Proximal Gradient Method

- Convex function : $f(x) = g(x) + h(x)$

  g: convex, cont. diff. $\quad$ h: convex but not diff. everywhere

- Can minimize $f(x)$ using sub-gradient algorithm, but rate of convergence is at best $O\left(\frac{1}{\sqrt{N}}\right)$, i.e.,

$$f_N^* - f^* \sim O\left(\frac{1}{\sqrt{N}}\right)$$

Proximal Gradient Iteration

Since g is cont. diff., if $h(x) = 0$, GD iteration would be

$$x_{k+1} = x_k - \alpha_k \nabla g(x_k)$$

With $h(x) \neq 0$, use "proximal" gradient iterate:

$$x_{k+1} = \arg\min_x \frac{1}{2\alpha_k} \| x - (x_k - \alpha_k \nabla g(x_k)) \|^2 + h(x)$$

penalty for deviating from $x_k - \alpha_k \nabla g(x_k)$

- Inherent Assumption: h is s.t. above min. is easy (closed-form).

Convergence Analysis:

$$x_{k+1} = \arg\min_x \frac{1}{2\alpha_k} \| x - (x_k - \alpha_k \nabla g(x_k)) \|^2 + h(x)$$

Write solution x_{k+1} as $x_k - \alpha_k d_k$ for convenience

- d_k is a "generalized" gradient at x_k.

$$\boxed{x_{k+1} = x_k - \alpha_k d_k}$$

Assume g has Lipschitz gradients, and use constant stepsize $\alpha_k = \alpha \ \forall k$. By Descent Lemma,

$$g(x_{k+1}) \leq g(x_k) - \alpha \nabla g(x_k)^T d_k + \frac{1}{2} L \alpha^2 \| d_k \|^2$$

Now, since $x_{k+1} = x_k - \alpha d_k$ minimizes

$$\frac{1}{2\alpha} \| x - (x_k - \alpha \nabla g(x_k)) \|^2 + h(x)$$

gradient $= \frac{1}{\alpha}(x - (x_k - \alpha \nabla g(x_k)))$ ← subgradient $v \in \partial h(x)$

$$\Rightarrow 0 = \frac{1}{\alpha}(x_{k+1} - (x_k - \alpha \nabla g(x_k))) + v, \text{ for some } v \in \partial h(x_{k+1})$$

i.e., $0 = (\nabla g(x_k) - d_k) + v$, for some $v \in \partial h(x_{k+1})$

i.e. $v = d_k - \nabla g(x_k)$, for some $v \in \partial h(x_{k+1})$

But by definition of $\partial h(x_{k+1})$,

$$h(x_{k+1}) + v^T (x - x_{k+1}) \leq h(x), \quad \forall x \in \mathbb{R}^n$$

i.e.,

$$h(x_{k+1}) + (d_k - \nabla g(x_k))^T (x - x_{k+1}) \leq h(x), \quad \forall x \in \mathbb{R}^n$$

i.e. $h(x_{k+1}) \leq (\nabla g(x_k) - d_k)^T (x - x_{k+1}) + h(x), \quad \forall x \in \mathbb{R}^n$ — (1)

Also, from earlier (descent lemma),

$$g(x_{k+1}) \leq g(x_k) - \alpha \nabla g(x_k)^T d_k + \frac{1}{2} L \alpha^2 \|d_k\|^2$$

$$\leq g(x) + \nabla g(x_k)^T (x_k - x) \quad \forall x \in \mathbb{R}^n$$

(convexity of g)

$$- \alpha \nabla g(x_k)^T d_k + \frac{1}{2} L \alpha^2 \|d_k\|^2 \quad \text{— (2)}$$

Combine (1), (2) to get

$$f(x_{k+1}) \leq f(x) + \nabla g(x_k)^T (\underbrace{x_k - x + x}_{=0} - x_{k+1} - \alpha d_k) + d_k^T (x_{k+1} - x) + \frac{1}{2} L \alpha^2 \|d_k\|^2$$

i.e.,

$$f(x_{k+1}) \leq f(x) + d_k^T (x_{k+1} - x) + \frac{1}{2} L \alpha^2 \|d_k\|^2$$

$$= f(x) + d_k^T (x_k - x) - \alpha \|d_k\|^2 + \frac{1}{2} L \alpha^2 \|d_k\|^2$$

($x_{k+1} = x_k - \alpha d_k$)

$$f(x_{k+1}) \leq f(x) + d_k^T(x_k - x) - \alpha \|d_k\|^2 + \frac{1}{2} L \alpha^2 \|d_k\|^2 \quad , \forall x \in \mathbb{R}^n$$

Set $\alpha < \frac{1}{L}$, i.e. $L < \frac{1}{\alpha}$. Then,

$$\boxed{f(x_{k+1}) \leq f(x) + d_k^T(x_k - x) - \frac{\alpha}{2} \|d_k\|^2, \forall x}$$

Setting $x = x_k$ on RHS, we obtain

$$f(x_{k+1}) \leq f(x_k) - \frac{\alpha}{2} \|d_k\|^2 \leq f(x_k)$$

i.e. the algorithm is a descent algorithm (if $\alpha < \frac{1}{L}$)

Also, setting $x = x^*$ on RHS in box, we obtain

$$f(x_{k+1}) \leq f(x^*) + d_k^T(x_k - x^*) - \frac{\alpha}{2} \|d_k\|^2$$

$$= f(x^*) - \frac{\alpha}{2} \left(\|d_k\|^2 - \frac{2}{\alpha} d_k^T (x_k - x^*) \right)$$

$$= f(x^*) - \frac{\alpha}{2} \left(\left\| \frac{x_k - x^*}{\alpha} - d_k \right\|^2 - \left\| \frac{x_k - x^*}{\alpha} \right\|^2 \right)$$

$$= f(x^*) - \frac{1}{2\alpha} \left(\|x_k - \alpha d_k - x^*\|^2 - \|x_k - x^*\|^2 \right)$$

$$= f(x^*) - \frac{1}{2\alpha} \left(\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 \right)$$

$$f(x_{k+1}) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)$$

Sum from $k=0$ to $N-1$ on both sides,

$$\sum_{k=0}^{N-1}\left(f(x_{k+1}) - f(x^*)\right) \leq \frac{1}{2\alpha}\left(\|x_0 - x^*\|^2 - \|x_{N+1} - x^*\|^2\right)$$

$$\leq \frac{1}{2\alpha}\|x_0 - x^*\|^2$$

But $f(x_N) \leq f(x_{N-1}) \leq \cdots \leq f(x_1)$

$$\Rightarrow \quad N\left(f(x_N) - f(x^*)\right) \leq \sum_{k=0}^{N-1} f(x_{k+1}) - f(x^*)$$

i.e. $$f(x_N) - f(x^*) \leq \frac{1}{2\alpha N}\|x_0 - x^*\|^2$$

$$\sim O\left(\frac{1}{N}\right)$$

Note The proximal gradient algorithm is only useful if we can easily solve

$$\min_x \frac{1}{2\alpha}\|x - x_k + \alpha \nabla g(x_k)\|^2 + h(x)$$

Example $h(x) = \|x\|_1 = \sum_{i=1}^{n} |x(i)|$

Let $z = x_k - \alpha \nabla g(x_k)$

$$\text{minimize}_x \quad \frac{1}{2\alpha}\sum_{i=1}^{n}\left(x(i) - z(i)\right)^2 + \sum_{i=1}^{n}|x(i)|$$

$$\underset{x}{\text{minimize}} \quad \sum_{i=1}^{n} \frac{1}{2\alpha}(x(i)-z(i))^2 + |x(i)|$$

$\equiv$ for each i, $\underset{x(i)}{\text{minimize}} \underbrace{\frac{(x(i)-z(i))^2}{2\alpha} + |x(i)|}_{C(x(i))}$

Case 1 $x(i) > 0$ optimal

$x(i) > 0 \Rightarrow C(x(i)) = \frac{(x(i)-z(i))^2}{2\alpha} + x(i)$

$\nabla C(x(i)) = 0 \Rightarrow \frac{1}{\alpha}(x(i)-z(i)) + 1 = 0$

$\Rightarrow x^*(i) = z(i) - \alpha$

This is solution if $z(i) > \alpha$

Case 2 $x(i) < 0$ is optimal

$x(i) < 0 \Rightarrow C(x(i)) = \frac{(x(i)-z(i))^2}{2\alpha} - x(i)$

$\nabla C(x(i)) = 0 \Rightarrow x^*(i) = z(i) + \alpha$

This is the solution if $z(i) < -\alpha$.

Case 3 $x(i) = 0$ is optimal

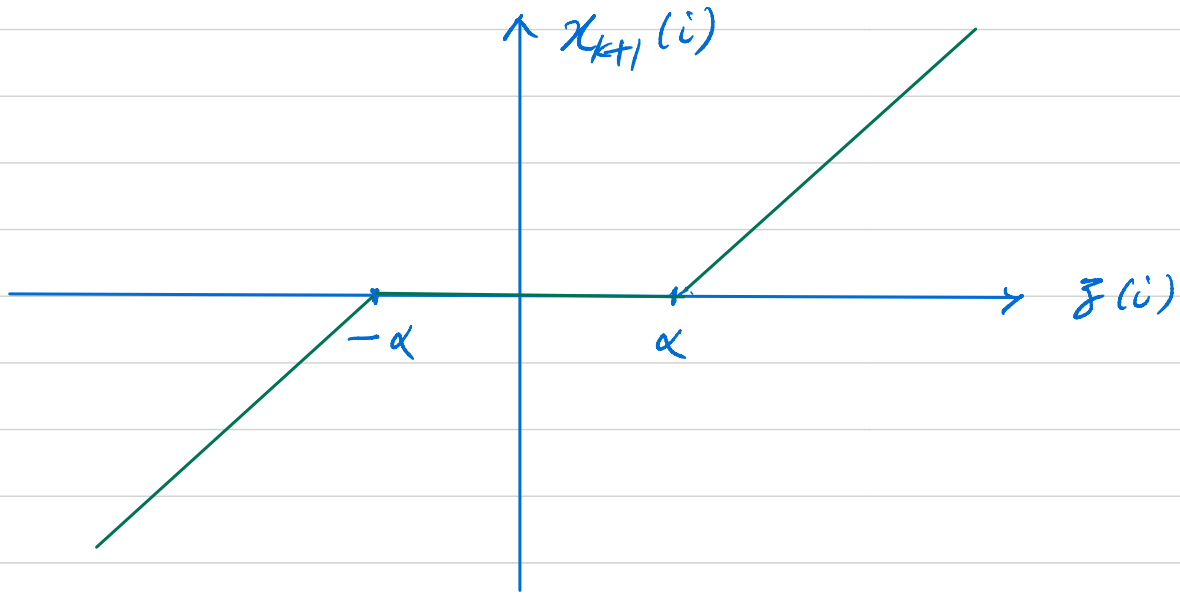
$\partial C(0) = -\frac{z(i)}{\alpha} + v, \quad v \in [-1, 1]$

$0 \in \partial C(0) \Leftrightarrow -\frac{z(i)}{\alpha} + v = 0$ for some $v \in [-1, 1]$

$\Leftrightarrow z(i) \in [-\alpha, \alpha]$.

Thus,

$$x_{k+1}(i) = x^*(i) = \begin{cases} z(i) - \alpha & \text{if } z(i) > \alpha \\ 0 & \text{if } -\alpha \le z(i) \le \alpha \\ z(i) + \alpha & \text{if } z(i) < -\alpha \end{cases}$$



$x_{k+1}(i)$ is obtained by "shrinking" $z(i)$ to 0 in α-neighborhood of 0.

$$x_{k+1}(i) = \mathcal{S}_\alpha(z(i))$$

↑ shrinkage operator

Problems of the form:

$$\underset{x}{\text{minimize}} \quad g(x) + \lambda \|x\|_1$$

↑ convex, cont. diff.

arise in many modern applications:
LASSO, compressed sensing, etc.