Accelerated Proximal-Gradient Method for Large Scale Convex Problems

Masoud Farivar

January 28, 2015

Outline

Proximal mapping

proximal mapping (or proximal operator) of a convex function h is

$$\text{prox}_t(x) = \underset{u}{\text{argmin}} \left(h(u) + \frac{1}{2t}\|u - x\|_2^2 \right)$$

Examples

$h(x) = 0 :\quad \text{prox}(x) = x$

$h(x) = I_C(x)$ (indicator function of C) : prox is projection on C

$$\text{prox}(x) = P_C(x) = \underset{u \in C}{\text{argmin}} \; \|u - x\|_2^2$$

$h(x) = \|x\|_1 :\quad \text{prox}_h$ is the soft thresholding (shrinkage) function

$$\text{prox}_t(x)_i = S_t(x_i) = (|x_i| - t)_+ \text{sign}(x_i) = \begin{cases} x_i - t & x_i \geq t \\ 0 & |x_i| \leq t \\ x_i + t & x_i \leq -t \end{cases}$$

Gradient Descent (Convergence)

Gradient Descent:

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

Definition. f is β-smooth when the gradient mapping ∇f is β-Lipschitz, i.e., $\forall x, y \in \mathbb{R}^n$

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$

Let f be a convex and β-smooth function on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$, one has

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \frac{1}{2\beta} \|\nabla f(x) - \nabla f(y)\|^2$$

Theorem

Assume that f is a continuously differentiable β-smooth and convex on $\mathbb{R}^n$. Then Gradient Descent with $\eta = \frac{1}{\beta}$ converges with O(1/t), i.e.,

$$f(x_k) - f^* \leq \frac{2\beta \|x_1 - x^*\|^2}{k+3}$$

Gradient Descent (Analysis)

Consider the following optimization problem,

$$\underset{x}{\text{minimize}} \quad g(x)$$

At iteration x_k we use a quadratic upper bound on g,

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \quad g(x_k) + \langle \nabla g(x_k), x - x_k \rangle + \frac{\beta}{2} \|x - x_k\|^2$$

We can equivalently write this as the quadratic optimization

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \quad \frac{1}{2} \|x - (x_k - \eta \nabla g(x_k))\|^2$$

where $\eta = \frac{1}{\beta}$. The yields the Gradient Descent algorithm:

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

First-order

The basic Gradient Descent has two disadvantages: 1) it can't be applied to optimize nondifferentiable functions, 2) slow convergence rate

Approaches to address these issues:

methods with improved convergence

- accelerated gradient method
- quasi-Newton methods
- conjugate gradient method

methods for nondifferentiable or constrained problems

- proximal gradient method
- subgradient method
- smoothing methods
- cutting-plane methods

Nesterov's Accelerated Gradient Descent

Consider the following sequences:

$$
\begin{aligned}
\lambda_s &= \frac{1+\sqrt{1+4\lambda_{s-1}^2}}{2}, \qquad \lambda_0 = 0, \\
\gamma_s &= \frac{1-\lambda_s}{\lambda_s+1}
\end{aligned}
$$

Nesterov's Accelerated Gradient Descent steps:

$$
\begin{aligned}
y_{s+1} &= x_s - \frac{1}{\beta}\nabla f\left(x_s\right), \qquad y_1 = x_1, \\
x_{s+1} &= \left(1-\gamma_s\right) y_{s+1} + \gamma_s y_s
\end{aligned}
$$

Nesterov's Accelerated Gradient Descent

Intuitively, Nesterov's Accelerated Gradient Descent performs a simple step of gradient descent to go from x_s to y_{s+1}:

$$y_{s+1} \quad = \quad x_s - \frac{1}{\beta} \nabla f\left(x_s\right)$$

and then it "slides" a little bit further than y_{s+1} in the direction given by the previous point y_s:

$$x_{s+1} \quad = \quad \left(1 - \gamma_s\right) y_{s+1} + \gamma_s y_s$$

Rate of convergence is $O(1/t^2)$ after t steps:

Theorem (Nesterov 1983)

Let f be a convex and β-smooth function, then Nesterov's Accelerated Gradient Descent satisfies:

$$f(y_t) - f(x^*) \leq \frac{2\beta \|x_1 - x^*\|^2}{t^2}$$

Composite Convex Optimization Problems

We consider composite optimization problems

$$\underset{x}{\text{minimize}} \quad f(x) := g(x) + h(x),$$

where g and h are convex but h is non-smooth.

Typically, g is a data-fitting term, and h is a regularizer,

The most well-studied example is ℓ_1-regularized least squares,

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|^2 + \lambda \|x\|_1.$$

Proximal-Gradient Method

Consider the following composite optimization problem,

$$\underset{x}{\text{minimize}} \quad g(x) + h(x)$$

At iteration x_k we use a quadratic upper bound on g,

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \quad g(x_k) + \langle \nabla g(x_k), x - x_k \rangle + \frac{\beta}{2} \|x - x_k\|^2 + h(x)$$

We can equivalently write this as the proximal quadratic optimization ($\eta = \frac{1}{\beta}$)

$$x_{k+1} = \underset{x}{\operatorname{argmin}} \quad \frac{1}{2} \|x - (x_k - \eta \nabla g(x_k))\|^2 + \eta h(x)$$

The solution is the proximal-gradient algorithm:

$$x_{k+1} = \operatorname{prox}_{\eta}[x_k - \eta \nabla g(x_k)]$$

Iterative Soft-Thresholding (ISTA)

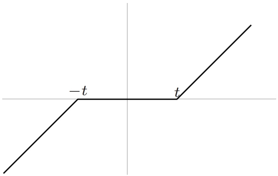
Consider lasso criterion

$$\text{minimize} \quad \frac{1}{2}\|y - Ax\|^2 + t\|x\|_1$$

Prox function is now

$$\text{prox}_t(x) = S_t(x) = (|x| - t)_+ \text{sign}(x).$$

where $S_t(x)$ is the soft-thresholding function discussed earlier:



FISTA

Let us now combine Nesterovs Accelerated Gradient Descentn with ISTA, i.e.,

$$\lambda_0 = 0, \;\; \lambda_k = \frac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2}, \text{ and } \gamma_k = \frac{1 - \lambda_k}{\lambda_{k+1}}.$$

Let $x_1 = y_1$ an arbitrary initial point, and

$$\begin{aligned} y_{k+1} &= \operatorname{prox}_\eta[x_k - \eta \nabla g(x_k)] \\ x_{k+1} &= (1 - \gamma_k) y_{k+1} + \gamma_k y_k. \end{aligned}$$

The convergence rate of FISTA is similar to Nesterovs Accelerated Gradient Descent: $O(1/k^2)$

References

[1] Beck, Amir, and Marc Teboulle, "A fast iterative shrinkage -thresholding algorithm for linear inverse problems." SIAM Journal on Imaging Sciences 2.1, 183-202, 2009.

[2] Schmidt, Mark, Nicolas L. Roux, and Francis R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization", Advances in neural information processing systems, 2011.

[3] Boyd, Stephen, and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2009.

[4] Course notes form Berkeley's EE227BT.