1 Convex Analysis

Main references:

- Vandenberghe (UCLA): EECS236C - Optimization methods for large scale systems,
  `http://www.seas.ucla.edu/~vandenbe/ee236c.html`
- Parikh and Boyd, Proximal algorithms, slides and note.
  `http://stanford.edu/~boyd/papers/prox_algs.html`
- Boyd, ADMM
  `http://stanford.edu/~boyd/admm.html`
- Simon Foucart and Holger Rauhut, Appendix B.

1.1 Motivations: Convex optimization problems

In applications, we encounter many constrained optimization problems. Examples

- 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:

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

- The constrained can be a convex set $\mathcal{C}$. That is

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

  we can define an 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 be reformulated as

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

- In abstract form, we encounter

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

  we can express it as

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

- For more applications, see Boyd's book.

A standard convex optimization problem can be formulated as

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

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

1.2 Convex 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 if lim $\inf_{x_n \to \bar{x}} f(x_n) \geq f(\bar{x})$.
  - The set epi$f := \{(x, \eta) | f(x) \leq \eta\}$ is called the epigraph of f.
  - Prop: 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 $x, y \in X$.
  - 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 function, 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 .

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

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.*

1.3 Gradients of convex functions

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:*

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

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. Hence

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

$\square$

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

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

(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): g is convex $\Leftrightarrow$

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

3. (b) $\Rightarrow$ (d): Define $f_x(z) = f(z) - \langle \nabla f(x), z \rangle$, $f_y(z) = f(z) - \langle \nabla f(y), z \rangle$. From (b), both $(L/2)\|z\|^2 - f_x(z)$ and $(L/2)\|z\|^2 - f_y(z)$ are convex, and $z = x$ minimizes f_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.

$\square$

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

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

$\square$

1.4 Strong convexity

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

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

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 $x, y \in dom f$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq m\|x - y\|^2$.

(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*

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

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

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

We take infimum in y then 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 $g(x)$:

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

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

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

□

1.5 Subdifferential

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 $\mathbf{x}$.

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

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

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

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

1.6 Proximal operator

Definition 1.1. *Given a convex function f, the proximal mapping of f is defined as*

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

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 $\mathcal{C}$.

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

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

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

Properties

- Let f be convex. Then

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

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:

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

Let $x^+ = \text{prox}_f(x) := \text{argmin}_z f(z) + \frac{1}{2}\|z - x\|^2$. 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

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

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

- The co-coercivity of prox_f implies that prox_f is Lipschitz continuous.

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

implies

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

1.7 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 $A \succ 0$, 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. $f(x) = \frac{1}{p}\|x\|^p$, $p \geq 1$, 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. $C = \{x | \langle Ax, x\rangle \leq 1\}$, 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

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

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

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

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.

□

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.

□

1.8 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

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

is a closed convex set in $\mathbb{R}^n$. A point $x \in D$ satisfying the constraints is called a feasible point. We assume $D \ne \emptyset$ 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. 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 $(\lambda, \mu) \in \mathrm{dom}\, g$ 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. $x \in D$, or equivalently, $f_i(x) \leq 0$ and $h_i(x) = 0$). Then for any $\lambda_i \geq 0$ and any μ_i, we have

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

This leads to

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

Hence

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

Hence

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

for all feasible pair (λ, μ) □

This is called weak duality property. Thus, the weak duality can also be read as

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

Definition 1.2. *(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 1.3. *When $d^* = p^*$, we say the strong duality holds.*

A sufficient condition for strong duality is the Slater condition: there exists a feasible x in relative interior of $\mathrm{dom} D$: $f_i(x) < 0$, $i = 1, ..., m$ 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,

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

The last line follows from

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

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 $i = 1, ..., p$, $\lambda_i \geq 0$ and $f_i(x^*) \leq 0$, we then get

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

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:*

$$\begin{aligned} f_i(x^*) \leq 0, &\quad i = 1, ..., m \\ \lambda_i^* \geq 0, &\quad i = 1, ..., m, \\ \lambda_i^* f_i(x^*) = 0, &\quad i = 1, ..., m \\ h_i(x^*) = 0, &\quad i = 1, ..., p \\ \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \mu_i^* \nabla g_i(x^*) = 0. \end{aligned}$$

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.

□

2 Optimization algorithms

2.1 Gradient Methods

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 method

- Forward method

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

  - 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:

$$t_k = \text{argmin}_t f(x - t\nabla f(x)).$$

- Backward method

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

Analysis for the fixed step size case

Proposition 2.15. *Suppose $f \in C^1$, convex and ∇f is Lipschitz with constant L. If the step size t satisfies $t \leq 1/L$, then the fixed-step size gradient descent method satisfies*

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

Remarks

- The sequence $\{x^k\}$ is bounded. Thus, it has convergent subsequence to some $\tilde{x}$ which is an optimal solution.
- 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:
$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$$
choose $y = x^+$ 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
$$f(x^*) \ge f(x) + \langle \nabla f(x), x^* - x\rangle$$
we get
$$\begin{aligned} f(x^+) &\le f(x) - \frac{t}{2}\|\nabla f(x)\|^2 \\ &\le f^* + \langle \nabla f(x), x - x^*\rangle - \frac{t}{2}\|\nabla f(x)\|^2 \\ &= f^* + \frac{1}{2t}\left(\|x - x^*\|^2 - \|x - x^* - t\nabla f(x)\|^2\right) \\ &= f^* + \frac{1}{2t}\left(\|x - x^*\|^2 - \|x^+ - x^*\|^2\right). \end{aligned}$$
4. Define $x^{i-1} = x$, $x^i = x^+$, sum this inequalities from $i = 1, ..., k$, we get
$$\begin{aligned} \sum_{i=1}^k \left(f(x^i) - f^*\right) &\le \frac{1}{2t}\sum_{i=1}^k \left(\|x^{i-1} - x^*\|^2 - \|x^i - x^*\|^2\right) \\ &= \frac{1}{2t}\left(\|x^0 - x^*\|^2 - \|x^k - x^*\|^2\right) \\ &\le \frac{1}{2t}\|x^0 - x^*\|^2 \end{aligned}$$
5. Since $f(x^i) - f^*$ is a decreasing sequence, we then get
$$f(x^k) - f^* \le \frac{1}{k}\sum_{i=1}^k \left(f(x^i) - f^*\right) \le \frac{1}{2kt}\|x^0 - x^*\|^2.$$

Proposition 2.16. *Suppose $f \in C^1$ and convex. The fixed-step size backward gradient method satisfies*
$$f(x^k) - f(x^*) \le \frac{1}{2kt}\|x^0 - x^*\|^2.$$
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

$$f(x^+) \leq f(x) - t\|\nabla f(x^+)\|^2$$

   Thus, $f(x^+) < f(x)$ unless $\nabla f(x^+) = 0$.

4. Take $z = x^*$, we obtain

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

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*

$$\|x^k - x^*\|^2 \leq c^k\|x^0 - x^*\|^2$$

$$f(x^k) - f(x^*) \leq \frac{c^k L}{2}\|x^0 - x^*\|^2,$$

where

$$c = 1 - t\frac{2mL}{m + L} < 1 \text{ if the step size } t \leq \frac{2}{m + L}.$$

Proof.

1. For $0 < t \leq 2/(m + L)$:

$$\begin{aligned} \|x^+ - x^*\|^2 &= \|x - t\nabla f(x) - x^*\|^2 \\ &= \|x - x^*\|^2 - 2t\langle \nabla f(x), x - x^* \rangle + t^2\|\nabla f(x)\|^2 \\ &\leq \left(1 - t\frac{2mL}{m + L}\right)\|x - x^*\|^2 + t\left(t - \frac{2}{m + L}\right)\|\nabla f(x)\|^2 \\ &\leq \left(1 - t\frac{2mL}{m + L}\right)\|x - x^*\|^2 = c\|x - x^*\|^2. \end{aligned}$$

   t is chosen so that $c < 1$. Thus, the sequence $x^k - x^*$ converges linearly with rate c.

2. From quadratic upper bound

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

   we get $f(x^k) - f(x^*)$ also converges to 0 with linear rate.

□

2.2 Subgradient 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:

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

Let $x^+ := \text{prox}_{tf}(x) := x - tG_t(x)$. From

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

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 $x^k = \text{prox}_{tf}(x^{k-1})$ 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,
$$f(z) \geq f(x^+) + \langle G_t(x), z - x^+\rangle = \langle G_t(x), z - x\rangle + t\|G_t(x)\|^2.$$

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 in 1980s.

2.5 Fixed point method

- The proximal point method can be viewed as a fixed point of the proximal map:
$$F(x) := \text{prox}_f(x).$$
- Let
$$G(x) = x - x^+ = (I - F)(x).$$
- Both F and G are firmly non-expansive, i.e.
$$\langle F(x) - F(y), x - y\rangle \geq \|F(x) - F(y)\|^2$$
$$\langle G(x) - G(y), x - y\rangle \geq \|G(x) - G(y)\|^2$$

Proof.
(1). $x^+ = \text{prox}_f(x) = F(x)$, $y^+ = \text{prox}_f(y) = F(y)$. $G(x) = x - x^+ \in \partial f(x^+)$. 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

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

(2). From $G = I - F$, we have

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

Theorem 2.3. *Assume F is firmly non-expansive. Let*

$$y^k = (1 - t_k)y^{k-1} + t_k F(y^{k-1}), \quad y^0 \text{ arbitrary}.$$

Suppose a fixed point y^* of F exists and

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

Then y^k converges to a fixed point of F.

Proof. 1. Let us define $G = (I - F)$. We have seen that G is also firmly non-expansive.

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

2. Suppose y^* is a fixed point of F, or equivalently, $G(y^*) = 0$. From firmly nonexpansive property of F and G, we get (with $y = y^{k-1}$, $y^+ = y^k$, $t = t_k$)

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

where $M = t_{min}(2 - t_{max})$.

3. Let us sum this inequality over k:

$$M \sum_{\ell=0}^{\infty} \|G(y^\ell)\|^2 \leq \|y^0 - y^*\|^2$$

This implies

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

and $\|y^k - y^*\|$ is non-increasing; hence y^k is bounded; and $\|y^k - y^*\| \to C$ as $k \to \infty$.

4. Since the sequence $\{y^k\}$ is bounded, any convergent subsequence, say $\bar{y}^k$, converges to $\bar{y}$ satisfying

$$G(\bar{y}) = \lim_{k\to\infty} G(\bar{y}^k) = 0,$$

by the continuity of G. Thus, any cluster point $\bar{y}$ of $\{y^k\}$ satisfies $G(\bar{y}) = 0$. Hence, by the previous argument with y^* replaced by $\bar{y}$, the sequence $\|y^k - \bar{y}\|$ is also non-increasing and has a limit.

5. We claim that there is only one limiting point of $\{y^k\}$. Suppose $\bar{y}_1$ and $\bar{y}_2$ are two cluster points of $\{y^k\}$. Then both sequences $\{\|y^k - \bar{y}_1\|\}$ and $\{\|y^k - \bar{y}_2\|\}$ 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 $y^{k_i^1} \to \bar{y}_1$ and $y^{k_i^2} \to \bar{y}_2$ 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 $\|y^k - \bar{y}_1\|$ and $\|y^k - \bar{y}_2\|$ we get

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

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

□

2.6 Proximal gradient method

This method is to 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

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 2.4. *The forward-backward method converges provided $Lt \le 1$.*

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

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

□

2.7 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
$$\sup_z P_D(z).$$
We shall solve this dual problem by proximal point method:

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

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

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

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

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

2.8 Alternating direction method of multipliers(ADMM)

Problem

$$\min f_1(x_1) + f_2(x_2) \text{ subject to } A_1x_1 + A_2x_2 - b = 0.$$

Assumptions

- f_i are closed and convex.

ADMM

- 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:

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

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

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

- 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 $(I + \partial h_1)^{-1} = A$ and $(I + \partial h_2)^{-1} = B$. These two operators are firmly nonexpansive. The Douglas-Rachford method is to find the fixed point of $y^k = Ty^{k-1}$.

$$T = I + A + B(2A - I).$$

2.9 Primal dual formulation

Consider

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

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

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

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,

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

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

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

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

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