Machine Learning II
A crash course on Optimization

Le Thi Khanh Hien
UMONS, thikhanhhien.le@umons.ac.be

Mons, March 2023

Table of content

1. Gradient descent method
2. Newton method
3. Proximal point algorithm

I. Gradient descent method

References:

- Jorge Nocedal and Stephen J. Wright, “Numerical Optimization”, Springer (2006).
- Y. Nesterov, “Lectures on Convex Optimization”, Springer Optimization and Its Applications book series, 2018.

Recap

General optimization problem

$$\begin{aligned} \min_{x} \quad & f(x) \\ s.t. \quad & x \in \mathcal{X}. \end{aligned}$$

- Convex optimization $\rightarrow$ a local minimizer is also a global minimizer.
- Fermat's optimality condition: Let $f \in \mathbb{E} \rightarrow \overline{\mathbb{R}}$ be a proper convex function. Then $\mathbf{x}^*$ is a minimizer to f if and only if $0 \in \partial f(\mathbf{x}^*)$.

How can we find optimal points?

Iterative Methods

To solve an optimization problem $x^* \in \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$, we start with some initial guess x^0, and then iteratively update x^k to produce a sequence $\{x^k\}_{k \geq 0}$ with the goal that the sequence converges to x^*, meaning $\|x^k - x^*\|$ for some norm $\|\cdot\|$ as $k \to \infty$.

Iterative Methods

To solve an optimization problem $x^* \in \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$, we start with some initial guess x^0, and then iteratively update x^k to produce a sequence $\{x^k\}_{k \geq 0}$ with the goal that the sequence converges to x^*, meaning $\|x^k - x^*\|$ for some norm $\|\cdot\|$ as $k \to \infty$.

Iterative Descent Methods

Consider the unconstrained optimization problem $\min_{x \in \mathbb{R}^n} f(x)$, where f is assumed to be continuously differentiable.

- If $\nabla f(x) = 0$: this is a candidate.
- If $\nabla f(x) \neq 0$: can we improve it? $f(x^{k+1}) < f(x^k)$?

Proposition 1.1

If $\nabla f(x)^\top d < 0$ then $\exists \delta$ such that $f(x + \alpha d) < f(x), \forall \alpha \in (0, \delta)$.

Proof. From the Taylor expansion

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x)^\top d + o(\alpha),$$

we have

$$f(x + \alpha d) - f(x) = \alpha(\nabla f(x)^\top d + o(\alpha)/\alpha)$$

Since $\lim_{\alpha \to 0} \frac{o(\alpha)}{\alpha} = 0$, there exists $\delta > 0$ such that $\left|\frac{o(\alpha)}{\alpha}\right| < -\nabla f(x)^\top d$ for all $\alpha \in (0, \delta)$.

Proposition 1.2

Suppose B is a positive definite matrix and $\nabla f(x) \neq 0$. Then $-B\nabla f(x)$ is a descent direction.

Proof.

Iterative Descent Methods for solving $\min_{x \in \mathbb{R}^n} f(x)$, where f is continuously differentiable.

$$x^{k+1} = x^k + \alpha_k d^k, \text{for } k = 0, 1, \ldots$$

- $\alpha_k > 0$: step-size
- d^k: descent direction.

Iterative Descent Methods for solving $\min_{x \in \mathbb{R}^n} f(x)$, where f is continuously differentiable.

$$x^{k+1} = x^k + \alpha_k d^k, \text{for } k = 0, 1, \ldots$$

- $\alpha_k > 0$: step-size
- d^k: descent direction.

Choice of direction

- Gradient descent $d^k = -\nabla f(x^k)$
- Diagonally scaled gradient descent $d^k = -B^k \nabla f(x^k)$, for some $B^k \succ 0$
- Newton direction $d^k = -(\nabla^2 f(x^k)^{-1} \nabla f(x^k)$ (suppose that $\nabla^2 f(x^k)^{-1} \succ 0$
- Modified Newton direction $d^k = -(\nabla^2 f(x^0)^{-1} \nabla f(x^k)$, for all k, or compute Newton direction once every m steps.

Iterative Descent Methods for solving $\min_{x\in\mathbb{R}^n} f(x)$, where f is continuously differentiable.

$$x^{k+1} = x^k + \alpha_k d^k, \text{for } k = 0, 1, \ldots$$

- $\alpha_k > 0$: step-size
- d^k: descent direction.

Choice of direction

- Gradient descent $d^k = -\nabla f(x^k)$
- Diagonally scaled gradient descent $d^k = -B^k \nabla f(x^k)$, for some $B^k \succ 0$
- Newton direction $d^k = -(\nabla^2 f(x^k)^{-1} \nabla f(x^k)$ (suppose that $\nabla^2 f(x^k)^{-1} \succ 0$
- Modified Newton direction $d^k = -(\nabla^2 f(x^0)^{-1} \nabla f(x^k)$, for all k, or compute Newton direction once every m steps.

Choice of step-size?

Gradient descent method for solving $\min_{x \in \mathbb{R}^n} f(x)$, where f is continuously differentiable.

Starting from an initial point x^0, update

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k),$$

where α_k is stepsize.

Gradient descent method for solving $\min_{x\in\mathbb{R}^n} f(x)$, where f is continuously differentiable.

Starting from an initial point x^0, update

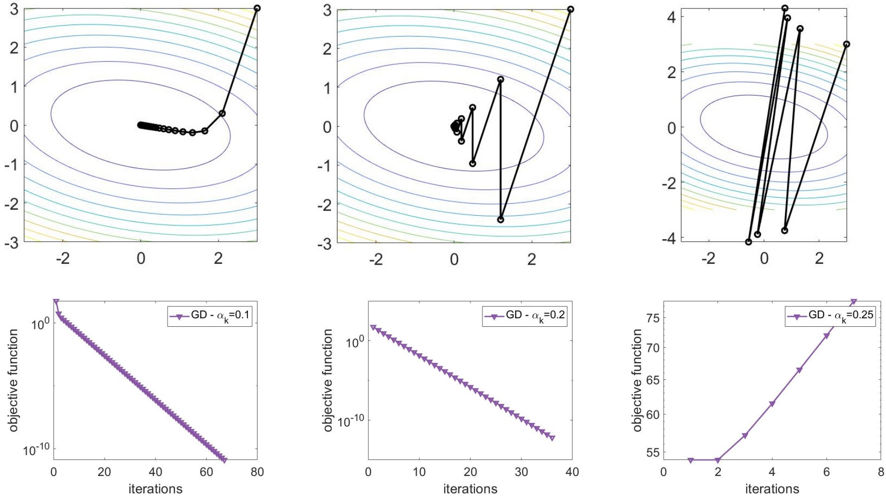
$$x^{k+1} = x^k - \alpha_k \nabla f(x^k),$$

where α_k is stepsize.

Example problem: $\min\limits_{x\in\mathbb{R}^2} f(x_1, x_2) = x_1^2 + x_1 x_2 + 4x_2^2.$

How do we choose a stepsize?

Example problem: $\min_{x \in \mathbb{R}^2} f(x_1, x_2) = x_1^2 + x_1 x_2 + 4x_2^2.$



How do we choose a stepsize?

- $\{\alpha_k\}$ is chosen in advance.
  - Choose $\alpha_k = \alpha$ for some constant $\alpha > 0$ (a constant stepsize). For example, if f is L-smooth then we can choose $0 < \alpha < \frac{2}{L}$.
  - Choose $\frac{\alpha}{\sqrt{k+1}}$ for some constant $\alpha > 0$.

- Backtracking line search. Fix two parameters $0 < \beta < 1$ and $0 < t \leq 0.5$. At iteration k: starting with $\alpha_k = 1$, while $f(x^k - \alpha_k \nabla f(x^k)) > f(x^k) - \alpha_k t \|\nabla f(x^k)\|_2^2$, shrink $\alpha_k = \beta \alpha_k$.

- Exact line search. Choose $\alpha_k = \arg\min_{s \geq 0} f(x^k - s\nabla f(x^k))$.

L-smooth function

Definition 1

A continuously differentiable function $f : \mathbb{E} \to \mathbb{R}$ is called an L-smooth function if

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|, \forall x, y \in \mathbb{E}.$$

Example

Show that f is L-smooth and determine L.

- $f(x) = \frac{1}{2}\|Ax - b\|_2^2$.

- (See Lab 2) Logistic regression loss $f(w) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-y^i \left\langle x^i, w\right\rangle\right)\right)$.

Proposition 1.3

L-smooth property of f implies the descent lemma

$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \forall x, y \in \mathbb{E}.$$

Proof. We have

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt$$

Hence,

$$\begin{aligned}
&f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\
&= \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\
&\le \int_0^1 L\|y - x\|^2 t dt = \frac{L}{2}\|y - x\|^2. \qquad \square
\end{aligned}$$

Proposition 1.3

L-smooth property of f implies the descent lemma

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \forall x, y \in \mathbb{E}.$$

Proof. We have

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt$$

Hence,

$$\begin{aligned} & f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ & = \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ & \leq \int_0^1 L\|y - x\|^2 t dt = \frac{L}{2}\|y - x\|^2. \end{aligned}$$

$\square$

There are a lot of other properties, see here:

`http://xingyuzhou.org/blog/notes/Lipschitz-gradient`.

Proposition 1.3

L-smooth property of f implies the descent lemma

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \forall x, y \in \mathbb{E}.$$

Proof. We have

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt$$

Hence,

$$\begin{aligned} & f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ & = \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \\ & \leq \int_0^1 L\|y - x\|^2 t dt = \frac{L}{2}\|y - x\|^2. \end{aligned}$$ □

There are a lot of other properties, see here:
`http://xingyuzhou.org/blog/notes/Lipschitz-gradient`.

Suppose f is twice continuously differentiable convex function.
f is L-smooth $\Longleftrightarrow x \mapsto \frac{L}{2}\|x\|^2 - f(x)$ is convex $\Longleftrightarrow \nabla^2 f(x) \preceq LI$

Convergence of GD for convex L-smooth function

Suppose f is convex and L-smooth. If $0 < \alpha < \frac{2}{L}$ then we have

- The sequence $\{x^k\}$ converges to a minimizer x^* of f.
- The following inequality holds for all $k \geq 0$

$$f(x^k) - f(x^*) \leq \frac{2(f(x^0) - f(x^*)\|x^0 - x^*\|^2}{2\|x^0 - x^*\|^2 + k\alpha(2 - L\alpha)(f(x^0) - f(x^*)}.$$

Corollary. If $\alpha = \frac{1}{L}$ then $f(x^k) - f(x^*) \leq \frac{2L\|x^0 - x^*\|^2}{k+4}$.

Convergence of GD for convex L-smooth function

Suppose f is convex and L-smooth. If $0 < \alpha < \frac{2}{L}$ then we have

- The sequence $\{x^k\}$ converges to a minimizer x^* of f.
- The following inequality holds for all $k \geq 0$

$$f(x^k) - f(x^*) \leq \frac{2(f(x^0) - f(x^*)\|x^0 - x^*\|^2}{2\|x^0 - x^*\|^2 + k\alpha(2 - L\alpha)(f(x^0) - f(x^*)}.$$

Corollary. If $\alpha = \frac{1}{L}$ then $f(x^k) - f(x^*) \leq \frac{2L\|x^0 - x^*\|^2}{k+4}$.

(See Lab 2) This is Theorem 2.1.14 in Y. Nesterov, Lectures on Convex Optimization, Springer Optimization and Its Applications book series, 2018. Read the proof and write it again.
https://tinyurl.com/2tu27k8b

Convergence of GD for strongly convex L-smooth function

Recall that f is convex and L-smooth. If there is a constant $\mu > 0$ such that $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex, we say f is μ-strongly convex. This condition is equivalent to

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2, \forall x, y.$$

The value $\kappa = \frac{L}{\mu} \geq 1$ is called the condition number of the function f.

Convergence of GD for strongly convex L-smooth function

Recall that f is convex and L-smooth. If there is a constant $\mu > 0$ such that $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex, we say f is μ-strongly convex. This condition is equivalent to

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2, \forall x, y.$$

The value $\kappa = \frac{L}{\mu} \geq 1$ is called the condition number of the function f.

Suppose f is μ-strongly convex and L-smooth. If $\alpha = \frac{2}{\mu+L}$, then

$$f(x^k) - f^* \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k}\|x^0 - x^*\|^2, \text{and} \quad \|x^k - x^*\|^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^{k}\|x^0 - x^*\|.$$

See Theorem 2.1.15 in Y. Nesterov, Lectures on Convex Optimization, Springer Optimization and Its Applications book series, 2018.

Iteration Complexity

II. Newton method and quasi-Newton method

References:

- Jorge Nocedal and Stephen J. Wright, “Numerical Optimization”, Springer (2006).
- Roger Fletcher, “Practical Methods of Optimization”, 2000.

Newton's method for systems of nonlinear equations.

Suppose $\phi : \mathbb{R}^n \to \mathbb{R}^n$, $\phi(x) = (\phi_1(x), \ldots, \phi_n(x))$. We need to solve $\phi(x) = 0$:

$$\begin{aligned}
\phi_1(x_1, \ldots, x_n) &= 0, \\
\phi_2(x_1, \ldots, x_n) &= 0, \\
&\vdots \\
\phi_n(x_1, \ldots, x_n) &= 0.
\end{aligned}$$

Key steps: Given an iterate $x^{(k)}$,

- **Linearization:** $\phi(x) \approx \phi(x^{(k)}) + J\phi(x^{(k)})(x - x^{(k)})$, where $J\phi(x)$ is the Jacobian of ϕ at x.
- **Solve** $\phi(x^{(k)}) + J\phi(x^{(k)})(x - x^{(k)}) = 0$ instead of $\phi(x) = 0$.

Newton's method:

Given an iterate $x^{(k)}$, update $x^{(k+1)} = x^{(k)} - J\phi(x^{(k)})^{-1}\phi(x^{(k)})$.

Newton's method for unconstrained optimization.

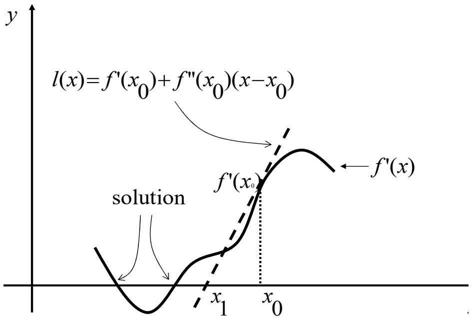
Let f be a twice differentiable function. Suppose we want to solve

$$\min_{x \in \mathbb{R}^n} f(x).$$

First-order optimality condition: $\nabla f(x) = 0$.

Pure Newton's method for solving $\nabla f(x) = 0$. Given an iterate $x^{(k)}$, update

$$x^{(k+1)} = x^{(k)} - \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}).$$



Newton's method

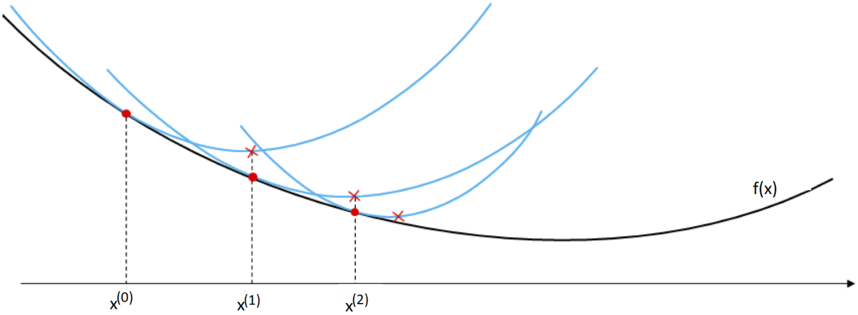
Another interpretation: quadratic Taylor approximation

$$f(x) \approx h(x)$$

$$:= f(x^{(k)}) + \nabla f(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2}(x - x^{(k)}) \nabla^2 f(x^{(k)})(x - x^{(k)}).$$

Minimizing $h(x)$ yields the update of Newton's method since

$$\nabla h(x) = \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})(x - x^{(k)}) = 0.$$

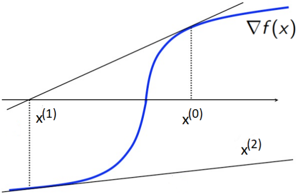


Some remarks.

Given an iterate $x^{(k)}$, update

$$x^{(k+1)} = x^{(k)} - \nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}).$$

- The Newton's direction is not defined when $\nabla^2 f(x^{(k)})$ is not invertible.
- The method can diverge when started far from a solution.



- If $\nabla^2 f(x^{(k)})$ is a positive definite matrix, then $d^{(k)} = -\nabla^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$ is a descent direction.

Local quadratic convergence of pure Newton's method.

Theorem 2.1

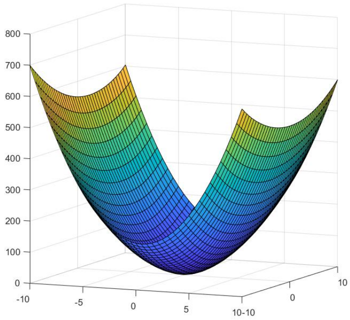
Let x^* be a minimizer ò f. Suppose that f is twice continuously differentiable in an open neighborhood of x^*. Suppose also that the Hessian $\nabla^2 f$ is Lipschitz continuous near x^* and $\nabla^2 f(x^*)$ is positive definite. If x_0 is sufficiently close to x^*, then the generated sequence $\{x^k\}$ is well defined and converges to x^* quadratically, that is,

$$\left\|x^{k+1} - x^*\right\| \leq C \left\|x^k - x^*\right\|^2,$$

where C is a constant.

An example

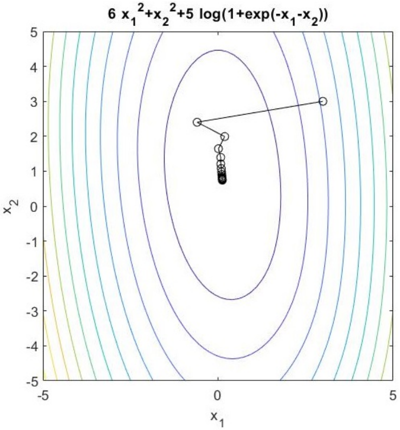
$$\min_x f(x) = 6x_1^2 + x_2^2 + 5\log(1 + e^{-x_1 - x_2}).$$



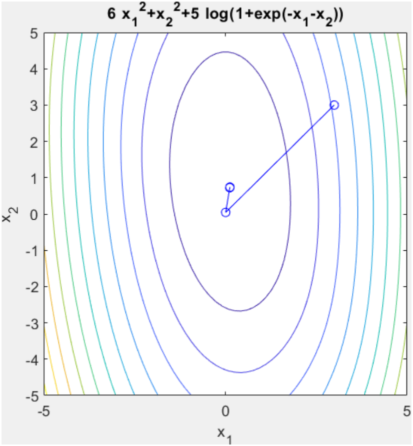
$$\nabla f(x) = \begin{bmatrix} 12x_1 - 5 + \frac{5}{1+e^{-x_1-x_2}} & 2x_2 - 5 + \frac{5}{1+e^{-x_1-x_2}} \end{bmatrix}^T$$

$$\nabla^2 f(x) = \begin{bmatrix} 12 + 5\frac{e^{-x_1-x_2}}{(1+e^{-x_1-x_2})^2} & 5\frac{e^{-x_1-x_2}}{(1+e^{-x_1-x_2})^2} \\ 5\frac{e^{-x_1-x_2}}{(1+e^{-x_1-x_2})^2} & 2 + 5\frac{e^{-x_1-x_2}}{(1+e^{-x_1-x_2})^2} \end{bmatrix}$$

An example



Gradient method



Newton's method

(Optional reading)

Definition of Newton decrement

$$\begin{aligned}
\delta(x^k) &:= \|d^k\|_{\nabla^2 f(x^k)} \\
&= \left(d^k \nabla^2 f(x^k) d^k\right)^{1/2} \\
&= \left(\nabla f(x^k)^\top \nabla^2 f(x^k)^{-1} \nabla f(x^k)\right)^{1/2}.
\end{aligned}$$

$\delta(x^k)^2/2$ is an approximate bound for the optimality gap

We have

$$\begin{aligned}
& f(x) - \min_y \left(f(y) + \nabla f(y)^T (x-y) + \frac{1}{2}(x-y)\nabla^2 f(y)(x-y)\right) \\
&= f(x) - \left(f(x) - \frac{1}{2}\nabla f(x^k)^T \nabla^2 f(x^k)^{-1} \nabla f(x^k)\right) \\
&= \frac{1}{2}\delta(x^k)^2.
\end{aligned}$$

Newton's method - full description

Initialize: Choosing initial point $x_0 \in \mathrm{dom} f$ and an error tolerance $\varepsilon > 0$.
for $k = 1, \ldots$ **do**
    Calculate $d^k = -\nabla f^2(x^k)^{-1} \nabla f(x^k)$ and $\delta_k^2 = \nabla f(x^k)^T \nabla^2 f(x^k)^{-1} \nabla f(x^k)$.
    if $\delta_k^2/2 \leq \varepsilon$ **then**
        stop
    end if
    Choose a stepsize α_k and update $x^{k+1} = x^k + \alpha_k d^k$.
end for

Algorithm 1: Newton's method

Damped Newton's method

Choose stepsize α_k by backtracking line search:

- Choose $0 < \sigma \leq 1/2$ and $0 < \beta < 1$.
- Start with $\alpha = 1$ and while

$$f(x^k + \alpha d^k) > f(x^k) + \sigma\alpha\nabla f(x^k)^\top d^k,$$

  we shrink $\alpha = \beta\alpha$.
- Let $\alpha_k = \alpha$.

(Optional reading)

Quasi-Newton method.

- Quasi-Newton's method replaces the Hessian $\nabla^2 f(x^k)$ by its approximation matrix G_k.
- G_k is a positive definite matrix so that the direction $d_k = -G_k^{-1}\nabla f(x^k)$ is a descent direction.
- Suppose we can calculate $\nabla f(x^{k-1})$, $\nabla f(x^k)$ and want to estimate $\nabla^2 f(x^k)$. Note that

$$\nabla f(x^{k-1}) - \nabla f(x^k) = \nabla^2 f(x^k)(x^{k-1} - x^k) + o\left(\left\|x^{k-1} - x^k\right\|\right).$$

  We want G_k to satisfy

$$\nabla f(x^{k-1}) - \nabla f(x^k) = G_k(x^{k-1} - x^k),$$

  or, equivalently, we want the following quasi-Newton condition to be satisfied

$$H_k(\nabla f(x^{k-1}) - \nabla f(x^k)) = x^{k-1} - x^k,$$

  where $H_k = G_k^{-1}$.

Denote $\gamma_k = \nabla f(x^k) - \nabla f(x^{k-1})$ and $\delta_k = x^k - x^{k-1}$. We rewrite the quasi-Newton condition as follows.

$$H_k \gamma_k = \delta_k. \tag{1}$$

Broyden's method (proposed by Charles George Broyden, 1967).
Idea: suppose H_k is a rank one correction of H_{k-1}, that is

$$H_k = H_{k-1} + auu^\top.$$

Then, from (1) we have

$$H_{k-1}\gamma_k + auu^\top \gamma_k = \delta_k$$

We can choose

$$u = \delta_k - H_{k-1}\gamma_k, \quad a = \frac{1}{\gamma_k^\top(\delta_k - H_{k-1}\gamma_k)}.$$

Update of H_k:

$$H_k = H_{k-1} + auu^\top = H_{k-1} + \frac{(\delta_k - H_{k-1}\gamma_k)(\delta_k - H_{k-1}\gamma_k)^\top}{\gamma_k^\top(\delta_k - H_{k-1}\gamma_k)}$$

Disadvantage: H_k may not be always positive. definite.

DFP Method (proposed by Davidon, Fletcher and Powell).
Idea: use rank two correction from H_{k-1}, that is

$$H_k = H_{k-1} + auu^\top + bvv^\top.$$

Then, from (1) we have

$$H_{k-1}\gamma_k + auu^\top\gamma_k + bvv^\top = \delta_k.$$

An obvious solution is

$$u = \delta_k, \quad v = H_{k-1}\gamma_k, \quad a = \frac{1}{u^\top\gamma_k}, \quad b = -\frac{1}{v^\top\gamma_k}.$$

Update of H_k:

$$H_k = H_{k-1} + \frac{\delta_k\delta_k^\top}{\delta_k^\top\gamma_k} - \frac{H_{k-1}\gamma_k\gamma_k^\top H_{k-1}}{(H_{k-1}\gamma_k)^\top\gamma_k}.$$

Note: If H_{k-1} is positive definite then H_k is also positive definite.

BFGS Method (proposed by Davidon, Fletcher and Powell).
The quasi-Newton condition can be rewritten as $\gamma_k = G_k \delta_k$.
Idea: Use the DFP formula to obtain G_k from G_{k-1}. This can be done by replacing H_k with G_k, H_{k-1} with G_{k-1} and swap γ_k and δ_k in the DFP formula for H_k

$$G_k = G_{k-1} + \frac{\gamma_k \gamma_k^\top}{\gamma_k^\top \delta_k} - \frac{G_{k-1}\delta_k \delta_k^\top G_{k-1}}{(G_{k-1}\delta_k)^\top \delta_k},$$

which implies

$$G_k^{-1} = G_{k-1}^{-1} + \left(1 + \frac{\gamma_k^\top G_{k-1}^{-1} \gamma_k}{\delta_k^\top \gamma_k}\right) \frac{\delta_k \delta_k^\top}{\delta_k^\top \gamma_k} - \frac{\delta_k \gamma_k^\top G_{k-1}^{-1} + G_{k-1}^{-1} \gamma_k \delta_k^\top}{\delta_k^\top \gamma_k}.$$

Hence

$$H_k = H_{k-1} + \left(1 + \frac{\gamma_k^\top H_{k-1} \gamma_k}{\delta_k^\top \gamma_k}\right) \frac{\delta_k \delta_k^\top}{\delta_k^\top \gamma_k} - \frac{\delta_k \gamma_k^\top H_{k-1} + H_{k-1} \gamma_k \delta_k^\top}{\delta_k^\top \gamma_k}.$$

III.Proximal point algorithm

References:

- N Parikh, S Boyd, “Proximal Algorithms”, Foundations and Trends in Optimization 1(3), 2014. Link
- Amir Beck, “First-Order Methods in Optimization”, MOS-SIAM Series on Optimization, 2017

Problem setting

We consider the following convex composite optimization problem

$$\min_{x \in \mathbb{E}} F(x) := f(x) + g(x), \tag{2}$$

where $f : \mathbb{E} \to \mathbb{R}$ is a differentiable convex function and $g : \mathbb{E} \to \overline{\mathbb{R}}$ is a proper lower-semicontinuous convex function.

Assumptions

- f is L-smooth, which is equivalent to $x \mapsto \frac{L}{2}\|x\|^2 - f(x)$ is convex.
- f is μ-strongly convex ($\mu \geq 0$).
- The optimal value F^* is attained at x^*.

Example

- General inverse problems: given a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $c \in \mathbb{R}^m$, solve

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - c\|_2^2 + \lambda R(x),$$

where $R(x)$ is a regularizer.

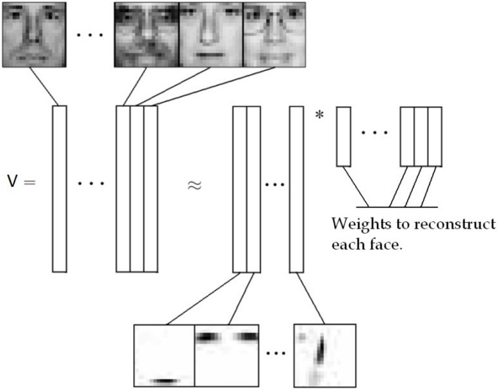
- l_1-regularized logistic regression

$$\min_{w \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y^i \left\langle x^i, w \right\rangle\right)\right) + \lambda \|w\|_1.$$

- Generalized constrained low-rank matrix factorization. Given a matrix $M \in \mathbb{R}_+^{m\times n}$ and an integer factorization rank $r > 0$, find

$$\min_{\substack{W \in \Omega_W \subseteq \mathbb{R}^{m\times r} \\ H \in \Omega_H \subseteq \mathbb{R}^{r\times n}}} f(M|WH) + R(W,H),$$

where $f(M|WH)$ is a cost function that measures the difference between M and WH, and R is a regularizer.



The basis elements extract facial features such as eyes, nose and lips.

Recall

Subdifferential of a convex function. Let $f : \mathbb{E} \to \overline{\mathbb{R}}$ be a proper convex function and $\bar{\mathbf{x}} \in \mathrm{dom} f$. A vector $v \in \mathbb{E}^*$ is called a subgradient of f at $\bar{\mathbf{x}}$ if

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \langle v, \mathbf{x} - \bar{\mathbf{x}} \rangle \quad \text{for all} \quad \mathbf{x} \in \mathbb{E}.$$

The subdifferential of f at $\bar{\mathbf{x}}$ is defined by

$$\partial f(\bar{\mathbf{x}}) := \{v \in \mathbb{E}^* |\ f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \langle v, \mathbf{x} - \bar{\mathbf{x}} \rangle \, \forall \mathbf{x} \in \mathbb{E}\}.$$

Fermat's optimality condition. Let $f \in \mathbb{E} \to \overline{\mathbb{R}}$ be a proper convex function. Then $\mathbf{x}^*$ is a minimizer to f if and only if $0 \in \partial f(\mathbf{x}^*)$.

Recall

Subdifferential of a convex function. Let $f : \mathbb{E} \to \overline{\mathbb{R}}$ be a proper convex function and $\bar{\mathbf{x}} \in \mathrm{dom} f$. A vector $v \in \mathbb{E}^*$ is called a subgradient of f at $\bar{\mathbf{x}}$ if

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \langle v, \mathbf{x} - \bar{\mathbf{x}} \rangle \quad \text{for all} \quad \mathbf{x} \in \mathbb{E}.$$

The subdifferential of f at $\bar{\mathbf{x}}$ is defined by

$$\partial f(\bar{\mathbf{x}}) := \{v \in \mathbb{E}^* | \ f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \langle v, \mathbf{x} - \bar{\mathbf{x}} \rangle \, \forall \mathbf{x} \in \mathbb{E}\}.$$

Fermat's optimality condition. Let $f \in \mathbb{E} \to \overline{\mathbb{R}}$ be a proper convex function. Then $\mathbf{x}^*$ is a minimizer to f if and only if $0 \in \partial f(\mathbf{x}^*)$.

(Optional reading)
Subgradient method
`https://stanford.edu/class/ee364b/lectures/subgrad_method_notes.pdf`

Let $\mathbb{E}$ be a Euclidean space.

Definition 2

Let $f : \mathbb{E} \to \overline{\mathbb{R}}$ be a proper l.s.c. (convex) function. We define the **proximal (prox) operator** $\mathrm{prox}_f : \mathbb{E} \to \mathbb{E}$ of f by

$$\mathrm{prox}_f(x) := \arg\min_{u \in \mathbb{E}} \left\{ f(u) + \frac{1}{2}\|u - x\|^2 \right\}.$$

Quiz 1

Do the convex function f and its proximal operator have the same domain?

Let $\mathbb{E}$ be a Euclidean space.

Definition 2

Let $f : \mathbb{E} \to \overline{\mathbb{R}}$ be a proper l.s.c. (convex) function. We define the **proximal (prox) operator** $\mathrm{prox}_f : \mathbb{E} \to \mathbb{E}$ of f by

$$\mathrm{prox}_f(x) := \arg\min_{u \in \mathbb{E}} \left\{ f(u) + \frac{1}{2}\|u - x\|^2 \right\}.$$

Quiz 1

Do the convex function f and its proximal operator have the same domain?

Proposition 3.1

Let $f : \mathbb{E} \to \overline{\mathbb{R}}$ be a proper l.s.c. convex function. Then prox_f is a well-defined mapping with full domain. Moreover, we have

$$u = \mathrm{prox}_f(x) \Leftrightarrow x - u \in \partial f(u)$$

and

$$\mathrm{prox}_f(x) = (\mathrm{Id} + \partial f)^{-1}(x) \quad \textit{for all} \quad x \in \mathbb{E}.$$

Quiz 2

Find prox_f *with* $f = \delta_D$*, where* D *is a closed convex set in* $\mathbb{E}$*.*

(A) $\operatorname{prox}_f(\mathbf{x}) = \mathbf{x}$.

(B) $\operatorname{prox}_f(\mathbf{x}) = \Pi_D(\mathbf{x})$.

(C) $\operatorname{prox}_f(\mathbf{x}) = D$.

Some Prox calculus rules

Proposition 3.2 (prox of separable functions)

Suppose that $f : \mathbb{E}_1 \times \ldots \times \mathbb{E}_s \to (-\infty, \infty]$ *satisfies the condition*

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_s) = \sum_{i=1}^{s} f_i(\mathbf{x}_i), \textit{ for any } \mathbf{x}_i \in \mathbb{E}_i,\ i = 1, \ldots, s.$$

Then for any $x_i \in \mathbb{E}_i$, $i = 1, \ldots, s$ *we have*

$$\mathrm{prox}_f(\mathbf{x}_1, \ldots, \mathbf{x}_s) = \mathrm{prox}_{f_1}(\mathbf{x}_1) \times \ldots \times \mathrm{prox}_{f_s}(\mathbf{x}_s),$$

Quiz 3

Suppose $\mathbb{E} = \mathbb{R}^n$. *Find* prox_f *with* $f(x) = t\|x\|_1$ *and* $t > 0$.

(A)

$$(\mathrm{prox}_f(\mathbf{x}))_i = \begin{cases} \mathbf{x}_i - t, & \text{if } \mathbf{x}_i < -t, \\ 0, & \text{if } |\mathbf{x}_i| \leq t, \\ \mathbf{x}_i + t, & \text{if } \mathbf{x}_i > t. \end{cases}$$

(B) $(\mathrm{prox}_f(\mathbf{x}))_i = \mathrm{sign}(\mathbf{x}_i)\left[t - |\mathbf{x}_i|\right]_+$.

(C) $(\mathrm{prox}_f(\mathbf{x}))_i = \mathrm{sign}(\mathbf{x}_i)\left[|\mathbf{x}_i| - t\right]_+$.

Proposition 3.3 (post-composition)

Let $g : \mathbb{E} \to (-\infty, \infty]$ *be a proper function. If* $f(\mathbf{x}) = \alpha g(\mathbf{x}) + a$*, with* $\alpha > 0$ *and* $a \in \mathbb{R}$*, then*

$$\mathrm{prox}_f(\mathbf{x}) = \mathrm{prox}_{\alpha g}(\mathbf{x}).$$

Proposition 3.4 (pre-composition)

Let $g : \mathbb{E} \to (-\infty, \infty]$ *be a proper function. Let* $\alpha \neq 0$ *and* $\mathbf{y} \in \mathbb{E}$*. Suppose* $f(\mathbf{x}) = g(\alpha \mathbf{x} + \mathbf{y})$*. Then we have*

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

(See Lab 2) Find prox_f with $f : \mathbb{R}^n \to \mathbb{R}, f(x) = t\|x\|_2^2$ and $t > 0$.

Proposition 3.3 (post-composition)

Let $g : \mathbb{E} \to (-\infty, \infty]$ *be a proper function. If* $f(\mathbf{x}) = \alpha g(\mathbf{x}) + a$, *with* $\alpha > 0$ *and* $a \in \mathbb{R}$, *then*

$$\mathrm{prox}_f(\mathbf{x}) = \mathrm{prox}_{\alpha g}(\mathbf{x}).$$

Proposition 3.4 (pre-composition)

Let $g : \mathbb{E} \to (-\infty, \infty]$ *be a proper function. Let* $\alpha \neq 0$ *and* $\mathbf{y} \in \mathbb{E}$. *Suppose* $f(\mathbf{x}) = g(\alpha \mathbf{x} + \mathbf{y})$. *Then we have*

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

(See Lab 2) Find prox_f with $f : \mathbb{R}^n \to \mathbb{R}, f(x) = t\|x\|_2^2$ and $t > 0$.

More formulas and codes:

http://proximity-operator.net/proximityoperator.html

Proximal gradient (PG) method

$\min_{x\in\mathbb{E}} \quad f(x) + g(x).$

Starting from an initial point x^0, update

$$x^{k+1} = \operatorname{prox}_{\lambda_k g}\Big(x^k - \lambda_k \nabla f(x^k)\Big).$$

Proximal gradient (PG) method

$\min_{x \in \mathbb{E}} \quad f(x) + g(x).$

Starting from an initial point x^0, update

$$x^{k+1} = \mathrm{prox}_{\lambda_k g}\Big(x^k - \lambda_k \nabla f(x^k)\Big).$$

Example

- Gradient method. When $g(x) = 0$

- Proximal point algorithm. When $f(x) = 0$

- Gradient projection method. When $g(x) = \delta_D(x)$

Quiz 4

PG method for solving

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-c\|_2^2 + \|x\|_1$$

has the update rule

(A) $(x^{k+1})_i = \mathrm{sign}\big((y^k)_i\big)\max\big\{|(y^k)_i| - \lambda_k, 0\big\}$, *for* $i = 1, \ldots, n$, *where* $y^k = x^k - \lambda_k A^\top(Ax^k - c)$.

(B) $(x^{k+1})_i = \mathrm{sign}\big((y^k)_i\big)\max\big\{|(y^k)_i| - \lambda_k, 0\big\}$, *for* $i = 1, \ldots, n$, *where* $y^k = x^k + \lambda_k A^\top(Ax^k - c)$.

(C) $x^{k+1} = \mathrm{prox}_{\|x\|_1}\big(x^k - \lambda_k A^\top(Ax^k - c)\big)$.

(Optional reading)

Let $x^+ = \mathrm{prox}_{\lambda g}\Big(x - \lambda \nabla f(x)\Big)$.

- x^+ minimizes g plus a simple quadratic local model of f around x.

$$\begin{aligned} x^+ &= \operatorname*{argmin}_u \lambda g(u) + \frac{1}{2}\|u - (x - \lambda \nabla f(x))\|^2 \\ &= \operatorname*{argmin}_u g(u) + f(x) + \langle \nabla f(x), u - x\rangle + \frac{1}{2\lambda}\|u - x\|^2. \end{aligned}$$

- Fixed point iteration. x^* is a solution of $\min_x f(x) + g(x)$ if and only if

$$\begin{aligned} 0 \in \nabla f(x^*) + \partial g(x^*) &\Leftrightarrow 0 \in \lambda \nabla f(x^*) + \lambda \partial g(x^*), \text{where } \lambda > 0 \\ &\Leftrightarrow (x^* - \lambda \nabla f(x^*)) \in (\mathrm{Id} + \lambda \partial g)(x^*) \\ &\Leftrightarrow x^* \in (\mathrm{Id} + \lambda \partial g)^{-1}(x^* - \lambda \nabla f(x^*)) \\ &\Leftrightarrow x^* = \mathrm{prox}_{\lambda g}(x^* - \lambda \nabla f(x^*)) \end{aligned}$$

We define gradient map as follows:

$$G_\lambda(x) = \frac{1}{\lambda}\Big(x - \mathrm{prox}_{\lambda g}\big(x - \lambda \nabla f(x)\big)\Big).$$

Then we have

$$\begin{aligned} x^+ &= \mathrm{prox}_{\lambda g}\big(x - \lambda \nabla f(x)\big) \\ &= x - \lambda G_\lambda(x). \end{aligned}$$

Note:

- $G_\lambda(x)$ is not a subgradient of $F = f + g$.
- We have $G_\lambda(x) = 0$ if and only if x minimizes $f(x) + g(x)$.
- We have $G_\lambda(x) - \nabla f(x) \in \partial g(x - tG_\lambda(x))$.

Convergence properties

Proposition 3.5

Suppose $0 < \lambda_k = \lambda \leq \frac{1}{L}$. *We have*

- **Property 1:** *PG algorithm is a descent method.*
- **Property 2:** $F(x^k) - F^* \leq \frac{1}{2\lambda} \left\|x^0 - x^*\right\|^2$.
- **Property 3:** $\left\|x^k - x^*\right\|^2 \leq \left(1 - \frac{\mu}{L}\right)^k \left\|x^0 - x^*\right\|^2$.

Proof of Property 1.

(Home reading)

Implications of assumptions

- L-smooth property of f implies

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \forall x, y \in \mathbb{E}.$$

- Convexity of $f(\cdot) - (\mu/2)\|\cdot\|^2$ implies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2, \forall x, y \in \mathbb{E}.$$

Substitute $y = x - \lambda G_\lambda(x)$ in these bounds we have

$$\frac{\mu\lambda^2}{2}\|G_\lambda(x)\|^2 \leq f\left(x - \lambda G_\lambda(x)\right) - f(x) + \lambda \langle \nabla f(x), G_\lambda(x) \rangle \leq \frac{L\lambda^2}{2}\|G_\lambda(x)\|^2$$

For all z we have

$$
\begin{aligned}
F\big(x - \lambda G_\lambda(x)\big) &\leq f(x) - \lambda \left\langle \nabla f(x), G_\lambda(x) \right\rangle + \frac{\lambda}{2} \left\| G_\lambda(x) \right\|^2 + g\big(x - \lambda G_\lambda(x)\big) \\
&\leq f(z) - \left\langle \nabla f(x), z - x \right\rangle - \frac{m}{2} \left\| z - x \right\|^2 \\
&\qquad - \lambda \left\langle \nabla f(x), G_\lambda(x) \right\rangle + \frac{\lambda}{2} \left\| G_\lambda(x) \right\|^2 + g\big(x - \lambda G_\lambda(x)\big) \\
&\leq f(z) - \left\langle \nabla f(x), z - x \right\rangle - \frac{m}{2} \left\| z - x \right\|^2 - \lambda \left\langle \nabla f(x), G_\lambda(x) \right\rangle \\
&\qquad + \frac{\lambda}{2} \left\| G_\lambda(x) \right\|^2 + g(z) - \left\langle G_\lambda(x) - \nabla f(x), z - x + \lambda G_\lambda(x) \right\rangle \\
&= f(z) + g(z) + \left\langle G_\lambda(x), x - z \right\rangle - \frac{\lambda}{2} \left\| G_\lambda(x) \right\|^2 - \frac{m}{2} \left\| z - x \right\|^2 .
\end{aligned}
$$

Let $x^+ = x - \lambda G_\lambda(x)$. Taking $z = x$, we have

$$
F(x^+) \leq F(x) - \frac{\lambda}{2} \left\| G_\lambda(x) \right\|^2 .
$$

Hence, PG algorithm is a descent method.

Proof of Property 2

Taking $z = x^*$ we have

$$\begin{aligned} F(x^+) - F(x^*) &\le \langle G_\lambda(x), x - x^* \rangle - \frac{\lambda}{2} \|G_\lambda(x)\|^2 - \frac{\mu}{2} \|x - x^*\|^2 \\ &= \frac{1}{2\lambda} \left(\|x - x^*\|^2 - \|x - x^* - \lambda G_\lambda(x)\|^2 \right) - \frac{\mu}{2} \|x - x^*\|^2 \qquad (3) \\ &= \frac{1}{2\lambda} \left((1 - \mu\lambda) \|x - x^*\|^2 - \left\|x^+ - x^*\right\|^2 \right). \end{aligned}$$

Hence,

$$F(x^+) - F(x^*) \le \frac{1}{2\lambda} \left(\|x - x^*\|^2 - \left\|x^+ - x^*\right\|^2 \right). \qquad (4)$$

Adding the Inequality (4) with $x = x^i$, $x^+ = x^{i+1}$ from $i = 0$ to $i = k - 1$, we have

$$\begin{aligned} \sum_{i=1}^{k} \left(F(x^k) - F^* \right) &\le \frac{1}{2\lambda} \sum_{i=0}^{k-1} \left(\left\|x^i - x^*\right\|^2 - \left\|x^{i+1} - x^*\right\|^2 \right) \\ &\le \frac{1}{2\lambda} \left\|x^0 - x^*\right\|^2. \end{aligned}$$

Since $F(x^i)$ is nonincreasing, we have

Distance to optimal set - Proof of Property 3

From Inequality (4) we have

$$\left\|x^{+} - x^{*}\right\|^{2} \leq \|x - x^{*}\|^{2}.$$

Hence, the distance to the optimal set does not increase. When $\lambda_k = \frac{1}{L}$, from Inequality (3) we have

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

Therefore,

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

This is linear convergence rate if f is strongly convex ($\mu > 0$).

Line search

- If L is not known, we can apply back-tracking line search: start at some $\lambda = \lambda^0$ and back-track $\lambda = C\lambda^0$, with $0 < C < 1$, until the following inequality holds

$$f\left(x - \lambda G_\lambda(x)\right) \le f(x) - \lambda \langle \nabla f(x), G_\lambda(x) \rangle + \frac{\lambda}{2} \left\| G_\lambda(x) \right\|^2 .$$

  (This inequality holds for $0 < \lambda \le \frac{1}{L}$.)
- The step size λ_i selected by the line search satisfies $\lambda_i \ge \lambda_{\min} = \min\left\{\lambda^0, \frac{C}{L}\right\}$.
- We obtain a similar $O(1/k)$ rate as for the case of using fixed step size

$$F(x^k) - F^* \le \frac{1}{2\sum_{i=0}^{k-1} \lambda_i} \left\| x^0 - x^* \right\|^2 \le \frac{1}{2k\lambda_{\min}} \left\| x^0 - x^* \right\|^2 .$$

- Distance to optimal set

$$\left\| x^k - x^* \right\|^2 \le (1 - m\lambda_{\min})^k \left\| x^0 - x^* \right\|^2 .$$

Quiz 5

Consider the following proximal gradient method for solving the convex composite problem $\min_x f(x) + g(x)$*: starting from an initial point* x^0*, update*

$$x^{k+1} = \mathrm{prox}_{\lambda_k g}\Big(x^k - \lambda_k \nabla f(x^k)\Big).$$

Suppose f *is L-smooth and we need to find an* ε*-optimal solution (that is,* x_ε^* *such that* $F(x_\varepsilon^*) - F^* \leq \varepsilon$*). Estimate the number of iterations of the PG method to obtain an* ε*-optimal solution:*

(A) $\frac{c}{\varepsilon^2}$, *where* c *is a constant.*

(B) $\frac{c}{\varepsilon}$, *where* c *is a constant.*

(C) $\frac{c}{\sqrt{\varepsilon}}$, *where* c *is a constant.*

Find more results in [Amir Beck, "First-Order Methods in Optimization", MOS-SIAM Series on Optimization, 2017].

- The generated sequence $\{x^k\}_{k\geq 0}$ converges to an optimal solution of Problem (2).
- $O(1/k)$ rate of convergence of the norm of the gradient mapping.
- Nonconvex case