Proximal gradient method:

Proximal gradient method

Suppose $g: \mathbb{R}^{n\times n} \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. For $\alpha > 0$ proximal operator of g at $\bar{X}$ is defined as

$$\text{prox}_g(\bar{X}) = \underset{X}{\text{argmin}}\ g(X) + \frac{1}{2\alpha}\|X - \bar{X}\|^2$$

Since g is a convex function $\theta(X) = g(X) + \frac{1}{2\alpha}\|X - \bar{X}\|^2$ is a strongly convex function. Hence $\min_X \theta(X)$ has unique finite solution, i.e. proximal operator is well defined.

Examples:

① $g(x) = c$.

Then $\text{prox}_{c,g}(\bar{x}) = \underset{x}{\text{argmin}}\ c + \frac{1}{2\alpha}(x - \bar{x})^2$

$= \bar{x}$

② $g(x) = c^T x$

If $x^* = \text{prox}_{c,g}(\bar{x})$

$= \underset{x}{\text{argmin}}\ c^T x + \frac{1}{2\alpha}\|x - \bar{x}\|^2$

then $c + \frac{1}{\alpha}(x^* - \bar{x}) = 0$

$\Rightarrow x^* - \bar{x} = -\alpha c$.

$\Rightarrow \text{prox}_{c,g}(\bar{x}) = \bar{x} - \alpha c^T x$

If $x^* = \text{prox}_{c,g}(\bar{x})$

then $\partial x^* + c + \frac{1}{\alpha}(x^* - \bar{x}) = 0$

$\Rightarrow x^* - \bar{x} = (\partial x^* + c)$

For a convex function g, $x^* = \text{prox}_{c,g}(\bar{x})$ iff $\frac{1}{\alpha}(\bar{x} - x^*) \in \partial g(x^*)$

$\Rightarrow$ Given that

$x^* = \text{prox}_{c,g}(\bar{x})$

$= \underset{x}{\text{argmin}}\ g(x) + \frac{1}{2\alpha}\|x - \bar{x}\|^2$

This implies

$0 \in \partial g(x^*) + \frac{1}{\alpha}(x^* - \bar{x})$

$\Rightarrow \frac{1}{\alpha}(\bar{x} - x^*) \in \partial g(x^*)$

If $x^* = \underset{x}{\text{argmin}}\ g(x)$ then $\text{prox}_{c,g}(x^*) = x^*$ and vice-versa.

$\Rightarrow$ Since $x^* = \underset{x}{\text{argmin}}\ g(x)$.

$0 \in \partial g(x^*)$

$\Rightarrow \frac{1}{\alpha}(x^* - x^*) \in \partial g(x^*)$

Form above stmt.,

$x^* = \text{prox}_{c,g}(x^*)$.

Suppose $g(x) = |x|$

For $\bar{x} \geq 0$

$\text{prox}_{c,g}(\bar{x}) = \underset{x}{\text{argmin}}\ x + \frac{1}{2\alpha}(x - \bar{x})^2$

$= \begin{cases} \bar{x} - \alpha & \text{if } \bar{x} \geq \alpha \\ 0 & \text{if } 0 \leq \bar{x} \leq \alpha \end{cases}$

For $\bar{x} < 0$,

$\text{prox}_{c,g}(\bar{x}) = \underset{x}{\text{argmin}} - x + \frac{1}{2\alpha}(x - \bar{x})^2$

$= \begin{cases} \bar{x} + \alpha & \text{if } \bar{x} \leq -\alpha \\ 0 & \text{if } -\alpha \leq \bar{x} \leq 0 \end{cases}$

Combining both cases:

$\text{prox}_{c,g}(\bar{x}) = \begin{cases} 0 & \text{if } -\alpha \leq \bar{x} \leq \alpha \\ \bar{x} - \alpha & \text{if } \bar{x} \geq \alpha \\ \bar{x} + \alpha & \text{if } \bar{x} \leq -\alpha \end{cases}$

If $g(x) = \|X\|_1 = \sum_{i,j=1}^{n} |X_{ij}|$

Suppose $X^* = \text{prox}_{c,g}(\bar{X})$

then from above result,

$X^*_{ij} = \begin{cases} \bar{X}_{ij} - \alpha & \text{if } \bar{X}_{ij} \geq \alpha \\ \bar{X}_{ij} + \alpha & \text{if } \bar{X}_{ij} \leq -\alpha \\ 0 & \text{if } -\alpha \leq \bar{X}_{ij} \leq \alpha \end{cases}$

Proximal Gradient Method

Consider the composite optimization problem:

$$\min_x f(x) + g(x)$$

where $f: \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable and $g: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper, convex but not necessarily differentiable.

If we want to generate a sequence $\{x^k\}$ by using linear approximation of f (factor to gradient descent) then

$$x^{k+1} = \underset{x}{\text{argmin}}\ g(x) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2\alpha}\|x - x^k\|^2$$

Claim: $x^{k+1} = \text{prox}_{c,g}(x^k - \alpha \nabla f(x^k))$

$\text{Proof}_{c,g}(x^k - \alpha \nabla f(x^k))$

$= \underset{x}{\text{argmin}}\ g(x) + \frac{1}{2\alpha}\|x - x^k + \alpha \nabla f(x^k)\|^2$

$= \underset{x}{\text{argmin}}\ g(x) + \frac{1}{2\alpha}\|x - x^k\|^2 + \nabla f(x^k)^T(x - x^k)$

$= x^{k+1}$

[Ignoring constant terms in 2nd equality].

If $x^* \in \underset{x}{\text{argmin}}\ f(x) + g(x)$

then $0 \in \{\nabla f(x^*) + \partial g(x^*)\}$

$\Rightarrow -\nabla f(x^*) \in \partial g(x^*)$ ①

Suppose $x = \text{prox}_{c,g}(x^* - \alpha \nabla f(x^*))$

then $\frac{1}{\alpha}(x^* - x^* + \alpha \nabla f(x^*)) \in \partial g(x^*)$ ②

Comparing ① & ②

$\bar{x} = x^*$

Hence if x^* is optimal sol^n of $\min_x f(x) + g(x)$

iff

$x^* = \text{prox}_{c,g}(x^* - \alpha \nabla f(x^*))$

ℓ_1/ℓ_2 regularization

Consider a data set $D = \{(a^1, y_1), \ldots, (a^m, y_m)\}$

If $y_i = h(a^i; x)$ then error is,

$\frac{1}{2}\sum_{i=1}^{m} \left(h(a^i; x) - y_i\right)^2 = f(x)$

In most cases we use a regularization function with f.

ℓ_2 - regularization

$$\min_x f(x) + \frac{\lambda}{2}\|x\|_2 = \theta(x)$$

Clearly θ is differentiable.

If x^* is optimal sol^n then

$\nabla f(x^*) + \lambda x^* = 0$

least square problem

$$\theta(x) = \frac{1}{2}\|Ax - b\|^2 + \frac{\lambda}{2}\|x\|^2.$$

Suppose x^* is optimal then

$A^T(Ax^* - b) + \lambda x^* = 0$

$\Rightarrow (A^TA + \lambda I)\, x^* = A^T b$

$\Rightarrow x^* = (A^TA + \lambda I)^{-1} A^T b$

One can observe that θ is a λ-strongly convex function.

$A^TA + \lambda I$ is differentiable even column vector of A is linearly dependent.

Similar result can be derived for logistic regression.

ℓ_1 regularization

Consider

$$\min_x f(x) + \lambda\|x\|_1 = \theta(x) \quad ①$$

Clearly θ is a composite function with $g(x) = \lambda\|x\|_1$.

② Can be written as

$$\min_x f(x) + \lambda\|x\|_1$$

Suppose we generate a sequence $\{x^k\}$ using proximal gradient method then

$x^{k+1} = \text{prox}_{\alpha g}\left(x^k - \alpha \nabla f(x^k)\right)$

Practice problem:

① Find $\text{prox}_{c,g}(\bar{x})$

i) $g(x) = \frac{1}{2}x^2 - 2x + 2x^2$

$\alpha = 0.5$, $\bar{x} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$

ii) $g(x) = x_1 x_2 x_1 + x_1 \log x_2$

$\alpha = 1$, $\bar{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

iii) $g(x) = \sum_{i=1}^{n} -\log(x_i)$

$\alpha = 0.5$, $\bar{x} = (1, 1, \ldots, 1)$

② Consider the problem

$$\min_x f(x) + g(x)$$

If $x^{k+1} = \underset{x}{\text{min}}\ g(x) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k)$

then show that

$x^{k+1} = \text{prox}_g\left(x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k)\right)$

③ Suppose

$g(x) = \max\{x_1 + x_2, 2x_1 - x_2\}$

Show that $\text{prox}_g\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}$

Hint: use $\partial - x^* \in \partial g(x^*)$

④ Consider the data set

$D = \{(-1, 0), (1, 2.5), (2, -3), (-2, 2.5)\}$

Suppose we want to find best fitting curve $y = \theta_1 x + \theta_2$ using

$\theta(x) = \frac{1}{2}\|A - b\|^2 + \lambda\|\theta_2\|$

where $\lambda = \left\lceil \frac{R}{10} - 5 \right\rceil$, R is last two digits of your roll number.

perform two iterations of proximal gradient method with $\theta^0\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\alpha_k = \frac{1}{10k}$

⑤ Define

$g(x) = \max\{x_1 + x_2 + x_3 + 1,\ x_1 + x_2 - x_3 + 1\}$

Show that

$\text{prox}_g\left(\begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

⑥ Define

$g(x) = \max\{-x_1 + x_2,\ -x_1 + x_2\}$

Show that

$\text{prox}_g\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$