IE 598: Big Data Optimization	**Fall 2016**

Lecture 3–4: Convex Functions – Aug 30 & Sep 1

Lecturer: Niao He *Scriber: Niao He*

Overview In these two lectures, we will introduce the concept of convex functions, and provide several ways to characterize convex functions, discuss some calculus that can be used to detect convexity of functions and compute the subgradients of convex function.

3.1 Definitions

Definition 3.1 (Convex function) *A function $f(x) : \mathbf{R}^n \to \mathbf{R}$ is convex if*

(i) $dom(f) \subseteq \mathbf{R}^n$ s a convex set;

(ii) $\forall x, y \in dom(f)$ and $\lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$.

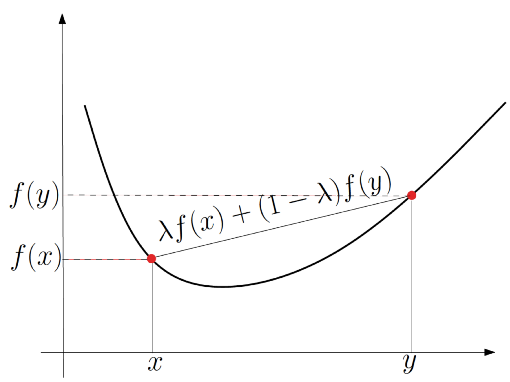

Figure 3.1: Example of convex function

A function is called strictly convex if (ii) holds with strict sign, i.e. $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$.

A function is called α-strictly convex if $f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex.

A function is called concave if $-f(x)$ is convex.

For example, a linear function is both convex and concave. Any norm is convex.

Remark 1 (Extended value function). Conventionally, we can think of f as an extended value function from $\mathbf{R}^n$ to $\mathbf{R} \cup \{+\infty\}$ by setting $f(x) = +\infty$ if $x \notin dom(f)$, the condition (ii) is equivalent as

$$\forall x, y, \forall \lambda \in [0,1], f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y).$$

Remark 2. (Slope inequality) What does convexity really mean? Let $z = \lambda x + (1-\lambda)y$, observe that $\|y-x\| : \|y-z\| : \|z-x\| = 1 : \lambda : (1-\lambda)$. Therefore

$$\begin{aligned}
& f(z) \leq \lambda f(x) + (1-\lambda) f(y) \\
\iff \quad & \frac{f(z)-f(x)}{1-\lambda} \leq f(y) - f(x) \leq \frac{f(y)-f(z)}{\lambda} \\
\iff \quad & \frac{f(z)-f(x)}{\|z-x\|} \leq \frac{f(y)-f(x)}{\|y-x\|} \leq \frac{f(y)-f(z)}{\|y-z\|}
\end{aligned}$$

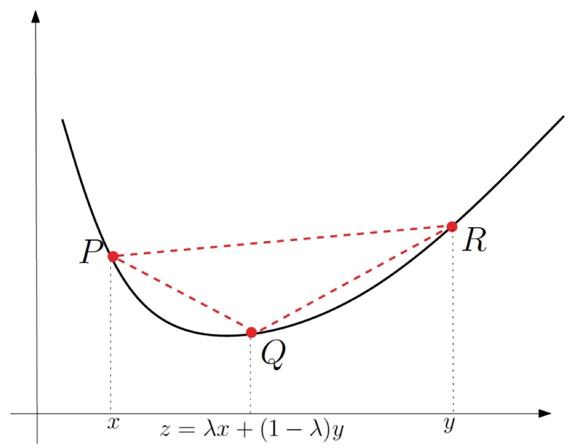

Figure 3.2: Slope PQ $\leq$ Slope PR $\leq$ Slope QR

3.2 Several Characterizations of Convex Functions

1. **Epigraph characterization**

   Proposition 3.2 *f is convex if and only if its epigraph*

   $$epi(f) := \{(x,t) \in \mathbf{R}^{n+1} : f(x) \leq t\}$$

   is a convex set.

   Proof: This can be verified by using the definition of convex function and convex set.

   - ($\Longrightarrow$) Suppose $(x, t_1), (y, t_2) \in epi(f)$, then $f(x) \leq t_1, f(y) \leq t_2$. For any $\lambda \in [0,1]$, by convexity of f, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \leq \lambda t_1 + (1-\lambda) t_2$. Thi implies that $\lambda \cdot (x, t_1) + (1-\lambda) \cdot (y, t_2) \in epi(f)$. Hence, $epi(f)$ is a convex set.
   - ($\Longleftarrow$) Let $x, y \in \mathbf{R}^n$, since $(x, f(x))$ and $(y, f(y))$ lie in $epi(f)$, by convexity of epigraph set, we have for any $\lambda \in [0,1]$, $(\lambda x + (1-\lambda)y, \lambda(x) + (1-\lambda) f(y)) \in epi(f)$. By definition, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$. Hence, function f is convex.

$\blacksquare$

2. **Level set characterization**

   Proposition 3.3 *If f is convex, then the level set for any $t \in \mathbf{R}$*

   $$C_t(f) = \{x \in dom(f) : f(x) \leq t\}$$

   is a convex set.

   For example, the unit norm ball $\{x : \|x\| \leq 1\}$ is a convex set since $\|\cdot\|$ is convex.

   Remark. The reverse is not true. A function with convex level set is not always convex. In fact, it is known as a quasi-convex function.

3. **One-dimensional characterization**

   Proposition 3.4 *f is convex if and only if its restriction on any line, i.e. function*

   $$\phi(t) := f(x + th)$$

   is convex on the axis for any x and h.

   Remark. Convexity is a one-dimensional property. In order to detect the convexity of a function, it all boils down to check the convexity of a one-dimensional function on the axis. From basic calculus, we already know that

   $$\begin{aligned}
   & \phi(t) \text{ is convex on } (a, b) \\
   \iff & \frac{\phi(s) - \phi(t_1)}{s - t_1} \leq \frac{\phi(t_2) - \phi(t_1)}{t_2 - t_1} \leq \frac{\phi(t_2) - \phi(s)}{t_2 - s}, \forall a < t_1 < s < t_2 < b && \text{(due to slope inequality)} \\
   \iff & \phi'(t_1) \leq \phi'(t_2), \forall a < t_1 < t_2 < b && \text{(if } \phi \text{ is differentiable)} \\
   \iff & \phi''(t) > 0, \forall a < t < b && \text{(if } \phi \text{ is twice-differentiable)}
   \end{aligned}$$

   Hence, if f is differentiable or twice-differentiable, we can characterize it by based on its first-order or second-order.

4. **First-order characterization for differentiable convex functions**

   Proposition 3.5 *Assume f is differentiable, then f is convex if and only if $dom(f)$ is convex and for any x, y,*

   $$f(x) \geq f(y) + \nabla f(y)^T (x - y). \tag{$\star$}$$

   Proof:

   - ($\Longrightarrow$) If f is convex, letting $z = (1-\epsilon)y + \epsilon x = y + \epsilon(x - y)$ with $\epsilon \in (0, 1)$, from the slop inequality, we have

     $$\frac{f(x) - f(y)}{\|x - y\|} \geq \frac{f(z) - f(y)}{\|z - y\|} = \frac{f(y + \epsilon(x - y)) - f(y)}{\epsilon\|x - y\|}.$$

     Hence, letting $\epsilon \to 0+$, we have

     $$f(x) - f(y) \geq \lim_{\epsilon \to 0+} \frac{f(y + \epsilon(x - y)) - f(y)}{\epsilon} = \nabla f(y)^T (x - y).$$

     Therefore, $f(y) \geq f(x) + \nabla f(x)^T (y - x)$.

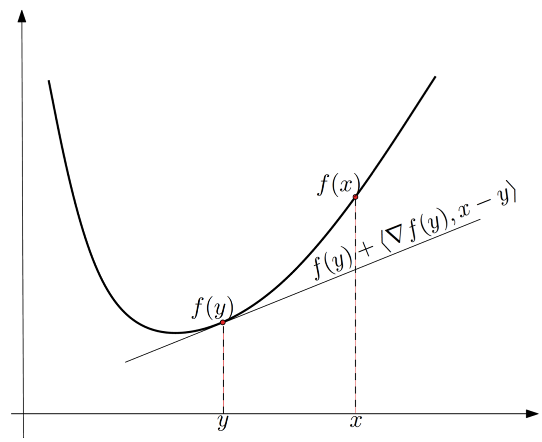

Figure 3.3: First-order condition

- ($\Longleftarrow$) If ($\star$) holds, letting $z = \lambda x + (1-\lambda) y$ for any $\lambda \in [0,1]$, we have
$$f(x) \geq f(z) + \nabla f(z)^T (x - z)$$
$$f(y) \geq f(z) + \nabla f(z)^T (y - z)$$
Adding the two inequalities with scalings λ and $(1-\lambda)$, it follows that
$$\lambda f(x) + (1-\lambda) f(y) \geq f(z) = f(\lambda x + (1-\lambda) y).$$
Hence, f is convex.

■

5. **Second-order characterization for twice-differentiable convex functions**

Proposition 3.6 *Assume f is twice-differentiable, then f is convex if and only if $dom(f)$ is convex and for any $x \in dom(f)$,*
$$\nabla^2 f(x) \succeq 0. \tag{$\star\star$}$$

Proof:

- ($\Longrightarrow$) If f is convex, then for any x, h, $\phi(t) = f(x + th)$ is convex on the axis. Hence, $\phi''(t) \geq 0, \forall t$. Particularly,
$$\phi''(0) = h^T \nabla^2 f(x) h \geq 0.$$
This implies that $\nabla^2 f(x) \succeq 0$.
- ($\Longleftarrow$) It suffices to show that every one-dimensional function $\phi(t) := f(x + t(y - x))$ is convex for any $x, y \in dom(f)$. The latter is indeed true because $\phi''(t) = (y-x)^T \nabla^2 f(x + t(y-x))(y-x) \geq 0$ due to ($\star\star$).

■

6. **Subgradient characterization for non-differentiable convex functions**

Proposition 3.7 *f is convex if and only if $\forall x \in int(dom(f))$, there exists g, such that*
$$f(x) \geq f(y) + g^T (x - y)$$
i.e. the subdifferential set is non-empty.

To be discussed in Section 3.5.

3.3 Calculus of Convex Functions

The following operators preserve the convexity of functions, which can be easily verified based on the definition.

1. **Taking conic combination**: If $f_\alpha(x), \alpha \in \mathcal{A}$ are convex functions and $\{\lambda_\alpha\}_{\alpha\in\mathcal{A}} \geq 0$, then

$$\sum_{\alpha\in\mathcal{A}} \lambda_\alpha f_\alpha(x)$$

   is also a convex function.

2. **Taking affine composition** If $f(x)$ is convex on $\mathbf{R}^n$, and $\mathcal{A}(y) : y \mapsto Ay + b$ is an affine mapping from $\mathbf{R}^k$ to $\mathbf{R}^n$, then

$$g(y) := f(Ay + b)$$

   is convex on $\mathbf{R}^k$.

   The proofs are straightforward and hence omitted.

3. **Taking superposition**:

   - If f is a convex function on $\mathbf{R}^n$ and $F(\cdot)$ is a convex and non-decreasing function on $\mathbf{R}$, then $g(x) = F(f(x))$ is convex.
   - More generally, if $f_i(x), i = 1, \ldots, m$ are convex on $\mathbf{R}^n$ and $F(y_1, \ldots, y_m)$ is convex and non-decreasing (component-wise) on $\mathbf{R}^m$, then

$$g(x) = F(f_1(x), \ldots, f_m(x))$$

     is convex.

   Proof: By convexity of f_i, we have

$$f_i(\lambda x + (1-\lambda)y) \leq \lambda f_i(x) + (1-\lambda) f_i(y), \forall i, \forall \lambda \in [0, 1].$$

   Hence, we have for any $\lambda \in [0, 1]$,

$$\begin{aligned} g(\lambda x + (1-\lambda)y) &= F(f_1(\lambda x + (1-\lambda)y), \ldots, f_m(\lambda x + (1-\lambda)y)) \\ &\leq F(\lambda f_1(x) + (1-\lambda) f_1(y), \ldots, \lambda f_m(x) + (1-\lambda) f_m(y)) && (\text{ by monotonicity of } F) \\ &\leq \lambda F(f_1(x), \ldots, f_m(x)) + (1-\lambda) F(f_1(x), \ldots, f_m(x)) && (\text{ by convexity of } F) \\ &= \lambda g(x) + (1-\lambda) g(y) && (\text{ by definition of } g) \end{aligned}$$

   ■

4. **Taking supremum**: If $f_\alpha(x), \alpha \in \mathcal{A}$ are convex, then

$$\sup_{\alpha\in\mathcal{A}} f_\alpha(x)$$

   is convex.

   Note that when $\mathcal{A}$ is finite, this can be considered as a special superposition with $F(y_1, \ldots, y_m) = \max(y_1, \ldots, y_m)$, which can be easily shown to be monotonic and convex.

Proof: We show that

$$\begin{aligned}
epi(\sup_{\alpha\in\mathcal{A}} f_\alpha) &= \{(x,t) : \sup_{\alpha\in\mathcal{A}} f_\alpha(x) \le t\} \\
&= \{(x,t) : f_\alpha(x) \le t, \forall \alpha \in \mathcal{A}\} \\
&= \cap_{\alpha\in\mathcal{A}}\{(x,t) : f_\alpha(x) \le t\} \\
&= \cap_{\alpha\in\mathcal{A}} epi(f_\alpha).
\end{aligned}$$

Since f_α is convex, $epi(f_\alpha)$ is therefore a convex set for any $\alpha \in \mathcal{A}$. Their intersection remains convex, i.e. $epi(\sup_{\alpha\in\mathcal{A}} f_\alpha)$ is a convex set, i.e. $\sup_{\alpha\in\mathcal{A}} f_\alpha(x)$ is convex. ■

5. **Partial minimization**: If $f(x,y)$ is convex in $(x,y) \in \mathbf{R}^n$ and C is a convex set, then

$$g(x) = \inf_{y\in C} f(x,y)$$

is convex.

Proof: Given any x_1, x_2, by definition, for any $\epsilon > 0$,

$$\begin{aligned}
\exists y_1 &: f(x_1,y_1) \le g(x_1) + \epsilon/2 \\
\exists y_2 &: f(x_2,y_2) \le g(x_2) + \epsilon/2
\end{aligned}$$

For any $\lambda \in [0,1]$, adding the two equations, we have

$$\lambda f(x_1,y_1) + (1-\lambda) f(x_2,y_2) \le \lambda g(x_1) + (1-\lambda) g(x_2) + \epsilon.$$

Invoking the convexity of $f(x,y)$, this implies

$$f(\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2) \le \lambda g(x_1) + (1-\lambda) g(x_2) + \epsilon.$$

Hence for any $\epsilon > 0$, $g(\lambda x_1 + (1-\lambda)x_2) \le \lambda g(x_1) + (1-\lambda)g(x_2) + \epsilon$. Letting $\epsilon \to 0$ leads to the convexity of g. ■

6. **Perspective function**: If $f(x)$ is convex , then the *perspective* of f

$$g(x,t) := t f(x/t)$$

is convex on its domain $dom(g) = \{(x,t) : x/t \in dom(f), t > 0\}$.

Proof: Observe that

$$(x,t,\tau) \in epi(g) \Longleftrightarrow t f(x/t) < \tau \Longleftrightarrow f(x/t) \le \tau/t \Longleftrightarrow (x/t, \tau/t) \in epi(f)$$

Define the perspective function $P : \mathbf{R}^n \times \mathbf{R}_{++} \times \mathbf{R} :\to \mathbf{R}^n \times \mathbf{R}, (x,t,\tau) \mapsto (x/t, \tau/t)$, then

$$epi(g) = P^{-1}(epi(f)).$$

Since f is convex, $epi(f)$ is a convex set. To show g is convex, it suffices to show that the inverse image of a convex set under the perspective function is convex.

Claim: If U is a convex set, then

$$P^{-1}(U) = \{(u,t) : u/t \in U, t > 0\}$$

is a convex set.
This is because if $(u,t) \in P^{-1}(U)$ and $(v,t) \in P^{-1}(U)$, for any $\lambda \in [0,1]$,

$$\frac{\lambda u + (1-\lambda)v}{\lambda t + (1-\lambda)s} = \mu \cdot \frac{u}{t} + (1-\mu) \cdot \frac{v}{s} \in U$$

where $\mu = \frac{\lambda t}{\lambda t + (1-\lambda)s} \in [0,1]$. Hence, $\lambda \cdot (u,t) + (1-\lambda) \cdot (v,s) \in P^{-1}(U)$. ■

3.4 Examples of Convex Functions

Example 1. *Simple univariate functions:*

- $x^2, x^4, \ldots$
- e^{ax} for any a
- $-\log(x)$
- $x\log(x)$

Example 2. *Multi-variate functions*:

- $\|\cdot\|$
- $\frac{1}{2}x^TQx + b^Tx + c$, when $Q \succeq 0$
- $\|Ax - b\|_2^2$
- $\max(a_1^Tx + b_1, \ldots, a_k^Tx - b_k)$
- relative entropy function $g(x,t) : \mathbf{R}^2_{++} \to \mathbf{R}, (x,t) \mapsto t\log(t) - t\log(x)$
- $\log(\sum_{i=1}^k e^{a_i^Tx+b_i})$

  Proof: It suffices to show that $f(x) = \log(\sum_{i=1}^n e^{x_i})$ is convex. Observe that any h, we have

  $$h^T\nabla^2 f(x)h = \frac{\sum_i e^{x_i}h_i^2}{\sum_i e^{x_i}} - \frac{(\sum_i e^{x_i}h_i)^2}{(\sum_i e^{x_i})^2}.$$

  Let $p_i = \frac{e^{x_i}}{\sum_i e^{x_i}}$, we have

  $$h^T\nabla^2 f(x)h = \sum_i p_ih_i^2 - (\sum_i p_ih_i)^2 \geq \sum_i p_ih_i^2 - \sum_i(\sqrt{p_i})^2\sum_i(\sqrt{p_i}h_i)^2 = \sum_i p_ih_i^2 - 1\cdot\sum_i p_ih_i^2 = 0.$$

  The first inequality is due to Cauchy-Schwarz inequality. Hence, $\nabla^2 f(x) \succeq 0$. ∎

- $-\log(det(X))$

  Proof: Let $f(X) = -\log(det(X))$, the domain $dom(f) = S^n_{++}$. Let $X, H \succ 0$, it is sufficient to show that $g(t) = f(X + tH)$ is convex on $dom(g) = \{t : X + tH \succ 0\}$. Since

  $$g(t) = -\log(det(X+tH)) = -\log(det(X^{1/2}(I+tX^{-1/2}HX^{-1/2})X^{1/2})) = -\sum_i \log(1+t\lambda_i) - \log(det(X))$$

  where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $X^{-1/2}HX^{-1/2}$. Note that for each i, $-\log(1 + t\lambda_i)$ is convex in t, so $g(t)$ is also convex. ∎

Example 3. Some distances:

- maximum distance to any set C: $d(x,C) := \max_{y\in C}\|x - y\|$
- minimum distance to a convex set C: $d(x,C) : \min_{y\in C}\|x - y\|$

Example 4. *Indicator and support functions*:

- indicator function of a convex set C: $I_C(x) := \begin{cases} 0, & x \in C \\ \infty, & x \notin C \end{cases}$
- support function of any set C (convex or not): $I_C^*(x) = \sup_{y \in C} x^T y$

3.5 Subgradients of Convex Functions

Definition 3.8 (Subgradient) *Let f be a convex function and $x \in dom(f)$, any vector g satisfying*

$$f(y) \geq f(x) + g^T(y - x)$$

is called a <u>*subgradient*</u> *of f at x.*

The set of all subgradients of f at x is called the <u>*subdifferential*</u>*, denoted as $\partial f(x)$.*

Example 1. If f is differentiable at $x \in dom(f)$, then $\nabla f(x)$ is the unique element of $\partial f(x)$.
Proof: Let $g \in \partial f(x)$, by definition, $\frac{f(x+td)-f(x)}{t} \geq g^T d, , \forall d$. Let $t \to 0$, we have $\nabla f(x)^T d \geq g^t d, \forall d$, which implies $\nabla f(x) = g$. ■

Example 2. Let $f(x) = |x|$, then $\partial f(0) = [-1, 1]$.
Proof: This is because $|x| \geq 0 + gx, \forall g \in [-1, 1]$. ■

Example 3. Let $f(x) = \|x\|_2$, then $\partial f(x) = \begin{cases} \frac{x}{\|x\|_2}, & x \neq 0 \\ \{g : \|g\|_2 \leq 1\}, & x = 0 \end{cases}$.
Proof: This is because $\|x\|_2 \geq 0 + g^T x, \forall \|g\|_2 \leq 1$. ■

Proposition 3.9 *If $\bar{x} \in int(dom(f))$, then $\partial f(\bar{x})$ is nonempty, closed, bounded, and convex.*

Proof:

- (Convexity and closedness): this is due to the fact that

$$\partial f(\bar{x}) = \cap_x \{g : f(x) \geq f(\bar{x}) + g^T(x - \bar{x})\}$$

  is a infinite system of linear inequalities. The sub-differentiable set can be treated as the intersection of halfspaces, hence is closed and convex.

- (Non-emptyness): applying the separation theorem on $(\bar{x}, f(\bar{x}))$ and $epi(f) = \{(x, t) : f(x) \leq t\}$, we have

$$\exists a, \beta, \text{s. t. } a^T \bar{x} + \beta f(\bar{x}) \leq a^T x + \beta t, \forall (x, t) \in epi(f).$$

  Claim: $\beta > 0$. We can first rule out $\beta \neq 0$ since $\bar{x} \in int(dom(f))$. We then rule out $\beta < 0$ by setting $x = \bar{x}$ and $t > f(\bar{x})$.
  Therefore, defining $g = \beta^{-1} a$, we have $f(x) \geq f(\bar{x}) + g^T(x - \bar{x})$, i.e. $g \in \partial(f)$.

- (Boundedness): if $\partial f(\bar{x})$ is unbounded, then there exist $s_k \in \partial f(\bar{x})$, such that $\|s_k\|_\infty \to \infty$ as $n \to \infty$. SInce $\bar{x} \in int(dom(f))$, there exits $\epsilon > 0$, such that $B(\bar{x}, \epsilon) = \{x : \|x - \bar{x}\| \le \epsilon\} \subset dom(f)$. Hence, letting $y_k = \bar{x} + \epsilon \frac{s_k}{\|s_k\|}$, we have $y_k \in B(\bar{x}, \epsilon)$, and

$$f(y_k) \ge f(\bar{x}) + s_k^T (y_k - \bar{x}) = f(\bar{x}) + \epsilon \|s_k\| \to \infty, \text{ as } k \to \infty.$$

However, every convex function can be shown to be continuous on its interior; it is Lipschitz continuous on any convex compact subset on the domain. This implies that $f(x)$ is bounded on the compact ball $B(\bar{x}, \epsilon)$, which leads to a contradiction.

■

Remark. The reverse is also true. If $\forall x \in int(dom(f))$, $\partial f(x)$ is nonempty, then f is convex.

Proof: Let $x, y \in dom(f)$, $z = \lambda x + (1 - \lambda) y \in int(dom(f))$, we have

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

Hence, $\lambda f(x) + (1 - \lambda) f(y) \ge f(z) = f(\lambda x + (1 - \lambda) y)$. ■

3.6 Calculus of Sub-differential

Determining the subdifferentiable set of a convex function at a given point is in general very difficult. That's why calculus of subdifferentiable sets is particularly important in convex analyis.

1. **Taking conic combination**: If $h(x) = \lambda f(x) + \mu g(x)$, where $\lambda, \mu \ge 0$ and f, g are both convex, then
$$\partial h(x) = \lambda \partial f(x) + \mu \partial g(x), \forall x \in int(dom(h)).$$
2. **Taking affine composition**: If $h(x) = f(Ax + b)$, where f is convex, then
$$\partial h(x) = A^T \partial f(Ax + b).$$
3. **Taking supremum**: If $h(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$ and each $f_\alpha(x)$ is convex, then
$$\partial h(x) \supseteq conv\{\partial f_\alpha(x) | \alpha \in \alpha(x)\}$$
where $\alpha(x) := \{\alpha : h(x) = f_\alpha(x)\}$.
4. **Taking superposition**: If $h(x) = F(f_1(x), \ldots, f_m(x))$, where $F(y_1, \ldots, y_m)$ is non-decreasing and convex, then
$$\partial h(x) \supseteq \left\{ \sum_{i=1}^m d_i \partial f_i(x) : (d_1, \ldots, d_m) \in \partial F(y_1, \ldots, y_m) \right\}.$$

Example 1. Let $h(x) = \max_{1 \le i \le n} (a_i^T x + b_i)$, then $a_k \in \partial h(x)$ if k is some index such that $h(x) = a_k^T x + b_k$.

Example 2. Let $h(x) = \mathbb{E}[F(x, \xi)]$ be a convex function, then $g(x) = \int G(x, \xi) p(\xi) d\xi \in \partial h(x)$ if $G(x, \xi) \in \partial F(x, \xi)$ for each ξ.

Example 3. Let $h(x) = \max_{y \in C} f(x, y)$ where $f(x, y)$ is convex in x for any y and C is closed, then $\partial f(x, y_*(x)) \subset \partial h(x)$, where $y_*(x) = argmax_{y \in C} f(x, y)$.

This is because if $g \in \partial f(x, y_*(x))$, we have

$$h(z) \geq f(z, y_*(x)) \geq f(x, y_*(x)) + g^T(z - x) = h(z) + g^T(z - x).$$

3.7 Other Properties of Convex Functions

Jensen's inequality. Let f be a convex function, then

$$f(\sum_i \lambda_i x_i) \leq \sum_i \lambda_i f(x_i)$$

as long as $\lambda_i \geq 0, \forall i$ and $\sum_i \lambda_i = 1$.

Moreover, let f be a convex function and X be a random variable, then

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Example . The Kullback-Liebler distance between two distributions is nonnegative: i.e.

$$KL(p||q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \geq 0$$

where $p_i \geq 0, q_i \geq 0, \sum_i p_i = \sum_i q_i = 1$.

Proof: Let $f(x) = -\log(x)$, f is convex, so

$$-\log(\sum_i p_i x_i) \leq -\sum_i p_i \log(x_i).$$

Plugging $x_i = q_i/p_i$, this leads to

$$0 = -\log(\sum_i q_i) \leq \sum_i p_i \log(p_i/q_i).$$

■

References

[BV04] Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[BN01] Ben-Tal, A., & Nemirovski, A. (2001). *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, (Vol. 2). SIAM.