CONVEX OPTIMIZATION

I-Liang Chern

Department of Mathematics
National Taiwan University

Fall, 2017

Contents

Chapter 1

Convex Analysis

Main references:

- Vandenberghe (UCLA): EECS236C - Optimization methods for large scale systems, `http://www.seas.ucla.edu/~vandenbe/ee236c.html`
- Y. Nesterov, Introductory Lectures on Convex Optimization, A Basic Course 1998.
- Parikh and Boyd, Proximal algorithms, slides and note. `http://stanford.edu/~boyd/papers/prox_algs.html` or Neal Parikh and Stephen Boyd, Proximal Algorithms, Foundations and Trend in Optimization Vol. 1, No. 3 (2013) 123?231.
- Boyd, ADMM `http://stanford.edu/~boyd/admm.html`
- Simon Foucart and Holger Rauhut, Appendix B.
- Ahmad Bazzi's youtube on convex optimization

1.1 Motivations: Convex optimization problems

Some examples of optimization problems In applications, we encounter many constrained optimization problems. Examples are

- Basis pursuit: exact sparse recovery problem

$$\min \|\mathbf{x}\|_1 \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{b}.$$

  or robust recovery problem

$$\min \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \le \epsilon.$$

- Image processing:

$$\min \|\nabla \mathbf{x}\|_1 \text{ subject to } \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \le \epsilon.$$

- Sometimes, the constraint can be described as a convex set $\mathcal{C}$. That is,

$$\min_x f_0(x) \text{ subject to } Ax \in \mathcal{C}.$$

Define the indicator function

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

We can rewrite the constrained minimization problem as a unconstrained minimization problem:

$$\min_x f_0(x) + \iota_{\mathcal{C}}(Ax).$$

This can also be reformulated as

$$\min_{x,y} f_0(x) + \iota_{\mathcal{C}}(y) \text{ subject to } Ax = y.$$

- In abstract form, we encounter the optimization problem:

$$\min f(x) + g(Ax)$$

This can can also be expressed as

$$\min f(x) + g(y) \quad \text{subject to} \quad Ax = y.$$

- For more applications, see Boyd's book.

A general form of convex optimization problems A standard convex optimization problem can be formulated as

$$\begin{aligned} &\min_{\mathbf{x} \in X} f_0(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y} \\ &\text{and} \qquad f_i(\mathbf{x}) \le b_i, \quad i = 1, ..., M \end{aligned}$$

Here, f_i's are convex. The space X is a Hilbert space. Here, we just take $X = \mathbb{R}^N$.

1.2 Convex sets

- **Convex set** A set $K \subset \mathbb{R}^N$ is called convex if for any $\mathbf{x}, \mathbf{y} \in K$, the line segment $(1-t)\mathbf{x} + t\mathbf{y} \in K$ for any $t \in [0,1]$. One can show that K is convex if and only if for any $\mathbf{x}_1, ..., \mathbf{x}_n \in K$, their convex combination $\sum_{i=1}^n t_i \mathbf{x}_i \in K$, where $t_i \in [0,1]$ and $\sum_i t_i = 1$.

- **Convex hull** Let $T \subset \mathbb{R}^N$. The convex hull conv(T) is defined to be the smallest convex set containing T. Indeed,

$$\text{conv}(T) = \left\{ \sum_{i=1}^n t_i \mathbf{x}_i | \mathbf{x}_i \in T,\ t_i \in [0,1],\ \sum_i t_i = 1 \right\}.$$

  The convex hull of an open (closed) set is open (closed).

- **Extreme points** of a convex set: a point $p \in K$ is called an extreme point of K if it does lie in the interior of a segment of two points of K. Every compact convex set is the convex hull of its extreme points.

- **Convex cone**: A set $K \in \mathbb{R}^n$ is a cone if $\mathbf{x} \in K$ implies $t\mathbf{x} \in K$ for all $t \geq 0$. If K is a cone and a convex set, we call it convex cone.

- **Dual cone**: for a cone $K \subset \mathbb{R}^N$, its dual cone is defined as

$$K^* = \{\mathbf{y} \in \mathbb{R}^N | \langle \mathbf{x}, \mathbf{y} \rangle \geq 0 \text{ for all } \mathbf{x} \in K\}.$$

- Examples:

  1. Second-order cone:

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{R}^{N+1} | \sqrt{\sum_{j=1}^N x_j^2} \leq x_{N+1} \right\}$$

- Hahn-Banach Theorem: Convex sets can be separated by hyperplanes. Given two convex sets $K_1, K_2 \subset \mathbb{R}^N$ whose interiors have empty intersection. Then there exists $\mathbf{w} \in \mathbb{R}^N$ and $\lambda \in \mathbb{R}$ such that

$$K_1 \subset \{\mathbf{x} | \langle \mathbf{x}, \mathbf{w} \rangle \leq \lambda\}$$
$$K_2 \subset \{\mathbf{x} | \langle \mathbf{x}, \mathbf{w} \rangle \geq \lambda\}$$

- Let $K \subset \mathbb{R}^N$ be a convex set. A point $\mathbf{x} \in K$ is called an extreme point of K if $\mathbf{x} = t\mathbf{y} + (1-t)\mathbf{z}$ for $\mathbf{y}, \mathbf{z} \in K$, then $\mathbf{y} = \mathbf{z} = \mathbf{x}$.

- Any compact convex set is the convex hull of its extreme points.

1.3 Convex functions

Goal: We want to extend theory of smooth convex analysis to non-differentiable convex functions.
Let X be a separable Hilbert space, $f : X \to (-\infty, +\infty]$ be a function.

- **Proper**: f is called proper if $f(x) < \infty$ for at least one x. The domain of f is defined to be: $dom f = \{x | f(x) < \infty\}$.
- **Lower Semi-continuity**: f is called lower semi-continuous (l.s.c.) if $\liminf_{x_n \to \bar{x}} f(x_n) \geq f(\bar{x})$. This definition is to guarantee that if $x_n \to \bar{x}$ and $f(x_n) \to \inf f(x)$, then $\bar{x}$ is a minimum.
  - The set epi$f := \{(x, \eta) | f(x) \leq \eta\}$ is called the epigraph of f.
  - Proposition: f is l.s.c. if and only if epi$f$ is closed. Sometimes, we call such f closed. (https://proofwiki.org/wiki/Characterization_of_Lower_Semicontinuity)
  - The indicator function $\iota_{\mathcal{C}}$ of a set $\mathcal{C}$ is closed if and only if $\mathcal{C}$ is closed.
- **Convex function**
  - f is called convex if $dom\ f$ is convex and Jensen's inequality holds: $f((1-\theta)x + \theta y) \leq (1-\theta)f(x) + \theta f(y)$ for all $0 \leq \theta \leq 1$ and any $x, y \in X$.
  - Proposition: f is convex if and only if epi$f$ is convex.
  - First-order condition: for $f \in C^1$, epi$f$ being convex is equivalent to $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ for all $x, y \in X$.
    Proof. If epi$f$ is convex, then by Hahn-Banach theorem, epi$f$ lies on one side of the tangent plane $\{(y, z) | z - f(x) - \langle \nabla f(x), y - x \rangle = 0\}$. This leads to $f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq 0$.
  - Second-order condition: for $f \in C^2$, Jensen's inequality is equivalent to $\nabla^2 f(x) \succeq 0$.
  - If f_α is a family of convex functions, then $\sup_\alpha f_\alpha$ is again a convex function.
- **Strictly convex**:
  - f is called strictly convex if the strict Jensen inequality holds: for $x \neq y$ and $t \in (0, 1)$,
    $$f((1-t)x + ty) < (1-t)f(x) + tf(y).$$
  - First-order condition: for $f \in C^1$, the strict Jensen inequality is equivalent to $f(y) > f(x) + \langle \nabla f(x), y - x \rangle$ for all $x, y \in X$.

- Second-order condition: for $f \in C^2$, ($\nabla^2 f(x) \succ 0) \Longrightarrow$ strict Jensen's inequality is equivalent to .

- Examples

  - $f(x) = |x|_p^p$, with $p \geq 1$. When $p > 1$, f is differentiable. However, $|x|_1$ is not differentiable at $x = 0$.
  - $f(x_1, x_2) = x_1^2$. The function is degenerate (minimum) at $\{(0, x_2) | x_2 \in \mathbb{R}\}$
  - Consider the underdetermined system:

$$Ax = b$$

where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. We assume $m < n$. The least square fit is to find $x^\dagger$ which

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

The functional $f(x)$ is a convex function. In particular, consider

$$f(x_1, x_2) = \frac{1}{2}(a_1 x_1 + a_2 x_2 - b)^2.$$

The minimizer is not unique.

  - Let $\Omega \subset \mathbb{R}^n$. $H_0^1(\Omega)$ be the Sobolev space, the completion of $C_0^1(\Omega)$ under the norm

$$\|u\|_1^2 := \int |u(x)|^2 + |\nabla u(x)|^2 \, dx.$$

The Dirichlet integral

$$D[u] := \int_\Omega |\nabla u(x)|^2 - u(x)\rho(x) \, dx$$

is convex in $u \in H_0^1(\Omega)$.

  - The Schmidt integral

$$\Phi[u] := \int k(x - y) u(x) u(y) \, dx \, dy$$

represents self-interaction of u with kernel $k(x)$.

  - Blurred image. Consider an observed image $z(x)$, $x \in \Omega \subset \mathbb{R}^2$. Suppose the observed image is blurred. An image deblurred problem is to recover a "true image" $u(x)$ operator Consider $u(x)$ from the blurred image z. An image model is

$$z = Ku + n$$

where

$$Ku(x) := \int k(x - y)u(y)\, dy.$$

is called a blur operator. Typical blur kernel is the Gaussian kernel

$$k(x) = \frac{1}{D} e^{-|x|^2/D}.$$

the function n is the Gaussian noise. $\|n\|_2^2 \le \epsilon$.
The image deblur problem is to minimize

$$f(u) = \alpha \|\nabla u\|_1 + \|Ku - z\|^2.$$

- Radon transform is an integral operator K.
- In support vector machine, given training set $(x_i, y_i) \in \mathbb{R}^{n+1}$, $i = 1, ..., N$, where $y_i = \pm 1$, we want to train a classifier which is a function $f(x)$ such that $f(x_i) \ge 1$ if $y_i = 1$ and $f(x_i) \le -1$ if $y_i = -1$. It is used to classify a new incident x. The function f has the form

  $$y = w^T x + b$$

  The parameters $w = (w_1, ..., w_n)^T$ and $b \in \mathbb{R}$ are the training parameters to be found. The training problem is to solve

  $$\min_w \|w\|, \quad \text{subject to } y_i(w^T x_i - b) \ge 1 \text{ for } i = 1, ..., N.$$

  The loss function is

  $$\ell(w) := \sum_{i=1}^{l} \max\left(1 - y_i(w^T \phi(x_i) + b), 0\right).$$

  This is a convex function.
- Let $\theta^* \in \mathbb{R}^p$ be a parameter to be estimated. The estimation is done by n independent measurements Y_i with outcomes y_i, $i = 1, ..., n$. It is modelled by the Poisson distribution:

  $$\mathbb{P}(Y_i = y_i | \theta^*) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \quad \lambda_i = \exp(-\langle a_i, \theta^* \rangle).$$

This means that $Y_1, ..., Y_n$ are independent random variables depending on $a_1, ..., a_n$ and parameter θ^*. Let $A = [a_1, ..., a_n]$ be a chosen measurement matrix. It can be deterministic or stochastic. Let us denote $(y_1, ..., y_n)^T = y$. Thus,

$$\mathbb{P}(Y = y|\theta) = \prod_i \mathbb{P}(Y_i = y_i|\theta) = C \exp\left(-f_n(\theta)\right),$$

where

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}[y_i\langle a_i, \theta\rangle + \exp(-\langle a_i, \theta\rangle)],$$

which is the loss function. It is a convex function.

Proposition 1.1. *A convex function $f : \mathbb{R}^N \to \mathbb{R}$ is continuous.*

See google proof.

Proposition 1.2. *Let $f : \mathbb{R}^N \to (-\infty, \infty]$ be convex. Then*

1. *a local minimizer of f is also a global minimizer;*
2. *the set of minimizers is convex;*
3. *if f is strictly convex, then the minimizer is unique.*

1.4 Gradients of convex functions

Definition 1.1. *Let X be a separable Hilbert space. An operator $F : X \to X$ is called monotone if*

$$\langle F(x) - F(y), x - y\rangle \geq 0, \quad \forall x, y \in X.$$

Proposition 1.3 (Monotonicity of $\nabla f(x)$)**.** *Suppose $f \in C^1$. Then f is convex if and only if $\mathrm{dom} f$ is convex and $\nabla f(x)$ is a monotone operator:*

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq 0.$$

Remark This implies that the directional derivative of f is nonnegative.

Proof. 1. ($\Rightarrow$) From convexity

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle, \quad f(x) \geq f(y) + \langle \nabla f(y), x - y\rangle.$$

Add these two, we get monotonicity of $\nabla f(x)$.

2. ($\Leftarrow$) Let $g(t) = f(x + t(y - x))$. Then $g'(t) = \langle \nabla f(x + t(y - x)), y - x\rangle \geq g'(0)$ by monotonicity (i.e. $\langle \nabla f(x + t(y - x)) - \nabla f(x), t(y - x)\rangle \geq 0$). Hence

$$f(y) = g(1) = g(0) + \int_0^1 g'(t)\, dt \geq g(0) + \int_0^1 g'(0)\, dt = f(x) + \langle \nabla f(x), y - x\rangle$$

$\square$

Remark The p-Laplacian with $p \geq 1$ is the gradient of the convex function

$$D_p[u] := \int_\Omega |\nabla u(x)|^p\, dx$$

It is a monotone operator.

Definition 1.2. *Let X be a Banach space. An operator $F : X \to X$ is called Lipschitz continuous with parameter L if*

$$\|F(x) - F(y)\| \leq L\|x - y\|, \quad \forall x, y \in X.$$

Example

- Consider a blur operator K with $\max |K(x)| < \infty$. Then Ku is Lipschitz.
- Consider the function: $f(x) = \frac{1}{2}\|Ax - b\|^2$, where $A \in \mathbb{R}^{m\times n}$ with $m \leq n$. The gradient of f is $F(x) := \nabla f(x) = A^*(Ax - b)$.

  $$\|F(x) - F(y)\| = \|A^*A(x - y)\| \leq \|A^*A\|\|x - y\|.$$

  One can show that $\|A^*A\| = \sigma_{\max}^2$, where $\sigma_{\max}$ is the maximum of the singular value of A.

Proposition 1.4. *Suppose f is convex and in C^1. The following statements are equivalent.*

(a) Lipschitz continuity of $\nabla f(x)$: there exists an $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in dom f.$$

(b) $g(x) := \frac{L}{2}\|x\|^2 - f(x)$ is convex.

(c) Quadratic upper bound

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

(d) *Co-coercivity*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Proof. 1. $(a) \Rightarrow (b)$:

$$\begin{aligned} & |\langle \nabla f(x) - \nabla f(y), x - y \rangle| \leq \|\nabla f(x) - \nabla f(y)\| \|x - y\| \leq L\|x - y\|^2 \\ \Leftrightarrow \quad & \langle \nabla g(x) - \nabla g(y), x - y \rangle = \langle L(x - y) - (\nabla f(x) - \nabla f(y)), x - y \rangle \geq 0 \end{aligned}$$

Therefore, $\nabla g(x)$ is monotonic and thus g is convex.

2. (b) $\Leftrightarrow$ (c):

$$\begin{aligned} & g \text{ is convex} \\ \Leftrightarrow \quad & g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle \\ \Leftrightarrow \quad & \frac{L}{2}\|y\|^2 - f(y) \geq \frac{L}{2}\|x\|^2 - f(x) + \langle Lx - \nabla f(x), y - x \rangle \\ \Leftrightarrow \quad & f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2. \end{aligned}$$

3. (b) $\Rightarrow$ (d): From (b), $(L/2)\|z\|^2 - f(z)$ is convex, so is $(L/2)\|z\|^2 - f_x(z)$, where $f_x(z) := f(z) - f(x) - \langle \nabla f(x), z - x \rangle$ with minimum at $z = x$. Thus from the proposition below

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = f_x(y) - f_x(x) \geq \frac{1}{2L}\|\nabla f_x(y)\|^2 = \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2.$$

Similarly, $z = y$ minimizes $f_y(z)$, we get

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2.$$

Adding these two together, we get the co-coercivity.

4. (d) $\Rightarrow$ (a): by Cauchy inequality.

□

Proposition 1.5. *Suppose f is convex and in C^1 with $\nabla f(x)$ being Lipschitz continuous with parameter L. Suppose x^* is a global minimum of f. Then*

$$\frac{1}{2L}\|\nabla f(x)\|^2 \leq f(x) - f(x^*) \leq \frac{L}{2}\|x - x^*\|^2.$$

Proof. 1. Right-hand inequality follows from quadratic upper bound.

2. Left-hand inequality follows by minimizing quadratic upper bound

$$f(x^*) = \inf_y f(y) \leq \inf_y \left(f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \right) = f(x) - \frac{1}{2L}\|\nabla f(x)\|^2.$$

□

1.5 Strong convexity

f is called strongly convex if $domf$ is convex and the strong Jensen inequality holds: there exists a constant $m > 0$ such that for any $x, y \in domf$ and $t \in [0, 1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{m}{2} t(1-t)\|x-y\|^2.$$

This definition is equivalent to the convexity of $g(x) := f(x) - \frac{m}{2}\|x\|^2$. This comes from the calculation

$$(1-t)\|x\|^2 + t\|y\|^2 - \|(1-t)x + ty\|^2 = t(1-t)\|x-y\|^2.$$

When$f \in C^2$, then strong convexity of f is equivalent to

$$\nabla^2 f(x) \succeq mI \quad \text{for any } x \in domf.$$

Proposition 1.6. *Suppose $f \in C^1$. The following statements are equivalent:*

(a) f is strongly convex, i.e. $g(x) = f(x) - \frac{m}{2}\|x\|^2$ is convex,

(b) for any $x, y \in domf$, $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq m\|x-y\|^2$.

(c) (quadratic lower bound):

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

Proposition 1.7. *If f is strongly convex, then f has a unique global minimizer x^* which satisfies*

$$\frac{m}{2}\|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{1}{2m}\|\nabla f(x)\|^2 \quad \textit{for all } x \in domf.$$

Proof. 1. For lelf-hand inequality, we apply quadratic lower bound

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^*\rangle + \frac{m}{2}\|x - x^*\|^2 = \frac{m}{2}\|x - x^*\|^2.$$

2. For right-hand inequality, quadratic lower bound gives

$$f(x^*) = \inf_y f(y) \geq \inf_y \left(f(x) + \langle \nabla f(x), y - x\rangle + \frac{m}{2}\|y - x\|^2 \right) \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|^2.$$

Here, we take infimum in y to get the left-hand inequality.

□

Proposition 1.8. *Suppose f is both strongly convex with parameter m and $\nabla f(x)$ is Lipschitz continuous with parameter L. Then f satisfies stronger co-coercivity condition*

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{mL}{m+L}\|x-y\|^2 + \frac{1}{m+L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Proof.

1. Consider $g(x) = f(x) - \frac{m}{2}\|x\|^2$. From strong convexity of f, we get $g(x)$ is convex.

2. From Lipschitz of f, we get g is also Lipschitz continuous with parameter $L - m$.

3. We apply co-coercivity to $g(x)$:

$$\langle \nabla g(x) - \nabla g(y), x - y\rangle \geq \frac{1}{L-m}\|\nabla g(x) - \nabla g(y)\|^2$$

$$\langle \nabla f(x) - \nabla f(y) - m(x-y), x - y\rangle \geq \frac{1}{L-m}\|\nabla f(x) - \nabla f(y) - m(x-y)\|^2$$

$$\left(1 + \frac{2m}{L-m}\right)\langle \nabla f(x) - \nabla f(y), x-y\rangle \geq \frac{1}{L-m}\|\nabla f(x) - \nabla f(y)\|^2 + \left(\frac{m^2}{L-m} + m\right)\|x-y\|^2.$$

□

1.6 Subdifferential

Definition 1.3. *Let f be convex. The subdifferential of f at a point x is a set defined by*

$$\partial f(x) = \{u \in X | (\forall y \in X)\ f(x) + \langle u, y - x\rangle \leq f(y)\}$$

$\partial f(x)$ is also called subgradients of f at x.

Remark Geometrically, the hyperplane $f(y) = f(x) + \langle u, y - x\rangle$ is a supported hyperplane of epi f at x.

Proposition 1. *(a) If f is convex and differentiable at* x, *then $\partial f(x) = \{\nabla f(x)\}$.*

(b) If f is convex, then $\partial f(x)$ is a closed convex set.

Examples

1. Let $f(x) = |x|$. Then $\partial f(0) = [-1, 1]$.

2. Let $\mathcal{C}$ be a closed convex set on $\mathbb{R}^N$. Then $\partial\mathcal{C}$ is locally rectifiable. Moreover,

$$\partial\iota_{\mathcal{C}}(x) = \{\lambda n \mid \lambda \geq 0,\ n \text{ is the unit outer normal of } \partial\mathcal{C} \text{ at } x\}.$$

Proposition 1.9. *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be convex and closed. Then x^* is a minimum of f if and only if $0 \in \partial f(x^*)$.*

Proposition 1.10. *The subdifferential of a convex function f is a set-valued monotone operator. That is, if $u \in \partial f(x)$, $v \in \partial f(y)$, then $\langle u - v, x - y\rangle \geq 0$.*

Proof. From

$$f(y) \geq f(x) + \langle u, y - x\rangle, \quad f(x) \geq f(y) + \langle v, x - y\rangle,$$

Combining these two inequalities, we get monotonicity. □

Proposition 1.11. *The following statements are equivalent.*

(1) *f is strongly convex (i.e. $f - \frac{m}{2}\|x\|^2$ is convex);*

(2) *(quadratic lower bound)*

$$f(y) \geq f(x) + \langle u, y - x\rangle + \frac{m}{2}\|x - y\|^2 \quad \textit{for any } x, y$$

where $u \in \partial f(x)$;

(3) *(Strong monotonicity of ∂f):*

$$\langle u - v, x - y\rangle \geq m\|x - y\|^2, \quad \textit{for any } x, y \textit{ with any } u \in \partial f(x), v \in \partial f(y).$$

1.7 Proximal operator

Definition 1.4. *Given a convex function f, the proximal mapping of f is defined as*

$$\textit{prox}_f(x) := \textit{arg min}_u \left(f(u) + \frac{1}{2}\|u - x\|^2 \right).$$

Since $f(u) + 1/2\|u - x\|^2$ is strongly convex in u, we get unique minimum. Thus, $\text{prox}_f(x)$ is well-defined.

Examples

- Let $\mathcal{C}$ be a convex set. Define indicator function $\iota_{\mathcal{C}}(x)$ as

$$\iota_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{otherwise} . \end{cases}$$

  Then $\text{prox}_{\iota_{\mathcal{C}}}(x)$ is the projection of x onto $\mathcal{C}$.

$$P_{\mathcal{C}} x \in \mathcal{C} \text{ and } (\forall z \in \mathcal{C}), \langle z - P_{\mathcal{C}}(x), x - P_{\mathcal{C}}(x)\rangle \le 0.$$

- $f(x) = \|x\|_1$: prox_f is the soft-thresholding:

$$\text{prox}_f(x)_i = \begin{cases} x_i - 1 & \text{if } x_i \ge 1 \\ 0 & \text{if } |x_i| \le 1 \\ x_i + 1 & \text{if } x_i \le -1 \end{cases}$$

Properties Let f be convex function.

- Proximal operator prox_f is a resolvent operator:

$$\text{prox}_f(x) = z = (I + \partial f)^{-1}(x).$$

  Let

$$z = \text{prox}_f(x) = \text{arg min}_u \left(f(u) + \frac{1}{2}\|u - x\|^2 \right)$$

  if and only if

$$0 \in \partial f(z) + z - x$$

  or

$$x \in z + \partial f(z).$$

  Sometimes, we express this as

$$\text{prox}_f(x) = z = (I + \partial f)^{-1}(x).$$

- Co-coercivity:

$$\langle \text{prox}_f(x) - \text{prox}_f(y), x - y\rangle \ge \|\text{prox}_f(x) - \text{prox}_f(y)\|^2.$$

  Let $x^+ = \text{prox}_f(x) := \text{arg min}_z f(z) + \frac{1}{2}\|z - x\|^2$. We have $x - x^+ \in \partial f(x^+)$. Similarly, $y^+ := \text{prox}_f(y)$ satisfies $y - y^+ \in \partial f(y^+)$. From monotonicity of ∂f, we get

$$\langle u - v, x^+ - y^+\rangle \ge 0$$

  for any $u \in \partial f(x^+)$, $v \in \partial f(y^+)$. Taking $u = x - x^+$ and $v = y - y^+$, we obtain co-coercivity.

- Non-expansive: The co-coercivity of prox_f implies that prox_f is 1-Lipschitz continuous, which is also called non-expansive.

$$\|\text{prox}_f(x) - \text{prox}_f(y)\|^2 \le |\langle x - y, \text{prox}_f(x) - \text{prox}_f(y)\rangle|$$

implies

$$\|\text{prox}_f(x) - \text{prox}_f(y)\| \le \|x - y\|.$$

1.8 Conjugate of a convex function

- For a function $f : \mathbb{R}^N \to (-\infty, \infty]$, we define its conjugate f^* by

$$f^*(y) = \sup_x \left(\langle x, y\rangle - f(x)\right).$$

Examples

1. $f(x) = \langle a, x\rangle - b, \quad f^*(y) = \sup_x(\langle y, x\rangle - \langle a, x\rangle + b) = \begin{cases} b & \text{if } y = a \\ \infty & \text{otherwise.} \end{cases}$

2. $f(x) = \begin{cases} ax & \text{if } x < 0 \\ bx & \text{if } x > 0. \end{cases}$, $a < 0 < b$.

$$f^*(y) = \begin{cases} 0 & \text{if } a < y < b \\ \infty & \text{otherwise.} \end{cases}$$

3. $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c$, where A is symmteric and non-singular, then

$$f^*(y) = \frac{1}{2}\langle y - b, A^{-1}(y - b)\rangle - c.$$

In general, if $A \succeq 0$, then

$$f^*(y) = \frac{1}{2}\langle y - b, A^\dagger(y - b)\rangle - c, \quad A^\dagger := (A^*A)^{-1}A^*$$

and dom $f^* =$ range $A + b$.

4. $f(x) = \frac{1}{p}\|x\|^p$, $p \ge 1$, then $f^*(u) = \frac{1}{p^*}\|u\|^{p^*}$, where $1/p + 1/p^* = 1$.

5. $f(x) = e^x$,

$$f^*(y) = \sup_x(xy - e^x) = \begin{cases} y \ln y - y & \text{if } y > 0 \\ 0 & \text{if } y = 0 \\ \infty & \text{if } y < 0 \end{cases}$$

6. $C = \{x | \langle Ax, x\rangle \le 1\}$, where A is s symmetric positive definite matrix. $\iota_C^* = \sqrt{\langle A^{-1}u, u\rangle}$.

Properties

- f^* is convex and l.s.c.
  Note that f^* is the supremum of linear functions. We have seen that supremum of a family of closed functions is closed; and supremum of a family of convex functions is also convex.

- Fenchel's inequality:
  $$f(x) + f^*(y) \geq \langle x, y \rangle.$$
  This follows directly from the definition of f^*:
  $$f^*(y) = \sup_x \left(\langle x, y \rangle - f(x) \right) \geq \langle x, y \rangle - f(x).$$
  This can be viewed as an extension of the Cauchy inequality
  $$\frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 \geq \langle x, y \rangle.$$

Proposition 1.12. *(1) $f^{**}(x)$ is closed and convex.*

(2) $f^{**}(x) \leq f(x)$.

(3) $f^{**}(x) = f(x)$ if and only if f is closed and convex.

Proof. 1. From Fenchel's inequality
$$\langle x, y \rangle - f^*(y) \leq f(x).$$
Taking sup in y gives $f^{**}(x) \leq f(x)$.

2. $f^{**}(x) = f(x)$ if and only if epif^{**} = epi$f$. We have seen $f^{**} \leq f$. This leads to eps $f \subset$ eps f^{**}. Suppose f is closed and convex and suppose $(x, f^{**}(x)) \notin$ epi$f$. That is $f^{**}(x) < f(x)$ and there is a strict separating hyperplane: $\{(z, s) : a(z - x) + b(s - f^{**}(x)) = 0\}$ such that
$$\left\langle \begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} z - x \\ s - f^{**}(x) \end{pmatrix} \right\rangle \leq c < 0 \quad \text{for all } (z, s) \in \text{epi} f$$
with $b \leq 0$.

3. If $b < 0$, we may normalize it such that $(a, b) = (y, -1)$. Then we have
$$\langle y, z \rangle - s - \langle y, x \rangle + f^{**}(x) \leq c < 0.$$

Taking supremum over $(z, s) \in \text{epi} f$,

$$\sup_{(z,s)\in\text{epi} f} (\langle y, z\rangle - s) = \sup_z (\langle y, z\rangle - f(z)) = f^*(y).$$

Thus, we get

$$f^*(y) - \langle y, x\rangle + f^{**}(x) \le c < 0.$$

This contradicts to Fenchel's inequality.

4. If $b = 0$, choose $\hat{y} \in \text{dom } f^*$ and add $\epsilon(\hat{y}, -1)$ to (a, b), we can get

$$\left\langle \begin{pmatrix} a + \epsilon\hat{y} \\ -\epsilon \end{pmatrix}, \begin{pmatrix} z - x \\ s - f^{**}(x) \end{pmatrix} \right\rangle \le c_1 < 0$$

Now, we apply the argument for $b < 0$ and get contradiction.

5. If $f^{**} = f$, then f is closed and convex because f^{**} is closed and convex no matter what f is.

□

Remark. When f is closed and convex, $f(x) = \sup_y(-f^*(y) + \langle y, x\rangle)$, the supremum of its linear supporting functions.

Proposition 1.13. *If f is closed and convex, then*

$$y \in \partial f(x) \Leftrightarrow x \in \partial f^*(y) \Leftrightarrow \langle x, y\rangle = f(x) + f^*(y).$$

Proof. 1.

$$\begin{aligned} y \in \partial f(x) &\Leftrightarrow f(z) \ge f(x) + \langle y, z - x\rangle \\ &\Leftrightarrow \langle y, x\rangle - f(x) \ge \langle y, z\rangle - f(z) \text{ for all } z \\ &\Leftrightarrow \langle y, x\rangle - f(x) = \sup_z (\langle y, z\rangle - f(z)) \\ &\Leftrightarrow \langle y, x\rangle - f(x) = f^*(y) \end{aligned}$$

2. For the equivalence of $x \in \partial f^*(x) \Leftrightarrow \langle x, y\rangle = f(x) + f^*(y)$, we use $f^{**}(x) = f(x)$ and apply the previous argument.

□

1.9 Method of Lagrange multiplier for constrained optimization problems

A standard convex optimization problem can be formulated as

$$\begin{aligned} &\inf_x f_0(x) \\ &\text{subject to} \quad f_i(x) \le 0, \quad i = 1, ..., m \\ &\text{and} \qquad\quad h_i(x) = 0 \quad i = 1, ..., p. \end{aligned}$$

We assume the domain

$$D := \bigcap_i \mathrm{dom} f_i \cap \bigcap_i \mathrm{dom} h_i$$

is a closed convex set in $\mathbb{R}^n$. A point $x \in D$ satisfying the constraints is called a *feasible point*. We assume $D \neq \emptyset$ and denote p^* the optimal value.

The method of Lagrange multiplier is to introduce augmented variables λ, μ and a Lagrangian so that the problem is transformed to a unconstrained optimization problem. Let us define the Lagrangian to be

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \mu_i h_i(x).$$

Here, λ and μ are the augmented variables, called the Lagrange multipliers or the dual variables.

Primal problem From this Lagrangian, we notice that

$$\sup_{\lambda \succeq 0} \left(\sum_{i=1}^m \lambda_i f_i(x) \right) = \iota_{\mathcal{C}_f}(x), \quad \mathcal{C}_f = \bigcap_i \{x | f_i(x) \le 0\}$$

and

$$\sup_{\mu} \left(\sum_{i=1}^p \mu_i h_i(x) \right) = \iota_{\mathcal{C}_h}(x), \quad \mathcal{C}_h = \bigcap_i \{x | h_i(x) = 0\}.$$

Hence

$$\sup_{\lambda \succeq 0, \mu} L(x, \lambda, \mu) = f_0(x) + \iota_{\mathcal{C}_f}(x) + \iota_{\mathcal{C}_h}(x)$$

Thus, the original optimization problem can be written as

$$p^* = \inf_{x \in D} \left(f_0(x) + \iota_{\mathcal{C}_f}(x) + \iota_{\mathcal{C}_h}(x) \right) = \inf_{x \in D} \sup_{\lambda \succeq 0, \mu} L(x, \lambda, \mu).$$

This problem is called the *primal problem*.

Dual problem From this Lagrangian, we define the dual function

$$g(\lambda, \mu) := \inf_{x \in D} L(x, \lambda, \mu).$$

This is an infimum of a family of concave closed functions in λ and μ, thus $g(\lambda, \mu)$ is a concave closed function. We assume that this minimization problem is much simpler than the original one. The dual problem is

$$d^* = \sup_{\lambda \succeq 0, \mu} g(\lambda, \mu).$$

This dual problem is the same as

$$\sup_{\lambda, \mu} g(\lambda, \mu) \quad \text{subject to } \lambda \succeq 0.$$

We refer $(\lambda, \mu) \in \text{dom } g$ with $\lambda \succeq 0$ as dual feasible variables. The primal problem and dual problem are connected by the following duality property.

Weak Duality Property

Proposition 2. *For any $\lambda \succeq 0$ and any μ, we have that*

$$g(\lambda, \mu) \le p^*.$$

In other words,

$$d^* \le p^*$$

Proof. Suppose x is feasible point (i.e. $x \in D$ and $f_i(x) \le 0$, $h_i(x) = 0$). Then for any $\lambda_i \ge 0$ and any μ_i, we have

$$\sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \mu_i h_i(x) \le 0.$$

This leads to

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \mu_i h_i(x) \le f_0(x).$$

Hence for any feasible pair $\lambda \succeq 0, \mu$,

$$g(\lambda, \mu) := \inf_{x \in D} L(x, \lambda, \mu) \le f_0(x) \text{ for all feasible } x.$$

Since $p^* = \inf\{f_0(x) | x \text{ feasible}\}$, we get

$$g(\lambda, \mu) \le p^*$$

for all feasible pair (λ, μ). Taking supremum over all feasible pair (λ, μ), we get $d^* \le p^*$. □

The property $d^* \le p^*$ is called weak duality property. It can also be read as

$$\boxed{\sup_{\lambda \succeq 0, \mu} \inf_{x \in D} L(x, \lambda, \mu) \le \inf_{x \in D} \sup_{\lambda \succeq 0, \mu} L(x, \lambda, \mu).}$$

Definition 1.5. *(a) A point x^* is called a primal optimal if it minimizes $\sup_{\lambda \succeq 0, \mu} L(x, \lambda, \mu)$.*

(b) A dual pair (λ^*, μ^*) with $\lambda^* \succeq 0$ is said to be a dual optimal if it maximizes $\inf_{x \in D} L(x, \lambda, \mu)$.

Strong duality

Definition 1.6. *When $d^* = p^*$, we say the strong duality holds.*

Counter-example that strong duality does not hold Consider

$$\min_{x, y > 0} e^{-x} \text{ subject to } x^2/y \le 0.$$

$D = \{(x, y) | y > 0\}$. Both $f_0(x, y) = e^{-x}$ and $f(x, y) = x^2/y$ are convex in D. The Lagrangian $L(x, y, \lambda) = e^{-x} + \lambda x^2/y$. The dual function is

$$g(\lambda) = \inf_{(x,y) \in D} L(x, y, \lambda) = \begin{cases} 0 & \text{if } \lambda \ge 0 \\ -\infty & \text{if } \lambda < 0 \end{cases}$$

We have $p^* = 1$ while $d^* = 0$.
Ref: `https://inst.eecs.berkeley.edu/~ee227a/fa10/login/l_dual_strong.html`

Slater condition A sufficient condition for strong duality is the Slater condition: there exists a feasible x in relative interior of D°, $f_i(x) < 0$, $i = 1, ..., m$ and $h_i(x) = 0$, $i = 1, ..., p$. Such a point x is called a strictly feasible point.

Theorem 1.1. *Suppose $f_0, ..., f_m$ are convex, $h(x) = Ax - b$, and assume the Slater condition holds: there exists $x \in D^\circ$ with $Ax - b = 0$ and $f_i(x) < 0$ for all $i = 1, ..., m$. Then the strong duality*

$$\sup_{\lambda \succeq 0, \mu} \inf_{x \in D} L(x, \lambda, \mu) = \inf_{x \in D} \sup_{\lambda \succeq 0, \mu} L(x, \lambda, \mu).$$

holds.

Proof. See pp. 234-236, Boyd's Convex Optimization.

Complementary slackness Suppose there exist x^*, $\lambda^* \succeq 0$ and μ^* such that x^* is the optimal primal point and (λ^*, μ^*) is the optimal dual point and the strong duality gap $p^* - d^* = 0$. In this case,

$$\begin{aligned} f_0(x^*) &= p^* = d^* = g(\lambda^*, \mu^*) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \mu_i^* h_i(x) \right) \\ &\le f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \mu_i^* h_i(x^*) \\ &\le f_0(x^*). \end{aligned}$$

The last line follows from

$$\sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \mu_i h_i(x) \le 0.$$

for any feasible pair (x, λ, μ). This leads to

$$\sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \mu_i^* h_i(x^*) = 0.$$

Since $h_i(x^*) = 0$ for $i = 1, ..., p$, $\lambda_i \ge 0$ and $f_i(x^*) \le 0$, we then get

$$\boxed{\lambda_i^* f_i(x^*) = 0 \quad \text{for all } i = 1, ..., m.}$$

This is called complementary slackness. It holds for any optimal solutions (x^*, λ^*, μ^*).

KKT condition

Proposition 1.14. *When f_0, f_i and h_i are differentiable, then the optimal points x^* to the primal problem and (λ^*, μ^*) to the dual problem satisfy the Karush-Kuhn-Tucker (KKT) condition:*

$$\begin{cases} f_i(x^*) \le 0, & i = 1, ..., m \\ \lambda_i^* \ge 0, & i = 1, ..., m, \\ \lambda_i^* f_i(x^*) = 0, & i = 1, ..., m \\ h_i(x^*) = 0, & i = 1, ..., p \end{cases}$$

$$\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p \mu_i^* \nabla g_i(x^*) = 0.$$

Remark. If $f_0, f_i, i = 0, ..., m$ are closed and convex, but may not be differentiable, then the last KKT condition is replaced by

$$0 \in \partial f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \partial f_i(x^*) + \sum_{i=1}^{p} \mu_i^* \partial g_i(x^*).$$

We call the triple (x^*, λ^*, μ^*) satisfies the optimality condition.

Theorem 1.2. *If f_0, f_i are closed and convex and h are affine. Then the KKT condition is also a sufficient condition for optimal solutions. That is, if $(\hat{x}, \hat{\lambda}, \hat{\mu})$ satisfies KKT condition, then $\hat{x}$ is primal optimal and $(\hat{\lambda}, \hat{\mu})$ is dual optimal, and there is zero duality gap.*

Proof.

1. From $f_i(\hat{x}) \leq 0$ and $h(\hat{x}) = 0$, we get that $\hat{x}$ is feasible.

2. From $\hat{\lambda}_i \geq 0$ and f_i being convex and h_i are linear, we get

$$L(x, \hat{\lambda}, \hat{\mu}) = f_0(x) + \sum_i \hat{\lambda}_i f_i(x) + \sum_i \hat{\mu}_i h_i(x)$$

   is also convex in x.

3. The last KKT condition states that $\hat{x}$ minimizes $L(x, \hat{\lambda}, \hat{\mu})$. Thus

$$\begin{aligned} g(\hat{\lambda}, \hat{\mu}) &= L(\hat{x}, \hat{\lambda}, \hat{\mu}) \\ &= f_0(\hat{x}) + \sum_{i=1}^{m} \hat{\lambda}_i f_i(\hat{x}) + \sum_{i=1}^{p} \hat{\mu}_i h_i(\hat{x}) \\ &= f_0(\hat{x}) \end{aligned}$$

   This shows that $\hat{x}$ and $(\hat{\lambda}, \hat{\mu})$ have zero duality gap and therefore are primal optimal and dual optimal, respectively.

□

Chapter 2

Minimizing $f(x)$

2.1 Gradient Descent Method

Cauchy, Polyak,

Assumptions

- $f \in C^1(\mathbb{R}^N)$ and convex
- $\nabla f(x)$ is Lipschitz continuous with parameter L
- Optimal value $f^* = \inf_x f(x)$ is finite and attained at x^*.

Gradient descent method

- Forward method:
$$\boxed{x^k = x^{k-1} - t_k \nabla f(x^{k-1})}$$
This is the forward Euler method to solve the ODE: $\dot{x} = -\nabla f(x)$.
  - Fixed step size: if t_k is constant
  - Backtracking line search: Choose $0 < \beta < 1$, initialize $t_k = 1$; take $t_k := \beta t_k$ until
  $$f(x - t_k \nabla f(x)) < f(x) - \frac{1}{2} t_k \|\nabla f(x)\|^2$$
  - Optimal line search:
  $$t_k = \text{arg min}_t f(x - t\nabla f(x)).$$

- Backward method

$$\boxed{x^k = x^{k-1} - t_k \nabla f(x^k).}$$

  This is the backward Euler method to solve the ODE: $\dot{x} = -\nabla f(x)$.

- The forward gradient method can be expressed as

$$x^k = \text{arg min}_x \left(f(x^{k-1}) + \langle \nabla f(x^{k-1}), x - x^{k-1} \rangle + \frac{t^k}{2} \|x - x^{k-1}\|^2 \right)$$

- The backward gradient method can be expressed as

$$x^k = \text{arg min}_x \left(f(x) + \frac{t^k}{2} \|x - x^{k-1}\|^2 \right)$$

Analysis for the fixed step size case

Proposition 2.15. *Suppose $f \in C^1$, convex and ∇f is Lipschitz with constant L. Suppose the optimal value $f^* := \inf_x f(x)$ is finite and attained at x^*. Consider the fixed-step size gradient descent method. If the step size t satisfies $t \le 1/L$, then the fixed-step size gradient descent method satisfies*

$$f(x^k) - f(x^*) \le \frac{1}{2kt} \|x^0 - x^*\|^2$$

Remarks

- If in addition f is strongly convex, then the sequence $\{x^k\}$ converges to the unique optimal solution x^* linearly.

Proof.

1. Let $x^+ := x - t\nabla f(x)$.

2. From quadratic upper bound:

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

   Choosing $y = x^+$ and $t < 1/L$, we get

$$f(x^+) \le f(x) + \left(-t + \frac{Lt^2}{2} \right) \|\nabla f(x)\|^2 \le f(x) - \frac{t}{2} \|\nabla f(x)\|^2.$$

3. From
$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle$$
we get
$$\begin{aligned} f(x^+) &\leq f(x) - \frac{t}{2}\|\nabla f(x)\|^2 \\ &\leq f^* + \langle \nabla f(x), x - x^* \rangle - \frac{t}{2}\|\nabla f(x)\|^2 \\ &= f^* + \frac{1}{2t}\left(\|x - x^*\|^2 - \|x - x^* - t\nabla f(x)\|^2\right) \\ &= f^* + \frac{1}{2t}\left(\|x - x^*\|^2 - \|x^+ - x^*\|^2\right). \end{aligned}$$

4. Define $x^{i-1} = x$, $x^i = x^+$, sum this inequalities from $i = 1, ..., k$, we get
$$\begin{aligned} \sum_{i=1}^{k}\left(f(x^i) - f^*\right) &\leq \frac{1}{2t}\sum_{i=1}^{k}\left(\|x^{i-1} - x^*\|^2 - \|x^i - x^*\|^2\right) \\ &= \frac{1}{2t}\left(\|x^0 - x^*\|^2 - \|x^k - x^*\|^2\right) \\ &\leq \frac{1}{2t}\|x^0 - x^*\|^2 \end{aligned}$$

5. Since $f(x^i) - f^*$ is a decreasing sequence, we then get
$$f(x^k) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}\left(f(x^i) - f^*\right) \leq \frac{1}{2kt}\|x^0 - x^*\|^2.$$

Proposition 2.16. *Suppose $f \in C^1$ and convex. The fixed-step size backward gradient method satisfies*
$$f(x^k) - f(x^*) \leq \frac{1}{2kt}\|x^0 - x^*\|^2.$$
Here, no assumption on Lipschitz continuity of $\nabla f(x)$ is needed.

Proof.

1. Define $x^+ = x - t\nabla f(x^+)$.

2. For any z, we have
$$f(z) \geq f(x^+) + \langle \nabla f(x^+), z - x^+ \rangle = f(x^+) + \langle \nabla f(x^+), z - x \rangle + t\|\nabla f(x^+)\|^2.$$

3. Take $z = x$, we get
$$f(x^+) \le f(x) - t\|\nabla f(x^+)\|^2$$
Thus, $f(x^+) < f(x)$ unless $\nabla f(x^+) = 0$.

4. Take $z = x^*$, we obtain
$$\begin{aligned} f(x^+) &\le f(x^*) + \langle \nabla f(x^+), x - x^* \rangle - t\|\nabla f(x^+)\|^2 \\ &\le f(x^*) + \langle \nabla f(x^+), x - x^* \rangle - \frac{t}{2}\|\nabla f(x^+)\|^2 \\ &= f(x^*) - \frac{1}{2t}\|x - x^* - t\nabla f(x^+)\|^2 + \frac{1}{2t}\|x - x^*\|^2 \\ &= f(x^*) + \frac{1}{2t}\left(\|x - x^*\|^2 - \|x^+ - x^*\|^2\right). \end{aligned}$$

Proposition 2.17. *Suppose f is strongly convex with parameter m and $\nabla f(x)$ is Lipschitz continuous with parameter L. Suppose the minimum of f is attended at x^*. Then the gradient method converges linearly, namely*
$$\|x^k - x^*\|^2 \le c^k\|x^0 - x^*\|^2$$
$$f(x^k) - f(x^*) \le \frac{c^k L}{2}\|x^0 - x^*\|^2,$$
where
$$c = 1 - t\frac{2mL}{m+L} < 1 \text{ if the step size } t \le \frac{2}{m+L}.$$

Proof. 1. For $0 < t \le 2/(m+L)$:
$$\begin{aligned} \|x^+ - x^*\|^2 &= \|x - t\nabla f(x) - x^*\|^2 \\ &= \|x - x^*\|^2 - 2t\langle \nabla f(x), x - x^* \rangle + t^2\|\nabla f(x)\|^2 \\ &\le \|x - x^*\|^2 - 2t\left(\frac{mL}{m+L}\|x - x^*\|^2 + \frac{1}{m+L}\|\nabla f(x)\|^2\right) + t^2\|\nabla f(x)\|^2 \\ &= \left(1 - t\frac{2mL}{m+L}\right)\|x - x^*\|^2 + t\left(t - \frac{2}{m+L}\right)\|\nabla f(x)\|^2 \\ &\le \left(1 - t\frac{2mL}{m+L}\right)\|x - x^*\|^2 = c\|x - x^*\|^2. \end{aligned}$$
t is chosen so that $c < 1$. Thus, the sequence $x^k - x^*$ converges linearly with rate c.

2. From quadratic upper bound
$$f(x^k) - f(x^*) \le \frac{L}{2}\|x^k - x^*\|^2 \le \frac{c^k L}{2}\|x^0 - x^*\|^2.$$
we get $f(x^k) - f(x^*)$ also converges to 0 with linear rate.

□

Example: least-squares method Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear map and $b \in \mathbb{R}^m$. We look for

$$\min_x \|Ax - b\|^2.$$

Suppose A^*A has eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2 > 0$ with normalized eigenvectors v_i, $i = 1, ..., r$. Suppose the kernel $N(A)$ is spanned by the orthonormal set $\{v_i | i = r+1, ..., n\}$. Then $\{v_1, ..., v_n\}$ form an orthonormal basis in $\mathbb{R}^n$. Let $u_i \in \mathbb{R}^m$ defined by $Av_i = \sigma_i u_i$, $i = 1, ..., r$. Then $\{u_1, ..., u_r\}$ is an orthonormal set in $R(A)$. We expand them to $u_{r+1}, ..., u_m$ to form an orthonormal basis in $\mathbb{R}^m$. We have

- $Av_i = \sigma_i u_i, \quad i = 1, ...r$
- $A^* u_i = \sigma_i v_i, \quad i = 1, ...r$
- $N(A) =< v_{r+1}, ..., v_n >, \quad R(A) =< u_1, ..., u_r >$
- $N(A^*) =< u_{r+1}, ..., u_m >, \quad R(A^*) =< v_1, ..., v_r >.$

The least-squares solution $x^\dagger$ satisfies the normal equation

$$A^*Ax = A^*b$$

If $b = \sum_{i=1}^m b_i u_i$, then

$$x^\dagger = \sum_{i=1}^r \frac{b_i}{\sigma_i} v_i.$$

and

$$\|Ax^\dagger - b\|^2 = \sum_{i=r+1}^m |b_i|^2.$$

The gradient of the map $f(x) = \frac{1}{2}\|Ax - b\|^2$ is

$$\nabla f(x) = A^*(Ax - b).$$

The gradient descent method gives

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

In terms of singular vectors, we have

$$x_i^k = x_i^{k-1} - t(\sigma_i^2 x_i^{k-1} - \sigma_i b_i), \quad i = 1, ..., r.$$

$$x_i^k = x_i^{k-1}, \quad i = r+1, ..., n,$$

where

$$x^k = \sum_{i=1}^{n} x_i^k v_i.$$

These give

$$x_i^k = x_i^0 \quad i = r+1, ..., n.$$

$$x_i^k \to \frac{b_i}{\sigma_i} \text{ as } k \to \infty, \quad i = 1, ...r.$$

Thus, $x^k \to x^*$, where

$$x^* = \sum_{i=1}^{n} x_i^* v_i = \sum_{i=1}^{r} \frac{b_i}{\sigma_i} v_i + \sum_{r+1}^{n} x_i^0 v_i.$$

We have

$$x_i^k - x_i^* = (1 - t\sigma_i^2)(x_i^{k-1} - x_i^*)v_i, \quad i = 1, ...r,$$

which gives the convergence

$$\|x^k - x^*\|^2 = \sum_{i=1}^{r} (1 - t\sigma_i^2)^{2k} |x_i^0 - x_i^*|^2,$$

provided

$$0 < t < \frac{2}{\sigma_1^2} = \frac{2}{L}.$$

Here, L is the Lipschitz parameter corresponding to $\nabla f(x) = A^*(Ax - b)$, which is exactly σ_1^2.

$$f(x^k) - f(x^*) = \frac{1}{2}\|Ax^k - Ax^*\|^2 = \sum_{i=1}^{r} \sigma_i^{2k} (1 - t\sigma_i^2)^{2k} |x_i^0 - x_i^*|^2.$$

2.2 Subgradient Descent Method

Assumptions

- f is closed and convex
- Optimal value $f^* = \inf_x f(x)$ is finite and attained at x^*.

Subgradient method

$$x^k = x^{k-1} - t_k v_{k-1}, \quad v_{k-1} \in \partial f(x^{k-1}).$$

t_k is chosen so that $f(x^k) < f(x^{k-1})$.

- This is a forward (sub)gradient method.
- It may not converge.
- If it converges, the optimal rate is

  $$f(x^k) - f(x^*) \le O(1/\sqrt{k}),$$

  which is very slow.

2.3 Proximal point method

Assumptions

- f is closed and convex
- Optimal value $f^* = \inf_x f(x)$ is finite and attained at x^*.

Proximal point method:

$$\boxed{x^k = \text{prox}_{tf}(x^{k-1}) = x^{k-1} - tG_t(x^{k-1})}$$

where

$$\text{prox}_{tf}(x) := \text{arg min}_z \left(tf(z) + \frac{1}{2}\|z - x\|^2 \right)$$

Let $x^+ := \text{prox}_{tf}(x) := x - tG_t(x)$. From the Euler-Lagrange equation, we get

$$G_t(x) \in \partial f(x^+).$$

Thus, we may view proximal point method is a backward subgradient method.

Proposition 2.18. *Suppose f is closed and convex and suppose an ptimal solution x^* of $\min f$ is attainable. Then the proximal point method $x^k = prox_{tf}(x^{k-1})$ with $t > 0$ satisfies*

$$f(x^k) - f(x^*) \le \frac{1}{2kt}\|x^0 - x^*\|.$$

Convergence proof:

1. Given x, let $x^+ := \text{prox}_{tf}(x)$. Let $G_t(x) := (x^+ - x)/t$. Then $G_t(x) \in \partial f(x^+)$. We then have, for any z,
$$f(z) \geq f(x^+) + \langle G_t(x), z - x^+ \rangle = f(x^+) + \langle G_t(x), z - x \rangle + t\|G_t(x)\|^2.$$

2. Take $z = x$, we get
$$f(x^+) \leq f(x) - t\|\nabla f(x^+)\|^2$$
Thus, $f(x^+) < f(x)$ unless $\nabla f(x^+) = 0$.

3. Take $z = x^*$, we obtain
$$\begin{aligned} f(x^+) &\leq f(x^*) + \langle G_t(x), x - x^* \rangle - t\|G_t(x)\|^2 \\ &\leq f(x^*) + \langle G_t(x), x - x^* \rangle - \frac{t}{2}\|G_t(x)\|^2 \\ &= f(x^*) + \frac{1}{2t}\|x - x^* - tG_t(x)\|^2 - \frac{1}{2t}\|x - x^*\|^2 \\ &= f(x^*) + \frac{1}{2t}\left(\|x^+ - x^*\|^2 - \|x - x^*\|^2\right). \end{aligned}$$

4. Taking $x = x^{i-1}$, $x^+ = x^i$, sum over $i = 1, ..., k$, we get
$$\sum_{i=1}^{k}(f(x^k) - f(x^*)) \leq \frac{1}{2t}\left(\|x^0 - x^*\| - \|x^k - x^*\|\right).$$
Since $f(x^k)$ is non-increasing, we get
$$k(f(x^k) - f(x^*)) \leq \sum_{i=1}^{k}(f(x^k) - f(x^*)) \leq \frac{1}{2t}\|x^0 - x^*\|.$$

2.4 Accelerated Proximal Point Method

The proximal point method is a first order method. With a small modification, it can be accelerated to a second order method. This is the work of Nesterov (1984). It was shown to be the best algorithm (Nesterov). The idea is to use an extrapolation from x^{k-1} to x^k. The acceleration algorithm reads

$$y^k = (\theta_k - 1)x^{k-1} + (2 - \theta_k)x^k, \quad x^{k+1} = \text{prox}_{tf}(y^k),$$

$$x_1 = x_0.$$

Here, the parameters θ and t will be chosen properly so that the slow convergence term will be cancelled. In fact, there is no constraint on t. The parameter θ_k is chosen as

$$\theta_k = \frac{2}{k+1}.$$

Then we have the following theorem

Theorem 2.3. *Assume f is closed and convex and the optimal value f^* is attainable. Then the above acceleration algorithm with $\theta_k = 2/(k+1)$ converges as*

$$f(x^k) - f^* \le \frac{\theta_k^2}{2t}\|x^0 - x^*\|^2.$$

Proof. From the extrapolation formulation

$$\begin{aligned} y^k &:= (\theta_k - 1)x^{k-1} + (2 - \theta_k)x^k \\ &= (1-\theta_k)x^k + (x^k + (\theta_k - 1)x^{k-1}) \\ &= (1-\theta_k)x^k + \theta_k v_k \end{aligned}$$

where

$$v^k := x^{k-1} + \frac{1}{\theta_{k-1}}(x^k - x^{k-1}).$$

Let us estimate the amount of decreasing of $f(x) - f^*$ in one step. Let us call x^k by x, x^{k+1} by x^+, v^k by v, v^{k+1} by v^+, y^k by y and θ_k by θ. We have

$$y = (1-\theta)x + \theta v,$$

$$x^+ = \text{prox}_{tf}(y),$$

$$v^+ = x + \frac{1}{\theta}(x^+ - x).$$

Let $G_t(x) := (x^+ - y)/t$. Then from $x^+ = \text{prox}_{tf}(y)$, we have $G_t(x) \in \partial f(x^+)$. Then for any z, we have

$$f(z) \ge f(x^+) + \langle G_t(x), z - x^+\rangle = f(x^+) + \frac{1}{t}\langle x^+ - y, z - x^+\rangle.$$

Thus,

$$f(x^+) \le f(z) + \frac{1}{t}\langle y - x^+, x^+ - z\rangle$$

We take $z = x^*$ and $z = x$, make a convex combination of these two inequalities with weights θ and $(1-\theta)$, we get

$$f(x^+) \le f^* + \frac{1}{t}\langle x^+ - y, x^* - x^+\rangle$$

$$f(x^+) \le \frac{1}{t}\langle x^+ - y, x - x^+\rangle$$

$$\begin{aligned} f(x^+) - f^* - (1-\theta)(f(x) - f^*) &= \frac{1}{t}\langle x^+ - y, \theta x^* + (1-\theta)x - x^+\rangle \\ &\le \frac{1}{t}\langle x^+ - y, \theta x^* + (1-\theta)x - x^+\rangle + \frac{1}{2t}\|x^+ - y\|^2 \\ &= \frac{1}{2t}\left(\|y - (1-\theta)x - \theta x^*\|^2 - \|x^+ - (1-\theta)x - \theta x^*\|^2\right) \\ &= \frac{\theta^2}{2t}\left(\|v - x^*\|^2 - \|v^+ - x^*\|^2\right). \end{aligned}$$

Now, we take $\theta_k = 2/(k+1)$, it satisfies

$$\theta_1 = 1, \quad \frac{1-\theta_k}{\theta_k^2} \le \frac{1}{\theta_{k-1}^2}, k \ge 2.$$

We have with $t_i = t$,

$$\frac{t_i}{\theta_i^2}\left(f(x^i) - f^*\right) + \frac{1}{2}\|v^i - x^*\|^2 \le \frac{(1-\theta_i)t_i}{\theta_i^2}\left(f(x^{i-1}) - f^*\right) + \frac{1}{2}\|v^{i-1} - x^*\|^2$$

Using $(1-\theta_i)/\theta_i^2 \le 1/\theta_{i-1}^2$, we obtain

$$\frac{t}{\theta_k^2}\left(f(x^k) - f^*\right) + \frac{1}{2}\|v^k - x^*\|^2 \le \frac{(1-\theta_1)t}{\theta_1^2}\left(f(x^0) - f^*\right) + \frac{1}{2}\|v^0 - x^*\|^2 = \frac{1}{2}\|x^0 - x^*\|^2.$$

This shows

$$f(x^k) - f^* \le \frac{\theta_k^2}{2t}\|x^0 - x^*\|^2 \le \frac{2}{t(k+1)^2}\|x^0 - x^*\|^2.$$

□

2.5 Mirror Descent Method

Vector-Covector view

1. The convergence rate of a gradient descent method depends on the inner product. In the gradient descent flow:
$$\dot{x} = -\nabla f(x),$$
the decay of f is
$$\frac{d}{dt}f(x(t)) = \nabla f(x) \cdot \dot{x} = -\|\nabla f(x(t))\|^2.$$
The rate depends on the inner product. We can change another inner product to speed up the convergence as the follows.

2. Let us use the following notation: $df_x(v)$ is the directional derivative of f at x in the direction v. We call v a tangent vector. The term df_x is called the differential of f at x. It is a linear functional on the tangent space at x. Let us call the tangent space V, its dual, the cotangent space V^*. Thus, $df_x \in V^*$. It is a co-vector.

3. We can associate V an inner product $\langle \cdot, \cdot \rangle$ (or a metric). In our case, $V = \mathbb{R}^n$ and the metric can be presented as $g_{ij} = \langle e_i, e_j \rangle$, where e_i is the unit vector in the x_i direction. In $V^* = \mathbb{R}^n$, we use $\{e^i\}$ as its dual basis. That is, $e^i(e_j) = \delta^i_j$.

4. With the inner product structure, the Riesz representation theorem states that for any functional $\alpha \in V^*$, there is a unique $\alpha^\# \in V$ such that

   $$\alpha(v) = \langle \alpha^\#, v \rangle.$$

   The operator $\alpha \mapsto \alpha^\#$ is 1-1,onto and linear. It is called the sharp operator, which maps a covector to a vector. Its inverse $\flat$, which maps V to V^*, is called a flat operator. Suppose $\alpha = \sum \alpha_i e^i$. Let us express $\alpha^\# = \alpha^{\#,i} e_i$. We want to find the expression of $\alpha^{\#,i}$. For any $v = \sum_j v^j e_j$, we have

   $$\alpha(v) = \alpha_i v^j e^i(e_j) = \alpha_i v^i = \langle \alpha^\#, v \rangle = g_{ij} \alpha^{\#,i} v^j.$$

   Let (g^{ij}) be the inverse matrix $(g_{ij})^{-1}$. We get

   $$\alpha^{\#,i} = g^{ij} \alpha_j.$$

5. The gradient $\nabla f(x)$ is defined to be

   $$\nabla f(x) := df_x^\#$$

   Note that

   $$\nabla f(x) = \sum_{i=1}^n g^{ij} \frac{\partial f(x)}{\partial x_j} e^i.$$

6. Using this metric, we have

   $$\frac{d}{dt} f(x) = \sum_i \frac{\partial f(x)}{\partial x^i} \dot{x}^i = -\sum_{ij} g^{ij} \frac{\partial f(x)}{\partial x^i} \frac{\partial f(x)}{\partial x^j}.$$

   Thus, the convergent rate of $f(x)$ depends on the choice of the metric g^{ij}.

7. The metric (g^{ij}) can be designed as a preconditioner to speed up the convergent rate.

8. In the above discussion, we should distinguish vector and covector. The basis in V is $\{e_i\}$ and its dual basis is $\{e^i\}$ in V^*. The correct way to write ∇f is

$$\nabla f = df_x^{\#} = \sum_{i=1}^{n} g^{ij} \frac{\partial f(x)}{\partial x_j} e^i.$$

It is equal to $(f_{x^1}, ..., f_{x^n})$ only because we choose $g^{ij} = \delta^{ij}$.

9. Another example to modify the gradient is to use the inverse of a Hessian. This leads to the Newton's method.

Mirror map and mirror descent algorithm

1. In the above discussion, all we need is a sharp operator. We can design a nonlinear sharp operator, called a mirror map.

2. The mirror map is determined by a strongly convex function $h : V \to \mathbb{R}$ with constant α. The differential $dh : x \mapsto dh_x$ is a map $V \to V^*$, where V is the tangent space, V^* the cotangent space. Since h is strongly convex, dh is 1-1 and onto.

3. Examples:
   - $h(x) = \frac{1}{2}\|x\|^2$. $dh_x = x$.
   - $h(x) = \sum_i (x_i \ln x_i - x_i)$. $dh_x = (\ln x_1, ..., \ln x_n)$.

4. The mirror descent algorithm is
   - $y^k = dh_{x_k}$
   - $y^{k+1} = y^k - t_k df_{x^k}$
   - $x^{k+1} = (dh)^{-1}(y^{k+1})$

Proximal point view The gradient descent

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

can be thought as

$$x^{k+1} = \text{arg min}_x \left(\langle \nabla f(x^k), x \rangle + \frac{1}{2}\|x - x^k\|^2 \right)$$

The last quadratic term is a regularization term. We can replace it by the Bregman divergence (distance): $D_h(x||x^k)$, where

$$D_h(y||x) := h(y) - h(x) - \langle \nabla h(x), y - x \rangle.$$

Then the proximal point method is

$$x^{k+1} = \text{arg min}_x \left(\langle \nabla f(x^k), x\rangle + D_h(x||x^k)\right)$$

Set the gradient to be zero at x^{k+1}, we get

$$t^k \nabla f(x^k) + \nabla h(x^{k+1}) - \nabla h(x^k) = 0.$$

This gives

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

or

$$x^{k+1} = (\nabla h)^{-1}\left(\nabla h(x^k) - t^k \nabla f(x^k)\right).$$

2.6 Fixed point method

The goal of this section is to show that a minimal sequence of a fixed point method converges.

Definition 2.7. *Let $\mathcal{X}$ be a Hilbert space. A mapping $T : \mathcal{X} \to \mathcal{X}$ is called nonexpansive if*

$$\|Tx - Ty\| \le \|x - y\|, \text{ for any } x, y \in \mathcal{X}.$$

It is called firmly nonexpansive if it satisfies one of the following two equivalent conditions:

$$\|Tx - Ty\|^2 \le \langle Tx - Ty, x - y\rangle \text{ for any } x, y \in \mathcal{X},$$
$$\|Tx - Ty\|^2 \le \|x - y\|^2 - \|(I - T)x - (I - T)y\|^2.$$

Remark T is nonexpansive $\Leftrightarrow$ $-T$ is nonexpansive. A firmly nonexpansive operator is also a nonexpansive operator.

Lemma 2.1. *T is nonexpansive if and only if ($F = (I + T)/2$ is firmly nonexpansive) or ($G := (I - T)/2$ is firmly nonexpansive.)*

Proof.

$$\begin{aligned}
&\|Tx - Ty\|^2 \le \|x - y\|^2 \\
\Leftrightarrow &\frac{1}{4}\|x - y\|^2 + \frac{1}{4}\|Tx - Ty\|^2 \le \frac{1}{2}\|x - y\|^2 \\
\Leftrightarrow &\frac{1}{4}\|x - y\|^2 + \frac{1}{4}\|Tx - Ty\|^2 \pm \frac{1}{2}\langle x - y, Tx - Ty\rangle \le \frac{1}{2}\|x - y\|^2 \pm \frac{1}{2}\langle x - y, Tx - Ty\rangle \\
\Leftrightarrow &\|\frac{1}{2}(I \pm T)x - \frac{1}{2}(I \pm T)y\|^2 \le \langle \frac{1}{2}(I \pm T)x - \frac{1}{2}(I \pm T)y, x - y\rangle.
\end{aligned}$$

□

Examples

1. $f : \mathcal{X} \to \mathbb{R}^*$ be a proper closed convex function and ∇f is Lipschitz continuous with Lipschitz constant L. Consider

$$F = I - t\nabla f.$$

Then F is nonexpansive provided $0 < t/L \leq 1$. In this case, the operator $G := (I - F)/2 = t/2\nabla f$ is a gradient operator.

2. Let $f : \mathcal{X} \to \mathbb{R}^*$ be a proper closed convex function. Let

$$F(x) := \text{prox}_f(x), \quad G = I - F.$$

Then both F and G are firmly nonexpansive. Further, $T = 2F - I$ is nonexpansive.

Proof. $x^+ = \text{prox}_f(x) = F(x)$, $y^+ = \text{prox}_f(y) = F(y)$. $G(x) = x - x^+ \in \partial f(x^+)$. From monotonicity of ∂f, we have

$$\langle G(x) - G(y), x^+ - y^+\rangle \geq 0.$$

This gives

$$\langle x^+ - y^+, x - y\rangle \geq \|x^+ - y^+\|^2.$$

That is

$$\langle F(x) - F(y), x - y\rangle \geq \|F(x) - F(y)\|^2.$$

The proof for $G = I - F$ being firmly nonexpansive follows from the Lemma above. □

3. Let $f : \mathcal{X} \to \mathbb{R}^*$ be closed convex and proper. We denote $\partial f = A$. Then A is a maximal monotone operator. Let

$$F_{tA} := I - tA, \quad , J_{tA} = \text{prox}_{tf} = (I + tA)^{-1}.$$

Solving $\min f(x)$ can be obtained by finding the time asymptotic limit of the ODE

$$\dot{x} + Ax = 0.$$

The ODE can be discreted by

- Forward Euler: $x^{k+1} = x^k - tA(x^k)$, that is $x^{k+1} = F_{tA}(x^k)$
- Backward Euler: $x^{k+1} = x^k - tA(x^{k+1})$, that is $x^{k+1} = J_{tA}(x^k)$

- Crank-Nicholson: $x^{k+1} - x^k = \frac{t}{2}\left(Ax^k + Ax^{k+1}\right)$. This is equivalent to

$$x^{k+1} = J_{tA/2} F_{tA/2} x^k.$$

We claim this is the same as the extraoplation (reflection):

$$x^{k+1} = R_{tA} x^k, \quad R_{tA} := 2J_{tA/2} - I.$$

This is because

$$(I + \frac{t}{2}A)(x^{k+1} + x^k) = 2x^k \Leftrightarrow (I + \frac{t}{2}A)x^{k+1} = (I - \frac{t}{2}A)x^k$$

Algorithm Now, we are given a nonexpansive map $T : \mathcal{X} \to \mathcal{X}$. Our goal is to construct an algorithm and to show it generates a weakly convergent sequence to a fixed point of T find fixed point of T. We consider the algorithm:

$$x^k = \left(1 - \frac{t_k}{2}\right) x^{k-1} + \frac{t_k}{2} T x^{k-1} = (1 - t_k)x^{k-1} + t_k F(x^{k-1}) = x^{k-1} - t_k G(x^{k-1}).$$

Here, $F = (I + T)/2$ and $G = (I - T)/2$. G plays the role as a gradient. We may think this is a general gradient descent algorithm.

Theorem 2.4. *Let $\mathcal{X}$ be a Hilbert space, T be a nonexpansive operator on $\mathcal{X}$. Suppose a fixed point x^* of T exists. Consider the algorithm:*

$$x^k := \left(1 - \frac{t_k}{2}\right) x^{k-1} + \frac{t_k}{2} T(x^{k-1}), \quad x^0 \textit{ arbitrary}$$

with

$$t_k \in [t_{min}, t_{max}], \quad 0 < t_{min} \le t_{max} < 2.$$

Then $\{x^k\}$ converges weakly to a fixed point of T.

Proof. 1. Let $F := (I + T)/2$, $G := (I - T)/2$. The algorithm can also be written as

$$x^k = x^{k-1} - t_k G(x^{k-1}).$$

We have seen that both F and G are firmly non-expansive. Further, (x^* is a fixed point of T) $\Leftrightarrow$ (x^* is a fixed point of F) $\Leftrightarrow$ ($G(x^*) = 0$).

2. From firmly nonexpansive property of F and G, we get (with $x = x^{k-1}$, $x^+ = x^k$, $t = t_k$)

$$\begin{aligned}
\|x^+ - x^*\|^2 - \|x - x^*\|^2 &= \|x^+ - x + x - x^*\|^2 - \|x - x^*\|^2 \\
&= 2\langle x^+ - x, x - x^*\rangle + \|x^+ - x\|^2 \\
&= 2\langle -tG(x), x - x^*\rangle + t^2\|G(x)\|^2 \\
&= 2\langle -t(G(x) - G(x^*)), x - x^*\rangle + t^2\|G(x)\|^2 \\
&\le -2t\|G(x) - G(x^*)\|^2 + t^2\|G(x)\|^2 \\
&= -t(2-t)\|G(x)\|^2 \\
&\le -M\|G(x)\|^2 \le 0,
\end{aligned}$$

where $M = t_{min}(2 - t_{max})$. We get that $\|x^k - x^*\|$ is non-increasing; hence $\{x^k\}$ is bounded; and $\|x^k - x^*\| \to C$ as $k \to \infty$.

3. Let us sum this inequality over k:

$$-\|x^0 - x^*\|^2 \le \sum_{\ell=0}^{\infty} \left(\|x^{\ell+1} - x^*\|^2 - \|x^\ell - x^*\|^2\right) \le -M \sum_{\ell=0}^{\infty} \|G(x^\ell)\|^2 \le 0.$$

$$\Rightarrow \quad M \sum_{\ell=0}^{\infty} \|G(x^\ell)\|^2 \le \|x^0 - x^*\|^2$$

This implies

$$\|G(x^k)\| \to 0 \quad \text{as } k \to \infty,$$

4. Since the sequence $\{x^k\}$ is bounded, it is weakly precompact. Suppose $\bar{x}^k$ be a subsequence of $\{x^k\}$ that converges to $\bar{x}$ weakly. We have that $\bar{x}^k \rightharpoonup \bar{x}$ and $\|G(\bar{x}^k)\| \to 0$. We claim that

$$G(\bar{x}) = 0.$$

This is a lemma due to Opial. Such property for G is called "demiclosedness."

Lemma 2.2. *Let F be nonexpansive in a Hilbert space $\mathcal{X}$. Let $G = I - F$. Suppose $x^n \rightharpoonup x$ and $G(x^n) \to 0$. Then $G(x) = 0$.*

From nonexpansion of F, we have

$$\begin{aligned}
\|x^n - x\|^2 \ge \|F(x^n) - F(x)\|^2 &= \| - x^n + F(x^n) + x^n - F(x)\|^2 \\
&= \|G(x^n)\|^2 - 2\langle G(x^n), x^n - F(x)\rangle + \|x^n - F(x)\|^2.
\end{aligned}$$

We take limit inf on both sides to get

$$\liminf \|x^n - x\|^2 \geq \liminf \|x^n - F(x)\|^2.$$

The right-hand side can be expressed as

$$\|x^n - F(x)\|^2 = \|x^n - x + x - F(x)\|^2 = \|x^n - x\|^2 + \|x - F(x)\|^2 + 2\langle x^n - x, x - F(x)\rangle.$$

Take liminf both sides, we get

$$\liminf \|x^n - x\|^2 \geq \liminf \|x^n - F(x)\|^2 \geq \|x - F(x)\|^2 + \liminf \|x^n - x\|^2,$$

This leads to $F(x) = x$, or equivalently $G(x) = 0$.

5. We claim that there is only one weak limiting point of $\{x^k\}$. Suppose $\bar{y}_1$ and $\bar{y}_2$ are two cluster points of $\{x^k\}$. Then by the previous argument, both sequences $\{\|x^k - \bar{y}_1\|\}$ and $\{\|x^k - \bar{y}_2\|\}$ are non-increasing and have limits. Since $\bar{y}_i$ are limiting points, there exist subsequences $\{k_i^1\}$ and $\{k_i^2\}$ such that $x^{k_i^1} \to \bar{y}_1$ and $x^{k_i^2} \to \bar{y}_2$ as $i \to \infty$. We can choose subsequences again so that we have

$$k_{i-1}^2 < k_i^1 < k_i^2 < k_{i+1}^1 \quad \text{for all } i$$

With this and the non-increasing of $\|x^k - \bar{y}_1\|$ and $\|x^k - \bar{y}_2\|$ we get

$$\|x^{k_{i+1}^1} - \bar{y}_1\| \leq \|x^{k_i^2} - \bar{y}_1\| \leq \|x^{k_i^1} - \bar{y}_1\| \to 0 \text{ as } i \to \infty.$$

On the other hand, $x^{k_i^2} \to \bar{y}_2$. Therefore, we get $\bar{y}_1 = \bar{y}_2$. This shows that there is only one limiting point, say x^*, and $x^k \to x^*$.

□

Remark When $t_k = 1$, we get the proximal point method.

Chapter 3

Minimizing $f(x) + g(x)$

Problem Minimize $h(x) := f(x) + g(x)$.

Assumptions:

- $g \in C^1$ convex, $\nabla g(x)$ Lipschitz continuous with parameter L
- f is closed and convex

Monotone inclusion problem Let $Ax = \partial f(x)$ and $Bx = \partial g(x)$. They are monotone operators because both f and g are convex and closed. The minimization problem is to solve

$$0 \in Ax + Bx.$$

Gradient flow formulation We want to find the equilibrium of the gradient flow

$$\dot{x} = -Ax - Bx.$$

We can derive numerical method for the above gradient flow. The basic idea is operator splitting. The operators associating with f are

- forward gradient descent operator: $F_{tA} := I - tA$,
- backward gradient descent operator $J_{tA} := (I + tA)^{-1}$.

Here, t is a small time-step size. In the case when f is an indicator function $f = \iota_C$, then

$$\text{prox}_{tf}(x) = \text{arg min}_{u \in C} \|u - x\|^2 = P_C(x),$$

where P_C is the projection onto C.

To reach the minimum of $f(x) + g(x)$, we apply the above forward or backward operators for f and g alternatively. We have

- Forward-forward method

$$\boxed{x^{n+1} = F_{tA}F_{tB}x^n}$$

- Forward-backward method (or called proximal gradient method)

$$\boxed{x^{n+1} = J_{tA}F_{tB}x^n}$$

- Backward-backward method

$$\boxed{x^{n+1} = J_{tA}J_{tB}x^n}$$

- Peaceman-Rachford algorithm: From J_A, we can define over-relaxation operator

$$R_A = 2J_A - I.$$

In the case when J_{tA} is a projection P_C, the operator R_Ax is a mirror image of x with respect to C. The Peaceman-Rachford algorithm is

$$\boxed{x^{n+1} = R_AR_B(x^n)}$$

- Douglas-Rachford algorithm

$$\boxed{x^{n+1} = \frac{1}{2}(I + R_AR_B)(x^n)}$$

The Douglas-Rachford method can also be written as

$$\begin{aligned} x^{n+1} &= (I - J_A - J_B + 2J_AJ_B)(x^n) \\ &= (J_A(2J_B - I) - J_B + I)(x^n) \end{aligned}$$

This can be written as

$$\begin{aligned} y^{n+1} &= J_Bx^n \\ z^{n+1} &= J_A(2y^{n+1} - x^n) \\ x^{n+1} &= x^n + z^{n+1} - y^{n+1} \end{aligned}$$

We can start from updating z first, then

$$\begin{aligned} z^{n+1} &= J_A(2y^n - x^n) \\ x^{n+1} &= x^n + z^{n+1} - y^n \\ y^{n+1} &= J_Bx^{n+1} \end{aligned}$$

By switching x- and y- updating, the above algorithm can also be written as

$$\begin{aligned} z^{n+1} &= J_A(2y^n - x^n) \\ y^{n+1} &= J_B(x^n + z^{n+1} - y^n) \\ x^{n+1} &= x^n + z^{n+1} - y^n \end{aligned}$$

In general, we have

$$\begin{aligned} T &:= (1-\alpha)I + \alpha R_A R_B, \quad 0 < \alpha \le 1; \\ R_A &:= (1-\alpha_A)I + \alpha_A J_{tA}, \quad 0 < \alpha_A \le 2, \\ R_B &:= (1-\alpha_B)I + \alpha_B J_{tB}, \quad 0 < \alpha_B \le 2. \end{aligned}$$

The Douglas-Rachford method can also be derived from the splitting of the ODE:

$$\dot{x} = -Ax - Bx.$$

In one step, it is approximated by

$$\begin{aligned} \frac{x^{k+1} - y^k}{t} &= -Ax^{k+1} - By^k \\ \frac{y^{k+1} - x^{k+1}}{t} &= -By^{k+1} + By^k \end{aligned}$$

If we call $tBy^k = u^k$. Then we can rewrite Douglas-Rachford method as

$$\begin{aligned} x^{k+1} &= (I + tA)^{-1}(y^k - u^k) \\ y^{k+1} &= (I + tB)^{-1}(x^{k+1} + u^k) \\ u^{k+1} &= u^k + x^{k+1} - y^{k+1}. \end{aligned}$$

By comparing with earlier formula

$$\begin{aligned} z^{n+1} &= J_A(y^n - (x^n - y^n)) \\ y^{n+1} &= J_B(z^{n+1} + (x^n - y^n)) \\ x^{n+1} &= x^n + z^{n+1} - y^n \end{aligned}$$

The last equation is

$$(x^{n+1} - y^{n+1}) = (x^n - y^n) + z^{n+1} - y^{n+1}$$

We see these two formulations are identical with $u \leftrightarrow (x - y)$ and $x \leftrightarrow z$.

This method can be viewed as a gradient flow below. We consider

$$\min f(x) + g(y) \quad \text{subject to } x = y.$$

The consider the Largrage method

$$L(x, y, u) := f(x) + g(y) + \langle u, x - y \rangle$$

The gradient flow is

$$\begin{aligned}
\dot{x} &= -Ax - u \\
\dot{y} &= -By + u \\
\dot{u} &= x - y.
\end{aligned}$$

3.1 Proximal gradient method

This is also known as the Forward-backward method

$$\boxed{x^k = \text{prox}_{tf}(x^{k-1} - t\nabla g(x^{k-1}))}$$

We can express prox_{tf} as $(I + t\partial f)^{-1}$. Therefore the proximal gradient method can be expressed as

$$x^k = (I + t\partial f)^{-1}(I - t\nabla g)x^{k-1}$$

Thus, the proximal gradient method is also called the forward-backward method.

Theorem 3.5. *The forward-backward method converges provided $Lt \leq 1$.*

Proof. 1. Given a point x, define

$$x' = x - t\nabla g(x), \quad x^+ = \text{prox}_{tf}(x').$$

Then

$$-\frac{x' - x}{t} = \nabla g(x), \quad -\frac{x^+ - x'}{t} \in \partial f(x^+).$$

Combining these two, we define a "gradient" $G_t(x) := -\frac{x^+ - x}{t}$. Then $G_t(x) - \nabla g(x) \in \partial f(x^+)$.

2. From the quadratic upper bound of g, we have

$$\begin{aligned}
g(x^+) &\leq g(x) + \langle \nabla g(x), x^+ - x \rangle + \frac{L}{2}\|x^+ - x\|^2 \\
&= g(x) + \langle \nabla g(x), x^+ - x \rangle + \frac{Lt^2}{2}\|G_t(x)\|^2 \\
&\leq g(x) + \langle \nabla g(x), x^+ - x \rangle + \frac{t}{2}\|G_t(x)\|^2,
\end{aligned}$$

The last inequality holds provided $Lt \leq 1$. Combining this with

$$g(x) \leq g(z) + \langle \nabla g(x), x - z \rangle$$

we get

$$g(x^+) \leq g(z) + \langle \nabla g(x), x^+ - z \rangle + \frac{t}{2}\|G_t(x)\|^2.$$

3. From first-order condition at x^+ of f

$$f(z) \geq f(x^+) + \langle p, z - x^+ \rangle \quad \text{for all } p \in \partial f(x^+).$$

Choosing $p = G_t(x) - \nabla g(x)$, we get

$$f(x^+) \leq f(z) + \langle G_t(x) - \nabla g(x), x^+ - z \rangle.$$

4. Adding the above two inequalities, we get

$$h(x^+) \leq h(z) + \langle G_t(x), x^+ - z \rangle + \frac{t}{2}\|G_t(x)\|^2$$

Taking $z = x$, we get

$$h(x^+) \leq h(x) - \frac{t}{2}\|G_t(x)\|^2.$$

Taking $z = x^*$, we get

$$\begin{aligned} h(x^+) - h(x^*) &\leq \langle G_t(x), x^+ - x^* \rangle + \frac{t}{2}\|G_t(x)\|^2 \\ &= \frac{1}{2t}\left(\|x^+ - x^* + tG_t(x)\|^2 - \|x^+ - x^*\|^2\right) \\ &= \frac{1}{2t}\left(\|x - x^*\|^2 - \|x^+ - x^*\|^2\right) \end{aligned}$$

□

3.2 Augmented Lagrangian Method

Problem

$$\min F_P(x) := f(x) + g(Ax)$$

Equivalent to the primal problem with constraint

$$\min f(x) + g(y) \quad \text{subject to} \quad Ax = y$$

Assumptions

- f and g are closed and convex.

Examples:

- $g(y) = \iota_{\{b\}}(y) = \begin{cases} 0 & \text{if } y = b \\ \infty & \text{otherwise} \end{cases}$
  The corresponding $g^*(z) = \langle z, b\rangle$.
- $g(y) = \iota_{\mathcal{C}}(y)$
- $g(y) = \|y - b\|^2$.

The Lagrangian is

$$L(x, y, z) := f(x) + g(y) + \langle z, Ax - y\rangle.$$

The primal function is

$$F_P(x) = \inf_y \sup_z L(x, y, z).$$

The primal problem is

$$\inf_x F_P(x) = \inf_x \inf_y \sup_z L(x, y, z).$$

The dual problem is

$$\begin{aligned}
\sup_z \inf_{x,y} L(x, y, z) &= \sup_z \left[\inf_x \left(f(x) + \langle z, Ax\rangle \right) + \inf_y \left(g(y) - \langle z, y\rangle \right) \right] \\
&= \sup_z \left[-\sup_x (\langle -A^* z, x\rangle - f(x)) - \sup_y (\langle z, y\rangle - g(y)) \right] \\
&= \sup_z \left(-f^*(-A^* z) - g^*(z) \right) = \sup_z \left(F_D(z) \right)
\end{aligned}$$

Thus, the dual function $F_D(z)$ is defined as

$$F_D(z) := \inf_{x,y} L(x, y, z) = -\left(f^*(-A^* z) + g^*(z) \right).$$

and the dual problem is

$$\sup_z F_D(z).$$

We shall solve this dual problem by proximal point method:

$$z^k = \text{prox}_{tF_D}(z^{k-1}) = \text{arg max}_u \left[-f^*(-A^T u) - g^*(u) - \frac{1}{2t}\|u - z^{k-1}\|^2 \right]$$

We have

$$\begin{aligned}
&\sup_u \left(-f^*(-A^T u) - g^*(u) - \frac{1}{2t}\|u - z\|^2 \right) \\
&= \sup_u \left(\inf_{x,y} L(x,y,u) - \frac{1}{2t}\|u - z\|^2 \right) \\
&= \sup_u \inf_{x,y} \left(f(x) + g(y) + \langle u, Ax - y \rangle - \frac{1}{2t}\|u - z\|^2 \right) \\
&= \inf_{x,y} \sup_u \left(f(x) + g(y) + \langle u, Ax - y \rangle - \frac{1}{2t}\|u - z\|^2 \right) \\
&= \inf_{x,y} \left(f(x) + g(y) + \langle z, Ax - y \rangle + \frac{t}{2}\|Ax - y\|^2 \right).
\end{aligned}$$

Here, the maximum $u = z + t(Ax - y)$. Thus, we define the augmented Lagrangian to be

$$L_t(x,y,z) := f(x) + g(y) + \langle z, Ax - y \rangle + \frac{t}{2}\|Ax - y\|^2$$

The augmented Lagrangian method is

$$\begin{aligned}
(x^k, y^k) &= \text{arg min}_{x,y} L_t(x, y, z^{k-1}) \\
z^k &= z^{k-1} + t(Ax^k - y^k)
\end{aligned}$$

Thus, the Augmented Lagrangian method is equivalent to the proximal point method applied to the dual problem:

$$\sup_z \left(-f^*(-A^* z) - g^*(z) \right).$$

3.3 Alternating direction method of multipliers (ADMM)

Problem

$$\min f_1(x_1) + f_2(x_2) \text{ subject to } A_1 x_1 + A_2 x_2 - b = 0.$$

Assumptions

- f_i are closed and convex.

Primal problem and dual problem Define the Lagrangian:

$$L(x_1, x_2, z) = f_1(x_1) + f_2(x_2) + \langle z, A_1 x_1 + A_2 x_2 - b \rangle.$$

The primal problem is

$$\inf_{x_1,x_2} \sup_z L(x_1,x_2,z).$$

The dual problem is

$$\begin{aligned}
\sup_z \inf_{x_1,x_2} L(x_1,x_2,z) &= \sup_z \left[\inf_{x_1}(f_1(x_1) + \langle z, A_1x_1\rangle) + \inf_{x_2}(f_2(x_2) + \langle z, A_2x_2\rangle) - \langle z, b\rangle\right] \\
&= \sup_z \left[(-f_1^*(A_1^*z) - \langle z, b\rangle) - f_2^*(A_2^*z)\right] \\
&= \sup_z \left[-h_1(z) - h_2(z)\right].
\end{aligned}$$

Now we solve this dual problem by proximal point method:

$$z^k = \text{prox}_{tF_D}(z^{k-1}) = \text{arg max}_u \left[-h_1(z) - h_2(z) - \frac{1}{2t}\|u - z^{k-1}\|^2\right]$$

We have

$$\begin{aligned}
&\sup_u \left(-f_1^*(-A_1^*u) - f_2^*(A_2^*u) - \langle u, b\rangle - \frac{1}{2t}\|u - z\|^2\right) \\
&= \sup_u \left(\inf_{x_1,x_2} L(x_1,x_2,u) - \frac{1}{2t}\|u - z\|^2\right) \\
&= \inf_{x_1,x_2} \left(f_1(x_1) + f_2(x_2) + \langle z, A_1x_1 + A_2x_2 - b\rangle + \frac{t}{2}\|A_1x_1 + A_2x_2 - b\|^2\right).
\end{aligned}$$

We thus define

$$L_t(x_1,x_2,z) := f_1(x_1) + f_2(x_2) + \langle z, A_1x_1 + A_2x_2 - b\rangle + \frac{t}{2}\|A_1x_1 + A_2x_2 - b\|^2.$$

ADMM:

$$\begin{aligned}
x_1^k &= \text{arg min}_{x_1} L_t(x_1, x_2^{k-1}, z^{k-1}) \\
&= \text{arg min}_{x_1} \left(f_1(x_1) + \frac{t}{2}\|A_1x_1 + A_2x_2^{k-1} - b + \frac{1}{t}z^{k-1}\|^2\right) \\
x_2^k &= \text{arg min}_{x_2} L_t(x_1^k, x_2, z^{k-1}) \\
&= \text{arg min}_{x_2} \left(f_2(x_2) + \frac{t}{2}\|A_1x_1^k + A_2x_2 - b + \frac{1}{t}z^{k-1}\|^2\right) \\
z^k &= z^{k-1} + t(A_1x_1^k + A_2x_2^k - b)
\end{aligned}$$

ADMM is the Douglas-Rachford method applied to the dual problem:

$$\max_z \left(-\langle b, z\rangle - f_1^*(-A_1^T z)\right) + \left(-f_2^*(-A_2^T z)\right) := -h_1(z) - h_2(z).$$

Douglas-Rachford method

$$\min h_1(z) + h_2(z)$$

$$\begin{aligned} z^k &= \text{prox}_{h_1}(y^{k-1}) \\ y^k &= y^{k-1} + \text{prox}_{h_2}(2z^k - y^{k-1}) - z^k. \end{aligned}$$

If we call $(I + \partial h_1)^{-1} = P_1$ and $(I + \partial h_2)^{-1} = P_2$. These two operators are firmly nonexpansive. They are sort of projections in the case when h_i are indicator functions. We also define the reflection operators $R_i = 2P_i - I$. The Douglas-Rachford method is to find the fixed point of $y^k = Ty^{k-1}$.

$$T = I - P_1 + P_2(2P_1 - I) = \frac{1}{2}(I + R_2 R_1).$$

3.4 Primal dual formulation

Consider

$$\inf_x \left(f(x) + g(Ax)\right)$$

Let

$$F_P(x) := f(x) + g(Ax)$$

Define $y = Ax$ consider $\inf_{x,y} f(x) + g(y)$ subject to $y = Ax$. Now, introduce method of Lagrange multiplier: consider

$$L_P(x, y, z) = f(x) + g(y) + \langle z, Ax - y\rangle$$

Then

$$F_P(x) = \inf_y \sup_z L_P(x, y, z)$$

The problem is

$$\inf_x \inf_y \sup_z L_P(x, y, z)$$

The dual problem is

$$\sup_z \inf_{x,y} L_P(x, y, z)$$

We find that

$$\inf_{x,y} L_P(x, y, z) = -f^*(-A^*z) - g^*(z). := F_D(z)$$

By assuming optimality condition, we have

$$\sup_z \inf_{x,y} L_P(x, y, z) = \sup_z F_D(z).$$

If we take $\inf_y$ first

$$\inf_y L_P(x,y,z) = \inf_y \left(f(x) + g(y) + \langle z, Ax - y\rangle\right) = f(x) + \langle z, Ax\rangle - g^*(z) := L_{PD}(x,z).$$

Then the problem is

$$\inf_x \sup_z L_{PD}(x,z).$$

On the other hand, we can start from $F_D(z) := -f^*(-A^*z) - g^*(z)$. Consider

$$L_D(z,w,x) = -f^*(w) - g^*(z) - \langle x, -A^*z - w\rangle$$

then we have

$$\sup_w \inf_x L_D(z,w,x) = F_D(z).$$

If instead, we exchange the order of inf and sup,

$$\sup_{z,w} L_D(z,w,x) = \sup_{z,w} \left(-f^*(w) - g^*(z) - \langle x, -A^*z - w\rangle\right) = f(x) + g(Ax) = F_P(x).$$

We can also take $\sup_w$ first, then we get

$$\sup_w L_D(z,w,x) = \sup_w \left(-f^*(w) - g^*(z) - \langle x, -A^*z - w\rangle\right) = f(x) - g^*(z) + \langle Ax, z\rangle = L_{PD}(x,z).$$

Let us summarize

$$\begin{aligned}
F_P(x) &= f(x) + g(Ax) \\
F_D(z) &= -f^*(-Az) - g^*(z) \\
L_P(x,y,z) &:= f(x) + g(y) + \langle z, Ax - y\rangle \\
L_D(z,w,x) &:= -f^*(w) - g^*(z) - \langle x, -A^*z - w\rangle \\
L_{PD}(x,z) &:= \inf_y L_P(x,y,z) = \sup_w L_D(z,w,x) = f(x) - g^*(z) + \langle z, Ax\rangle \\
F_P(x) &= \sup_z L_{PD}(x,z) \\
F_D(z) &= \inf_x L_{PD}(x,z)
\end{aligned}$$

By assuming optimality condition, we have

$$\inf_x \sup_z L_{PD}(x,z) = \sup_z \inf_x L_P(x,z).$$