Proximal Gradient Method. Proximal operator

Seminar

Optimization for ML. Faculty of Computer Science. HSE University

Regularized / Composite Objectives

Many nonsmooth problems take the form

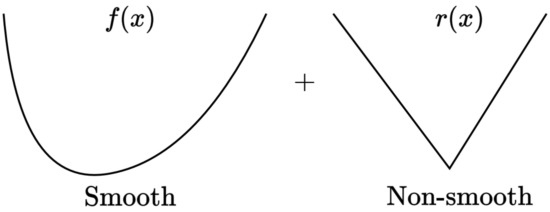
$$\min_{x \in \mathbb{R}^n} \varphi(x) = f(x) + r(x)$$

- **Lasso, L1-LS, compressed sensing**

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2, r(x) = \lambda\|x\|_1$$

- **L1-Logistic regression, sparse LR**

$$f(x) = -y \log h(x) - (1-y)\log(1-h(x)), r(x) = \lambda\|x\|_1$$



Non-smooth convex optimization lower bounds

convex (non-smooth)	strongly convex (non-smooth)
$f(x_k) - f^* \sim \mathcal{O}\left(\dfrac{1}{\sqrt{k}}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\dfrac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* \sim \mathcal{O}\left(\dfrac{1}{k}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\dfrac{1}{\varepsilon}\right)$

Non-smooth convex optimization lower bounds

convex (non-smooth)	strongly convex (non-smooth)
$f(x_k) - f^* \sim \mathcal{O}\left(\dfrac{1}{\sqrt{k}}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\dfrac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* \sim \mathcal{O}\left(\dfrac{1}{k}\right)$ $k_\varepsilon \sim \mathcal{O}\left(\dfrac{1}{\varepsilon}\right)$

- Subgradient method is optimal for the problems above.
- One can use Mirror Descent (a generalization of the subgradient method to a possiby non-Euclidian distance) with the same convergence rate to better fit the geometry of the problem.
- However, we can achieve standard gradient descent rate $\mathcal{O}\left(\frac{1}{k}\right)$ (and even accelerated version $\mathcal{O}\left(\frac{1}{k^2}\right)$) if we will exploit the structure of the problem.

Proximal operator

Proximal operator

For a convex set $E \in \mathbb{R}^n$ and a convex function $f : E \rightarrow \mathbb{R}$ operator $\text{prox}_f(x)$ s.t.

$$\text{prox}_f(x) = \underset{y \in E}{\text{argmin}} \left[f(y) + \frac{1}{2} ||y - x||_2^2 \right]$$

is called **proximal operator** for function f at point x

From projections to proximity

Let $\mathbb{I}_S$ be the indicator function for closed, convex S. Recall orthogonal projection $\pi_S(y)$

From projections to proximity

Let $\mathbb{I}_S$ be the indicator function for closed, convex S. Recall orthogonal projection $\pi_S(y)$

$$\pi_S(y) := \arg\min_{x \in S} \frac{1}{2}\|x - y\|_2^2.$$

From projections to proximity

Let $\mathbb{I}_S$ be the indicator function for closed, convex S. Recall orthogonal projection $\pi_S(y)$

$$\pi_S(y) := \arg\min_{x \in S} \frac{1}{2}\|x - y\|_2^2.$$

With the following notation of indicator function

$$\mathbb{I}_S(x) = \begin{cases} 0, & x \in S, \\ \infty, & x \notin S, \end{cases}$$

Rewrite orthogonal projection $\pi_S(y)$ as

$$\pi_S(y) := \arg\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x - y\|^2 + \mathbb{I}_S(x).$$

From projections to proximity

Let $\mathbb{I}_S$ be the indicator function for closed, convex S. Recall orthogonal projection $\pi_S(y)$

$$\pi_S(y) := \arg\min_{x \in S} \frac{1}{2}\|x - y\|_2^2.$$

With the following notation of indicator function

$$\mathbb{I}_S(x) = \begin{cases} 0, & x \in S, \\ \infty, & x \notin S, \end{cases}$$

Rewrite orthogonal projection $\pi_S(y)$ as

$$\pi_S(y) := \arg \min_{x \in \mathbb{R}^n} \frac{1}{2}\|x - y\|^2 + \mathbb{I}_S(x).$$

Proximity: Replace $\mathbb{I}_S$ by some convex function!

$$\text{prox}_r(y) = \text{prox}_{r,1}(y) := \arg\min \frac{1}{2}\|x - y\|^2 + r(x)$$

Proximal Gradient Method

Proximal Gradient Method Theorem

Consider the proximal gradient method

$$x_{k+1} = \text{prox}_{\alpha r}\left(x_k - \alpha \nabla f(x_k)\right)$$

for the criterion $\phi(x) = f(x) + r(x)$ s.t.: 1. f is convex, differentiable with Lipschitz gradients; 1. r is convex and prox-friendly. Then Proximal Gradient Method with fixed step size $\alpha = \frac{1}{L}$ converges with rate $O(\frac{1}{k})$

ISTA and FISTA

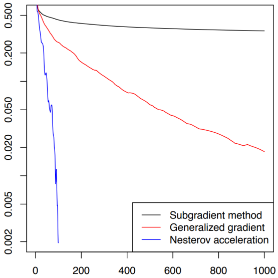
Methods for solving problems involving $L1$ regularization (e.g. Lasso).

ISTA (Iterative Shrinkage-Thresholding Algorithm)

- Step:

$$x_{k+1} = \text{prox}_{\alpha\lambda||\cdot||_1}\left(x_k - \alpha\nabla f(x_k)\right)$$

- Convergence: $O(\frac{1}{k})$



FISTA (Fast Iterative Shrinkage-Thresholding Algorithm)

- Step:

$$x_{k+1} = \text{prox}_{\alpha\lambda||\cdot||_1}\left(y_k - \alpha\nabla f(y_k)\right),$$

$$t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2},$$

$$y_{k+1} = x_{x+1} + \frac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k)$$

- Convergence: $O(\frac{1}{k^2})$

Problem 1. ReLU in prox

Question

Find the $\text{prox}_f(x)$ for $f(x) = \lambda \max(0, x)$:

$$\text{prox}_{\lambda \max(0,\cdot)}(x) = \underset{y \in \mathbb{R}}{\text{argmin}} \left[\frac{1}{2} ||y - x||^2 + \lambda \max(0, y) \right]$$

Problem 2. Grouped l_1-regularizer

Question

Find the $\text{prox}_f(x)$ for $f(x) = ||x||_{1/2} = \sum_{g=0}^{G} ||x_g||_2$ where $x \in \mathbb{R}^n = [\underbrace{x_1, x_2}_{1}, \ldots, \underbrace{\ldots}_{g}, \ldots, \underbrace{x_{n-2}, x_{n-1}, x_n}_{G}]$:

$$\text{prox}_{||x||_{1/2}}(x) = \underset{y \in \mathbb{R}}{\text{argmin}} \left[\frac{1}{2}||y - x||_2^2 + \sum_{g=0}^{G} ||y_g||_2 \right]$$

Linear Least Squares with L_1-regularizer

Proximal Methods Comparison for Linear Least Squares with L_1-regularizer Open in Colab.