AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

February 4, 2026

Outline

Proximal Gradient Descent

Proximal Operator

Convergence Analysis

Composite Model in ML

Composite Model: let f be convex smooth, h convex proper l.s.c.,

$$\min_{x \in \mathbb{R}^d} \quad F(x) := f(x) + h(x)$$

Sparsity:

$$\min_{w \in \mathbb{R}^d} \quad \frac{1}{2n}\|Xw - y\|_2^2 + \lambda\|w\|_1$$

Low-rank:

$$\min_{X \in \mathbb{R}^{d_1 \times d_2}} \quad \frac{1}{2}\|M - X\|_F^2 + \lambda\|X\|_*$$

where $\|X\|_* := \sum_{i=1}^{\min\{d_1,d_2\}} \sigma_i(X)$ is the nuclear norm of matrix X.

Constraints:

$$\min_{w \in \mathbb{R}^d} \quad f(w) + \iota_{\mathcal{C}}(w), \quad \iota_{\mathcal{C}}(w) = \begin{cases} 0, & w \in \mathcal{C}, \\ +\infty, & w \notin \mathcal{C}. \end{cases}$$

Unified Framework Revisited

Reformulation of GD: by first-order optimality condition,

$$
\begin{aligned}
x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\
&= \arg\min_{x \in \mathbb{R}^d} \left\{ \underbrace{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle}_{\text{first-order approximation}} + \underbrace{\frac{1}{2\alpha_k} \|x - x_k\|_2^2}_{\text{proximal term}} \right\}
\end{aligned}
$$

Reformulation of Projected GD:

$$
\begin{aligned}
x_{k+1} &= \Pi_{\mathcal{C}}(x_k - \alpha_k \nabla f(x_k)) \\
&= \arg\min_{x \in \mathbb{R}^d} \left\{ \tfrac{1}{2} \|x - (x_k - \alpha_k \nabla f(x_k))\|_2^2 + \alpha_k \mathbb{1}_{\mathcal{C}}(x) \right\} \\
&= \arg\min_{x \in \mathbb{R}^d} \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 + \mathbb{1}_{\mathcal{C}}(x) \right\}
\end{aligned}
$$

Proximal Operator

Definition: for any convex function h,

$$\text{prox}_h(x) := \arg\min_z \left\{ h(z) + \frac{1}{2}\|z - x\|_2^2 \right\}.$$

Algorithm:

- Projected GD: $x_{k+1} = \text{prox}_{\alpha_k \mathbb{1}_{\mathcal{C}}}(x_k - \alpha_k \nabla f(x_k))$.
- Proximal GD: using a more general h instead of $\mathbb{1}_{\mathcal{C}}$:

$$x_{k+1} = \text{prox}_{\alpha_k h}(x_k - \alpha_k \nabla f(x_k)).$$

Proximal GD: one way to solve composite model

- reduces to GD if $h = 0$;
- reduces to proximal point method if $f = 0$.

Outline

Proximal Gradient Descent

Proximal Operator

Convergence Analysis

Why Using Proximal Operators

Several Reasons:

- well-defined under very general conditions, including nonsmooth convex function, or even regular nonconvex functions;
- can be evaluated efficiently for many widely used functions;
- conceptually and mathematically simple, and unify many well-known optimization algorithms.

Examples:

- $h = \mathbb{1}_{\mathcal{C}}$, then $\mathrm{prox}_h(x) = \Pi_{\mathcal{C}}(x)$ (projection);
- $h(x) = \lambda\|x\|_1$, then $\mathrm{prox}_h(x) = S_\lambda(x)$ (soft-thresholding);
- $h(x) = \frac{\lambda}{2}x^\top Ax + b^\top x$, then $\mathrm{prox}_h(x) = (A + \lambda I)^{-1}(x - b)$;

Basic Rules of Proximal Operators

Affine Transformation: if $f(x) = ag(x) + b$, then $\mathrm{prox}_f(x) = \mathrm{prox}_{ag}(x)$.

Affine Sum: if $f(x) = g(x) + a^\top x + b$, then $\mathrm{prox}_f(x) = \mathrm{prox}_g(x - a)$.

Quadratic Sum: if $f(x) = g(x) + \frac{\rho}{2}\|x - a\|_2^2$, then

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

Scaling and Translation: if $f(x) = g(ax + b)$ with $a \neq 0$, then

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

Proof of Scaling and Translation

Definition: $\mathrm{prox}_f(x) = \underset{z}{\mathrm{argmin}} \left(g(az+b) + \frac{1}{2}\|z-x\|^2 \right)$.

Change of Variable: let $u = az + b \iff z = \frac{u-b}{a}$.

$$
\begin{aligned}
u^* &= \underset{u}{\mathrm{argmin}} \left(g(u) + \frac{1}{2} \left\| \frac{u-b}{a} - x \right\|^2 \right) \\
&= \underset{u}{\mathrm{argmin}} \left(g(u) + \frac{1}{2a^2} \|u - (ax+b)\|^2 \right)
\end{aligned}
$$

Rescaling: multiply objective by constant $a^2 > 0$:

$$
\begin{aligned}
u^* &= \underset{u}{\mathrm{argmin}} \left(a^2 g(u) + \frac{1}{2} \|u - (ax+b)\|^2 \right) \\
&= \mathrm{prox}_{a^2 g}(ax+b).
\end{aligned}
$$

Conclusion: Recover z^* using $z^* = \frac{u^*-b}{a}$.

More Properties of Proximal Operators

Affine Composition: if $f(x) = g(Ax + b)$, then hard to express $\text{prox}_f(x)$ in $\text{prox}_g(x)$ in general. However, if $AA^\top = \alpha^{-1}I$, then

$$\text{prox}_f(x) = x - \alpha A^\top \Big(Ax + b - \text{prox}_{\alpha^{-1}g}(Ax + b)\Big)$$

Norm Composition: if $f(x) = g(\|x\|_2)$ with g increasing over $\text{dom}(g) = [0, \infty)$, then

$$\text{prox}_f(x) = \begin{cases} \text{prox}_g(\|x\|_2) \frac{x}{\|x\|_2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Nonexpansiveness: similar to non-expansiveness of projection.

- Firmly Nonexpansive:

$$\langle \text{prox}_h(x_1) - \text{prox}_h(x_2), x_1 - x_2 \rangle \geq \|\text{prox}_h(x_1) - \text{prox}_h(x_2)\|_2^2.$$

- Nonexpansive (1-Lipschitz):

$$\|\text{prox}_h(x_1) - \text{prox}_h(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Proof of Norm Composition

Using Definition:

$$\mathsf{prox}_f(x) = \arg\min_{z\in\mathbb{R}^d} \left\{ g(\|z\|_2) + \tfrac{1}{2}\|z\|_2^2 - \langle z, x\rangle + \tfrac{1}{2}\|x\|_2^2 \right\}.$$

Double Minimization:

$$\mathsf{prox}_f(x) = \arg\min_{\alpha\geq 0} \min_{\|z\|_2=\alpha} \left\{ g(\alpha) + \tfrac{1}{2}\alpha^2 - \langle z, x\rangle + \tfrac{1}{2}\|x\|_2^2 \right\}$$

Cauchy Schwarz: equality attains at $z^* = \alpha^* \frac{x}{\|x\|_2}$, where

$$\begin{aligned} \alpha^* &= \arg\min_{\alpha\geq 0} \left\{ g(\alpha) + \tfrac{1}{2}\alpha^2 - \alpha\|x\| + \tfrac{1}{2}\|x\|_2^2 \right\} \\ &= \arg\min_{\alpha\geq 0} \left\{ g(\alpha) + \tfrac{1}{2}(\alpha - \|x\|_2)^2 \right\} \end{aligned}$$

Conclusion: $\alpha^* = \mathsf{prox}_g(\|x\|_2)$ and $\mathsf{prox}_f(x) = \mathsf{prox}_g(\|x\|_2)\frac{x}{\|x\|_2}$ if $x \neq 0$.
Since g is increasing, $\mathsf{prox}_f(0) = 0$ is trivial.

Proof of Nonexpansiveness

Optimality Condition: let $z_i = \mathsf{prox}_h(x_i)$ for $i = 1, 2$, then $x_1 - z_1 \in \partial h(z_1)$ and $x_2 - z_2 \in \partial h(z_2)$.

Convexity: add two inequalities

$$\begin{cases} h(z_2) \geq h(z_1) + \langle x_1 - z_1, z_2 - z_1 \rangle \\ h(z_1) \geq h(z_2) + \langle x_2 - z_2, z_1 - z_2 \rangle \end{cases}$$

Conclusion: from $\langle x_1 - z_1 - x_2 + z_2, z_1 - z_2 \rangle \geq 0$, we obtain firmly nonexpansiveness; by Cauchy-Schwarz, we obtain Nonexpansiveness.

Outline

Proximal Gradient Descent

Proximal Operator

Convergence Analysis

Monotonicity of Objective

Assumptions: we only assume f is ℓ-smooth, and h is proper closed convex, so F can be nonsmooth.

Monotonicity: if f is also convex and $\alpha_k = 1/\ell$, then $f(x_{k+1}) \leq f(x_k)$ and $\|x_{k+1} - x^*\|_2 \leq \|x_k - x^*\|_2$.

- different from subgradient method, for which objective value might be non-monotonic.

Fundamental Inequality: let $y^+ = \text{prox}_{\frac{1}{\ell}h}(y - \frac{1}{\ell}\nabla f(y))$,

$$F(y^+) - F(x) \leq \frac{\ell}{2}(\|x - y\|_2^2 - \|x - y^+\|_2^2) - B_f(x, y),$$

where and B_f denotes the Bregman divergence induced by f.

Proof: for $f(x_{k+1}) \leq f(x_k)$, taking $x = y = x_k$; for $\|x_{k+1} - x^*\|_2 \leq \|x_k - x^*\|_2$, taking $x = x^*$ and $y = x_k$.

Proof of the Fundamental Inequality

Auxiliary Function:

$$\phi(z) := f(y) + \langle \nabla f(y), z - y \rangle + \frac{\ell}{2}\|z - y\|_2^2 + h(z).$$

Strong Convexity: $\phi(z)$ is ℓ-strongly convex.

Optimality: $0 \in \partial\phi(y^+)$, so combining with strong convexity,

$$\phi(x) \geq \phi(y^+) + \frac{\ell}{2}\|x - y^+\|_2^2.$$

Smoothness: f is ℓ-smooth, so using quadratic Taylor bound,

$$\begin{aligned}\phi(y^+) &= f(y) + \langle \nabla f(y), y^+ - y \rangle + \frac{\ell}{2}\|y^+ - y\|_2^2 + h(y^+) \\ &\geq f(y^+) + h(y^+) = F(y^+).\end{aligned}$$

Conclusion: adding two inequalities and using definition of ϕ and B_f,

$$F(x) - B_f(x, y) + \frac{\ell}{2}\|x - y\|_2^2 = \phi(x) \geq F(y^+) + \frac{\ell}{2}\|x - y^+\|_2^2.$$

Convergence of Proximal GD

Convergence of proximal GD for nonconvex smooth f

Suppose f is ℓ-smooth, h is proper closed convex, and F has a minimizer x^*. If $\alpha_k = 1/\ell$, then

$$\min_{i=0,\ldots,k-1} \text{dist}(0, \partial F(x_{i+1}))^2 \leq \frac{8\ell(F(x_0) - F(x^*))}{k}.$$

Unified Proof:

- set $x = y = x_k$, then

$$F(x_{k+1}) - F(x_k) \leq -\frac{\ell}{2}\|x_{k+1} - x_k\|_2^2.$$

- using optimality condition and ℓ-smooth,

$$\begin{aligned} \text{dist}(0, \partial F(x_{k+1})) &\leq \|\nabla f(x_{k+1}) + s_{k+1}\| \qquad (s_{k+1} \in \partial h(x_{k+1})) \\ &\leq \|\nabla f(x_{k+1}) - \nabla f(x_k)\| + \|\nabla f(x_k) + s_{k+1}\| \\ &\leq 2\ell\|x_{k+1} - x_k\|. \end{aligned}$$

- using telescoping and taking minimum yields the final result.

Convergence of Proximal GD

Convergence of proximal GD for convex smooth f

Suppose f is convex and ℓ-smooth, h is proper closed convex, and F has a minimizer x^*. If $\alpha_k = 1/\ell$, then

$$F(x_k) - F(x^*) \le \frac{\ell \|x_0 - x^*\|_2^2}{2k}.$$

Unified Proof:

- set $x = x^*$ and $y = x_k$, by convexity $B_f(x^*, x_k) \ge 0$,

$$\begin{aligned} F(x_{k+1}) - F(x^*) &\le \tfrac{\ell}{2}\|x_k - x^*\|_2^2 - \tfrac{\ell}{2}\|x_{k+1} - x^*\|_2^2 - B_f(x^*, x_k) \\ &\le \tfrac{\ell}{2}\|x_k - x^*\|_2^2 - \tfrac{\ell}{2}\|x_{k+1} - x^*\|_2^2. \end{aligned}$$

- Using telescoping and monotonicity:

$$k(F(x_k) - F(x^*)) \le \sum_{i=0}^{k-1} (F(x_{i+1}) - F(x^*)) \le \tfrac{\ell}{2}\|x_0 - x^*\|_2^2.$$

Comparison with Subgradient Methods: $O(1/k)$ v.s. $O(1/\sqrt{k})$.

Convergence of Proximal GD

Convergence of proximal GD for strongly convex smooth f

Suppose f is μ-strongly convex and ℓ-smooth, h is proper closed convex. If $\alpha_k = 1/\ell$, then

$$\|x_k - x^*\|_2^2 \leq \left(1 - \frac{\mu}{\ell}\right)^k \|x_0 - x^*\|_2^2.$$

Unified Proof:

- set $x = x^*$ and $y = x_k$, and by strong convexity $B_f(x^*, x_k) \geq \frac{\mu}{2}\|x^* - x_{k+1}\|_2^2$,

$$0 \leq F(x_{k+1}) - F(x^*) \leq \frac{\ell-\mu}{2}\|x_k - x^*\|_2^2 - \frac{\ell}{2}\|x_{k+1} - x^*\|_2^2.$$

- Using the recursion below to conclude

$$\|x_{k+1} - x^*\|_2^2 \leq \left(1 - \frac{\mu}{\ell}\right) \|x_k - x^*\|_2^2.$$

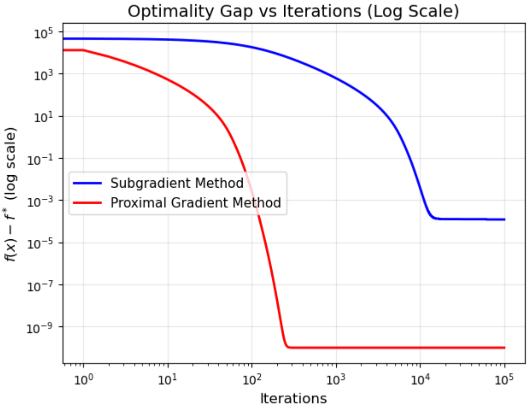
Comparison with Subgradient Methods: linear v.s. $O(1/k)$.

Numerical Experiments

Problem: $\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \|\mathbf{x}\|_1$.

Setup: $A \in \mathbb{R}^{2000\times 1000}$, i.i.d. Gaussian, $\ell = \lambda_{\max}(\mathbf{A}^\top \mathbf{A})$.

Step sizes: Subgradient: $\alpha_k = \frac{c}{\sqrt{k}}$; Proximal GD: $\alpha_k = \frac{1}{\ell}$



Acceleration

ISTA: proximal GD applied to LASSO is also called Iterative Shrinkage Thresholding Algorithm (ISTA).

FISTA: NAG can be applied in the proximal GD setting to accelerate algorithm. To initialize, set $x_0 = y_0$ and $t_0 = 1$.

$$\begin{cases} x_{k+1} = \mathsf{prox}_{\alpha_k h}(y_k - \alpha_k \nabla f(y_k)) \\ t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} \\ y_{k+1} = x_{k+1} + \frac{t_k - 1}{t_{k+1}}(x_{k+1} - x_k) \end{cases}$$

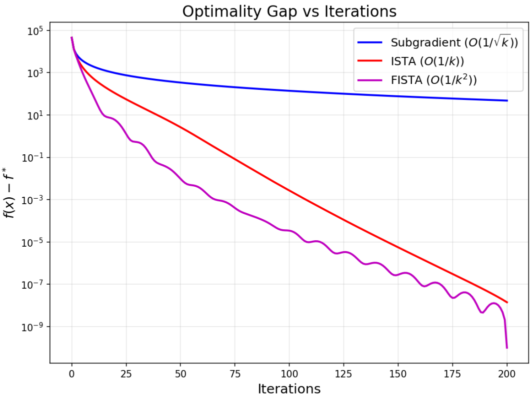
Originally proposed to accelerate ISTA in LASSO problem (Fast ISTA, [Beck and Teboulle, 2009]), but people also call it FISTA even when applied to general problem.

Numerical Experiments

Problem: $\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \|\mathbf{x}\|_1$.

Setup: $A \in \mathbb{R}^{2000\times 1000}$, i.i.d. Gaussian, $\ell = \lambda_{\max}(\mathbf{A}^\top \mathbf{A})$.

Step sizes: Subgradient: $\alpha_k = \frac{c}{\sqrt{k}}$; ISTA/FISTA: $\alpha_k = \frac{1}{\ell}$



Acceleration

Convergence of FISTA for convex smooth f

Suppose f is convex and ℓ-smooth, h is proper closed convex. If $\alpha_k = 1/\ell$, then using FISTA, we have

$$F(x_k) - F(x^*) \le \frac{2L\|x_0 - x^*\|_2^2}{(k+1)^2}.$$

Unified Proof:

- fundamental inequality $F(y^+) \le F(x) + \frac{\ell}{2}(\|x-y\|_2^2 - \|x-y^+\|_2^2)$
- $z_k := t_k x_{k+1} - (t_k - 1)x_k$, then $y_{k+1} = (1 - \frac{1}{t_{k+1}})x_{k+1} + \frac{1}{t_{k+1}} z_k$.
- Lyapunov function $\mathcal{E}_k := t_k^2(F(x_k) - F(x^*)) + \frac{\ell}{2}\|z_k - x^*\|_2^2$
- $t_k \ge \frac{k+2}{2}$
- $\mathcal{E}_0 \le \frac{\ell}{2}\|x_0 - x^*\|_2^2$ because $z_0 = x_1$ and applying fundamental inequality with $x = x^*$ and $y_0 = x_0$.

Proof of Lyapunov Function

Proof: we want to show $\mathcal{E}_{k+1} \leq \mathcal{E}_k$.

Choose a comparison point $\bar{x} := (1 - \frac{1}{t_{k+1}})x_{k+1} + \frac{1}{t_{k+1}}x^*$, then

$$F(\bar{x}) - F^* \leq \left(1 - \frac{1}{t_{k+1}}\right)(F(x_{k+1}) - F^*)$$

Apply Fundamental inequality with $x = \bar{x}$ and y_{k+1}, x_{k+2}

$$F(x_{k+2}) \leq F(\bar{x}) + \frac{\ell}{2}(\|\bar{x} - y_{k+1}\|_2^2 - \|\bar{x} - x_{k+2}\|_2^2)$$

Show $\|\bar{x} - y_{k+1}\|_2^2 = \frac{1}{t_{k+1}^2}\|z_k - x^*\|_2^2$ and $\|\bar{x} - x_{k+2}\|_2^2 = \frac{1}{t_{k+1}^2}\|z_{k+1} - x^*\|_2^2$

Combine the above three and use $t_{k+1}(t_{k+1} - 1) = t_k^2$.

Wrap-up and Next Week

Week 5: Key takeaways

- Proximal gradient descent (Proximal GD)
- Proximal operator and its properties
- Convergence analysis of Proximal GD
- FISTA and its convergence analysis

Thank you for your attendance!

Questions?

Next week: Spring Festival!

References

Beck, A. and Teboulle, M. (2009).
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM journal on imaging sciences, 2(1):183–202.