

Prox-Gradient Method

Sanjay Lall and Stephen Boyd

EE104

Stanford University

Prox-gradient method

Minimizing composite functions

- ▶ want to minimize $F(\theta) = f(\theta) + g(\theta)$ (called a *composite function*)
- ▶ f is differentiable, but g need not be
- ▶ example: minimize $\mathcal{L}(\theta) + \lambda r(\theta)$, with $r(\theta) = \|\theta\|_1$
- ▶ we'll see idea of gradient method extends directly to composite functions

Selective linearization

- ▶ at iteration k , linearize f *but not* g

$$\hat{F}(\theta; \theta^k) = f(\theta^k) + \nabla f(\theta^k)^T(\theta - \theta^k) + g(\theta)$$

- ▶ want $\hat{F}(\theta; \theta^k)$ small, but with θ near θ^k
- ▶ choose θ^{k+1} to minimize $\hat{F}(\theta; \theta^k) + \frac{1}{2h^k} \|\theta - \theta^k\|_2^2$, with $h^k > 0$
- ▶ same as minimizing

$$g(\theta) + \frac{1}{2h^k} \|\theta - (\theta^k - h^k \nabla f(\theta^k))\|^2$$

- ▶ for many 'simple' functions g , this minimization can be done analytically
- ▶ this iteration from θ^k to θ^{k+1} is called *prox-gradient step*

Prox-gradient iteration

► prox-gradient iteration has two parts:

1. *gradient step*: $\theta^{k+1/2} = \theta^k - h^k \nabla f(\theta^k)$

2. *prox step*: choose θ^{k+1} to minimize $g(\theta) + \frac{1}{2h^k} \|\theta - \theta^{k+1/2}\|_2^2$

($\theta^{k+1/2}$ is an intermediate iterate, in between θ^k and θ^{k+1})

► step 1 handles differentiable part of objective, *i.e.*, f

► step 2 handles second part of objective, *i.e.*, g

Proximal operator

- ▶ given function $q : \mathbf{R}^d \rightarrow \mathbf{R}$, and $\kappa > 0$,

$$\mathbf{prox}_{q,\kappa}(v) = \underset{\theta}{\operatorname{argmin}} \left(q(\theta) + \frac{1}{2\kappa} \|\theta - v\|_2^2 \right)$$

is called the *proximal operator* of q at v , with parameter κ

- ▶ the prox-gradient step can be expressed as

$$\theta^{k+1} = \mathbf{prox}_{g,h^k}(\theta^{k+1/2}) = \mathbf{prox}_{g,h^k}(\theta^k - h^k \nabla f(\theta^k))$$

- ▶ hence the name prox-gradient iteration

How to choose step length

- ▶ same as for gradient, but using $F(\theta) = f(\theta) + g(\theta)$
- ▶ a simple scheme:
 - ▶ if $F(\theta^{k+1}) \geq F(\theta^k)$, set $h^{k+1} = h^k/2$, $\theta^{k+1} = \theta^k$ (a *rejected step*)
 - ▶ if $F(\theta^{k+1}) < F(\theta^k)$, set $h^{k+1} = 1.2h^k$ (an *accepted step*)
- ▶ reduce step length by half if it's too long; increase it 20% otherwise

Stopping criterion

- ▶ stopping condition for prox-gradient method:

$$\left\| \nabla f(\theta^{k+1}) - \frac{1}{h^k}(\theta^{k+1} - \theta^{k+1/2}) \right\|_2 \leq \epsilon$$

- ▶ analog of $\|\nabla f(\theta^{k+1})\|_2 \leq \epsilon$ for gradient method
- ▶ $-\frac{1}{h^k}(\theta^{k+1} - \theta^{k+1/2})$ is a surrogate for the gradient for g (which need not be differentiable)

Prox-gradient method summary

choose an initial $\theta^1 \in \mathbf{R}^d$ and $h^1 > 0$ (e.g., $\theta^1 = 0$, $h^1 = 1$)

for $k = 1, 2, \dots, k^{\max}$

1. gradient step. $\theta^{k+1/2} = \theta^k - h^k \nabla f(\theta^k)$
2. prox step. $\theta^{\text{tent}} = \operatorname{argmin}_{\theta} \left(g(\theta) + \frac{1}{2h^k} \|\theta - \theta^{k+1/2}\|_2^2 \right)$
3. if $F(\theta^{\text{tent}}) < F(\theta^k)$,
 - (a) set $\theta^{k+1} = \theta^{\text{tent}}$, $h^{k+1} = 1.2h^k$
 - (b) quit if $\left\| \nabla f(\theta^{k+1}) - \frac{1}{h^k} (\theta^{k+1} - \theta^{k+1/2}) \right\|_2 \leq \epsilon$
4. else set $h^k := 0.5h^k$ and go to step 1

Prox-gradient method convergence

- ▶ prox-gradient method finds a stationary point
 - ▶ suitably defined for non-differentiable functions
 - ▶ assuming some technical conditions hold

- ▶ for *convex problems*
 - ▶ prox-gradient method is *non-heuristic*
 - ▶ for any starting point θ^1 , $F(\theta^k) \rightarrow F^*$ as $k \rightarrow \infty$

- ▶ for *non-convex problems*
 - ▶ prox-gradient method is *heuristic*
 - ▶ we can (and often do) have $F(\theta^k) \not\rightarrow F^*$

Prox-gradient for regularized ERM

Prox-gradient for sum squares regularizer

- ▶ let's apply prox-gradient method to $F(\theta) = \mathcal{L}(\theta) + \lambda\|\theta\|_2^2$
 - ▶ $f(\theta) = \mathcal{L}(\theta)$
 - ▶ $g(\theta) = \lambda\|\theta\|_2^2 = \lambda\theta_1^2 + \dots + \lambda\theta_d^2$
- ▶ in prox step, we need to minimize $\lambda\theta_i^2 + \frac{1}{2h^k}(\theta_i - \theta_i^{k+1/2})^2$ over θ_i
- ▶ solution is $\theta_i = \frac{1}{1+2\lambda h^k}\theta_i^{k+1/2}$
- ▶ so prox step just shrinks the gradient step $\theta^{k+1/2}$ by the factor $\frac{1}{1+2\lambda h^k}$
- ▶ prox-gradient iteration:
 1. gradient step: $\theta^{k+1/2} = \theta^k - h^k \nabla \mathcal{L}(\theta^k)$
 2. prox step: $\theta^{k+1} = \frac{1}{1+2\lambda h^k}\theta^{k+1/2}$

Prox-gradient for ℓ_1 regularizer

▶ let's apply prox-gradient method to $F(\theta) = \mathcal{L}(\theta) + \lambda\|\theta\|_1$

▶ $f(\theta) = \mathcal{L}(\theta)$

▶ $g(\theta) = \lambda\|\theta\|_1 = \lambda|\theta_1| + \dots + \lambda|\theta_d|$

▶ in prox step, we need to minimize $\lambda|\theta_i| + \frac{1}{2h^k}(\theta_i - \theta_i^{k+1/2})^2$ over θ_i

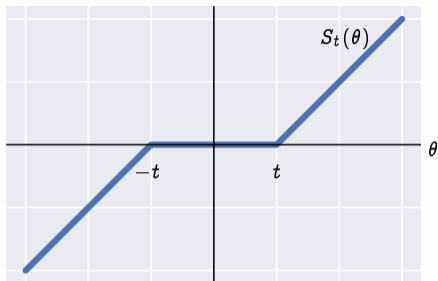
▶ solution is

$$\theta_i^{k+1} = \begin{cases} \theta_i^{k+1/2} - 2\lambda h^k & \theta_i^{k+1/2} > 2\lambda h^k \\ 0 & |\theta_i^{k+1/2}| \leq 2\lambda h^k \\ \theta_i^{k+1/2} + 2\lambda h^k & \theta_i^{k+1/2} < -2\lambda h^k \end{cases}$$

▶ called *soft threshold function*

▶ sometimes written as $\theta_i^{k+1} = S_{2\lambda h^k}(\theta_i^{k+1/2}) = \text{sign}(\theta_i)(|\theta_i| - 2\lambda h^k)_+$

Soft threshold function



► prox-gradient iteration for regularized ERM with ℓ_1 regularization:

1. gradient step: $\theta^{k+1/2} = \theta^k - h^k \nabla \mathcal{L}(\theta^k)$

2. prox step: $\theta^{k+1} = S_{2\lambda h^k}(\theta_i^{k+1/2})$

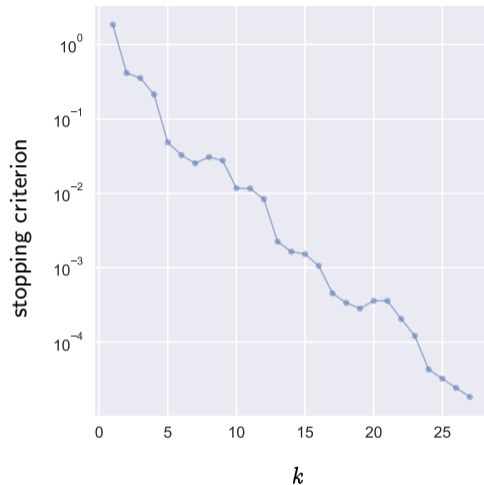
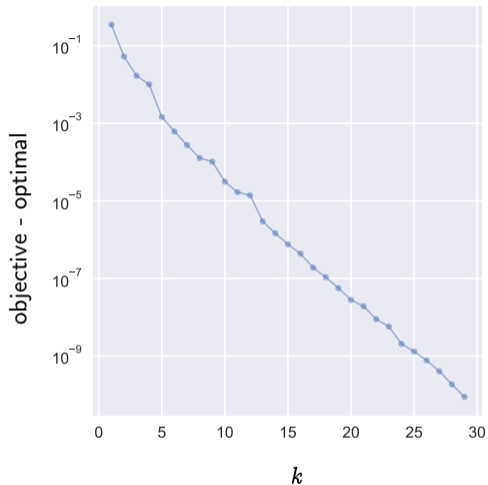
► the soft threshold step shrinks all coefficients

► and sets the small ones to zero

Prox-gradient step for nonnegative regularizer

- ▶ let's apply prox-gradient method to $F(\theta) = \mathcal{L}(\theta) + r(\theta)$, where $r(\theta) = 0$ for $\theta \geq 0$, ∞ otherwise
 - ▶ $f(\theta) = \mathcal{L}(\theta)$
 - ▶ $g(\theta) = q(\theta_1) + \dots + q(\theta_d)$
- ▶ in prox step, we need to minimize $q(\theta_i) + \frac{1}{2h^k}(\theta_i - \theta_i^{k+1/2})^2$ over θ_i
- ▶ solution is $\theta_i = \left(\theta_i^{k+1/2}\right)_+$
- ▶ so prox step just replaces the gradient step $\theta_i^{k+1/2}$ with its positive part
- ▶ prox-gradient iteration:
 1. gradient step: $\theta^{k+1/2} = \theta^k - h^k \nabla \mathcal{L}(\theta^k)$
 2. prox step: $\theta^{k+1} = \left(\theta^{k+1/2}\right)_+$

Example



- ▶ synthetic data, $n = 500$, $d = 200$
- ▶ lasso (square loss, ℓ_1 regularization), $\lambda = 0.1$