Lecture 7. Proximal Gradient Methods

Bao Wang
Department of Mathematics
Scientific Computing and Imaging Institute
University of Utah
Math 5750/6880, Fall 2021

So far, we have formulated training machine learning models as

$$\min f(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i(\boldsymbol{x}) + R(\boldsymbol{x})$$

where $\boldsymbol{x}$ is the parameter of the machine learning model, $\mathcal{L}_i(\boldsymbol{x})$ is the loss of the ith training instance, and $R(\boldsymbol{x})$ is the regularization term.

How to find the optimal $\boldsymbol{x}^*$ if $R(x)$ is not differentiable everywhere, e.g. ℓ_1-regularization?

Subgradient methods or proximal gradient methods.

Recap: Subgradient method

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \boldsymbol{g}^t, \tag{1}$$

where $\boldsymbol{g}^t$ is any subgradient of f at $\boldsymbol{x}^t$.

We say $\boldsymbol{g}$ is a subgradient of f at the point $\boldsymbol{x}$ if

$$f(\boldsymbol{z}) \geq \underbrace{f(\boldsymbol{x}) + \boldsymbol{g}^\top(\boldsymbol{z} - \boldsymbol{x})}_{\text{a linear under-estimate of } f}, \quad \forall \boldsymbol{z}. \tag{2}$$

The set of all subgradients of f at $\boldsymbol{x}$ is called the subdifferential of f at $\boldsymbol{x}$, denoted by $\partial f(\boldsymbol{x})$.

Recap: Convergence of subgradient methods

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

Subgradient method is very slow!

Proximal gradient descent for composite functions

Consider the composite model

$$\min_{\boldsymbol{x}} F(\boldsymbol{x}) := f(\boldsymbol{x}) + h(\boldsymbol{x}), \;\; \boldsymbol{x} \in \mathbb{R}^n,$$

let $F^{opt} := \min_{\boldsymbol{x}} F(\boldsymbol{x})$ be the optimal cost.

1. ℓ_1 regularized minimization for promoting sparsity (e.g., lasso)

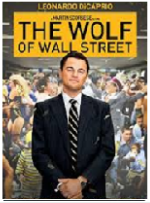
$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \|\boldsymbol{x}\|_1$$

2. nuclear norm (sum of the singular values) regularized minimization for promoting low-rank structure (Netflix competition)

$$\min_{\boldsymbol{X}} f(\boldsymbol{X}) + \|\boldsymbol{X}\|_*$$

Matrix completion

Movies



Users

Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?
Sam	5	?	?	?

Recommender system through matrix completion! (User privacy?)

Nuclear norm?

The rank is the L0 norm of the spectrum (the vector of singular values) and so the best approximation is the L1 norm of the spectrum. However, for a positive semidefinite matrix, the singular values are the eigenvalues and the sum of eigenvalues is the trace.

The connection between the rank and the nuclear norm can be understood as follows. The convex hull of all rank-1 matrices with norm equal to 1 is the nuclear ball. This is why the tightest convex approximation to the rank functional is the nuclear norm.

A proximal view of gradient descent

We note that the gradient descent iteration

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^{t} - \eta_t \nabla f(\boldsymbol{x}^{t})$$

can be written as

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

Motivation. GD can be considered as find the optimal solution of the linear approximation of $f(\boldsymbol{x}^{t})$, and the linear approximation is accurate when $\boldsymbol{x}$ and $\boldsymbol{x}^{t}$ is close to each other.

We note that the gradient descent iteration

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)$$

can be written as

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

$$\Leftrightarrow$$

$$\boldsymbol{x}^{t+1} = \arg\min_{\boldsymbol{x}} \Big\{ \frac{1}{2}\|\boldsymbol{x} - (\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t))\|_2^2 \Big\}.$$

Proximal gradient algorithm

Define the proximal operator

$$prox_h(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + h(\boldsymbol{z}) \right\}$$

for any convex function h.

This allows one to express GD update as (set $h(\boldsymbol{z}) = 0$),

$$\boldsymbol{x}^{t+1} = prox_0(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)). \tag{3}$$

One can generalize (3) to accommodate more general h,

$$\boldsymbol{x}^{t+1} = prox_{\eta_t h}(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t)).$$

The proximal gradient algorithm alternates between gradient updates on f and proximal minimization on h, and it will be useful if $prox_h$ is inexpensive.

Consider the composite model

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

Proximal gradient descent

for $k = 0, 1, \cdots$

$$\boldsymbol{x}^{t+1} = prox_{\eta_t h}(\boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t))$$

Proximal mapping/operator

The proximal operator is define by

$$prox_h(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + h(\boldsymbol{z}) \right\}.$$

> 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 conceptually and mathematically simple, and covers many well-known optimization algorithms.

Example (ℓ_1 norm)

If $h(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$, then $(prox_{\lambda h}(\boldsymbol{x}))_i = \psi_{st}(x_i; \lambda)$ (soft-thresholding) where

$$\psi_{st}(x; \lambda) = \begin{cases} x - \lambda & \text{if } x \geq \lambda \\ x + \lambda & \text{if } x \leq -\lambda \\ 0 & \text{else} \end{cases}$$

Why?

$$prox_{\lambda\|\boldsymbol{x}\|_1}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{z}\|_1 \right\} = \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{z}\|_2^2 - \langle \boldsymbol{z}, \boldsymbol{x} \rangle + \lambda\|\boldsymbol{z}\|_1 \right\}$$

Note that

$$\arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{z}\|_2^2 - \langle \boldsymbol{z}, \boldsymbol{x} \rangle + \lambda\|\boldsymbol{z}\|_1 \right\} = \sum_i \mathcal{L}_i,$$

where

$$\mathcal{L}_i := \frac{1}{2} z_i^2 - z_i x_i + \lambda |z_i|.$$

If $x_i > 0$, then we must have $z_i \geq 0$, otherwise, let $z_i^* < 0$ minimizes $\mathcal{L}_i$, then $-z_i^*$ enables even smaller $\mathcal{L}_i$.

If $x_i < 0$, then we must have $z_i \leq 0$.

If $x_i > 0$, since $z_i \geq 0$, then we have

$$\mathcal{L}_i = -x_i z_i + \frac{1}{2} z_i^2 + \lambda z_i,$$

$$\frac{\partial \mathcal{L}}{\partial z_i} = 0 \Rightarrow -x_i + z_i + \lambda = 0 \Rightarrow z_i = x_i - \lambda.$$

Here, we require the RHS is positive (we require $z_i \geq 0$), i.e., $x_i \geq \lambda$.

If $x_i < 0$, since $z_i \leq 0$, then we have

$$\mathcal{L}_i = -x_i z_i + \frac{1}{2} z_i^2 - \lambda z_i,$$

$$\frac{\partial \mathcal{L}}{\partial z_i} = 0 \Rightarrow -x_i + z_i - \lambda = 0 \Rightarrow z_i = x_i + \lambda.$$

Here, we require the RHS is negative (we require $z_i \leq 0$), i.e., $x_i \leq -\lambda$.

Finally, let us consider the case when $-\lambda < x_i < \lambda$, our goal is

$$\arg\min \mathcal{L}_i := -x_i z_i + \frac{1}{2} z_i^2 + \lambda |z_i|.$$

1. $z_i = 0 \Rightarrow \mathcal{L}_i = 0$
2. $z_i > 0 \Rightarrow \mathcal{L}_i = -x_i z_i + \frac{1}{2} z_i^2 + \lambda z_i$ and the minimum is obtained when $z_i = 1 - \lambda$, in this case we have

$$\mathcal{L}_i = -x_i(1-\lambda) + \frac{1}{2}(1-\lambda)^2 + \lambda(1-\lambda) > 0$$

3. $z_i < 0 \Rightarrow \mathcal{L}_i = -x_i z_i + \frac{1}{2} z_i^2 - \lambda z_i$ and the minimum is obtained when $z_i = 1 + \lambda$, in this case we have

$$\mathcal{L}_i = -x_i(1+\lambda) + \frac{1}{2}(1+\lambda)^2 + \lambda(1+\lambda) > 0.$$

Basic rules

If $f(\boldsymbol{x}) = ag(\boldsymbol{x}) + b$ with $a > 0$, then

$$prox_f(\boldsymbol{x}) = prox_{ag}(\boldsymbol{x}).$$

Basic rules, Affine addition

If $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \boldsymbol{a}^\top \boldsymbol{x} + b$, then

$$prox_f(\boldsymbol{x}) = prox_g(\boldsymbol{x} - \boldsymbol{a})$$

Basic rules, Quadratic addition

If $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x} - \boldsymbol{a}\|_2^2$, then

$$prox_f(\boldsymbol{x}) = prox_{\frac{1}{1+\rho}g}\left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\boldsymbol{a}\right)$$

Proof.

$$\begin{aligned}
prox_f(\boldsymbol{x}) &= \arg\min_{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + g(\boldsymbol{z}) + \frac{\rho}{2}\|\boldsymbol{z} - \boldsymbol{a}\|_2^2\right\} \\
&= \arg\min_{\boldsymbol{x}}\left\{\frac{1+\rho}{2}\|\boldsymbol{z}\|_2^2 - \langle \boldsymbol{z}, \boldsymbol{x} + \rho\boldsymbol{a}\rangle + g(\boldsymbol{z})\right\} \\
&= \arg\min_{\boldsymbol{x}}\left\{\frac{1}{2}\|\boldsymbol{z}\|_2^2 - \frac{1}{1+\rho}\langle \boldsymbol{z}, \boldsymbol{x} + \rho\boldsymbol{a}\rangle + \frac{1}{1+\rho}g(\boldsymbol{z})\right\} \\
&= \arg\min_{\boldsymbol{x}}\left\{\frac{1}{2}\left\|\boldsymbol{z} - \left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\boldsymbol{a}\right)\right\|_2^2 + \frac{1}{1+\rho}g(\boldsymbol{z})\right\} \\
&= prox_{\frac{1}{1+\rho}g}\left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\boldsymbol{a}\right)
\end{aligned}$$

Basic rules, Scaling and translation

If $f(\boldsymbol{x}) = g(a\boldsymbol{x} + \boldsymbol{b})$ with $a \neq 0$, then

$$prox_f(\boldsymbol{x}) = \frac{1}{a}\Big(prox_{a^2 g}(a\boldsymbol{x} + \boldsymbol{b}) - \boldsymbol{b}\Big)$$

Why?

$$
\begin{aligned}
prox_f(\boldsymbol{x}) &= \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2 + g(a\boldsymbol{z} + b) \right\} \\
&= \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2} \| \frac{\boldsymbol{z}' - b}{a} - \boldsymbol{x}\|_2^2 + g(\boldsymbol{z}') \right\} \text{ (Let } \boldsymbol{z}' = a\boldsymbol{z} + b) \\
&= \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2} \|\boldsymbol{z}' - (a\boldsymbol{x} + b)\|_2^2 + a^2 g(\boldsymbol{z}') \right\}
\end{aligned}
$$

Next, consider

$$
\boldsymbol{z}'^* = \arg\min_{\boldsymbol{z}'} \left\{ \frac{1}{2} \|\boldsymbol{z}' - (a\boldsymbol{x} + b)\|_2^2 + a^2 g(\boldsymbol{z}') \right\} = prox_{a^2 g}(a\boldsymbol{x} + b).
$$

Moreover, we have $\boldsymbol{z}^* = \frac{\boldsymbol{z}'^* - \boldsymbol{b}}{a}$, thus

$$
prox_f(\boldsymbol{x}) = \frac{1}{a}\Big(prox_{a^2 g}(a\boldsymbol{x} + \boldsymbol{b}) - \boldsymbol{b} \Big).
$$

Basic rules, Orthogonal mapping

If $f(\boldsymbol{x}) = g(\boldsymbol{Qx})$ with $\boldsymbol{Q}$ orthogonal ($\boldsymbol{QQ}^\top = \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}$), then

$$prox_f(\boldsymbol{x}) = \boldsymbol{Q}^\top prox_g(\boldsymbol{Q}^\top \boldsymbol{x})$$

$$
\begin{aligned}
prox_f(\boldsymbol{x}) &= \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|^2 + f(\boldsymbol{z}) \right\} \\
&= \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|^2 + g(\boldsymbol{Q}\boldsymbol{z}) \right\} \\
&= \arg\min_{\boldsymbol{z}} \left\{ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{Q}^\top\boldsymbol{z}'\|_2^2 + g(\boldsymbol{z}') \right\}
\end{aligned}
$$

Let $\boldsymbol{z}'^* = \arg\min_{\boldsymbol{z}'} \left\{ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{Q}^\top\boldsymbol{z}'\|_2^2 + g(\boldsymbol{z}') \right\} = prox_g(\boldsymbol{Q}^\top\boldsymbol{x})$ and we have $\boldsymbol{z}^* = \boldsymbol{Q}^\top\boldsymbol{z}'^*$, therefore

$$
prox_f(\boldsymbol{x}) = \boldsymbol{Q}^\top prox_g(\boldsymbol{Q}^\top\boldsymbol{x})
$$

Basic rules, Orthogonal affine mapping

If $f(\boldsymbol{x}) = g(\boldsymbol{Qx} + \boldsymbol{b})$ with $\underbrace{\boldsymbol{Q}\boldsymbol{Q}^\top = \alpha^{-1}\boldsymbol{I}}_{\text{does not require } \boldsymbol{Q}^\top\boldsymbol{Q} = \alpha^{-1}\boldsymbol{I}}$, then

$$prox_f(\boldsymbol{x}) = \left(\boldsymbol{I} - \alpha\boldsymbol{Q}^\top\boldsymbol{Q}\right)\boldsymbol{x} + \alpha\boldsymbol{Q}^\top\left(prox_{\alpha^{-1}g}(\boldsymbol{Qx} + \boldsymbol{b}) - \boldsymbol{b}\right)$$

Basic rules, Norm composition

If $f(\boldsymbol{x}) = g(\|\boldsymbol{x}\|_2)$ with $domain(g) = [0, \infty)$, then

$$prox_f(\boldsymbol{x}) = prox_g(\|\boldsymbol{x}\|_2)\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} \ \forall \boldsymbol{x} \neq 0$$

Proof. Observe that

$$
\begin{aligned}
\min_{\boldsymbol{z}} \left\{ f(\boldsymbol{z}) + \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{x}\|_2^2 \right\} &= \min_{\boldsymbol{z}} \left\{ g(\|\boldsymbol{z}\|_2) + \frac{1}{2}\|\boldsymbol{z}\|_2^2 - \boldsymbol{z}^\top \boldsymbol{x} + \frac{1}{2}\|\boldsymbol{x}\|_2^2 \right\} \\
&= \min_{\alpha \geq 0} \min_{\|\boldsymbol{z}\|_2 = \alpha} \left\{ g(\alpha) + \frac{1}{2}\alpha^2 - \boldsymbol{z}^\top \boldsymbol{x} + \frac{1}{2}\|\boldsymbol{x}\|_2^2 \right\} \\
&\underbrace{=}_{Cauchy-Schwarz} \min_{\alpha \geq 0} \left\{ g(\alpha) + \frac{1}{2}\alpha^2 - \alpha\|\boldsymbol{x}\|_2 + \frac{1}{2}\|\boldsymbol{x}\|_2^2 \right\} \\
&= \min_{\alpha \geq 0} \left\{ g(\alpha) + \frac{1}{2}(\alpha - \|\boldsymbol{x}\|_2)^2 \right\}
\end{aligned}
$$

From the above calculation, we know the optimal point is

$$
\alpha^* = prox_g(\|\boldsymbol{x}\|_2) \ \ and \ \ \boldsymbol{z}^* = \alpha^* \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} = prox_g(\|\boldsymbol{x}\|_2)\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2},
$$

thus concluding proof.

Convergence analysis

Lemma 5. [Cost monotonicity] Suppose f is convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$F(\boldsymbol{x}^{t+1}) \leq F(\boldsymbol{x}^{t}).$$

Remark. This is different from subgradient methods (for which objective value might be non-monotonic in t). Also, constant stepsize rule is recommended when f is convex and smooth.

Lemma 6. (key lemma) Let $\boldsymbol{y}^+ = prox_{\frac{1}{L}h}\Big(\boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y})\Big)$, then

$$F(\boldsymbol{y}^+) - F(\boldsymbol{x}) \leq \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 - \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}^+\|_2^2 - \underbrace{g(\boldsymbol{x},\boldsymbol{y})}_{\geq 0 \ \text{by convexity}}$$

where $g(\boldsymbol{x},\boldsymbol{y}) := f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle$.

Take $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{x}^t$ and hence $\boldsymbol{y}^+ = \boldsymbol{x}^{t+1}$ to complete the proof of Lemma 5.

Proof of Lemma 6. Define $\phi(\boldsymbol{z}) = f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{z} - \boldsymbol{y} \rangle + \frac{L}{2}\|\boldsymbol{z} - \boldsymbol{y}\|_2^2 + h(\boldsymbol{z})$. It is easily seen that $\boldsymbol{y}^+ = \arg\min_{\boldsymbol{z}} \phi(\boldsymbol{z})$. Two important properties:

1. Since $\phi(\boldsymbol{z})$ is L-strongly convex, one has

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

2. From smoothness,

$$\phi(\boldsymbol{y}^+) = \underbrace{f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{y}^+ - \boldsymbol{y} \rangle + \frac{L}{2}\|\boldsymbol{y}^+ - \boldsymbol{y}\|_2^2}_{\text{upper bound on } f(\boldsymbol{y}^+) \text{ (L-smoothness)}} + h(\boldsymbol{y}^+) \geq f(\boldsymbol{y}^+) + h(\boldsymbol{y}^+) = F(\boldsymbol{y}^+).$$

Proof of Lemma 6 (cont'd). Taken collectively, these yield

$$\phi(\boldsymbol{x}) \geq F(\boldsymbol{y}^{+}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^{+}\|_2^2,$$

which together with the definition of $\phi(\boldsymbol{x})$ gives

$$\underbrace{f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + h(\boldsymbol{x})}_{= f(\boldsymbol{x}) + h(\boldsymbol{x}) - g(\boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x}) - g(\boldsymbol{x}, \boldsymbol{y})} + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \geq F(\boldsymbol{y}^{+}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}^{+}\|_2^2$$

which finishes the proof.

Lemma 7. Suppose f is convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2 \le \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2.$$

Proof. From Lemma 6, taking $\boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{y} = \boldsymbol{x}^t$ (and hence $\boldsymbol{y}^+ = \boldsymbol{x}^{t+1}$) yields

$$\underbrace{F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^*)}_{\ge 0} + \underbrace{g(\boldsymbol{x}, \boldsymbol{y})}_{\ge 0} \le \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^{t+1}\|_2^2$$

which immediately concludes the proof.

Remark. Proximal gradient iterates are not only monotonic w.r.t. cost, but also monotonic in estimation error.

Theorem. [Convergence of proximal gradient methods for convex problems] Suppose f is convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$F(\boldsymbol{x}^t) - F^{opt} \leq \frac{L\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{2t}.$$

Proof. With Lemma 6 in mind, set $\boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{y} = \boldsymbol{x}^t$ to obtain

$$\begin{aligned} F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^*) &\le \frac{L}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 - \underbrace{g(\boldsymbol{x}^*, \boldsymbol{x}^t)}_{\ge 0 \text{ by convexity}} \\ &\le \frac{L}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \end{aligned}$$

Apply it recursively and add up all inequalities to get

$$\sum_{k=0}^{t-1}\Big(F(\boldsymbol{x}^{k+1}) - F(\boldsymbol{x}^*)\Big) \le \frac{L}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2.$$

This combines with monotonicity of $F(\boldsymbol{x}^t)$ (cf. Lemma 6) yields

$$F(\boldsymbol{x}^t) - F(\boldsymbol{x}^*) \le \frac{\frac{L}{2}\|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2}{t}.$$

Theorem. [Convergence of proximal gradient methods for strongly convex problems]
Suppose f is μ-strongly convex and L-smooth. If $\eta_t \equiv 1/L$, then

$$\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^t \|\boldsymbol{x}^0 - \boldsymbol{x}^*\|_2^2.$$

Proof. Taking $\boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{y} = \boldsymbol{x}^t$ (and hence $\boldsymbol{y}^+ = \boldsymbol{x}^{t+1}$) in Lemma 6 gives

$$
\begin{aligned}
F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^*) &\leq \frac{1}{L}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^{t+1}\|_2^2 - \underbrace{g(\boldsymbol{x}^*, \boldsymbol{x}^t)}_{\geq \frac{\mu}{2}\|\boldsymbol{x}^* - \boldsymbol{x}^t\|_2^2} \\
&\leq \frac{L-\mu}{2}\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2 - \frac{L}{2}\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2.
\end{aligned}
$$

This taken collectively with $F(\boldsymbol{x}^{t+1}) - F(\boldsymbol{x}^*) \geq 0$ yields

$$
\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\|_2^2 \leq (1 - \frac{\mu}{L})\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2.
$$

Applying it recursively concludes the proof.