# Convex Optimization

## Part 3: Proximal gradient method

Namhoon Lee

POSTECH

12 Oct 2022

# Admin

Midterm exam

- (When) Monday 24 October 2022 between 9:30-10:45am
  - Decided to keep the original plan due to conflict of multiple students
- Bring your student ID card
- All materials up until proximal gradient method
- Closed book; only bring pen/pensil/eraser

Assignments

- (A1) results will be released on Tuesday
- (A2) results will be released on Thursday
- If you have any questions, please contact TAs (and set up a meeting if needed).

# Composite function

Consider minimizing a composite function $f$

$$f(x) = g(x) + h(x)$$

- $g$ convex, differentiable
- $h$ convex, not necessarily differentiable
- $f$ convex, non-differentiable; can use subgradient method but slow

# Proximal gradient method

Proximal gradient method: start with some initial point $x_1$ and repeat

$$x_{t+1} = \text{prox}_{\eta h}(x_t - \eta \nabla g(x_t))$$

proximal mapping

$$\text{prox}_{\eta h}(x) = \arg\min_u \; h(u) + \frac{1}{2\eta}\|u - x\|_2^2$$

- step size $\eta$ controls relative importance between $h$ and proximal term
- faster than subgradient method for minizing $f$ (in terms of convergence rate)
- $h(\cdot) = 0$ reduces to gradient descent
- $h(\cdot) = \mathcal{I}_{\mathbb{C}}(\cdot)$ reduces to projected subgradient method

# Interpretation

Proximal gradient step

$$x^+ = \mathrm{prox}_{\eta h}(x - \eta \nabla g(x))$$

from definition of proximal mapping

$$\begin{aligned} x^+ &= \arg\min_u \; h(u) + \frac{1}{2\eta}\|u - (x - \eta\nabla g(x))\|_2^2 \\ &= \arg\min_u \; h(u) + \frac{1}{2\eta}\|u - x\|_2^2 + \nabla g(x)^\top (u - x) + \frac{\eta}{2}\|\nabla g(x)\|_2^2 \\ &= \arg\min_u \; h(u) + \underbrace{g(x) + \nabla g(x)^\top (u - x) + \frac{1}{2\eta}\|u - x\|_2^2}_{\text{quadratic approx. at } x} \end{aligned}$$

$x^+$ minimizes $h(u)$ plus a simple quadratic model of $g(u)$ around $x$

# Why prox-grad?

$\mathrm{prox}_{\eta h}(\cdot)$ only depends on $h$.

- useful when $h$ is simple, with inexpensive prox-operator

$\mathrm{prox}_{\eta h}(\cdot)$ has a closed-form solution for many important functions $h$.

- some examples later

prox-grad can be faster than subgrad

- as if minimizing smooth function (or only $g$)
- but in terms of convergence rate rather than actual computations

# Proximal mapping

Proximal mapping (or prox-operator) of $x$ given convex $h$

$$\mathrm{prox}_h(x) = \arg\min_u \; h(u) + \frac{1}{2}\|u - x\|_2^2$$

*i.e.*, find $u$ that minimizes $h$ while staying close to $x$

some properties

- from optimality conditions of minimization in the definition

$$\begin{aligned} u = \mathrm{prox}_h(x) \quad &\iff \quad x - u \in \partial h(u) \\ &\iff \quad h(z) \geq h(u) + (x-u)^\top (z-u) \text{ for all } z \end{aligned}$$

- firm nonexpansiveness (co-coercivity with constant $1$)

$$(\operatorname{prox}_h(x) - \operatorname{prox}_h(y))^\top (x - y) \geq \| \operatorname{prox}_h(x) - \operatorname{prox}_h(y) \|_2^2$$

  if $u = \operatorname{prox}_h(x), v = \operatorname{prox}_h(y)$, then

$$x - u \in \partial h(u), \quad y - v \in \partial h(v)$$

  combining this with monotonicity of subdifferential gives

$$(x - u - y + v)^\top (u - v) \geq 0$$

- nonexpansiveness (Lipschitz continuity with constant $1$)

$$\| \operatorname{prox}_h(x) - \operatorname{prox}_h(y) \|_2 \leq \|x - y\|_2$$

  follows from firm nonexpansiveness and Cauchy-Schwarz inequality

# Examples

$h(x) = 0$

$$x^+ = x - \eta \nabla g(x)$$

- gradient method

$h(x) = \mathcal{I}_{\mathbb{C}}(x)$ – indicator function

$$x^+ = \mathrm{P}_{\mathbb{C}}(x - \eta \nabla g(x))$$

where

$$\mathrm{P}_{\mathbb{C}}(x) = \arg\min_{u \in \mathbb{C}} \frac{1}{2}\|u - x\|_2^2$$

- projected gradient method
- can be used for constrained optimization

$h(x) = \|x\|_1$

$$x^+ = \mathrm{S}_\eta(x - \eta \nabla g(x))$$

where

$$(\mathrm{S}_\eta(u))_i = \begin{cases} u_i - \eta & \text{for } u_i \geq \eta \\ 0 & \text{for } |u_i| \leq \eta \\ u_i + \eta & \text{for } u_i \leq \eta \end{cases}$$

or

$$(\mathrm{S}_\eta(u))_i = (|u_i| - \eta)_+ \operatorname{sign}(u_i)$$

- "soft-threshold" (shrinkage) operation

# Constrained optimization

Consider optimization problem with convex constraint

$$\min_x f(x) \quad \text{subject to} \quad x \in \mathbb{C}$$

- Simply applying subgradient method may not give a feasible solution.

# Projection

- for $x \in \operatorname{int} \mathbb{C}$ the projection is $x$ itself
- for $x \notin \operatorname{int} \mathbb{C}$ the projection is where the Euclidean ball (centered at $x$) touches
- for $\mathbb{C}$ convex the projection is unique; otherwise not

# Projected subgradient method

convex constrained optimization

$$\begin{aligned} &\min\ f(x) \\ &\text{s.t. } x \in \mathbb{C} \end{aligned}$$

projected subgradient method

$$x_{t+1} = \mathrm{P}_{\mathbb{C}}(x_t - \eta g_t)$$

where $g_t \in \partial f(x_t)$ and $\mathrm{P}_{\mathbb{C}}(\cdot)$ is the projection onto $\mathbb{C}$

$$\mathrm{P}_{\mathbb{C}}(x) = \arg\min_{u \in \mathbb{C}} \frac{1}{2}\|u - x\|_2^2$$

notice the projection step is equivalent to

$$\mathrm{P}_{\mathbb{C}}(x) = \arg\min_{u} \frac{1}{2}\|u - x\|_2^2 + \mathcal{I}_{\mathbb{C}}(u)$$

- $\min_{x \in \mathbb{C}} f(x) \;\equiv\; \min_x f(x) + \mathcal{I}_{\mathbb{C}}(x)$

then one can apply subgradient method

- recall subgradient for indicator function becomes normal cone

$$\partial \mathcal{I}_{\mathbb{C}}(x) = \mathcal{N}_{\mathbb{C}}(x) \quad \text{for } x \in \mathbb{C}$$

# Convergence analysis

**Theorem (projected subgradient method)**

*For $G$-Lipschitz convex $f$ projected subgradient method with step size $\eta$ satisfies*

$$f(x_t, \textit{best}) - f(x^*) \leq \frac{R^2}{2\eta T} + \frac{G^2\eta}{2}$$

*where* $R = \|x_1 - x^*\|_2$.

- converegence rate same as subgradient method $\mathcal{O}(1/\epsilon^2)$

Proof.

The distance between two points becomes smaller after projection onto a convex set. Also, projection of a point that is already in the set is the point itself. Using these and following the proof step for the subgradient method, one can show that the projected subgradient descent has the same convergence rate as the subgradient method.

$$\begin{aligned}
\|x_{t+1} - x^*\| &\leq \|y_{t+1} - x^*\|^2 \\
&= \|x_t - \eta g_t - x^*\|^2 \\
&= \|x_t - x^*\|^2 - 2\eta\langle g_t, x_t - x^*\rangle + \eta^2\|g_t\|^2 \\
&\leq \|x_t - x^*\|^2 - 2\eta(f(x_t) - f(x^*)) + \eta^2 G^2
\end{aligned}$$

After telescoping is what is obtained for subgradient method. □

# ISTA and FISTA (Beck and Teboulle 2009)

Recall lasso

$$f(\beta) = \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{\text{smooth}} + \underbrace{\lambda\|\beta\|_1}_{\text{non-smooth}}$$

- non-smooth, subgradient method, $\mathcal{O}(1/\sqrt{t})$ convergence rate
- composite function of smoooth and non-smooth; proximal gradient method

proximal mapping

$$\begin{aligned}\operatorname{prox}_{\eta h}(\beta) &= \arg\min_u \frac{1}{2\eta}\|u-\beta\|_2^2 + \lambda\|u\|_1 \\ &= \mathrm{S}_{\lambda\eta}(\beta)\end{aligned}$$

which is the soft-thresholding operator

proximal gradient update

$$\beta^+ = \mathrm{S}_{\lambda\eta}(\beta + \eta X^\top (y - X\beta))$$

- "iterative shrinkage-thresholding algorithm" or ISTA
- $\mathcal{O}(1/t)$ convergence rate

accelerated proximal gradient method

$$\gamma_{t+1} = \frac{1 + \sqrt{1 + 4\gamma_t^2}}{2}$$
$$y_{t+1} = x_t + \left(\frac{\gamma_t - 1}{\gamma_{t+1}}\right)(x_t - x_{t-1})$$
$$x_{t+1} = \mathrm{prox}_{\eta h}(y_{t+1})$$

- "fast iterative shrinkage-thresholding algorithm" or FISTA
- make use of the idea of (Nesterov) momentum
- $\mathcal{O}(1/t^2)$ convergence rate

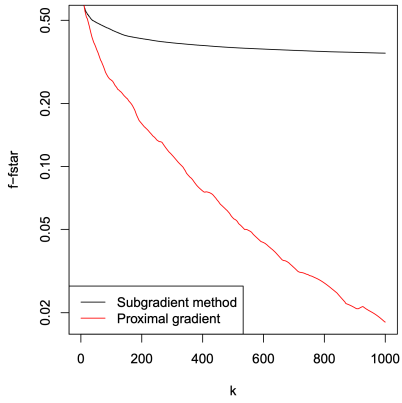


Figure: prox-grad (ISTA) vs. sub-grad; from R. Tibshirani

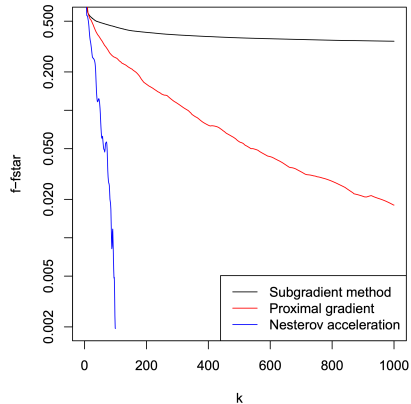


Figure: accelerated prox-grad (FISTA) vs. sub-grad; from R. Tibshirani

# Gradient map

Define gradient map

$$G_\eta(x) = \frac{1}{\eta}\Big(x - \operatorname{prox}_{\eta h}\left(x - \eta \nabla g(x)\right)\Big)$$

Then rewrite the proximal gradient update (easier to analyse)

$$\begin{aligned} x^+ &= \operatorname{prox}_{\eta h}(x - \eta \nabla g(x)) \\ &= x - \eta G_\eta(x) \end{aligned}$$

- $G_\eta(x)$ is not a (sub)gradient of $f$
- from the subgradient definition of prox-operator

$$G_\eta(x) \in \nabla g(x) + \partial h(x - \eta G_\eta(x))$$

- $G_\eta(x) = 0$ if and only if $x$ is a minimizer of $f$

# Bound on proximal gradient update

Consider minimizing composite function $f$

$$\min f(x) = g(x) + h(x)$$

where $g$ is $\alpha$-strongly convex and $\beta$-smooth

proximal gradient update

$$x^+ = x - \eta G_\eta(x)$$

## Lemma (bound on proximal gradient update)

*For $f$ $\alpha$-strongly convex and $\beta$-smooth, proximal gradient update satisfies*

$$f(x - \eta G_\eta(x)) \leq f(z) + G_\eta(x)^\top (x - z) - \frac{\eta}{2}\|G_\eta(x)\|_2^2 - \frac{\alpha}{2}\|x - z\|_2^2$$

*for all $z$ where $G$ is the gradient map.*

Proof.

$$
\begin{aligned}
f(x-\eta G_\eta(x)) &= g(x-\eta G_\eta(x)) + h(x-\eta G_\eta(x)) \\
&\overset{(A)}{\leq} g(x) - \eta \nabla g(x)^\top G_\eta(x) + \frac{\eta}{2}\|G_\eta(x)\|_2^2 + h(x-\eta G_\eta(x)) \\
&\overset{(B)}{\leq} g(z) - \nabla g(z)^\top (z-x) - \frac{\alpha}{2}\|z-x\|_2^2 - \eta \nabla g(x)^\top G_\eta(x) + \frac{\eta}{2}\|G_\eta(x)\|_2^2 \\
&\quad + h(x-\eta G_\eta(x)) \\
&\overset{(C)}{\leq} g(z) - \nabla g(z)^\top (z-x) - \frac{\alpha}{2}\|z-x\|_2^2 - \eta \nabla g(x)^\top G_\eta(x) + \frac{\eta}{2}\|G_\eta(x)\|_2^2 \\
&\quad + h(z) - (G_\eta(x) - \nabla g(x))^\top (z - x + \eta G_\eta(x)) \\
&= g(z) + h(z) + G_\eta(x)^\top (x-z) - \frac{\eta}{2}\|G_\eta(x)\|_2^2 - \frac{\alpha}{2}\|x-z\|_2^2
\end{aligned}
$$

- (A) $g$ $\beta$-smooth and $\eta \leq 1/\beta$
- (B) $g$ $\alpha$-strongly convex
- (C) $h$ convex and $G_\eta(x) - \nabla g(x) \in \partial h(x - \eta G_\eta(x))$

# Progress in one iteration

Lemma with $z = x$

$$f(x^+) \leq f(x) - \frac{\eta}{2}\|G_\eta(x)\|_2^2$$

- proximal gradient method is a descent method

Lemma with $z = x^*$

$$\begin{aligned} f(x^+) - f^* &\leq G_\eta(x)^\top (x - x^*) - \frac{\eta}{2}\|G_\eta(x)\|_2^2 - \frac{\alpha}{2}\|x - x^*\|_2^2 \\ &= \frac{1}{2\eta}\Big(\|x - x^*\|_2^2 - \|x - x^* - \eta G_\eta(x)\|_2^2\Big) - \frac{\alpha}{2}\|x - x^*\|_2^2 \\ &= \frac{1}{2\eta}\Big((1 - \alpha\eta)\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\Big) \end{aligned}$$

# Convergence analysis

## Theorem (smooth case)

*Proximal gradient method for fixed step size $\eta \leq 1/\beta$ satisfies*

$$f(x_{T+1}) - f^* \leq \frac{R^2}{2\eta T}$$

*where* $R = \|x_1 - x^*\|_2$.

## Proof.

Substitute $x = x_t$, $x^+ = x_{t+1}$ and set for $\alpha = 0$. Sum for $T$ iterations and use proximal gradient method is a descent method. □

Theorem (smooth and strongly convex case)

*Proximal gradient method for fixed step size $\eta \leq 1/\beta$ satisfies*

$$\|x_{t+1} - x^*\|_2^2 \leq R^2(1 - \eta\alpha)^t$$

*where $R = \|x_1 - x^*\|_2$ and $\eta\alpha \leq 1$.*

Proof.

Substitute $x = x_t$, $x^+ = x_{t+1}$. Use that $f(x_{t+1}) - f(x^*) \geq 0$. □

# Summary

Proximal gradient method

- Useful when minimizing convex composite functions $f = g + h$ where $h$ is non-differentiable but simple
- Less general but faster than subgradient method
- Convergence properties similar to standard gradient method

Any questions?

# References I

 Beck, Amir and Marc Teboulle (2009). "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". In: *SIAM journal on imaging sciences* 2.1, pp. 183–202.