Optimization for Machine Learning
机器学习中的优化方法

陈 程

华东师范大学 软件工程学院

chchen@sei.ecnu.edu.cn

Final project

- Projects will be evaluated based on a combination of:
  - presentation (40%) at Tuesday of the 18th week
  - report (60%), deadline: Tuesday of the 19th week
- Projects can either be individual or in teams of size up to 3 students.

Plagiarism is forbidden!

Final project

Types of projects:

- optimization in application
- methodology projects
- survey projects
- a new algorithm

Review

condition	stepsize	convergence rate	iteration complexity
convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$
strongly convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\left(1-\frac{1}{\kappa}\right)^t\right)$	$O(\kappa \log \frac{1}{\varepsilon})$

Table: Convergence Properties of GD & PGD

	stepsize	convergence rate	iteration complexity
convex	$\eta_t \approx \frac{1}{\sqrt{t}}$	$O\left(\frac{1}{\sqrt{t}}\right)$	$O(\frac{1}{\varepsilon^2})$
strongly convex	$\eta_t \approx \frac{1}{t}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

Table: Convergence Properties of Subgradient Descent

Outline

Composite problems

$$min_{\mathbf{x}} F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$$

- f is convex and smooth
- h is convex (may not be differentiable)
- Let $F^* = \min_{\mathbf{x}} F(\mathbf{x})$ be the optimal value

Example: ℓ_1 regularized minimization:

$$\min_{\mathbf{x}} f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1$$

use ℓ_1 regularization to promote sparsity

A proximal view of of gradient descent

We first revisit gradient descent

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$$

$$\Updownarrow$$

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}} \left\{ \underbrace{f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle}_{\text{first-order approximation}} + \underbrace{\frac{1}{2\eta_t} \|\mathbf{x} - \mathbf{x}_t\|_2^2}_{\text{proximal term}} \right\}$$

By the optimality condition, $\mathbf{x}_{t+1}$ is the point where $f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle$ and $-\frac{1}{2\eta_t} \|\mathbf{x} - \mathbf{x}_t\|_2^2$ have the same slope.

How about projected gradient descent?

$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t))$$

$$\Updownarrow$$

$$\begin{aligned}
\mathbf{x}_{t+1} &= \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{1}{2\eta_t} \|\mathbf{x} - \mathbf{x}_t\|_2^2 + \mathbb{1}_{\mathcal{C}}(\mathbf{x}) \right\} \\
&= \arg\min_{\mathbf{x}} \left\{ \frac{1}{2} \|\mathbf{x} - (\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t))\|_2^2 + \eta_t \mathbb{1}_{\mathcal{C}}(\mathbf{x}) \right\}
\end{aligned}$$

where

$$\mathbb{1}_{\mathcal{C}}(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathcal{C} \\ +\infty, & \text{otherwise} \end{cases}$$

Proximal operator (近端算子)

Define the proximal operator

$$\text{prox}_h(\mathbf{x}) \triangleq \arg\min_{\mathbf{z}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 + h(\mathbf{z}) \right\}$$

for any convex function h.

Then, the update of projected gradient descent is

$$\mathbf{x}_{t+1} = \text{prox}_{\eta_t \mathbb{1}_{\mathcal{C}}}(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t))$$

Proximal operator (近端算子)

Define the proximal operator

$$\mathrm{prox}_h(\mathbf{x}) \triangleq \arg\min_{\mathbf{z}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 + h(\mathbf{z}) \right\}$$

for any convex function h.

Then, the update of projected gradient descent is

$$\mathbf{x}_{t+1} = \mathrm{prox}_{\eta_t \mathbb{1}_{\mathcal{C}}}(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t))$$

Proximal gradient descent (近端梯度下降法)

In each iteration, the proximal gradient descent method for composite objective function $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$ computes

$$\mathbf{x}_{t+1} = \mathrm{prox}_{\eta_t h}(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)).$$

- alternates between gradient updates on f and proximal minimization on h
- useful if the prox_h can be efficiently computed

Proximal gradient descent (近端梯度下降法)

In each iteration, the proximal gradient descent method for composite objective function $F(\mathbf{x}) = f(\mathbf{x}) + h(\mathbf{x})$ computes

$$\mathbf{x}_{t+1} = \text{prox}_{\eta_t h}(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)).$$

- alternates between gradient updates on f and proximal minimization on h
- useful if the prox_h can be efficiently computed

Outline

Proximal operator

$$\operatorname{prox}_h(\mathbf{x}) \triangleq \arg\min_{\mathbf{z}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 + h(\mathbf{z}) \right\}$$

- well-defined under very general conditions (including nonsmooth convex functions)
- can be evaluated efficiently for many widely used functions (in particular, regularizers)
- this abstraction is mathematically simple but covers many well-known optimization algorithms

Example: indicator functions

If $h(\mathbf{x}) = \mathbb{1}_{\mathcal{C}}$ is the "indicator" function

$$h(\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{x} \in \mathcal{C} \\ +\infty, & \text{otherwise} \end{cases}$$

then

$$\operatorname{prox}_h(\mathbf{x}) = \arg\min_{\mathbf{z} \in \mathcal{C}} \|\mathbf{z} - \mathbf{x}\|_2^2 \quad \text{(Euclidean projection)}$$

Example: ℓ_1 Norm

If $h(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, then

$$(\text{prox}_{\lambda h}(\mathbf{x}))_i = \psi_{st}(x_i; \lambda) \quad \text{soft-thresholding}$$

where

$$\psi_{st}(x) = \begin{cases} x - \lambda, & \text{if } x > \lambda \\ x + \lambda, & \text{if } x < -\lambda \\ 0, & \text{otherwise} \end{cases}$$

Example: ℓ_1 Norm

If $h(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$, then

$$(\text{prox}_{\lambda h}(\mathbf{x}))_i = \psi_{st}(x_i; \lambda) \quad \text{soft-thresholding}$$

where

$$\psi_{st}(x) = \begin{cases} x - \lambda, & \text{if } x > \lambda \\ x + \lambda, & \text{if } x < -\lambda \\ 0, & \text{otherwise} \end{cases}$$

Basic rules of proximal operator

- **affine addition:** if $f(\mathbf{x}) = g(\mathbf{x}) + \mathbf{a}^\top \mathbf{x} + b$, then

$$\mathrm{prox}_f(\mathbf{x}) = \mathrm{prox}_g(\mathbf{x} - \mathbf{a})$$

- **quadratic addition:** if $f(\mathbf{x}) = g(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{a}\|_2^2$, then

$$\mathrm{prox}_f(\mathbf{x}) = \mathrm{prox}_{\frac{1}{1+\rho}g}\left(\frac{1}{1+\rho}\mathbf{x} - \frac{\rho}{1+\rho}\mathbf{a}\right)$$

- **scaling and translation:** if $f(\mathbf{x}) = g(a\mathbf{x} + b)$, then

$$\mathrm{prox}_f(\mathbf{x}) = \frac{1}{a}\left(\mathrm{prox}_{a^2 g}(a\mathbf{x} + b) - b\right)$$

Basic rules of proximal operator

- **affine addition:** if $f(\mathbf{x}) = g(\mathbf{x}) + \mathbf{a}^\top \mathbf{x} + b$, then

$$\mathrm{prox}_f(\mathbf{x}) = \mathrm{prox}_g(\mathbf{x} - \mathbf{a})$$

- **quadratic addition:** if $f(\mathbf{x}) = g(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{a}\|_2^2$, then

$$\mathrm{prox}_f(\mathbf{x}) = \mathrm{prox}_{\frac{1}{1+\rho}g}\left(\frac{1}{1+\rho}\mathbf{x} - \frac{\rho}{1+\rho}\mathbf{a}\right)$$

- **scaling and translation:** if $f(\mathbf{x}) = g(a\mathbf{x} + b)$, then

$$\mathrm{prox}_f(\mathbf{x}) = \frac{1}{a}\left(\mathrm{prox}_{a^2 g}(a\mathbf{x} + b) - b\right)$$

Basic rules of proximal operator

- **affine addition:** if $f(\mathbf{x}) = g(\mathbf{x}) + \mathbf{a}^\top \mathbf{x} + b$, then

$$\text{prox}_f(\mathbf{x}) = \text{prox}_g(\mathbf{x} - \mathbf{a})$$

- **quadratic addition:** if $f(\mathbf{x}) = g(\mathbf{x}) + \frac{\rho}{2} \|\mathbf{x} - \mathbf{a}\|_2^2$, then

$$\text{prox}_f(\mathbf{x}) = \text{prox}_{\frac{1}{1+\rho} g}\left(\frac{1}{1+\rho}\mathbf{x} - \frac{\rho}{1+\rho}\mathbf{a}\right)$$

- **scaling and translation:** if $f(\mathbf{x}) = g(a\mathbf{x} + b)$, then

$$\text{prox}_f(\mathbf{x}) = \frac{1}{a}\left(\text{prox}_{a^2 g}(a\mathbf{x} + b) - b\right)$$

Basic rules of proximal operator

- **norm composition:** if $f(\mathbf{x}) = g(\|\mathbf{x}\|_2)$ with $\mathrm{dom}\ g = [0, \infty)$, then

$$\mathrm{prox}_f(\mathbf{x}) = \mathrm{prox}_g(\|\mathbf{x}\|_2)\frac{\mathbf{x}}{\|\mathbf{x}\|_2}, \ \forall \mathbf{x} \neq \mathbf{0}$$

Nonexpansiveness of proximal operators

- **(firm nonexpansiveness)**

$$\langle \mathrm{prox}_h(\mathbf{x}_1) - \mathrm{prox}_h(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2 \rangle \geq \|\mathrm{prox}_h(\mathbf{x}_1) - \mathrm{prox}_h(\mathbf{x}_2)\|_2^2$$

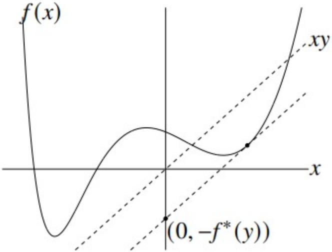
- **(nonexpansiveness)**

$$\|\mathrm{prox}_h(\mathbf{x}_1) - \mathrm{prox}_h(\mathbf{x}_2)\|_2 \leq \|\mathbf{x}_1 - \mathbf{x}_2\|_2$$

Conjugate functions (共轭函数)

The conjugate of a function f is

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \operatorname{dom} f} \{\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})\}$$



Fenchel's inequality: $f(\mathbf{x}) + f^*(\mathbf{y}) \geq \langle \mathbf{y}, \mathbf{x} \rangle$

Conjugate Functions

Property: If f is convex and closed. Then

- $\mathbf{y} \in \partial f(\mathbf{x}) \iff \mathbf{x} \in \partial f^*(\mathbf{y})$
- $f^{**} = f$

Examples:

- **Indicator function:**

$$f(\mathbf{x}) = \mathbb{1}_{\mathcal{C}}(\mathbf{x}), \qquad f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathcal{C}} \langle \mathbf{x}, \mathbf{y} \rangle$$

- **Norm:**

$$f(\mathbf{x}) = \|\mathbf{x}\|, \qquad f^*(\mathbf{y}) = \begin{cases} 0, & \|\mathbf{y}\|_* \leq 1 \\ +\infty, & \|\mathbf{y}\|_* > 1 \end{cases}$$

where $\|\mathbf{y}\|_* = \sup_{\|\mathbf{x}\| \leq 1} \langle \mathbf{x}, \mathbf{y} \rangle$ is the dual norm.

Moreau Decomposition

Suppose f is closed and convex. Then

$$\mathbf{x} = \text{prox}_f(\mathbf{x}) + \text{prox}_{f^*}(\mathbf{x})$$

Example: prox of support function

For any closed and convex set $\mathcal{C}$, the support function is defined as $S_{\mathcal{C}}(\mathbf{x}) = \sup_{\mathbf{z} \in \mathcal{C}} \langle \mathbf{x}, \mathbf{z} \rangle$. Then

$$\mathrm{prox}_{S_{\mathcal{C}}}(\mathbf{x}) = \mathbf{x} - \mathcal{P}_{\mathcal{C}}(\mathbf{x})$$

Examples

- ℓ_∞ **norm:**

$$\text{prox}_{\|\cdot\|_\infty}(\mathbf{x}) = \mathbf{x} - \mathcal{P}_{\mathcal{B}_{\|\cdot\|_1}}(\mathbf{x})$$

where $\mathcal{B}_{\|\cdot\|_1} = \{\mathbf{z} | \|\mathbf{z}\|_1 \leq 1\}$ is unit ℓ_1 ball.

- **max function:** Let $g(\mathbf{x}) = \{x_1, \ldots, x_n\}$, then

$$\text{prox}_g(\mathbf{x}) = \mathbf{x} - \mathcal{P}_\Delta(\mathbf{x})$$

where $\Delta = \{\mathbf{z} \in \mathbb{R}^n_+ | \mathbf{1}^\top \mathbf{z} = 1\}$ is probability simplex.

Outline

Convergence for Convex Problems

Suppose f is convex and L-smooth. The proximal graident descent with stepsize $\eta_t = 1/L$ obeys

$$F(\mathbf{x}_t) - F(\mathbf{x}^*) \leq \frac{L \left\| \mathbf{x}_0 - \mathbf{x}^* \right\|_2^2}{2t}.$$

- Achieves better iteration complexity ($O(1/\varepsilon)$) than subgradient method ($O(1/\varepsilon^2)$).

Convergence for Strongly Convex Problems

Suppose f is μ-strongly convex and L-smooth. The proximal graident descent with stepsize $\eta_t = 1/L$ obeys

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \le \left(1 - \frac{\mu}{L}\right)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 .$$

- Achieves linear convergence $O(\kappa \log \frac{1}{\varepsilon})$.

Summary

condition	stepsize	convergence rate	iteration complexity
convex	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$
strongly convex	$\eta_t = \frac{1}{L}$	$O\left(\left(1-\frac{1}{\kappa}\right)^t\right)$	$O(\kappa \log \frac{1}{\varepsilon})$

Table: Convergence Properties of Proximal Gradient Descent

condition	stepsize	convergence rate	iteration complexity
convex	$\eta_t \approx \frac{1}{\sqrt{t}}$	$O\left(\frac{1}{\sqrt{t}}\right)$	$O(\frac{1}{\varepsilon^2})$
strongly convex	$\eta_t \approx \frac{1}{t}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

Table: Convergence Properties of Subgradient Descent