# Proximal Gradient Descent: convergence [2025 update]

Friday, March 7, 2025 10:44 AM

$$\min_x \quad F(x) := \underbrace{g(x)}_{\text{smooth}} + \underbrace{h(x)}_{\text{easy proximity operator}}$$

Includes gradient descent as special case (but doesn't analyze strongly convex case)

Assume $\nabla g$ 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: $x_{k+1} = \text{prox}_h(\overbrace{x_k - \nabla g(x_k)}^{\tilde{x}})$ ... 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

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

Property of prox

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

i.e. if $y = x_{k+1}$, ie. (Fermat's rule) $0 \in y - \tilde{x} + \partial h(y)$

$y = \text{prox}_h(\tilde{x} := x_k - \nabla g(x_k))$ so $\boxed{\tilde{x} - y \in \partial h(y) \text{ if } y = \text{prox}_h(\tilde{x})}$ $(*)$

Key inequality (via descent lemma)

Since $g$ is 1-Lipschitz, the descent lemma says

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

hence

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

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

$= \text{prox}_h(\tilde{x})$

# Proximal Gradient Descent: convergence (p. 2)

Friday, March 7, 2025 10:44 AM

use $\forall z,\ g(z) \geq g(x) + \langle \nabla g(x), z-x \rangle$

$$F(y) \leq \underbrace{g(x) + \langle \nabla g(x), -G(x) \rangle}_{\leq\ g(z) + \langle \nabla g(x), x-z \rangle} + \underbrace{h(\overbrace{x-G(x)}^{y})}_{\leq\ h(z) + \langle v, y-z \rangle} + \frac{1}{2}\|G(x)\|^2$$

via convexity and definition of subgradients

$h(z) + \langle v, y-z \rangle$ where $v \in \partial h(y)$

well, note that $v = G(x) - \nabla g(x) \in \partial h(y)$

since $G(x) = x - \mathrm{prox}_h(\overbrace{x - \nabla g(x)}^{\tilde{x}})$, i.e.,

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

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

combine to form $f(z)$

⚠ SUBTLE... Go slowly!

so...

$$F(y) \leq g(z) + \langle \nabla g(x), x-z \rangle + h(z) + \langle G(x) - \nabla g(x), \underbrace{y}_{=x-G(x)} - z \rangle - \langle \nabla g(x), G(x) \rangle + \frac{1}{2}\|G(x)\|^2$$

$$= F(z) + \langle \cancel{\nabla g(x)}, x-z \rangle + \langle G(x) - \cancel{\nabla g(x)}, x-z \rangle - \langle G(x) - \cancel{\nabla g(x)}, G(x) \rangle - \langle \cancel{\nabla g(x)}, G(x) \rangle + \frac{1}{2}\|G(x)\|^2$$

cancel cancel

$$= F(z) + \langle G(x), x-z \rangle - \langle G(x), G(x) \rangle + \frac{1}{2}\|G(x)\|^2$$

$$= F(z) + \langle G(x), x-z \rangle - \frac{1}{2}\|G(x)\|^2$$

i.e. $\forall z$

$$\boxed{F(y) \leq F(z) + \langle G(x), x-z \rangle - \frac{1}{2}\|G(x)\|^2 \qquad \bigstar}$$

:

# Proximal Gradient Descent: convergence (p. 3)

Friday, March 7, 2025 10:44 AM

Recall $F = g + h$

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

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 $z = x^* \in \operatorname{argmin} F(x)$

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

complete-the-square

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

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

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

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

telescopes!

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

$$\le \tfrac{1}{2}\|x_0 - x^*\|^2.$$

since a descent method,

$$\boxed{F(x_K) - F^* \le \frac{1}{K}\sum_{k=1}^{K} F(x_k) - F^*}$$

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

or if we "undo" our scaling by $1/L$,

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

i.e. $O(K)$ convergence rate

PICTURE