Prox-Gradient Method

Jong-Han Kim

EE787 Machine learning
Kyung Hee University

Prox-gradient method

Minimizing composite functions

- want to minimize $F(\theta) = f(\theta) + g(\theta)$ (called *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 $\hat{F}(\theta; \theta^k) + \frac{1}{2h^k}||\theta - \theta^k||^2$, 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 $g(\theta) + \frac{1}{2h^k} ||\theta - \theta^{k+1/2}||^2$

  ($\theta^{k+1/2}$ 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 $q : \mathbf{R}^d \to \mathbf{R}$, and $\kappa > 0$,

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

  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 $F(\theta^{k+1}) > F(\theta^k)$, set $h^{k+1} = h^k/2$, $\theta^{k+1} = \theta^k$ (a *rejected step*)
  - if $F(\theta^{k+1}) \leq F(\theta^k)$, 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:

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

- analog of $\|\nabla f(\theta^{k+1})\| \leq \epsilon$ for gradient method
- second term $-\frac{1}{h^k}(\theta^{k+1} - \theta^{k+1/2})$ serves the purpose of a 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 $k = 1, 2, \ldots, k^{\max}$

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

Prox-gradient method convergence

- prox-gradient method always 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, $F(\theta^k) \to F^\star$ as $k \to \infty$

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

Prox-gradient for regularized ERM

Prox-gradient for sum squares regularizer

- let's apply prox-gradient method to $F(\theta) = \mathcal{L}(\theta) + \lambda||\theta||_2^2$
  - $f(\theta) = \mathcal{L}(\theta)$
  - $g(\theta) = \lambda||\theta||_2^2 = \lambda\theta_1^2 + \cdots + \lambda\theta_d^2$
- 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

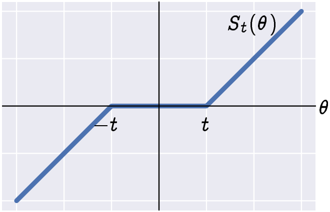
- let's apply prox-gradient method to $F(\theta) = \mathcal{L}(\theta) + \lambda||\theta||_1$
  - $f(\theta) = \mathcal{L}(\theta)$
  - $g(\theta) = \lambda||\theta||_1 = \lambda|\theta_1| + \cdots + \lambda|\theta_d|$
- in prox step, we need to minimize $\lambda|\theta_i| + \frac{1}{2h^k}(\theta_i - \theta_i^{k+1/2})^2$ over θ_i
- solution is

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

- called *soft threshold function*
- sometimes written as

$$\begin{aligned} \theta_i^{k+1} = S_{\lambda h^k}(\theta_i^{k+1/2}) &= \mathrm{sign}(\theta_i^{k+1/2})(|\theta_i^{k+1/2}| - \lambda h^k)_+ \\ &= (\theta_i^{k+1/2} - \lambda h^k)_+ - (-\theta_i^{k+1/2} - \lambda h^k)_+ \end{aligned}$$

Soft threshold function

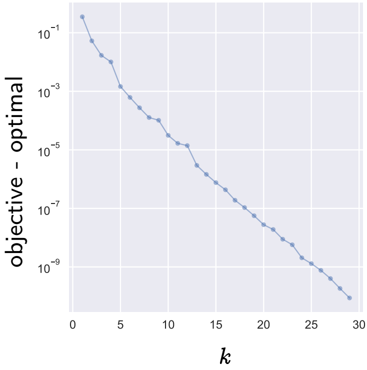


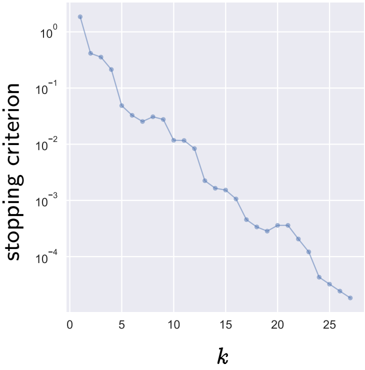
- prox-gradient iteration for regularized ERM with ℓ_1 regularization:
  1. gradient step: $\theta^{k+1/2} = \theta^k - h^k \nabla \mathcal{L}(\theta^k)$
  2. prox step: $\theta_i^{k+1} = S_{\lambda h^k}(\theta_i^{k+1/2}) \quad$ for $i = 1, \ldots, d$.
- the soft threshold step shrinks all coefficients
- 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
  - $f(\theta) = \mathcal{L}(\theta)$
  - $g(\theta) = q(\theta_1) + \cdots + q(\theta_d)$
- 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), $\lambda = 0.1$