \neq i} a_j \beta_j.$$

Expand:

$$\frac{1}{2} \left\| a_i \beta_i - r_{-i} \right\|^2 = \frac{1}{2} \left\| a_i \right\|^2 \beta_i^2 - (a_i^\top r_{-i}) \beta_i + \text{const.}$$

Subproblem:

$$\min_{\beta_i \in \mathbb{R}} \ \frac{\left\| a_i \right\|^2}{2} \beta_i^2 - (a_i^\top r_{-i}) \beta_i + \lambda |\beta_i|.$$

Optimality: $0 \in \left\| a_i \right\|^2 \beta_i - a_i^\top r_{-i} + \lambda \, \partial |\beta_i|$.

Soft-thresholding Update:

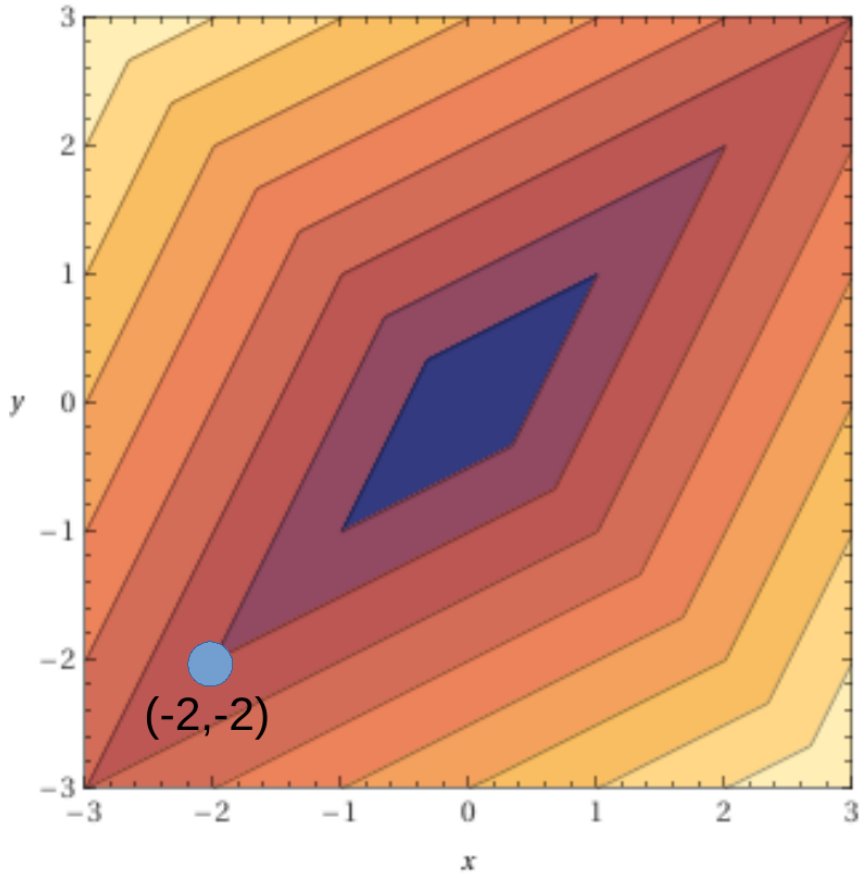
$$\beta_i^+ \ = \ \frac{1}{\left\| a_i \right\|^2} \mathcal{S}_\lambda(a_i^\top r_{-i}), \qquad \mathcal{S}_\lambda(u) := \operatorname{sign}(u) \, \max(|u| - \lambda, \, 0).$$

When CD Fails

Counterexample: consider

$$f(x, y) = |x + y| + 3|x - y|.$$

Contour:



When CD Fails

Counterexample: consider

$$f(x, y) = |x + y| + 3|x - y|.$$

Start at $(-2, -2)$**.** CD gets stuck.

- Fix $y = -2$, minimize over x: solution $x^+ = -2$.
- Fix $x = -2$, minimize over y: solution $y^+ = -2$.

But $(-2, -2)$ is not the global minimum. The global min is $(0, 0)$.

Why CD fails? CD can only move along d coordinate axes. If none of them is a descent direction, CD stops.

Conclusion: CD needs smooth + separable nonsmooth. Lasso satisfies this; $|x + y| + 3|x - y|$ does not.

Convergence of Cyclic CD

Cyclic CD: sweep $i_k = 1, 2, \ldots, d, 1, 2, \ldots$ in order.

Convergence of Cyclic CD

Suppose f is convex, ∇f is continuous, and for every x and every i, the function $\xi \mapsto f(x_{-i}, \xi)$ has a unique minimizer and is monotone between x_i and that minimizer. Then cyclic CD converges to a global minimizer.

Why unique minimizer?

- Without uniqueness, CD can stall at a non-stationary point, exactly what happened in the $|x + y| + 3|x - y|$ example.
- Lasso satisfies the uniqueness condition when A has no zero columns: the univariate subproblem is strictly convex in the quadratic part and the ℓ_1 term is convex.

Proof Sketch

Define the intermediate iterates

$$z_i^{(k)} := (x_1^{(k+1)}, \ldots, x_i^{(k+1)}, x_{i+1}^{(k)}, \ldots, x_d^{(k)}).$$

Monotonicity. By construction,

$$f(x^{(k)}) \geq f(z_1^{(k)}) \geq \cdots \geq f(z_{d-1}^{(k)}) = f(x^{(k+1)}).$$

So $f(x^{(k)}) \to f(\bar{x})$ for any limit point $\bar{x}$.

Contradiction. Suppose $z_1^{(k_j)} - x^{(k_j)} \not\to 0$. Set

$$s_1^{(k_j)} = \frac{z_1^{(k_j)} - x^{(k_j)}}{\left\| z_1^{(k_j)} - x^{(k_j)} \right\|}.$$

A subsequence converges to some $\bar{s}_1 \neq 0$. For any $\varepsilon \in [0, 1]$,

$$f(z_1^{(k_j)}) \leq f(x^{(k_j)} + \varepsilon \bar{\gamma} s_1^{(k_j)}) \leq f(x^{(k_j)}).$$

Taking $j \to \infty$: $f(\bar{x}) = f(\bar{x} + \varepsilon \bar{\gamma} \bar{s}_1)$ for all $\varepsilon \in [0, 1]$, contradicting uniqueness of the coordinate-wise minimizer.

Optimality. We have $\nabla_i f(\bar{x}) = 0$ for all i, so $\nabla f(\bar{x}) = 0$.

Randomized Coordinate Gradient Descent

Idea: sample the coordinate at random. RCGD is a special case of SGD.

Randomized CGD. For $t = 1, \ldots, T$:

- Sample $i_t \sim P_t$ from a distribution on $[d]$.
- Update $x^{t+1} = x^t - \eta_t \nabla_{i_t} f(x^t)\, e_{i_t}$.

Unbiasedness. For uniform P_t, let $\tilde{g}(x) := d \cdot \nabla_{i_t} f(x)\, e_{i_t}$. Then

$$\mathbb{E}[\tilde{g}(x)] = \sum_{i=1}^{d} \frac{1}{d} \cdot d \nabla_i f(x)\, e_i = \nabla f(x).$$

Variance.

$$\mathbb{E}\left[\|\tilde{g}(x)\|^2\right] = d\, \|\nabla f(x)\|^2 .$$

RCGD vs. vanilla SGD: variance shrinks as $\nabla f(x_t) \to 0$. No need for diminishing step sizes or variance reduction.

Importance Sampling

Coordinate-wise Smoothness: $f : \mathbb{R}^d \to \mathbb{R}$ is coordinate-wise smooth if there exist $\ell_1, \ldots, \ell_d > 0$ with

$$|\nabla_i f(x + ue_i) - \nabla_i f(x)| \leq \ell_i |u|, \quad \forall\, i, x, u.$$

Relation to global smoothness: if f is C^2,

$$\lambda_{\max}(\nabla^2 f(x)) \leq \sum_{i=1}^{d} \ell_i.$$

Typically $L \leq \sum_i \ell_i$, often much smaller.

RCGD with importance sampling:

$$x^{t+1} = x^t - \frac{1}{\ell_{i_t}} \nabla_{i_t} f(x^t)\, e_{i_t}, \text{ with } P(i_t = i) = \frac{\ell_i}{\sum_{j=1}^{d} \ell_j}.$$

Idea. Put more effort into directions with larger curvature ℓ_i. The step size $1/\ell_{i_t}$ also adapts automatically to the coordinate-wise conditioning.

Convergence Rates of RCGD

Convergence of RCGD under convexity

Let f be convex and coordinate-wise smooth. Let x_T be generated by RCGD with importance sampling. Then

$$\mathbb{E}[f(x_T)] - f(x^\star) \leq \frac{2R(x_1)^2 \sum_{i=1}^d \ell_i}{T-1},$$

where $R(x_1) := \sup\{\|x - x^\star\| : f(x) \leq f(x_1)\}$.

Convergence of RCGD under strong convexity

If f is additionally μ-strongly convex, then

$$\mathbb{E}[f(x_T)] - f(x^\star) \leq \left(1 - \frac{1}{\kappa}\right)^T (f(x_1) - f(x^\star)),$$

where $\kappa = \frac{1}{\mu}\sum_{i=1}^d \ell_i$.

Outline

Sparse PCA

PCA looks for the direction along which the data varies the most.

- For $u \in \mathbb{R}^p$, $\|u\| = 1$, the projection of data x onto u is $u^\top x$.
- If the data covariance is S, then the variance after projection is
$$\mathrm{Var}(u^\top x) = u^\top S u$$
- The first principal component solves
$$\max_{\|u\|_2=1} \; u^\top S u.$$
- For k components, this becomes
$$\max_{U \in \mathbb{R}^{p\times k}} \; \mathrm{tr}(U^\top S U) \qquad \text{s.t.} \qquad U^\top U = I_k.$$

Intuition:

- Large variance means the projected data are well spread out.
- Small variance means the projected data almost collapse.
- Keeping large-variance directions preserves more information.

Sparse PCA

Naive Sparse PCA:

$$\max_{U} \ \mathrm{tr}(U^\top S U) \qquad \text{s.t.} \qquad U^\top U = I_k, \ \|U\|_0 \le s.$$

This is nonconvex and NP-hard.

Set $X := UU^\top$, then $\mathrm{tr}(U^\top S U) = \mathrm{tr}(SX)$, and X is a rank-k orthogonal projector:

$$X = X^\top, \qquad X \succeq 0, \qquad X^2 = X, \qquad \mathrm{tr}(X) = k.$$

The nonconvex part is $X^2 = X$, i.e., eigenvalues of X are in $\{0, 1\}$.

Fantope: convex hull of this projector set

$$\mathcal{F}_k := \{X \in \mathbb{R}^{p \times p} : 0 \preceq X \preceq I, \ \mathrm{tr}(X) = k\}.$$

Sparse PCA via ADMM

Convex Relaxation of Sparse PCA:

$$\max_{X} \ \langle S, X \rangle - \lambda \|X\|_1 \quad \text{s.t.} \quad X \in \mathcal{F}_k.$$

ADMM splitting. Introduce $X = Z$:

$$\max_{X,Z} \ \langle S, X \rangle - \lambda \|Z\|_1 \quad \text{s.t.} \quad X \in \mathcal{F}_k, \ \ X = Z.$$

ADMM updates.

- X-update: Fantope projection, i.e., eigendecompose $Z^k - U^k + S/\rho$, clip eigenvalues to $[0, 1]$ with constraint $\sum \sigma_i = k$.
- Z-update: entrywise soft-threshold $\mathcal{S}_{\lambda/\rho}(X^{k+1} + U^k)$.
- Dual update: $U^{k+1} = U^k + X^{k+1} - Z^{k+1}$.

Caveat: the bottleneck is eigen-decomposition $O(p^3)$.

Robust PCA via ADMM

Recall the Robust PCA formulation from the motivation:

$$\min_{L,S} \ \|L\|_* + \lambda \|S\|_1 \quad \text{s.t.} \quad L + S = M.$$

Why not CD or ISTA?

- Two nonsmooth regularizers coupled by a linear constraint.
- Nuclear norm has no coordinate-wise structure.

ADMM updates:

$$L^{k+1} = \mathrm{SVT}_{1/\rho}(M - S^k + Y^k/\rho) \qquad \text{(singular value threshold)}$$

$$S^{k+1} = \mathcal{S}_{\lambda/\rho}(M - L^{k+1} + Y^k/\rho) \qquad \text{(soft-threshold)}$$

$$Y^{k+1} = Y^k + \rho(M - L^{k+1} - S^{k+1}) \qquad \text{(dual update)}$$

Caveat: the bottleneck is SVT, roughly the same as eigen-decomposition.

Graphical LASSO

Model: use a graph to describe dependence among random variables.

- each node is a variable $X_1, \ldots, X_p$;
- an edge means a direct dependence;
- no edge means a conditional independence.

For Gaussian data

$$x_1, \ldots, x_n \sim \mathcal{N}(0, \Sigma), \qquad C := \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top,$$

the relevant parameter is the precision matrix $\Theta := \Sigma^{-1}$.

Precision matrix: for Gaussian distributions,

$$\Theta_{ij} = 0 \quad \Longleftrightarrow \quad X_i \perp X_j \mid X_{-\{i,j\}}.$$

So zeros in Θ mean missing edges in the conditional independence graph.

Goal: estimate a sparse precision matrix, hence a sparse graph.

Graphical LASSO and ADMM

For $x \sim \mathcal{N}(0, \Sigma)$ with $\Theta = \Sigma^{-1}$, the Gaussian density is

$$p(x; \Theta) \propto \det(\Theta)^{1/2} \exp\left(-\frac{1}{2} x^\top \Theta x\right).$$

So for samples $x_1, \ldots, x_n$, the negative log-likelihood is equivalent to

$$-\ell(\Theta) \equiv -\log\det\Theta + \langle C, \Theta \rangle .$$

To encourage a sparse graph, add an ℓ_1 penalty:

$$\min_{\Theta \succ 0} \; -\log\det\Theta + \langle C, \Theta \rangle + \lambda \|\Theta\|_1 .$$

ADMM splitting: introduce Z with $\Theta = Z$,

$$\min_{\Theta \succ 0,\, Z} \; -\log\det\Theta + \langle C, \Theta \rangle + \lambda \|Z\|_1 \qquad \text{s.t.} \qquad \Theta - Z = 0.$$

Graphical LASSO and ADMM

ADMM updates:

- Θ-update: eigen-decomposition of $\rho(Z^k - U^k) - C$, then solve a scalar quadratic for each eigenvalue;
- Z-update: entrywise soft-thresholding $Z^{k+1} = \mathcal{S}_{\lambda/\rho}(\Theta^{k+1} + U^k)$;
- U-update: dual ascent.

Θ-update:

- KKT stationarity gives
$$-\Theta^{-1} + C + \rho(\Theta - Z^k + U^k) = 0 \iff \rho\Theta - \Theta^{-1} = \rho(Z^k - U^k) - C.$$
- Eigen-decomposition gives
$$\rho(Z^k - U^k) - C = Q\,\mathrm{diag}(d_i)Q^\top,$$
- Solve the 1-d quadratic equation
$$\Theta^{k+1} = Q\,\mathrm{diag}(t_i)Q^\top, \qquad \rho t_i^2 - d_i t_i - 1 = 0.$$

Outline

Block CD for Group LASSO

Formulation: given groups $B_1, \ldots, B_m$ partitioning $[d]$,

$$\min_{\beta} \ \frac{1}{2} \|A\beta - b\|^2 + \lambda \sum_{g=1}^{m} \|\beta_{B_g}\|_2 .$$

Meaning:

- the regularizer is block-separable;
- promotes group sparsity, i.e., β_{B_g} either entirely zero or nonzero;

Fix all blocks except β_{B_g}. Define the partial residual

$$r_{-g} := b - \sum_{h \neq g} A_{B_h} \beta_{B_h}.$$

Then the block subproblem becomes

$$\min_{\beta_{B_g}} \ \frac{1}{2} \|A_{B_g}\beta_{B_g} - r_{-g}\|^2 + \lambda \|\beta_{B_g}\|_2 .$$

Closed-Form Block Update

Assume $A_{B_g}^\top A_{B_g} = cI$, e.g., columns in block B_g are orthogonal with equal norm. Let $v_g := A_{B_g}^\top r_{-g}$, then

$$\boxed{\beta_{B_g}^+ = \frac{1}{c}\left(1 - \frac{\lambda}{\|v_g\|_2}\right)_+ v_g}.$$

Block soft-thresholding

- If $\|v_g\|_2 \leq \lambda$, then $\beta_{B_g}^+ = 0$, so the entire block is removed.
- If $\|v_g\|_2 > \lambda$, then $\beta_{B_g}^+$ keeps the direction of v_g and shrinks its magnitude.

This is the vector analogue of scalar soft-thresholding.

If $A_{B_g}^\top A_{B_g} = cI$ doesn't hold, then no close form solution in general.

Iterative Hard Thresholding (IHT)

Setting. $\min_x f(x)$ s.t. $\|x\|_0 \leq s$, with f smooth convex.

IHT Algorithm:

- Initialize $x^0 = 0$.
- For $k = 0, 1, 2, \ldots,$

$$x^{k+1} = H_s(x^k - \eta \nabla f(x^k)),$$

where $H_s(v)$ keeps only the top-s entries of v in magnitude.

Interpretation: Projected gradient descent onto the sparse set

$$H_s(v) \;=\; \arg \min_{\|x\|_0 \leq s} \|x - v\|^2 .$$

The sparse set is not convex, but the projection has a trivial closed form.

Convergence: under RIP condition, IHT converges linearly to the true sparse signal [Blumensath and Davies, 2009].

Nonconvex Regularizer: SCAD

For sparse estimation (scalar case), LASSO solves

$$\min_{\beta} \ \frac{1}{2}(\beta - z)^2 + \lambda|\beta|,$$

The solution is soft-thresholding:

$$\hat{\beta} = \mathcal{S}_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+.$$

So:

- if $|z| \leq \lambda$, then $\hat{\beta} = 0$: this creates sparsity;
- if $|z| > \lambda$, then

  $$\hat{\beta} = z - \lambda \ \text{sign}(z),$$

  so the coefficient survives, but is still pulled toward zero.

Consequence:

- Near zero, this is desirable.
- For large true coefficients, this creates estimation bias.

Nonconvex Regularizer: SCAD/MCP

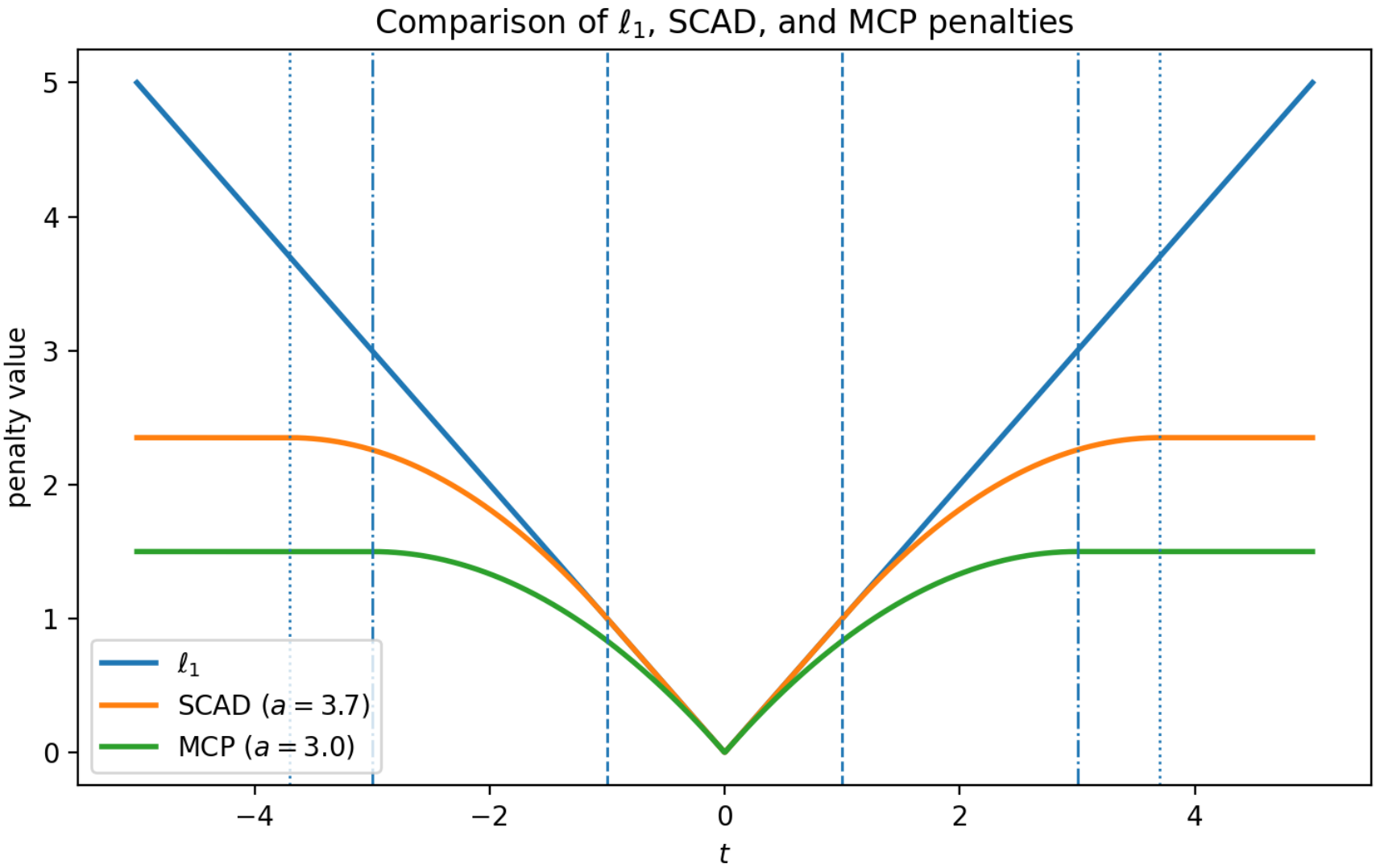
SCAD penalty: with $a > 2$, typically $a = 3.7$ [Fan and Li, 2001],

$$p_\lambda^{\mathrm{SCAD}}(t) = \begin{cases} \lambda|t|, & |t| \le \lambda, \\ -\dfrac{t^2 - 2a\lambda|t| + \lambda^2}{2(a-1)}, & \lambda < |t| \le a\lambda, \\ \dfrac{(a+1)\lambda^2}{2}, & |t| > a\lambda, \end{cases}$$

MCP penalty: with $a > 1$, typically $a = 3$ [Zhang, 2010],

$$p_\lambda^{\mathrm{MCP}}(t) = \begin{cases} \lambda|t| - \dfrac{t^2}{2a} & \text{if } |t| \le a\lambda, \\ \dfrac{a\lambda^2}{2} & \text{if } |t| > a\lambda, \end{cases}$$

Nonconvex Regularizer: SCAD/MCP



Nonconvex Regularizer: SCAD/MCP

For a scalar penalty $p_\lambda(|x|)$, the proximal step solves nonconvex problem

$$\min_x \ \frac{1}{2}(x-u)^2 + p_\lambda(|x|).$$

MCP thresholding rule:

$$T_{\mathrm{MCP}}(u) = \begin{cases} 0, & |u| \le \lambda, \\ \dfrac{a}{a-1}(|u|-\lambda)\ \mathrm{sign}(u), & \lambda < |u| \le a\lambda, \\ u, & |u| > a\lambda. \end{cases}$$

Interpretation:

- small $|u|$: set to zero;
- medium $|u|$: shrink, but less than soft-thresholding;
- large $|u|$: keep unchanged.

So MCP keeps sparsity near zero, but removes bias for large signals.

Nonconvex Regularizer: SCAD/MCP

SCAD thresholding rule:

$$T_{\mathrm{SCAD}}(u) = \begin{cases} 0, & |u| \le \lambda, \\ \mathrm{sign}(u)(|u| - \lambda), & \lambda < |u| \le 2\lambda, \\ \dfrac{(a-1)u - a\lambda\ \mathrm{sign}(u)}{a-2}, & 2\lambda < |u| \le a\lambda, \\ u, & |u| > a\lambda. \end{cases}$$

SCAD vs. MCP:

- both are exactly zero for small $|u|$;
- both reduce shrinkage for medium $|u|$;
- both return u for large $|u|$, so large coefficients are nearly unbiased.
- SCAD has a smoother transition region; MCP uses a simpler two-stage shrinkage rule.

Wrap-up and Next Week

Week 12: Key takeaways

- general sparse optimization setup, ℓ_1-relaxation;
- coordinate descent (CD), CGD, RGCD, importance sampling;
- sparse PCA, Robust PCA, graphical LASSO;
- block CD for group LASSO, IHT, nonconvex regularizer SCAD/MCP.

Thank you for your attendance!

Questions?

Next week (Week 13): min-max problem and bilevel optimization

References I


Blumensath, T. and Davies, M. E. (2009).
Iterative hard thresholding for compressed sensing.
Applied and computational harmonic analysis, 27(3):265–274.


Candes, E. J. (2008).
The restricted isometry property and its implications for compressed sensing.
Comptes rendus mathematique, 346(9-10):589–592.


Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011).
Robust principal component analysis?
Journal of the ACM (JACM), 58(3):1–37.


Fan, J. and Li, R. (2001).
Variable selection via nonconcave penalized likelihood and its oracle properties.
Journal of the American statistical Association, 96(456):1348–1360.

References II


Friedman, J. H., Hastie, T., and Tibshirani, R. (2010).
Regularization paths for generalized linear models via coordinate descent.
Journal of statistical software, 33:1–22.


Zhang, C.-H. (2010).
Nearly unbiased variable selection under minimax concave penalty.
The Annals of Statistics, 38(2):894–942.