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 *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 $\theta$ near $\theta^k$

- choose $\theta^{k+1}$ 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 $\theta^k$ to $\theta^{k+1}$ 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 $\theta^{k+1}$ to minimize $g(\theta) + \frac{1}{2h^k} \|\theta - \theta^{k+1/2}\|^2$

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

- step 1 handles differentiable part of objective, i.e., $f$
- step 2 handles second part of objective, i.e., $g$
Proximal operator

Given function \( q : \mathbb{R}^d \to \mathbb{R} \), and \( \kappa > 0 \),

\[
\text{prox}_{q, \kappa}(v) = \arg\min_{\theta} \left( q(\theta) + \frac{1}{2\kappa} \| \theta - v \|^2 \right)
\]

is called the \textit{proximal operator} of \( q \) at \( v \), with parameter \( \kappa \).

The prox-gradient step can be expressed as

\[
\theta^{k+1} = \text{prox}_{g, h_k} (\theta^{k+1/2}) = \text{prox}_{g, h_k} (\theta^k - h_k \nabla f(\theta^k))
\]

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

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

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

2. prox step. $\theta^{\text{tent}} = \arg\min_{\theta} \left( g(\theta) + \frac{1}{2h^k} \|\theta - \theta^{k+1/2}\|^2 \right)$

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

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 $\theta^1$, $F(\theta^k) \to F^*$ as $k \to \infty$

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

- solution is $\theta_i = \frac{1}{1 + 2\lambda h_k} \theta_i^{k+1/2}$

- so prox step just shrinks the gradient step $\theta_i^{k+1/2}$ by the factor $\frac{1}{1 + 2\lambda h_k}$

- prox-gradient iteration:
  1. gradient step: $\theta_i^{k+1/2} = \theta_i^k - h_k \nabla \mathcal{L}(\theta_i^k)$
  2. prox step: $\theta_i^{k+1} = \frac{1}{1 + 2\lambda h_k} \theta_i^{k+1/2}$
Prox-gradient for $\ell_1$ 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 $\theta_i$
- 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 $\ell_1$ regularization:

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

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

- 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 \( \theta \geq 0 \), \( \infty \) 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 \( \theta_i \).

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, $\ell_1$ regularization), $\lambda = 0.1$