

Lasso: Algorithms and Extensions



Yuxin Chen

Princeton University, Fall 2020

Outline

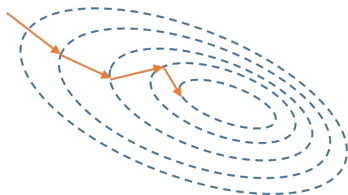
- Proximal operators
- Proximal gradient methods for lasso and its extensions
- Nesterov's accelerated algorithm

Proximal operators

Gradient descent

$$\text{minimize}_{\beta \in \mathbb{R}^p} f(\beta)$$

where $f(\beta)$ is convex and differentiable



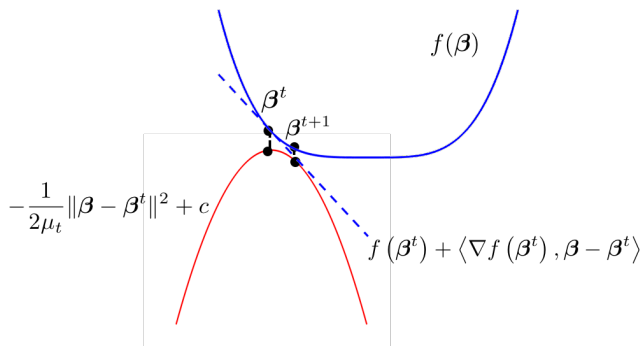
Algorithm 10.1 Gradient descent

for $t = 0, 1, \dots$:

$$\beta^{t+1} = \beta^t - \mu_t \nabla f(\beta^t)$$

where μ_t : step size / learning rate

A proximal point of view of GD



$$\beta^{t+1} = \arg \min_{\beta} \left\{ \underbrace{f(\beta^t) + \langle \nabla f(\beta^t), \beta - \beta^t \rangle}_{\text{linear approximation}} + \underbrace{\frac{1}{2\mu_t} \|\beta - \beta^t\|^2}_{\text{proximal term}} \right\}$$

- When μ_t is small, β^{t+1} tends to stay close to β^t

Proximal operator

If we define the proximal operator

$$\text{prox}_h(\mathbf{b}) := \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|\boldsymbol{\beta} - \mathbf{b}\|^2 + h(\boldsymbol{\beta}) \right\}$$

for any convex function h , then one can write

$$\boldsymbol{\beta}^{t+1} = \text{prox}_{\mu_t f_t}(\boldsymbol{\beta}^t)$$

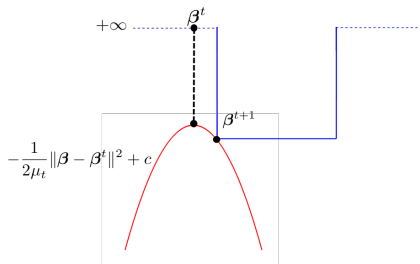
where $f_t(\boldsymbol{\beta}) := f(\boldsymbol{\beta}_t) + \langle \nabla f(\boldsymbol{\beta}_t), \boldsymbol{\beta} - \boldsymbol{\beta}_t \rangle$

Why consider proximal operators?

$$\text{prox}_h(\mathbf{b}) := \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|\boldsymbol{\beta} - \mathbf{b}\|^2 + h(\boldsymbol{\beta}) \right\}$$

- It is well-defined under very general conditions (including nonsmooth convex functions)
- The operator can be evaluated efficiently for many widely used functions (in particular, regularizers)
- This abstraction is conceptually and mathematically simple, and covers many well-known optimization algorithms

Example: characteristic functions



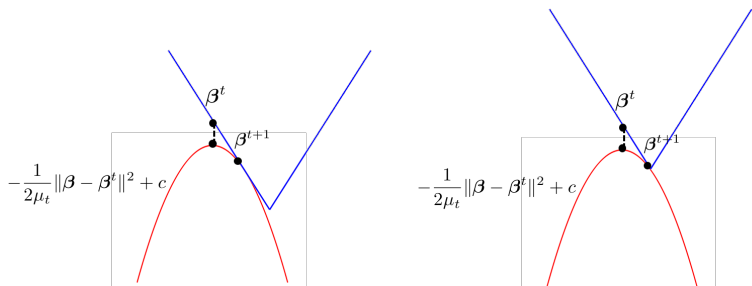
- If h is characteristic function

$$h(\beta) = \begin{cases} 0, & \text{if } \beta \in \mathcal{C} \\ \infty, & \text{else} \end{cases}$$

then

$$\text{prox}_h(\mathbf{b}) = \arg \min_{\beta \in \mathcal{C}} \|\beta - \mathbf{b}\|_2 \quad (\text{Euclidean projection})$$

Example: ℓ_1 norm



- If $h(\beta) = \|\beta\|_1$, then

$$\text{prox}_{\lambda h}(\mathbf{b}) = \psi_{\text{st}}(\mathbf{b}; \lambda)$$

where soft-thresholding $\psi_{\text{st}}(\cdot)$ is applied in an entry-wise manner.

Example: ℓ_2 norm

$$\text{prox}_h(\mathbf{b}) := \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|\boldsymbol{\beta} - \mathbf{b}\|^2 + h(\boldsymbol{\beta}) \right\}$$

- If $h(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|$, then

$$\text{prox}_{\lambda h}(\mathbf{b}) = \left(1 - \frac{\lambda}{\|\mathbf{b}\|} \right)_+ \mathbf{b}$$

where $a_+ := \max\{a, 0\}$. This is called *block soft thresholding*.

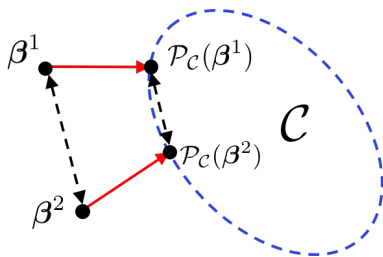
Example: log barrier

$$\text{prox}_h(\mathbf{b}) := \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2} \|\boldsymbol{\beta} - \mathbf{b}\|^2 + h(\boldsymbol{\beta}) \right\}$$

- If $h(\boldsymbol{\beta}) = -\sum_{i=1}^p \log \beta_i$, then

$$(\text{prox}_{\lambda h}(\mathbf{b}))_i = \frac{b_i + \sqrt{b_i^2 + 4\lambda}}{2}$$

Nonexpansiveness of proximal operators

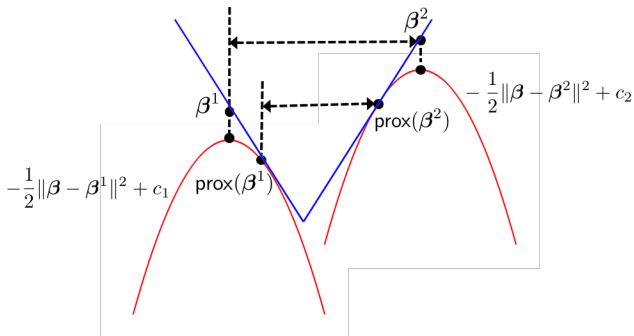


Recall that when $h(\beta) = \begin{cases} 0, & \text{if } \beta \in \mathcal{C} \\ \infty & \text{else} \end{cases}$, $\text{prox}_h(\beta)$ is Euclidean projection \mathcal{P}_C onto \mathcal{C} , which is nonexpansive:

$$\|\mathcal{P}_C(\beta^1) - \mathcal{P}_C(\beta^2)\| \leq \|\beta^1 - \beta^2\|$$

Nonexpansiveness of proximal operators

Nonexpansiveness is a property for general $\text{prox}_h(\cdot)$



Fact 10.1 (Nonexpansiveness)

$$\|\text{prox}_h(\beta^1) - \text{prox}_h(\beta^2)\| \leq \|\beta^1 - \beta^2\|$$

- In some sense, proximal operator behaves like projection

Proof of nonexpansiveness

Let $z^1 = \text{prox}_h(\beta^1)$ and $z^2 = \text{prox}_h(\beta^2)$. Subgradient characterizations of z^1 and z^2 read

$$\beta^1 - z^1 \in \partial h(z^1) \quad \text{and} \quad \beta^2 - z^2 \in \partial h(z^2)$$

The claim would follow if

$$(\beta^1 - \beta^2)^\top (z^1 - z^2) \geq \|z^1 - z^2\|^2 \quad (\text{together with Cauchy-Schwarz})$$

$$\iff (\beta^1 - z^1 - \beta^2 + z^2)^\top (z^1 - z^2) \geq 0$$

$$\iff \begin{cases} h(z^2) \geq h(z^1) + \underbrace{\langle \beta^1 - z^1, z^2 - z^1 \rangle}_{\in \partial h(z^1)} \\ h(z^1) \geq h(z^2) + \underbrace{\langle \beta^2 - z^2, z^1 - z^2 \rangle}_{\in \partial h(z^2)} \end{cases}$$

Proximal gradient methods

Optimizing composite functions

$$\text{(Lasso)} \quad \underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \underbrace{\frac{1}{2} \|\mathbf{X}\beta - \mathbf{y}\|^2}_{:=f(\beta)} + \underbrace{\lambda \|\beta\|_1}_{:=g(\beta)} = f(\beta) + g(\beta)$$

where $f(\beta)$ is differentiable, and $g(\beta)$ is non-smooth

- Since $g(\beta)$ is non-differentiable, we cannot run vanilla gradient descent

Proximal gradient methods

One strategy: replace $f(\beta)$ with linear approximation, and compute the proximal solution

$$\beta^{t+1} = \arg \min_{\beta} \left\{ f(\beta^t) + \langle \nabla f(\beta^t), \beta - \beta^t \rangle + g(\beta) + \frac{1}{2\mu_t} \|\beta - \beta^t\|^2 \right\}$$

The optimality condition reads

$$\mathbf{0} \in \nabla f(\beta^t) + \partial g(\beta^{t+1}) + \frac{1}{\mu_t} (\beta^{t+1} - \beta^t)$$

which is equivalent to optimality condition of

$$\begin{aligned} \beta^{t+1} &= \arg \min_{\beta} \left\{ g(\beta) + \frac{1}{2\mu_t} \left\| \beta - (\beta^t - \mu_t \nabla f(\beta^t)) \right\|^2 \right\} \\ &= \text{prox}_{\mu_t g} (\beta^t - \mu_t \nabla f(\beta^t)) \end{aligned}$$

Proximal gradient methods

Alternate between gradient updates on f and proximal minimization on g

Algorithm 10.2 Proximal gradient methods

for $t = 0, 1, \dots$:

$$\beta^{t+1} = \text{prox}_{\mu_t g} \left(\beta^t - \mu_t \nabla f(\beta^t) \right)$$

where μ_t : step size / learning rate

Projected gradient methods

When $g(\beta) = \begin{cases} 0, & \text{if } \beta \in \underbrace{\mathcal{C}}_{\text{convex}} \\ \infty, & \text{else} \end{cases}$ is characteristic function:

$$\begin{aligned}\beta^{t+1} &= \mathcal{P}_{\mathcal{C}} \left(\beta^t - \mu_t \nabla f(\beta^t) \right) \\ &:= \arg \min_{\beta \in \mathcal{C}} \left\| \beta - (\beta^t - \mu_t \nabla f(\beta^t)) \right\|\end{aligned}$$

This is a first-order method to solve the constrained optimization

$$\begin{aligned}\text{minimize}_{\beta} & \quad f(\beta) \\ \text{s.t.} & \quad \beta \in \mathcal{C}\end{aligned}$$

Proximal gradient methods for lasso

For lasso: $f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2$ and $g(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$,

$$\begin{aligned}\text{prox}_g(\boldsymbol{\beta}) &= \arg \min_{\mathbf{b}} \left\{ \frac{1}{2}\|\boldsymbol{\beta} - \mathbf{b}\|^2 + \lambda\|\mathbf{b}\|_1 \right\} \\ &= \psi_{\text{st}}(\boldsymbol{\beta}; \lambda)\end{aligned}$$

$$\begin{aligned}\implies \quad \boldsymbol{\beta}^{t+1} &= \psi_{\text{st}}\left(\boldsymbol{\beta}^t - \mu_t \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}^t - \mathbf{y}); \mu_t \lambda\right) \\ &\quad \text{(iterative soft thresholding)}\end{aligned}$$

Proximal gradient methods for group lasso

Sometimes variables have a natural group structure, and it is desirable to set all variables within a group to be zero (or nonzero) simultaneously

$$\text{(group lasso)} \quad \underbrace{\frac{1}{2} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2}_{:=f(\boldsymbol{\beta})} + \lambda \underbrace{\sum_{j=1}^k \|\boldsymbol{\beta}_j\|}_{:=g(\boldsymbol{\beta})}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^{p/k}$ and $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k \end{bmatrix}$.

$$\text{prox}_g(\boldsymbol{\beta}) = \psi_{\text{bst}}(\boldsymbol{\beta}; \lambda) := \left[\left(1 - \frac{\lambda}{\|\boldsymbol{\beta}_j\|} \right)_+ \boldsymbol{\beta}_j \right]_{1 \leq j \leq k}$$

$$\implies \boldsymbol{\beta}^{t+1} = \psi_{\text{bst}}(\boldsymbol{\beta}^t - \mu_t \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}^t - \mathbf{y}); \mu_t \lambda)$$

Proximal gradient methods for elastic net

Lasso does not handle highly correlated variables well: if there is a group of highly correlated variables, lasso often picks one from the group and ignore the rest.

- Sometimes we make a compromise between lasso and ℓ_2 penalties

$$\text{(elastic net)} \quad \underbrace{\frac{1}{2} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2}_{:=f(\boldsymbol{\beta})} + \lambda \underbrace{\left\{ \|\boldsymbol{\beta}\|_1 + (\gamma/2) \|\boldsymbol{\beta}\|_2^2 \right\}}_{:=g(\boldsymbol{\beta})}$$

$$\text{prox}_{\lambda g}(\boldsymbol{\beta}) = \frac{1}{1 + \lambda\gamma} \psi_{\text{st}}(\boldsymbol{\beta}; \lambda)$$

$$\implies \boldsymbol{\beta}^{t+1} = \frac{1}{1 + \mu_t \lambda \gamma} \psi_{\text{st}}\left(\boldsymbol{\beta}^t - \mu_t \mathbf{X}^\top (\mathbf{X}\boldsymbol{\beta}^t - \mathbf{y}); \mu_t \lambda\right)$$

- soft thresholding followed by multiplicative shrinkage

Interpretation: majorization-minimization

$$f_{\mu_t}(\boldsymbol{\beta}, \boldsymbol{\beta}^t) := \underbrace{f(\boldsymbol{\beta}^t) + \langle \nabla f(\boldsymbol{\beta}^t), \boldsymbol{\beta} - \boldsymbol{\beta}^t \rangle}_{\text{linearization}} + \underbrace{\frac{1}{2\mu_t} \|\boldsymbol{\beta} - \boldsymbol{\beta}^t\|^2}_{\text{trust region penalty}}$$

majorizes $f(\boldsymbol{\beta})$ if $0 < \mu_t < \frac{1}{L}$, where L is Lipschitz constant¹ of $\nabla f(\cdot)$

Proximal gradient descent is a majorization-minimization algorithm

$$\boldsymbol{\beta}^{t+1} = \underbrace{\arg \min_{\boldsymbol{\beta}}}_{\text{minimization}} \left\{ \underbrace{f_{\mu_t}(\boldsymbol{\beta}, \boldsymbol{\beta}^t) + g(\boldsymbol{\beta})}_{\text{majorization}} \right\}$$

¹This means $\|\nabla f(\boldsymbol{\beta}) - \nabla f(\mathbf{b})\| \leq L\|\boldsymbol{\beta} - \mathbf{b}\|$ for all $\boldsymbol{\beta}$ and \mathbf{b}

Convergence rate of proximal gradient methods

Theorem 10.2 (fixed step size; Nesterov '07)

Suppose g is convex, and f is differentiable and convex whose gradient has Lipschitz constant L . If $\mu_t \equiv \mu \in (0, 1/L)$, then

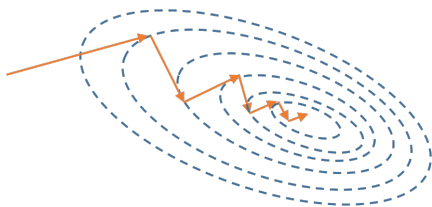
$$f(\beta^t) + g(\beta^t) - \min_{\beta} \{f(\beta) + g(\beta)\} \leq O\left(\frac{1}{t}\right)$$

- Step size requires an upper bound on L
- May prefer backtracking line search to fixed step size
- **Question:** can we further improve the convergence rate?

Nesterov's accelerated gradient methods

Nesterov's accelerated method

Problem of gradient descent: zigzagging



Nesterov's idea: include a momentum term to avoid overshooting

Nesterov's accelerated method

Nesterov's idea: include a momentum term to avoid overshooting

$$\begin{aligned}\beta^t &= \text{prox}_{\mu_t g} \left(\mathbf{b}^{t-1} - \mu_t \nabla f \left(\mathbf{b}^{t-1} \right) \right) \\ \mathbf{b}^t &= \beta^t + \underbrace{\alpha_t \left(\beta^t - \beta^{t-1} \right)}_{\text{momentum term}} \quad (\text{extrapolation})\end{aligned}$$

- A simple (**but mysterious**) choice of extrapolation parameter

$$\alpha_t = \frac{t-1}{t+2}$$

- Fixed size $\mu_t \equiv \mu \in (0, 1/L)$ or backtracking line search
- Same computational cost per iteration as proximal gradient

Convergence rate of Nesterov's accelerated method

Theorem 10.3 (Nesterov '83, Nesterov '07)

Suppose f is differentiable and convex and g is convex. If one takes $\alpha_t = \frac{t-1}{t+2}$ and a fixed step size $\mu_t \equiv \mu \in (0, 1/L)$, then

$$f(\beta^t) + g(\beta^t) - \min_{\beta} \{f(\beta) + g(\beta)\} \leq O\left(\frac{1}{t^2}\right)$$

In general, this rate cannot be improved if one only uses gradient information!

Numerical experiments (for lasso)

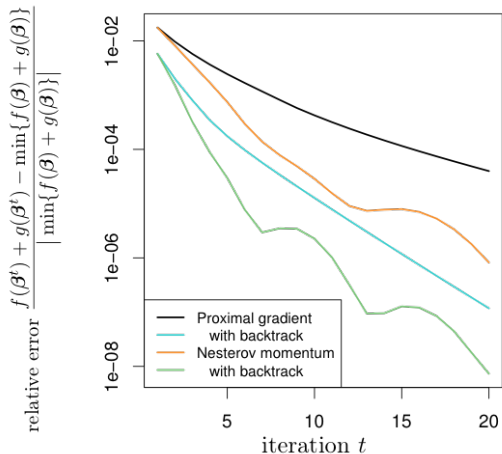


Figure credit: Hastie, Tibshirani, & Wainwright '15

Reference

- "*Proximal algorithms*," Neal Parikh and S. Boyd, *Foundations and Trends in Optimization*, 2013.
- "*Convex optimization algorithms*," D. Bertsekas, *Athena Scientific*, 2015.
- "*Convex optimization: algorithms and complexity*," S. Bubeck, *Foundations and Trends in Machine Learning*, 2015.
- "*Statistical learning with sparsity: the Lasso and generalizations*," T. Hastie, R. Tibshirani, and M. Wainwright, 2015.
- "*Model selection and estimation in regression with grouped variables*," M. Yuan and Y. Lin, *Journal of the royal statistical society*, 2006.
- "*A method of solving a convex programming problem with convergence rate $O(1/k^2)$* ," Y. Nesterov, *Soviet Mathematics Doklady*, 1983.

Reference

- "*Gradient methods for minimizing composite functions*," Y. Nesterov, *Technical Report*, 2007.
- "*A fast iterative shrinkage-thresholding algorithm for linear inverse problems*," A. Beck and M. Teboulle, *SIAM journal on imaging sciences*, 2009.