LASSO Notes

Justin_Ds

LASSO

LASSO (Least Absolute Shrinkage and Selection Operator) estimates $\boldsymbol{\beta}^\star$ by solving the following convex optimization problem:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where:

- $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$: residual sum of squares (RSS). 残差平方和
- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{d} |\beta_j|$: ℓ_1-norm penalty.
- $\lambda > 0$: tuning parameter that controls the trade-off between goodness of fit and sparsity. 拟合"好坏"和"稀疏性" balance
- Interpretation as MAP estimator with a Laplace prior on $\boldsymbol{\beta}^\star$
- Questions:
  - How to compute LASSO estimate?
  - What is the statistical properties of LASSO?

Not prof

Chatgpt

LASSO 贝叶斯解释

How to compute LASSO: proximal gradient method

LASSO 可看作 MAP of prior func of $p(\beta_i^*) = \frac{\tau}{2} e^{-\tau|\beta^*|}$

$\beta_i^* \sim$ density of Laplace distribution

$\tau = \frac{\lambda}{2\sigma}$

A more general class of convex optimization

Consider unconstrained convex optimization problem of the form 一类凸优化问题

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}) := f(\boldsymbol{x}) + h(\boldsymbol{x})$$

where

- $f(\boldsymbol{x})$: a differentiable, convex function 可微 凸函数
- $h(\boldsymbol{x})$: a convex, potentially non-differentiable function (e.g., ℓ_1-norm). 不确定可微性
- Example: LASSO can be viewed as taking

$$f(\boldsymbol{x}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2, \qquad h(\boldsymbol{x}) = \lambda\|\boldsymbol{\beta}\|_1.$$

可微凸　　凸但不可微

Issue: gradient descent (GD) does not work (due to non-smoothness)

在0处不可微，不能用梯度下降

处理非光滑函数，对h(x)通过近端映射(proximal mapping)

A Proximal View of Gradient Descent

近端梯度法

- To motivate proximal gradient methods, we first revisit gradient descent for $\min_{\boldsymbol{x}} f(\boldsymbol{x})$, where $f(\cdot)$ is convex and smooth
- Gradient descent update: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t)$ 对f以梯度下降
- This is equivalent to

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}} \left\{ \underbrace{f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x} - \boldsymbol{x}_t \rangle}_{\text{first-order approximation at } \boldsymbol{x}_t} + \underbrace{\frac{1}{2\eta} \|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2}_{\text{proximal term}} \right\}$$

近端项

一阶近似项

- Heuristics: search for $\boldsymbol{x}_{t+1}$ that 在x_t处一阶泰勒展开(线性近似)
  - aim to minimize $f(\cdot)$ (through minimizing first-order approximation)
  - remains close to $\boldsymbol{x}_t$ such that first-order approximation at $\boldsymbol{x}_t$ is valid (enforced by proximal term)
- Benefit: minimizing a quadratic function, admits simple solution (i.e., GD)

加入对步长的限制，且最小化f(·)

Proximal gradient method: algorithm

Consider an iterative algorithm: starting from $\boldsymbol{x}_t$, update

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}} \left\{ \underbrace{f(\boldsymbol{x}_t) + \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x} - \boldsymbol{x}_t \rangle}_{\text{first-order approximation at } \boldsymbol{x}_t} + h(\boldsymbol{x}) + \underbrace{\frac{1}{2\eta} \|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2}_{\text{proximal term}} \right\}$$

- Define proximal operator

$$\text{prox}_h(\boldsymbol{v}) = \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} \left\{ h(\boldsymbol{x}) + \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{v}\|_2^2 \right\}$$

- If this proximal operator is easy to compute, then we can express

$$\boldsymbol{x}_{t+1} = \text{prox}_{\eta h}(\boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t))$$

- alternates between gradient updates on f and proximal minimization on h

Proximal view of GD: (活泼梯度下降法)

proof: $l(x) = f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta}\|x - x_t\|_2^2$

$= \frac{1}{2\eta} x^T x - \frac{1}{\eta}\langle x, x_t \rangle + \langle \nabla f(x_t), x \rangle + \text{const}$

$\nabla l(x) = \frac{1}{\eta} x - \frac{1}{\eta} x_t + \nabla f(x_t)$

$= \frac{1}{\eta}(x - x_t + \eta \cdot \nabla f(x_t)) = 0$

$\Rightarrow x = x_t - \eta \nabla f(x_t)$

x here is minimizer the $l(x)$

Proximal gradient method [近端梯度法☆]

$l(x) = f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + h(x) + \frac{1}{2\eta}\|x - x_t\|_2^2$

$= h(x) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta}\|x - x_t\|_2^2 + \text{const}$

$= h(x) + \frac{1}{2\eta}\|x - x_t + \eta \nabla f(x_t)\|_2^2 - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2 + \text{const}$

$= h(x) + \frac{1}{2\eta}\|x - x_t + \eta \nabla f(x_t)\|_2^2 + \text{const}$

Define the proximal operator: $\mathbb{R}^d \to \mathbb{R}^d$

$\text{prox}_h(v) = \underset{x \in \mathbb{R}^d}{\text{argmin}} \left\{ h(x) + \frac{1}{2}\|x - v\|_2^2 \right\}$

从优化问题的角度看，prox算子是对函数 h 与一个二次惩罚项的联合最小化问题的解。当 η 固定时，这一问题是强凸的，且在一般的适度条件下有唯一解。

几何直观上：$\text{prox}_{\eta h}(y)$ 是在点 y 附近寻找一个点 x，该点在尽量靠近 y 的同时，使 $h(x)$ 的值更小，从而达到折中。这正对应了不可微项的"修正"作用。

如果 $h(x)$ 是如 $\lambda\|x\|_1$ 这样的 L1 正则项，则 $\text{prox}_{\eta h}$ 的闭式解为软阈值算子，从而产生稀疏解。这也是 Proximal Gradient 被广泛用于 Lasso 等稀疏优化问题的原因。

Notice that

$l(x) \propto \frac{1}{2}\|x - (x_t - \eta \nabla f(x_t))\|_2^2 + \eta h(x) + \text{const}$

The minimizer of $l(x)$ is given by

$\text{prox}_{\eta h}(x_t - \eta \cdot \nabla f(x_t))$

Proximal gradient method: properties

Proximal gradient algorithm: for $t = 1, 2, \ldots$

$$\boldsymbol{x}_{t+1} = \mathsf{prox}_{\eta h}(\boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t))$$

(凸且L光滑)

- fast convergence when f is convex and L-smooth: take $\eta = 1/L$,

$$F(\boldsymbol{x}_t) - F^\star \leq \frac{L}{2t}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$$

收敛速度 $O(\frac{1}{t})$

- exponential convergence when f is μ-strongly convex (μ-强凸指数收敛)

$$\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2^2 \leq (1 - \mu/L)^t \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$$

- when $h(\boldsymbol{x}) = 0$ when $\boldsymbol{x} \in \mathcal{A}$ and $h(\boldsymbol{x}) = \infty$ otherwise, this gives the projected gradient descent for $\min_{\boldsymbol{x} \in \mathcal{A}} f(\boldsymbol{x})$:

$$h(x) = \begin{cases} 0, & x \in A \\ \infty, & x \notin A \end{cases}$$

投影梯度下降

$$\boldsymbol{x}_{t+1} = \mathcal{P}_{\mathcal{A}}(\boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t))$$

投影回可行域A

- Recommended reading material: Lecture 5 of the course Large-Scale Optimization for Data Science

Application to LASSO

- LASSO:

$$f(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \qquad \text{and} \qquad h(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$$

- The proximal operator admits closed-form expression 软阈值

$$\text{prox}_h(\boldsymbol{v}) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \frac{1}{2}\|\boldsymbol{\beta} - \boldsymbol{v}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\} = \text{shrink}_\lambda(\boldsymbol{v})$$ 封闭形解

where $\text{shrink}_\lambda(\cdot)$ applies entrywise shrinkage to $\boldsymbol{v}$ towards zero:

$$[\text{shrink}_\lambda(\boldsymbol{v})]_j = \begin{cases} v_j - \lambda, & \text{if } v_j \geq \lambda, \\ v_j + \lambda, & \text{if } v_j \leq -\lambda, \\ 0, & \text{otherwise.} \end{cases}$$

几何含义：当 v_j 不足大，该坐标置零 ⇒ 稀疏；v_j 足够大，v_j 进入更新坐标 β_j

- Proximal gradient algorithm for LASSO:

$$\boldsymbol{\beta}_{t+1} = \text{shrink}_{\eta\lambda}\left(\boldsymbol{\beta}_t - 2\eta\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}_t + 2\eta\boldsymbol{X}^\top\boldsymbol{Y}\right)$$

$\beta_t - \eta\nabla f(\beta_t) = \beta_t - \eta(2X^TX\beta_t - 2X^TY) = \beta_t - 2\eta X^TX\beta_t + 2\eta X^TY$

Solve Lasso: $prox_h(v) = \arg\min_\beta \{\underbrace{\frac{1}{2}\|\beta - v\|_2^2 + \lambda\|\beta\|_1}_{\ell(\beta)}\}$

问题分解为一维优化问题:

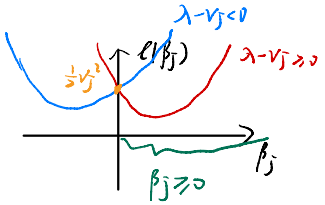
$\ell(\beta) = \sum_{j=1}^d \underbrace{\frac{1}{2}(\beta_j - v_j)^2 + \lambda|\beta_j|}_{\ell_j(\beta_j)} = \sum_{j=1}^d \ell_j(\beta_j)$

To minimize $\ell_j(\beta_j) = \frac{1}{2}(\beta_j - v_j)^2 + \lambda|\beta_j|$

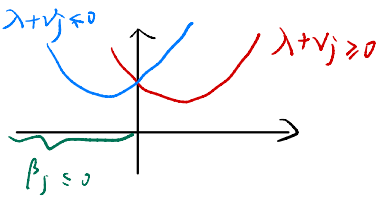
$$= \begin{cases} \frac{1}{2}(\beta_j - v_j)^2 + \lambda\beta_j, & \text{if } \beta_j \geq 0 \\ \frac{1}{2}(\beta_j - v_j)^2 - \lambda\beta_j, & \text{if } \beta_j \leq 0 \end{cases}$$

$$= \begin{cases} \frac{1}{2}\beta_j^2 + (\lambda - v_j)\beta_j + \frac{1}{2}v_j^2, & \text{if } \beta_j \geq 0 \\ \frac{1}{2}\beta_j^2 - (\lambda + v_j)\beta_j + \frac{1}{2}v_j^2, & \text{if } \beta_j \leq 0 \end{cases}$$

Within $\beta_j \geq 0$, $\ell_j(\beta_j)$ is minimized at $\begin{cases} \lambda - v_j & \text{if } \lambda - v_j \geq 0 \\ 0, & \text{if } \lambda - v_j \leq 0 \end{cases}$



Within $\beta_j \leq 0$, $\ell_j(\beta_j)$ is minimized at $\begin{cases} 0, & \text{if } \lambda + v_j \geq 0 \\ \lambda + v_j, & \text{if } \lambda + v_j \leq 0 \end{cases}$



$\hat{\beta}_j = \arg\min_\beta \ell(\beta_j)$

$$\Rightarrow \hat{\beta}_j = \begin{cases} \lambda - v_j, & v_j > \lambda \\ 0, & -\lambda \leq v_j \leq \lambda \\ \lambda + v_j, & v_j < -\lambda \end{cases} = S_\lambda(v_j) = \text{sign}(v_j)\max(|v_j| - \lambda, 0)$$

对非光滑凸函数, 定义次梯度 (Subgradient)

for convex function f: define g is a subgradient at point x, if for $\forall y$,

$$f(y) \geq f(x) + g^T(y - x)$$

Example: $f(x) = |x|$, $\partial f(0) = [-1, 1]$

check: 0 is minimizer of $f \Leftrightarrow 0 \in \partial f(x)$

Pf: "$\Rightarrow$" if x is minimizer, $\forall y \in \mathbb{R}^d$

$$f(y) \geq f(x) = f(x) + \langle 0, y - x\rangle \Rightarrow 0 \in \partial f(x)$$

"$\Leftarrow$" if $0 \in \partial f(x)$, $\forall y \in \mathbb{R}^d$, $f(y) \geq f(x) + \langle 0, y - x\rangle = f(x)$ □

Lasso 回归问题可写为:

$$\min_{\beta \in \mathbb{R}^d} f(\beta) + h(\beta)$$

where $f(\beta) = \|Y - X\beta\|_2^2 \triangleq RSS$

$h(\beta) = \lambda\|\beta\|_1 = \lambda\sum_{j=1}^d |\beta_j|$, L_1-N

if $\ell(\beta) = \sum \frac{1}{2\eta}(\beta_j - v_j)^2 + \lambda|\beta_j|$

$\ell_j(\beta_j) = \frac{1}{2\eta}(\beta_j^2 - 2v_j\beta_j + v_j^2) \pm \lambda\beta_j$

$$= \begin{cases} \frac{1}{2\eta}\beta_j^2 - (\lambda - \frac{v_j}{\eta})\beta_j + \frac{1}{2\eta}v_j^2 \\ \frac{1}{2\eta}\beta_j^2 - (\lambda + \frac{v_j}{\eta})\beta_j + \frac{1}{2\eta}v_j \end{cases}$$

$v_j - \eta\lambda$

Statistical properties of LASSO

Setup

LASSO:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\},$$

- Independent, sub-Gaussian noise $\|\varepsilon_i\|_{\psi_2} \le \sigma$ 独立次高斯噪声
- Sparsity: $n \gg s\log d$ 稀疏性，记 $S=\{j:\beta_j^*\neq 0\}$，$|S|=s \ll d$
- Theory-informed tuning parameter selection: 真参数支撑集

$$\lambda \asymp \sigma\sqrt{n\log d}$$

为了良好的特征选择能力和估计精度

- Question: 成比例
  - Does LASSO recover the support of $\beta^\star$? model selection consistency
  - Does LASSO provide reliable estimate for $\beta^\star$? $\sqrt{\frac{s\log d}{n}}$ 量级

Optimality condition

The optimality condition for unconstrained convex optimization

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x})$$

- if f is smooth: $\nabla f(\widehat{\boldsymbol{x}}) = \mathbf{0}$
- in general (when f might not be smooth): $\mathbf{0} \in \partial f(\widehat{\boldsymbol{x}})$

Here $\partial f(\boldsymbol{x}) \subseteq \mathbb{R}^d$ is the **subgradient** of the confex function f at $\boldsymbol{x}$:

$$\boldsymbol{g} \in \partial f(\boldsymbol{x}) \quad \Longleftrightarrow \quad f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^\top(\boldsymbol{y}-\boldsymbol{x}) \quad \text{for all} \quad \boldsymbol{y} \in \mathbb{R}^d$$

Check (in homework):

- if f is smooth at $\boldsymbol{x}$: $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$
- the optimality condition for LASSO is: for each $1 \leq j \leq d$

$$\left[\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}^\top\widehat{\boldsymbol{\beta}})\right]_j \quad \begin{cases} = \lambda \cdot \mathsf{sign}(\widehat{\beta}_j) & \text{if} \quad \widehat{\beta}_j \neq 0 \\ \in [-\lambda, \lambda] & \text{if} \quad \widehat{\beta}_j = 0 \end{cases}$$