Definition 2.1 *Convex set: a set $C \subseteq \mathbb{R}^n$ is a convex set if for any $x, y \in C$, we have*

$$tx + (1-t)y \in C, \ \textit{for all } 0 \le t \le 1$$

Definition 2.2 *Convex combination of $x_1, ..., x_k \in \mathbb{R}^n$: any linear combination*

$$\theta_1 x_1 + ... + \theta_k x_k, \ \textit{with } \theta_i \ge 0, \ \textit{and } \sum_{i=1}^k \theta_i = 1$$

Definition 2.3 *Convex hull of set C: all convex combinations of elements in C.*

This is always a convex set (and is the smallest convex set that contains C).

Definition 2.4 *Cone: a set $C \subseteq \mathbb{R}^n$ is a cone if for any $x \in C$, we have $tx \in C$ for all $t \ge 0$*

Definition 2.5 *Convex cone: a cone that is also convex, i.e.,*

$$x_1, x_2 \in C \implies t_1 x_1 + t_2 x_2 \in C \ \textit{for all } t_1, t_2 \ge 0$$

The set of all conic combination of points in C is called the **conic hull** of C

Convex Hull
$\mathrm{conv}(\mathcal{X}) = \{\boldsymbol{x} \mid \boldsymbol{x} = \sum_{i=1}^K \lambda_i \boldsymbol{x}_i, \lambda_i \ge 0, \sum_{i=1}^K \lambda_i = 1\}$
Conic Hull
$\mathrm{cone}(\mathcal{X}) = \{\boldsymbol{x} \mid \boldsymbol{x} = \sum_{i=1}^K \lambda_i \boldsymbol{x}_i, \lambda_i \ge 0\}$

The conic hull of a set C collects all conic combinations of points in C, and is the smallest *convex* cone containing C.

Combinations

	$\boldsymbol{z} = a\boldsymbol{x} + b\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$
linear	$a, b \in \mathbb{R}$
conic	$a, b \ge 0$
affine	$a + b = 1$
convex	$a + b = 1, a, b \ge 0$

The Tangent(Polar) Cone and Normal Cone

➢ *Normal Cone to set C at point x* :
$N_C(x) = \{v \in V \mid \langle v, y - x\rangle \le 0, \forall y \in C\} = (C - x)^\circ$
Even if C is not convex this cone is a convex cone

➢ *Polar Cone to any cone C* :
$C^\circ = \{v \in V \mid \langle v, y\rangle \le 0, \forall y \in C\}$
For general sets C, the tangent cone need not be convex.

2.2.2 Examples of convex sets

- Empty set, point, line.
- Norm ball: $\{x : \|x\| \le r\}$, for given norm $\|\cdot\|$, radius r.
- Hyperplane: $\{x : a^T x = b\}$, for given a, b.
- Halfspace: $\{x : a^T x \le b\}$.
- Affine space: $\{x : Ax = b\}$, for given A, b.
- Polyhedron: $\{x : Ax \le b\}$, where $\le$ is interpreted componentwise. The set $\{x : Ax \le b, Cx = d\}$ is also a polyhedron.
- Simplex: special case of polyhedra, given by $\mathrm{conv}\{x_0, ..., x_k\}$, where these points are affinely independent. The canonical example is the probability simplex,

1.4.4. Polytopes
$\mathcal{X} = \{\boldsymbol{x} \mid \boldsymbol{a}_1^\intercal \boldsymbol{x} \le b1, \boldsymbol{a}_2^\intercal \boldsymbol{x} \le b2 \ldots\} = \{\boldsymbol{x} \mid \boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}\}$

2.2.3 Examples of convex cones

- Norm cone: $\{(x, t) : \|x\| \le t\}$, for given norm $\|\cdot\|$. It is called second-order cone under the l_2 norm $\|\cdot\|_2$.
- Normal cone: given any set C and point $x \in C$, the normal cone is

  $$\mathcal{N}_C(x) = \{g : g^T x \ge g^T y, \ \text{for all } y \in C\}$$

  This is always a convex cone, regardless of C.
- Positive semidefinite cone:

  $$\mathbb{S}^n_+ = \{X \in \mathbb{S}^n : X \succeq 0\}$$

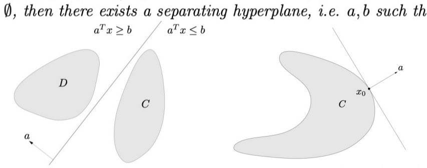
  where $X \succeq 0$ means that X is positive semidefinite ($\mathbb{S}^n$ is the set of $n \times n$ symmetric matrices).

Theorem 1.4 *For a convex optimization problem any local optima is a global optima.*
Theorem 1.5 *The set of optimal solutions to a convex optimization problem is a convex set.*

2.2.4 Key properties of convex sets

Theorem 1.6 (Separating Hyperplane) *If C and D are non-empty convex sets which are disjoint, i.e. $C \cap D = \emptyset$, then there exists a separating hyperplane, i.e. a, b such that,*

$a^T x \le b, \ \textit{for all } x \in C,$
$a^T x \ge b, \ \textit{for all } x \in D.$



Theorem 1.7 (Supporting Hyperplane) *If C is a non-empty convex set, and $x_0 \in boundary(C)$, then there is a vector a such that,*

$$a^T (x - x_0) \le 0, \ \textit{for all } x \in C.$$

2.2.5 Operations preserving convexity

2.2.5.1 Operations

- Intersection: the intersection of convex sets is convex.
- Scaling and translation: if C is convex, then $aC + b = \{ax + b : x \in C\}$ is convex for any a, b.
- Affine images and preimages: if $f(x) = Ax + b$ and C is convex, then $f(C) = \{f(x) : x \in C\}$ is convex and if D is convex, then $f^{-1}(D) = \{x : f(x) \in D\}$ is convex. Compared to scaling and translation this operation also has rotation and dimension reduction.
- Perspective images and preimages: the perspective function is $P : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}^n$ (where $\mathbb{R}_{++}$ denotes positive reals),

  $$P(x, z) = x/z$$

  for $z > 0$. If $C \subseteq \mathrm{dom}(P)$ is convex then so is $P(C)$, and if D is convex then so is $P^{-1}(D)$.
- Linear-fractional images and preimages: the perspective map composed with an affine function,

  $$f(x) = \frac{Ax + b}{c^T x + d}$$

  is called a linear-fractional function, defined on $c^T x + d > 0$. If $C \subseteq \mathrm{dom}(f)$ is convex then so is $f(C)$, and if D is convex then so is $f^{-1}(D)$.

f is l.s.c. if epi f is closed.

f is convex if dom$f$ is convex and
1. $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$
2. $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \ge 0$
3. $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle$
4. $\nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{0}$, if f is twice differentiable
5. epi f is convex

f is α-strongly convex if dom$f$ is convex and
1. $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) - \frac{\alpha}{2}\lambda(1-\lambda)\|x - y\|_2^2$
2. $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \ge \alpha\|x - y\|_2^2$
3. $f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|x - y\|_2^2$
4. $f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex
5. $\nabla^2 f(\boldsymbol{x}) \succeq \alpha \boldsymbol{I}$, if f is twice differentiable

f is L-Lipschitz gradient (L-smooth) if f is differentiable and
1. $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$
2. $\left|f(y) - f(x) - \langle \nabla f(x), y - x\rangle\right| \le \frac{L}{2}\|y - x\|_2^2$
3. $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \ge \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2$
4. $\nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$, if f is twice differentiable

f is L-Lipschitz Hessian if f is twice differentiable and
1. $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L\|x - y\|$
2. $|f(x) - f(y) - \langle \nabla f(x), y - x\rangle - \langle \nabla^2 f(x)(y - x), y - x\rangle| \le \frac{L}{6}\|y - x\|_2^3$

f is α-strongly convex and β-smooth
1. $\langle \boldsymbol{x} - \boldsymbol{y}, \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\rangle \ge \dfrac{\alpha\beta}{\alpha + \beta}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{\alpha+\beta}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2$

Note: strongly convex implies strictly convex, which subsequently implies convex. In equation format:

$$\textit{strongly convex} \Rightarrow \textit{strictly convex} \Rightarrow \textit{convex}$$

1. A function is convex iff the univariate functions $g(t) = f(x + tv)$ are convex for any $v \in \mathbb{R}^d$, and for any $x \in \mathrm{dom}(f)$.

2.2 More Examples of Convex Functions

1. $\exp(ax)$ is convex for any a over $\mathbb{R}$.
2. $\log x$ is concave on $\mathbb{R}_{++}$.
3. $a^T x + b$ is convex (and concave).
4. The least squares loss $\|Ax - b\|^2$ is convex (for any A, b).
5. Any norm is convex, i.e. $\|x\|$ is a convex function.
6. The spectral norm, and the trace norm of a matrix are convex, i.e. $\|X\|_{\mathrm{op}} = \sigma_1(X)$, $\|X\|_{\mathrm{tr}} = \sum_{i=1}^d \sigma_i(X)$ where $\sigma_i(X)$ denotes the i-th singular value of X.
7. **Convex Indicators:** If C is a convex set, then the indicator function (which is defined on the extended reals):
   $$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C. \end{cases}$$

2.3.3 Key properties of convex functions

- A function is convex if and only if its restriction to any line is convex
- Epigraph characterization: a function f is convex if and only if its epigraph

  $$epi(f) = (x, t) \in dom(f) \times \mathbb{R} : f(x) \le t$$

  is a convex set.
- Convex sublevel sets: if f is convex, then its sublevel sets

  $$x \in dom(f) : f(x) \le t$$

  are convex, for all $t \in \mathbb{R}$. The converse is not true.

- Jensen's inequality: if f is convex, and X is a random variable supported on $dom(f)$, then $f(\mathbb{E}[\mathbb{X}]) \le \mathbb{E}[f(x)]$.
- Long-sum-exp function: $g(x) = log(\sum_{i=1}^k e^{a_i^T x + b_i})$ for fixed a_i, b_i. This is often called the soft max, since it smoothly approximates $\max_{i=1,...,k}(a_i^T x + b_i)$.

2.5 Operations which Preserve Convexity

1. **Non-negative Linear Combination:** Suppose $f_1, \ldots, f_m$ are convex, then so is $\sum_{i=1}^m a_i f_i$ for any $a_1, \ldots, a_m \ge 0$.
2. **Pointwise Max:** If the collection of functions f_s for $s \in S$ are convex, then so is $g(x) = \sup_{s \in S} f_s(x)$.
3. **Partial Minimization:** If $g(x, y)$ is a convex function, and C is a convex set, then $f(x) = \min_{y \in C} g(x, y)$ is a convex function.

- Affine composition: if f is convex, then $g(x) = f(Ax + b)$ is convex.
- General composition: suppose $f = hg$, where $g : \mathbb{R}^n \to \mathbb{R}, h : \mathbb{R} \to \mathbb{R}, f : \mathbb{R}^n \to \mathbb{R}$. Then:
  (1) f is convex if h is convex and nondecreasing, g is convex
  (2) f is convex if h is convex and nonincreasing, g is concave
  (3) f is concave if h is concave and nondecreasing, g is concave
  (4) f is convex if h is convex and nonincreasing, g is convex
  Note: To memorize this, think of the chain rule when $n = 1$:

  $$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$$

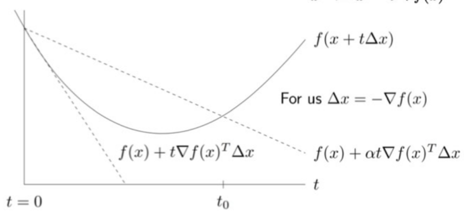
Backtracking Line Search

1. First, fix parameters $0 < \beta < 1$ and $0 < \alpha \le \frac{1}{2}$
2. At each iteration (of gradient descent), start with $t = t_{\mathrm{init}}$ (something relatively large), and while

   $$f(x - t\nabla f(x)) > f(x) - \alpha t\|\nabla f(x)\|_2^2$$

   shrink $t := \beta t$. Else, perform the gradient descent update

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



GD on Smooth Functions

All assume objective function f is twice differentiable and β-smooth

(1) f is β-smooth

The main descent Lemma
For any step-size $\eta < \frac{2}{\beta}$, the GD algorithm is a descent algorithm,
For any $\eta \le \frac{1}{\beta}$,
$$f(x^{t+1}) \le f(x^t) - \frac{\eta}{2}\|\nabla f(x^t)\|_2^2 \qquad (3\text{-}1)$$

1. If $\|\nabla f(x^t)\|_2 > 0$ then we have strict descent, i.e. $f(x^{t+1}) < f(x^t)$.
2. Furthermore, if the gradient is large (in norm) then an iteration of GD decreases the function by a large amount.
3. Just by smoothness (no convexity), we already see that GD doesn't suffer from the "bouncing around" problem

Theorem 3.3 *Let x *be any minimizer of f, then GD with step-size $\frac{1}{\beta}$ has the property that within k iterations it will reach a point x such that*

The main theorem $\quad \|\nabla f(x)\|_2 \le \sqrt{\frac{2\beta}{k}(f(x^0) - f(x^*))}.$

Dimension-free: The result (the error) goes doesn't depend on dimension d.

(2) f is β-smooth, and convex

Theorem 3.4 *Let x^* be any minimizer of f, then GD with step-size $\frac{1}{\beta}$ has the property that after k iterations it will reach a point x^k such that*

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

1. It is worth noting that now GD will find a point as good as the best point x^* However, the guarantee is still much slower than the one we derived earlier for quadratics. To obtain ϵ-error we need to take roughly $1/\epsilon$ steps.

We say that gradient descent has convergence rate $O(1/k)$, i.e., it finds ϵ-suboptimal point in $O(1/\epsilon)$ iterations. We read this by saying that after k iterations, the gap between the criterion and where we are goes down by $1/k$.

(3) f is β-smooth, α-strongly convex

Theorem 4.1 *Let x^* denote the minimizer of f, then after k iterations the GD iterate x^k satisfies,*

$$\|x^k - x^*\|_2^2 \le \left(1 - \frac{1}{\kappa}\right)^k \|x^0 - x^*\|_2^2. \qquad \kappa = \frac{\beta}{\alpha}.$$

Gradient Descent convergence rate under strong convexity is $O(\gamma^k)$, i.e., it finds ϵ-suboptimal point in $O(\log(1/\epsilon))$ iterations. Exponentially fast!

Introduction to subgradients

Definition 6.5 (Subgradient) *g is a **subgradient** of a convex function f at x if*

$$f(y) \ge f(x) + g^T(y - x) \qquad \forall y$$

- Always exists in the relative interior of the dom(f).
- If f is indeed differentiable at x, then $g = \nabla f(x)$ uniquely.
- This definition is universal - can hold for non-convex functions too. However, it could be possible tha g doesn't exist.

Some properties of the subdifferential:

- For convex f, $\partial f(x) \neq \emptyset$. However, for concave f, $\partial f(x) = \emptyset$.
- $\partial f(x)$ is closed and convex for any f.
- Since the subgradient is unique at points of differentiability, $\partial f(x) = \{\nabla f(x)\}$ when f is differentiable at x.
- $\partial f(x)$ is singleton, then f is differentiable at x and $\nabla f(x)$ is that only element of $\partial f(x)$.

Indicator Function: $f(x) = \mathbb{I}_C(x) = \begin{cases} \infty, & \text{if } x \notin C \\ 0, & \text{if } x \in C \end{cases} \Longrightarrow \partial f(x) = N_C(x)$

Optimality conditions (Lecture Note 2)

For $\min_{x \in C} f(x)$, where f is a convex function, C is a convex set.
What can I say about the solution x^* ?

(1) Unconstrained Case

$C = \mathbb{R}^d$, $\text{dom}(f) = \mathbb{R}^d$

Theorem 2-1

x^* is optimal, if and only if $0 \in \partial f(x^*)$

(2) Constrained, differentiable case

Theorem 2-2

A feasible point x^* is optimal, if and only if $\nabla f(x^*)^T(y - x^*) \geq 0$ for $\forall\, y \in C$

$$\Updownarrow$$

$-\nabla f(x^*) \in N_C(x^*) \iff -\nabla f(x^*)(y - x^*) \leq 0$

(3) Constrained case (General)

Theorem 2-3

A feasible point x^* is optimal, if and only if $0 \in \partial f(x^*) + N_C(x^*)$, for $\forall\, y \in C$

Lipschitz Function (bound for subgradient)

Assume objective function f is G-Lipschitz: (f is convex)

$$|f(x) - f(y)| \leq G\|x - y\|_2$$

Then all subgradients will have bounded l2 norm: for $\forall g \in \partial f(x)$, $\|g\|_2 \leq G$

Theorem 4-1 Convergence for subgradient methods

Suppose f is convex and G-Lipschitz, then

$$f(x^{best}) - f(x^*) \leq \frac{\|x^0 - x^*\|_2^2 + G^2 \sum_{t=0}^{k-1} \eta_t^2}{2\sum_{t=0}^{k-1} \eta_t}$$

① For any sequence of step size, satisfies two conditions in 2-1, we will have

$f(x^{best}) - f(x^*) \to 0$, as $k \to \infty$ convergence

② If step size is chosen to be a constant $\eta = \frac{R}{G\sqrt{k}}$

$f(x^{best}) - f(x^*) \leq \frac{GR}{\sqrt{k}}$ to get to ε error, we need $\frac{1}{\varepsilon^2}$ iterations

(1) GD, β smooth + convex $\qquad f(x^k) - f(x^*) \lesssim \frac{1}{k}$

(2) GD, β smooth + α strongly convex $\qquad f(x^k) - f(x^*) \lesssim (1 - \frac{1}{k})^k$

(3) Subgradient GD, G Lipschitz $\qquad f(x^{best}) - f(x^*) \lesssim \frac{1}{\sqrt{k}}$

The main takeaways from the above result are that the subgradient method is slow, but optimal for the class of convex, Lipschitz functions. However, GD is (potentially) sub-optimal for both smooth functions (it gets $1/k$ instead of $1/k^2$ rates), and for smooth and strongly convex functions where the dependence on κ is better in the lower bound (which has dependence on $\sqrt{\kappa}$ instead of κ).

Projected Gradient Descent

$$y^{t+1} = x^t - \eta \nabla f(x^t)$$
$$x^{t+1} = P_C(y^{t+1}).$$

$$x^{t+1} = P_C(y^{t+1}) := \arg\min_{x \in C} \frac{1}{2}\|x - y^{t+1}\|_2^2.$$

Theorem 6.1 *Suppose that f is convex and G-Lipschitz, and define x^{best} to be the best iterate seen so far and choose step-size η_t in each round, then we have the guarantee:*

$$f(x^{best}) - f(x^*) \leq \frac{\|x^0 - x^*\|_2^2 + G^2 \sum_{t=0}^{k-1} \eta_t^2}{2\sum_{t=0}^{k-1} \eta_t}.$$

Proximal Gradient Descent

convex, function g and a potentially non-smooth convex function h.

$$\min_{x \in \mathbb{R}^d} g(x) + h(x).$$

For a convex function f the proximal operator is defined to be:

$$\text{prox}_f(v) = \underset{x}{\text{argmin}}\left(f(x) + \tfrac{1}{2}\|x - v\|_2^2\right)$$

① compute $y^{t+1} = x^t - \eta_t \nabla g(x^t)$

② compute by solving

$$x^{t+1} = \underset{x \in \mathbb{R}^d}{\text{argmin}}\left[h(x) + \frac{1}{2\eta_t}\|x - y^{t+1}\|_2^2\right]$$
$$= \text{prox}_{\eta_t h}(y^{t+1})$$

Optimality Condition for proximal GD

if $u = \text{prox}_{\eta, h}(x) = \underset{u}{\text{argmin}}\ \frac{1}{2}\|x - u\|_2^2 + h(u)$

then $0 \in u - x + \eta\, \partial h(u)$

$\Rightarrow$ Define Gradient Mapping:

$$G_\eta(x) = \frac{1}{\eta}\left[x - \text{prox}_{\eta h}(x - \eta \nabla g(x))\right]$$
$$\Downarrow$$
$$x^{t+1} = x^t - \eta_t \underbrace{G_{\eta_t}(x^t)}_{\text{direction}}$$

Lemma 6.2

$G_\eta(x^*) = 0 \iff 0 \in \nabla g(x^*) + \partial h(x^*)$

Theorem 7.6 *After k iterations the proximal method, for convex h, achieves the guarantee:*

$$h(x^k) - h(x^*) \leq \frac{\|x^0 - x^*\|_2^2}{2\eta k}.$$

Stochastic Gradient Descent $\mathbb{E}_\xi[g(x,\xi)] = \nabla f(x)$.

SGD for Lipschitz Convex Functions

$$\mathbb{E} f\left(\frac{1}{k}\sum_{t=1}^{k} x^t\right) - f(x^*) \leq \frac{RG}{\sqrt{k}}.$$

It achieves the same rate of convergence as a function of ***k*** but each iteration of SGD faster than sub-gradient method.

SGD for Strongly Convex Functions

fixed step-size $\eta < 1/\alpha$,

$$\mathbb{E}\|x^k - x^*\|_2^2 \leq (1 - \alpha\eta)^k \|x^0 - x^*\|_2^2 + \frac{\eta G^2}{\alpha}.$$

For $\eta_t = \frac{1}{\alpha(t+1)}$

$$\mathbb{E} f\left(\frac{1}{k}\sum_{t=1}^{k} x^t\right) - f(x^*) \leq \frac{G^2(1 + \log k)}{2\alpha k}$$

Mirror Descent

$\Rightarrow$ use subgradient descent, $f(x^{best}) - f(x^*) \leq \frac{\sqrt{d}}{\sqrt{k}}$

$\Downarrow$ improved

$\Rightarrow$ use mirror descent, $f(x^{best}) - f(x^*) \lesssim \sqrt{\frac{\log d}{k}}$

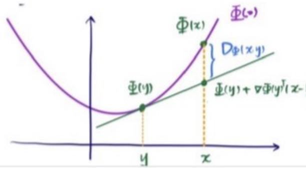
mirror map Φ: differentiable, α-strongly convex, w.r.t. $\|\cdot\|$

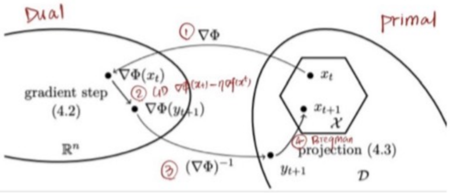
$$\Downarrow$$

$$\Phi(y) \geq \Phi(x) + \nabla\Phi^T(x)(y - x) + \frac{\alpha}{2}\|x - y\|^2$$

Bregman Divergence

$$D_\Phi(x, y) = \Phi(x) - (\Phi(y) + \nabla\Phi(y)^T(x - y))$$





The KKT conditions are necessary for optimality under strong duality, and always sufficient.

The Lagrange dual function g(u, v) is always concave

Slaters's condition, is a sufficient condition for strong duality to hold.

For a differentiable function f, we cannot use ∂f(x) = {∇f(x)} unless f is convex when applying the stationarity conditions.

$$\min_x f(x)$$
subject to $h_i(x) \leq 0 \quad i \in \{1, \dots, m\}$
$\ell_j(x) = 0, \quad j \in \{1, \dots, r\}$. , we can define our (Lagrange) dual problem as:

$$f(x) \geq f(x) + \sum_{j=1}^{r} u_j \ell_j(x) + \sum_{i=1}^{m} v_j h_j(x) := L(x, u, v).$$

$$p^* = \min_{x \text{ feasible}} f(x) \geq \min_x L(x, u, v) := g(u, v)$$

$$\max_{u,v} g(u, v)$$
subject to $v \geq 0$.

Dual is always concave maximization

$$g(u, v) = \min_x \left[f(x) + \sum_{j=1}^{r} u_j \ell_j(x) + \sum_{i=1}^{m} v_j h_j(x)\right]$$

Slater's Condition

$$\min_x f(x)$$
subject to $h_i(x) \leq 0 \quad i \in \{1, \dots, m\}$
$\ell_j(x) = 0, \quad j \in \{1, \dots, r\}$.

Slater's Theorem: Suppose that there exists a point $x_0 \in$ relative $\text{int}(\mathcal{D})$ such that,

$$\ell_j(x_0) = 0, \quad j \in \{1, \dots, r\}$$
$$h_i(x_0) \leq 0, \quad i \in \{1, \dots, k\}$$
$$h_i(x_0) < 0, \quad i \in \{k+1, \dots, m\},$$

KKT Conditions and Optimality

Recall that for the problem

$$\min_x \quad f(x)$$
subject to $\quad h_i(x) \leq 0,\ i = 1, \dots, m$
$\ell_j(x) = 0,\ j = 1, \dots, r$

the KKT conditions are

- $0 \in \partial_x\left(f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)\right)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \leq 0,\ \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \geq 0$ for all i (dual feasibility)

KKT Without Convexity

$$0 \in \partial f(\widehat{x}) + \sum_{j=1}^{r} \widehat{u}_j \partial \ell_j(\widehat{x}) + \sum_{i=1}^{m} \widehat{v}_j \partial h_j(\widehat{x}).$$

Fenchel Conjugate

$$f^*(y) = \sup_{x \in \text{dom}(f)} (y^T x - f(x)).$$

Fenchel's Inequality

$$f^*(y) \geq x^T y - f(x)$$

Properties:

- Conjugate of conjugate f^{**} satisfies $f^{**} \leq f$
- If f is closed and convex, then $f^{**} = f$
- If f is closed and convex, then for any x, y,

$$x \in \partial f^*(y) \iff y \in \partial f(x)$$
$$\iff f(x) + f^*(y) = x^T y$$

- If $f(u, v) = f_1(u) + f_2(v)$, then

$$f^*(w, z) = f_1^*(w) + f_2^*(z)$$

Conjugate Function

The **conjugate function** of f is

$$f^\star(\mathbf{y}) = \sup_{\mathbf{x} \in \text{dom}(f)} \left\{\mathbf{y}^T\mathbf{x} - f(\mathbf{x})\right\}$$

also called **Legendre-Fenchel transformation**.

Fenchel's inequality

$f(\mathbf{x}) + f^*(\mathbf{y}) \geq \mathbf{x}^T\mathbf{y}, \forall \mathbf{x}, \mathbf{y}$

Lemma 12.5

(1) **Duality**: If f is **lower semi-continuous (l.s.c.)** and convex, then $f^{\star\star} = f$.
Function f is l.s.c. if $f(\mathbf{x}) \leq \liminf_{t \to \infty} f(\mathbf{x}_t)$ for $\mathbf{x}_t \to \mathbf{x}$.

(2) **Fenchel's inequality**: $\mathbf{x}^T\mathbf{y} \leq f(\mathbf{x}) + f^\star(\mathbf{y})$.

(3) If f and g are l.s.c. and convex, then $(f+g)^\star(\mathbf{x}) = \inf_{\mathbf{y}}\left\{f^\star(\mathbf{y}) + g^\star(\mathbf{x} - \mathbf{y})\right\}$.

(4) If f is μ-strongly convex, then $f^\star$ is differentiable and $\frac{1}{\mu}$-smooth.

Examples

- Quadratic: $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}$ where $\mathbf{Q} > 0$, $f^\star(y) = \frac{1}{2}\mathbf{y}^T\mathbf{Q}^{-1}\mathbf{y}$
- Negative entropy: $f(\mathbf{x}) = \sum_{i=1}^{n} x_i \log(x_i)$, $f^\star(\mathbf{y}) = \sum_{i=1}^{n} e^{y_i - 1}$
- Negative logarithm: $f(\mathbf{x}) = -\sum_{i=1}^{n} \log(x_i)$, $f^\star(\mathbf{y}) = -\sum_{i=1}^{n} \log(-y_i) - n$.
- Norm: $f(\mathbf{x}) = \|\mathbf{x}\|$, $f^\star(\mathbf{y}) = \begin{cases} 0, & \|\mathbf{y}\|_* \leq 1 \\ +\infty, & \|\mathbf{y}\|_* > 1 \end{cases}$

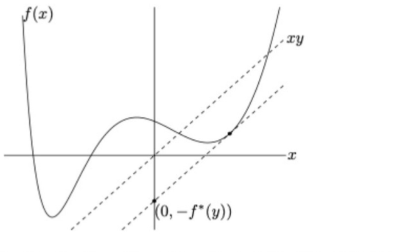


Figure 3.8 A function $f : \mathbf{R} \to \mathbf{R}$, and a value $y \in \mathbf{R}$. The conjugate function $f^*(y)$ is the maximum gap between the linear function yx and $f(x)$, as shown by the dashed line in the figure. If f is differentiable, this occurs at a point x where $f'(x) = y$.