Proximal Gradient Descent: convergence

Friday, March 12, 2021 6:40 AM

$$\min_x \; f(x) := \underset{\nwarrow \text{ smooth}}{g(x)} + \underset{\nwarrow \text{ easy proximity operator}}{h(x)}$$

Assume ∇g is Lipschitz continuous

WLOG let Lipschitz constant be $L = 1$ for simplicity

(ie., redefine $\tilde{f}(x) = \frac{1}{L} \cdot f(x)$)

Assume $g, h \in \Gamma_0(\mathbb{R}^n)$

Algorithm: $x_{k+1} = \text{prox}_h(x_k - \nabla g(x_k))$ ... or if $L \neq 1$, use stepsize $t = 1/L$

$$x_{k+1} = \text{prox}_{t \cdot h}(x_k - t \nabla g(x_k))$$

Analysis:

introduce the gradient map $G(x) = x - \text{prox}_h(x - \nabla g(x))$

ex: $h(x) = 0 \Rightarrow \text{prox}_h(y) = y$ so $G(x) = \nabla g(x)$

thus the algo. can be written as

$$x_{k+1} = x_k - G(x) \quad \text{... looks like gradient descent.}$$

Property of prox

let $w = \text{prox}_h(x) = \underset{w}{\text{argmin}} \; \frac{1}{2}\|w - x\|^2 + h(w)$

ie. (Fermat's rule) $0 \in w - x + \partial h(w)$

so $\boxed{x - w \in \partial h(w) \text{ if } w = \text{prox}_h(x)}$ $(*)$

Key inequality (via descent lemma)

Since g is 1-Lipschitz, the descent lemma says

$$g(y) \leq g(x) + \langle \nabla g(x), y - x \rangle + \frac{1}{2}\|y - x\|^2$$

hence

$$f(y) = g(y) + \underline{h(y)} \leq g(x) + \underline{h(y)} + \langle \nabla g(x), y - x \rangle + \frac{1}{2}\|y - x\|^2$$

So, thinking of x as x_k, and $y = x - G(x)$, this means

$$f(y) \le \underbrace{g(x)}_{\le} + \underbrace{h(\overbrace{x-G(x)}^{y})}_{\le} + \langle \nabla g(x), G(x) \rangle + \tfrac{1}{2}\|G(x)\|^2$$

via convexity and definition of subgradients

$$g(z) + \langle \nabla g(x), x - z \rangle \qquad h(z) + \langle v, y - z \rangle$$

where $v = G(x) - \nabla g(x) \in \partial h(y)$

since $G(x) = x - \text{prox}_h(x - \nabla g(x))$, i.e.,

$y = x - G(x) = \text{prox}_h(x - \nabla g(x))$ so via $(*)$

$$(x - \nabla g(x)) - (x - G(x)) \in \partial h(x - G(x)) = \partial h(y)$$

combine to form $f(z)$

Simplifying:

$$\forall z \quad f(y) \le f(z) + \langle G(x), x - z \rangle - \tfrac{1}{2}\|G(x)\|^2 \qquad \star$$

recall $x_{k+1} = y = x - G(x)$, $x = x_k$

if $z = x \implies f(y) \le f(x) - \tfrac{1}{2}\|G(x)\|^2$, i.e., a descent method.

if $z = x^* \in \text{argmin } f(x)$

$$\implies f(y) - f^* \le \langle G(x), x - x^* \rangle - \tfrac{1}{2}\|G(x)\|^2$$

$$= \tfrac{1}{2}\left(\|x - x^*\|^2 - \|x - x^* - G(x)\|^2\right)$$

$$= \tfrac{1}{2}\left(\|x - x^*\|^2 - \|y - x^*\|^2\right)$$

For $y = x_{k+1}$, $x = x_k$, now sum for $k = 1, \dots, K$

$$\underbrace{\sum_{k=1}^{K} f(x_k) - f^*} \le \tfrac{1}{2} \sum_{k=1}^{K} \|x_{k-1} - x^*\|^2 - \|x_k - x^*\|^2$$

telescopes!

since a descent method,

$$\le \tfrac{1}{2}\left(\|x_0 - x^*\|^2 - \underbrace{\|x_K - x^*\|^2}_{\text{unknown, but } \ge 0}\right)$$

$$f(x_K) - f^* \le \frac{1}{K}\sum_{k=1}^{K} f(x_k) - f^* \le \tfrac{1}{2}\|x_0 - x^*\|^2$$

$$\implies f(x_K) - f^* \le \frac{1}{2K}\|x_0 - x^*\|^2$$

i.e., if $L \ne 1$,

$$\boxed{f(x_k) - f^* \le \frac{L}{2K}\|x_0 - x^*\|^2}$$