Convexity theory

Victor Kitov

Table of Contents

Convex sets

Definition 1

Set X is convex if $\forall x, y \in X$, $\forall \alpha \in (0,1)$:

$$\alpha x + (1-\alpha)y \in X$$

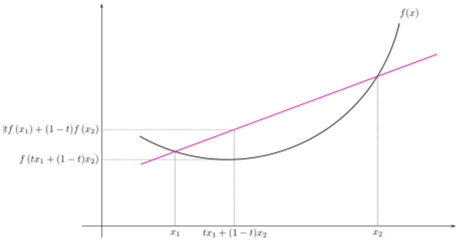
We will suppose that all functions, considered in this lecture will be defined on convex sets.

Convex functions[1]

Definition 2

Function $f(x)$ is **convex** on a set X if $\forall\alpha \in (0,1]$, $x_1 \in X$, $x_2 \in X$:

$$f\left(\alpha x_1 + (1-\alpha)\,x_2\right) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$$



[1] Using norm axioms, prove that any norm will be a convex function.

Multivariate and univariate convexity

Theorem 1

Let $f:\mathbb{R}^D\to\mathbb{R}$. $f(x)$ is convex <=> $g(\alpha)=f(x+\alpha v)$ is 1-D convex for $\forall x,v\in\mathbb{R}^D$ and $\forall\alpha\in\mathbb{R}$ such that $x+\alpha v\in dom(f)$.

=> Take $\forall x,v\in\mathbb{R}^D$ and $\forall\alpha_1,\alpha_2,\beta\in\mathbb{R}$. Using convexity of f:

$$\begin{aligned} g(\beta\alpha_1+(1-\beta)\alpha_2) &= f(x+v(\beta\alpha_1+(1-\beta)\alpha_2)) \\ &= f(\beta(x+\alpha_1 v)+(1-\beta)(x+\alpha_2 v)) \\ &\le \beta f(x+\alpha_1 v)+(1-\beta)f(x+\alpha_2 v)=\beta g(\alpha_1)+(1-\beta)g(\alpha_2) \end{aligned}$$

so $g(\alpha)$ is convex.

<= Take $\forall x,y\in dom(f)$ and $\forall\alpha\in(0,1)$. Then using convexity of $g(\alpha)=f(x+\alpha(y-x))$:

$$\underbrace{g(\alpha)}_{f((1-\alpha)x+\alpha y)} = g(0\cdot(1-\alpha)+1\cdot\alpha)\le(1-\alpha)\underbrace{g(0)}_{f(x)}+\alpha\underbrace{g(1)}_{f(y)}$$

Properties

Theorem 2

Suppose $f(x)$ is twice differentiable on $dom(f)$. Then the following properties are equivalent:

1. $f(x)$ *is convex*
2. $f(y) \geq f(x) + \nabla f(x)^T(y - x) \quad \forall x, y \in dom(f)$
3. $\nabla^2 f(x) \succeq 0 \quad \forall x \in dom(f)$

We will prove theorem 2 by proving that $1 \Leftrightarrow 2$ and $2 \Leftrightarrow 3$.

Proof 1=>2

By definition of convexity $\forall\lambda \in (0,1)$, $x, y \in dom(f)$:

$$f(\lambda y + (1-\lambda)x) \le \lambda f(y) + (1-\lambda)f(x) = \lambda(f(y) - f(x)) + f(x) \Rightarrow$$

$$f(y) - f(x) \ge \frac{f(x + \lambda(y-x)) - f(x)}{\lambda}$$

In the limit $\lambda \downarrow 0$:

$$f(y) - f(x) \ge \nabla f^T(x)(y-x)$$

Here we used Taylor's expansion

$$f(x + \lambda(y-x)) = f(x) + \nabla f(x)^T \lambda(y-x) + o(\lambda \|y-x\|)$$

Proof 2=>1

Take $\forall x, y \in dom(f)$. Apply property 2 to x, y and $z = \lambda x + (1 - \lambda)y$. We get

$$f(x) \geq f(z) + \nabla f^T(z)(x - z) \quad (1)$$

$$f(y) \geq f(z) + \nabla f^T(z)(y - z) \quad (2)$$

Multiplying 1 by λ and 2 by $(1 - \lambda)$ and adding, we get

$$\lambda f(x) + (1 - \lambda)f(y) \geq f(z) + \nabla f^T(z)(\lambda x + (1 - \lambda)y - z)$$
$$= f(z) = f(\lambda x + (1 - \lambda)y)$$

Proof 2=>3, 1 dimensional case

Take $\forall x, y \in dom(f)$, $y > x$. Following property 2, we have:

$$f(y) \geq f(x) + f'(x)(y - x)$$
$$f(x) \geq f(y) + f'(y)(x - y)$$

So

$$f'(x)(y - x) \leq f(y) - f(x) \leq f'(y)(y - x)$$

After dividing by $(y - x)^2$ we get

$$\frac{f'(y) - f'(x)}{y - x} \geq 0 \quad \forall x, y, x \neq y$$

Taking $y \to x$ we get

$$f''(x) \geq 0 \quad \forall x \in dom(f)$$

Proof 3=>2

By mean value version of Taylor theorem we get for some $z \in [x, y]$:

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)(y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x) \\ &\geq f(x) + \nabla f(x)(y-x) \end{aligned}$$

since $\nabla^2 f(z) \succcurlyeq 0 \quad \forall z$ by condition 3.

2=>3, 1 dimensional case

For any $x, y, \lambda \in [0,1]$ by Taylor expansion we get:

$$\begin{aligned} f(x+\lambda(y-x)) &= f(x)+f'(x)\lambda(y-x)+\frac{1}{2}f''(x)\lambda^2(y-x)^2+o\left(\lambda^3\right) \\ &\geq f(x)+f'(x)(y-x) \end{aligned}$$

In the limit $\lambda \to 0$ we get $f''(x) \geq 0$.

Proof 2=>3 for D-dimensional case

From theorem 1 convexity of $f(x)$ is equivalent to convexity of $g(\alpha) = f(x + \alpha v)$ $\forall x, v \in \mathbb{R}^D$ and $\alpha \in \mathbb{R}$ such that $z = x + \alpha v \in dom(f)$. From property 3 this is equivalent to

$$g''(\alpha) = v^T \nabla^2 f(x + \alpha v) v \geq 0$$

Because z and v are arbitrary, last condition is equivalent to $\nabla^2 f(x) \succcurlyeq 0$.

Optimality for convex functions

Theorem 3

Suppose convex function $f(x)$ satisfies $\nabla f(x^*) = 0$ for some x^*. Then x^* is the global minimum of $f(x)$.

Proof. Since $f(x)$ is convex, then from condition 2 of theorem 2$\forall x, y \in dom(f)$:

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

Taking $y = x^*$ we have

$$f(x) \geq f(x^*) + \nabla f^T(x^*)(x - x^*) = f(x^*)$$

Since x was arbitrary, x^* is a global minimum. □

Optimality for convex functions[3]

Comments on theorem (3):

- $\nabla f(x^*) = 0$ is necessary condition for local minimum. Together with convexity it becomes sufficient condition.
- $\nabla f(x^*) = 0$ without convexity is not sufficient for any local optimality.

Properties of minimums of convex function defined on convex set[2]:

- Set of global minimums is convex
- Local minimum is global minimum

[2] Prove them

[3] Prove that global minimums of convex function (defined on convex set) form a convex set.

Jensen's inequality

Theorem 4

For any convex function $f(x)$ and random variable X it holds that

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

Proof. For simplicity consider differentiable[4] $f(x)$. From property 2 of theorem 2 $\forall x, y \in \text{dom}(f)$:

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

By taking $x = X$ and $y = \mathbb{E}X$, obtain

$$f(X) \geq f(\mathbb{E}X) + \nabla f^T(\mathbb{E}X)(\mathbb{E}X - X)$$

After taking expectation of both sides, we get

$$\mathbb{E}f(X) \geq f(\mathbb{E}X) + \nabla f^T(\mathbb{E}X)(\mathbb{E}X - \mathbb{E}X) = f(\mathbb{E}X)$$

□

[4] for general proof consider sub-derivatives, which always exist.

Alternative proof of Jensen's inequality

- Convexity => by induction for $\forall K = 2, 3, \ldots$ and $\forall p_k \geq 0 : \sum_{k=1}^{K} p_k = 1$

$$\sum_{k=1}^{K} f(p_k x_k) \leq \sum_{k=1}^{K} p_k f(x_k) \tag{3}$$

- For r.v. X_K with $P(X_K = x_i) = p_i$ (3) becomes

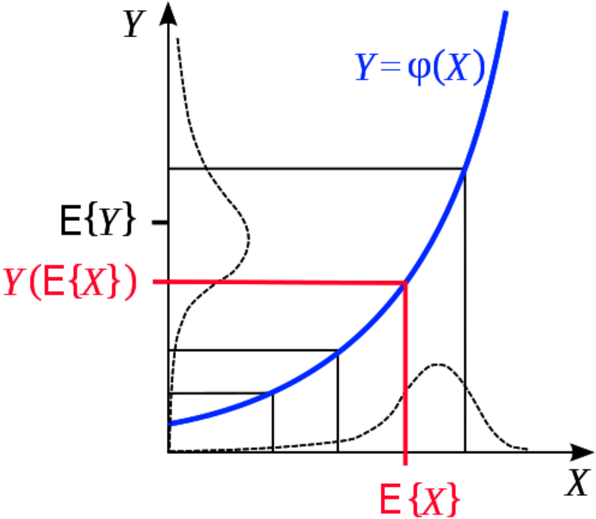
$$f(\mathbb{E}X_K) \leq \mathbb{E}f(X_K) \tag{4}$$

- For arbitrary X we may consider $X_K \uparrow X$. In the limit $K \to \infty$ (4) becomes[5]

$$f(\mathbb{E}X) \leq \mathbb{E}f(X)$$

[5] Strictly speaking you need to prove continuity of f and $\mathbb{E}$ here.

Illustration of Jensen's inequality



Generating convex functions[6]

- Any norm is convex
- If $f(\cdot)$ and $g(\cdot)$ are convex, then
  - $f(x)+g(x)$ is convex
  - $F(x)=f(g(x))$ is convex for non-decreasing $f(\cdot)$
  - $F(x)=\max\{f(x),g(x)\}$ is convex
- These properties can be extrapolated on any number of functions.
- If $f(x)$ is convex, $x\in\mathbb{R}^D$, then for all $\alpha>0$, $Q\in\mathbb{R}^{DxD}$, $Q\succcurlyeq 0$, $B\in\mathbb{R}^{KxD}$, $c\in\mathbb{R}^K$, $K=1,2,...$ the following functions are also convex:
  - $\alpha f(x)$ is convex
  - B^Tx+c
  - $x^TQx+Bx+c$,
  - $F(x)=f(Bx+c)$, for $x\in\mathbb{R}^D$,

[6] Prove these properties.

Exercises

Are the following functions convex?

- $f(x) = |x|$
- $f(x) = \|x\|_1 + \|x\|_2^2$
- $f(x) = (3x_1 - 5x_2)^2 + (4x_1 - 2x_2)^2$
- $x \ln x, -\ln x, -x^p$ for $x > 0$, $p \in (0, 1)$.
- x^p, $p > 1$.
- $\ln(1 + e^{-x})$, $[1 - x]_+$
- $F(w) = \sum_{n=1}^{N} [1 - w^T x_n]_+ + \lambda \sum_{d=1}^{D} |w_d|$
- $F(w) = \sum_{n=1}^{N} \ln(1 + e^{-w^T x_n}) + \lambda \sum_{d=1}^{D} w_d^2$

Exercises

Suppose $f(x)$ and $g(x)$ are convex. Can the following functions be non-convex?

- $f(x) - g(x)$, $f(x)g(x)$, $f(x)/g(x)$, $|f(x)|$, $f^2(x)$, $\min\{f(x), g(x)\}$

Suppose $f(x)$ is convex, $f(x) \geq 0 \ \forall x \in \text{dom}(f), k \geq 1$. Can $g(x) = f^k(x)$ be non-convex?

Table of Contents

Strictly convex functions[7]

Definition 3

Function $f(x)$ is **strictly convex** on a set X if
$\forall\alpha\in(0,1]$, $x_1,x_2\in X$, $x_1\neq x_2$:

$$f\left(\alpha x_1+(1-\alpha)\,x_2\right)<\alpha f(x_1)+(1-\alpha)f(x_2)$$

[7] Prove that global minimum of strictly convex function defined on convex set is unique.

Criterion for strict convexity

Theorem 5

*Function $f(x)$ is **strictly convex** <=> $\forall x, y \in \mathrm{dom}(f)$, $x \neq y$:*

$$f(y) > f(x) + \nabla f(x)^T (y - x) \quad (5)$$

<= The same as proof 2=>1 for theorem 2 with replacement $\geq \to >$.

Criterion for strict convexity

=> Using property 2 of theorem 2 we have

$$\forall x, z: \quad f(z) \geq f(x) + \nabla f(x)^T (z - x) \tag{6}$$

Suppose (5) does not hold, so
$\exists y: f(y) = f(x) + \nabla f(x)^T (y - x)$. It follows that

$$\nabla f(x)^T (y - x) = f(y) - f(x) \tag{7}$$

Consider $u = \alpha x + (1 - \alpha) y$ for $\forall \alpha \in (0, 1)$. Using (6) and (7):

$$\begin{aligned} f(u) = f(\alpha x + (1 - \alpha) y) \geq f(x) + \nabla f(x)^T (u - x) \\ = f(x) + \nabla f(x)^T (\alpha x + (1 - \alpha) y - x) \\ = f(x) + \nabla f(x)^T (1 - \alpha)(y - x) \\ = f(x) + (1 - \alpha)(f(y) - f(x)) = (1 - \alpha) f(y) + \alpha f(x) \end{aligned}$$

- Obtained inequality $f(\alpha x + (1 - \alpha) y) \geq (1 - \alpha) f(y) + \alpha f(x)$ contradicts strict convexity. So (6) should hold as strict inequality (5).

Jensen's inequality

Theorem 6

*For strictly convex function $f(x)$ **equality** in Jensen's inequality*

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

holds <=> $X = \mathbb{E}X$ with probability 1.

Proof. 1) Consider $X \neq \mathbb{E}X$ with probability 1:
From theorem (5) $\forall x \neq y \in \mathrm{dom}(f)$:

$$f(x) > f(y) + \nabla f^T(y)(y - x)$$

By taking $x = X$ and $y = \mathbb{E}X$, obtain

$$f(X) > f(\mathbb{E}X) + \nabla f^T(\mathbb{E}X)(\mathbb{E}X - X)$$

After taking expectation of both sides, we get

$$\mathbb{E}f(X) > f(\mathbb{E}X) + \nabla f^T(\mathbb{E}X)(\mathbb{E}X - \mathbb{E}X) = f(\mathbb{E}X)$$

Jensen's inequality

2) Consider case $X = \mathbb{E}X$ with probability 1.
In this case with probability 1

$$f(X) = f(\mathbb{E}X)$$

which after taking expectation becomes

$$\mathbb{E}f(X) = \mathbb{E}f(\mathbb{E}X) = f(\mathbb{E}X)$$

Properties of strictly convex functions[10]

Properties of minimums of strictly convex function defined on convex set[8]:

- Global minimum is unique.
- If $\nabla^2 f(x) \succ 0 \, \forall x \in \text{dom}(f)$, then $f(x)$ is strictly convex
    - proof: use mean value version of Taylor theorem and strict convexity criterion (5).
    - strict convexity does not imply $\nabla^2 f(x) \succ 0 \, \forall x \in \text{dom}(f)$[9]

[8] Prove them

[9] Think of an example.

[10] Prove that global minimums of convex function (defined on convex set) form a convex set.

Table of Contents

Concave functions

Definition 4

Function $f(x)$ is **concave** on a set X if
$\forall\alpha\in(0,1]$, $x_1\in X$, $x_2\in X$:

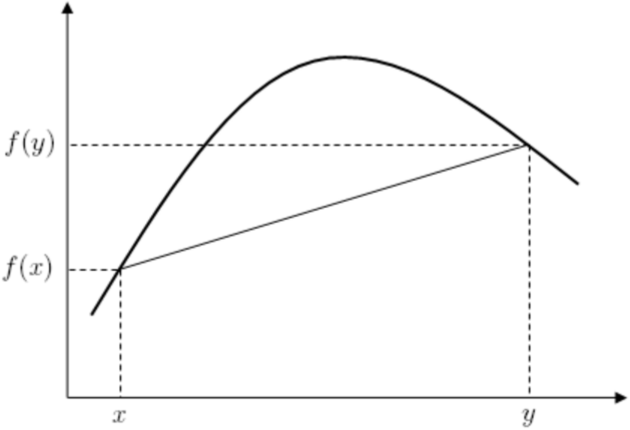
$$f\left(\alpha x_1+(1-\alpha)\,x_2\right)\geq\alpha f(x_1)+(1-\alpha)f(x_2)$$

Definition 5

Function $f(x)$ is **strictly concave** on a set X if
$\forall\alpha\in(0,1]$, $x_1,x_2\in X$, $x_1\neq x_2$:

$$f\left(\alpha x_1+(1-\alpha)\,x_2\right)>\alpha f(x_1)+(1-\alpha)f(x_2)$$

Concave function example



Properties of concave functions

- $f(x)$ is convex $\iff -f(x)$ is concave
- Differentiable function $f(x)$ is **concave** $<=> \forall x, y \in \text{dom}(f)$:

$$f(y) \le f(x) + \nabla f(x)^T (y - x)$$

- Twice differentiable function $f(x)$ is **concave** $<=> \forall x \in \text{dom}(f)$:$\nabla^2 f(x) \succcurlyeq 0$
- Global maximums of concave function on convex set form a convex set.
- Local maximum of a concave function is global
- $\nabla f(x^*) = 0 <=> x^*$ is global maximum.
- Jensen's inequality: for random variable X and concave $f(x)$:

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

  - equality is achieved $<=>$ f is linear on $\{x : P(X = x) > 0\}$.
    - this holds when $X = \mathbb{E}X$ with probability 1.

Properties of strictly concave functions

- $f(x)$ is strictly convex $\iff -f(x)$ is strictly concave
- Differentiable function $f(x)$ is **concave** $<=>\forall x, y \in \text{dom}(f),\ x \neq y$:
$$f(y) < f(x) + \nabla f(x)^T (y - x)$$
- $\forall x \in \text{dom}(f)$:$\nabla^2 f(x) \succ 0 => f(x)$ is strictly concave.
- Global maximum of strictly concave function on a convex set is unique.
- Jensen's inequality: for random variable X, and strictly concave $f(x)$:
$$\mathbb{E}[f(X)] < f(\mathbb{E}X)$$
when $X \neq \mathbb{E}X$ with some probability>0.
- When $X = \mathbb{E}X$ with probability 1 $\mathbb{E}[f(X)] = f(\mathbb{E}X)$

Table of Contents

Kullback-Leibler divergence

- Kullback-Leibler divergence between 2 probability discrete distributions[11]

$$KL(P||Q) := \sum_i P_i \ln \frac{P_i}{Q_i}$$

- Kullback-Leibler divergence between 2 probability density functions[12]:

$$KL(P||Q) := \int P(x) \ln \frac{P(x)}{Q(x)} dx$$

[11] Suppose $P(i,j) = P_1(i)P_2(j)$ and $Q(i,j) = Q_1(i)Q_2(j)$. Show that $KL(P||Q) = KL(P_1||Q_1) + KL(P_2||Q_2)$

[12] Show that KL diveregence is invariant to reparamtrization $x \to y(x)$

Kullback-Leibler divergence

- Properties of $KL(P||Q)$:
  - defined only for distributions P, Q such that $P_i = 0 \Rightarrow Q_i = 0$
  - $KL(P||Q) \neq KL(Q||P)$
  - symmetrical version:
    $KL_{sym}(P||Q) := \frac{1}{2}\left(KL(P||Q) + KL(Q||P)\right)$
  - $KL(P||Q) \geq 0 \; \forall P, Q$

Non-negativity of KL

- $KL(P||Q) \geq 0 \, \forall P, Q$
  Proof: Consider r.v. U such that $P(U_i = \frac{Q_i}{P_i}) = P_i$

$$KL(P||Q) = \sum_i P_i \ln \frac{P_i}{Q_i} = \sum_i P_i \left(-\ln \frac{Q_i}{P_i} \right) = \mathbb{E}\left(-\ln U\right)$$
$$\geq \{\text{convexity of } -\ln(\cdot) + \text{Yensen's inequality}\}$$
$$\geq -\ln \mathbb{E} U = -\ln \sum_i P_i \frac{Q_i}{P_i} = -\ln \sum_i Q_i = -\ln 1 = 0$$

- $KL(P||Q) = 0$ is achieved $\Leftrightarrow P_i = Q_i \, \forall i$.
  Proof: $KL(P||Q) = 0 \Leftrightarrow U \equiv const = c$ with probability 1 which gives

$$\frac{P_i}{Q_i} = c \Leftrightarrow P_i = cQ_i \quad \forall i \tag{8}$$

Summing (8) by i we obtain $1 = c$, so $P_i = Q_i \, \forall i$.

Connection of KL and maximum likelihood

Consider discrete r.v. $\xi \in \{1, 2, ...K\}$. Suppose we estimate probabilities $p(\xi = i)$ with distribution $q_\theta(i)$ parametrized by θ. We observe N independent trials of ξ, $\#[\xi = i] = N_i$. Maximum likelihood estimate for θ gives:

$$\widehat{\theta} = \arg\max_\theta \prod_{n=1}^{N} \prod_{i=1}^{K} q_\theta(i)^{\mathbb{I}[\xi_n = i]} = \arg\max_\theta \prod_{i=1}^{K} q_\theta(i)^{N_i}$$

$$= \arg\max_\theta \left(\prod_{i=1}^{K} q_\theta(i)^{N_i} \right)^{1/N} = \arg\max_\theta \prod_{i=1}^{K} q_\theta(i)^{N_i/N}$$

$$= \arg\max_\theta \sum_{i=1}^{K} \frac{N_i}{N} \ln q_\theta(i) = \{p(i) := \frac{N_i}{N}, \sum_{i=1}^{K} p(i) \ln p(i) = const(\theta)\}$$

$$= \arg\min_\theta \left\{ \sum_{i=1}^{K} p(i) \ln p(i) - \sum_{i=1}^{K} p(i) \ln q_\theta(i) \right\} = \arg\min_\theta KL(p||q(\theta))$$