Probability and Stochastic Processes I - Lecture 19

Michael Evans
University of Toronto
http://www.utstat.utoronto.ca/mikevans/stac62/STAC622023.html

2023

Jensen's Inequality

Definition III.7.2 A set $C \subset R^k$ is *convex* if whenever $\mathbf{x}_1, \mathbf{x}_2 \in C$ and $\alpha \in [0,1]$, then $\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in C$. A function $f : C \to R^1$ is *convex* if C is convex and for every $\alpha \in [0,1]$,then

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2)$$

and f is *concave* if $f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \geq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2)$. ■

- $L(\mathbf{x}_1, \mathbf{x}_2) = \{\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 : \alpha \in [0,1]\}$ is the *line segment* joining $\mathbf{x}_1$ and $\mathbf{x}_2$

- if $f : C \to R^1$ is convex then $-f$ is concave and conversely

- **fact**: if $f : C \to R^1$ is defined on open convex $C \subset R^k$, then f is convex whenever the Hessian matrix

$$\left(\frac{\partial^2 f(x_1, \ldots, x_k)}{\partial x_i \partial x_j}\right) \in R^{k \times k}$$

is positive semidefinite for every $\mathbf{x} \in C$

Exercise III.7.5 (i) Prove the line segment $L(\mathbf{x}_1, \mathbf{x}_2)$ is convex.
(ii) Prove $[\mathbf{a}, \mathbf{b}] \subset R^k$ is convex. What about $(\mathbf{a}, \mathbf{b}], (\mathbf{a}, \mathbf{b}), [\mathbf{a}, \mathbf{b})$?
(iii) Prove $B_r(\boldsymbol{\mu}) \subset R^k$ is convex.
(iv) Prove $E_r(\boldsymbol{\mu}, \Sigma)$ is convex (hint: use $E_r(\boldsymbol{\mu}, \Sigma) = \boldsymbol{\mu} + \Sigma^{1/2} B_r(\mathbf{0})$.
(v) Prove that the affine function $f : R^k \to R^1$ given by $f(\mathbf{x}) = a + \mathbf{c}'\mathbf{x}$ for constants $a \in R^1, \mathbf{c} \in R^k$ is convex on R^k.
(vi) Prove that $f(x) = -\log x$ is convex on $C = (0, \infty)$.
(vii) If $\Sigma \in R^{k \times k}$ is positive semidefinite, then prove $f(x) = \mathbf{x}'\Sigma\mathbf{x}$ is convex on R^k.

Example III.7.2 - suppose $P_{\mathbf{X}}(\{\mathbf{x}_1, \mathbf{x}_2\}) = 1$ with $P_{\mathbf{X}}(\{\mathbf{x}_1\}) = \alpha_1$, $P_{\mathbf{X}}(\{\mathbf{x}_2\}) = 1 - \alpha_1$

- then $L(\mathbf{x}_1, \mathbf{x}_2)$ is convex and note $P_{\mathbf{X}}(L(\mathbf{x}_1, \mathbf{x}_2)) = 1$ $(L(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{B}^k)$

- suppose $f : L(\mathbf{x}_1, \mathbf{x}_2) \to R^1$ is convex

- then for this simple context Jensen's inequality is immediate

$$E(f(\mathbf{X})) = \alpha_1 f(\mathbf{x}_1) + (1 - \alpha_1) f(\mathbf{x}_2) \geq f(\alpha_1 \mathbf{x}_1 + (1 - \alpha_1)\mathbf{x}_2) = f(E(\mathbf{X}))$$

- geometrically consider the line segment

$$\{\alpha(\mathbf{x}_1, f(\mathbf{x}_1)) + (1-\alpha)(\mathbf{x}_2, f(\mathbf{x}_2)) : \alpha \in [0,1]\}$$

in R^{k+1} and convexity of f on the line segment implies the line segment lies above the graph

$$\{(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2, f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2)) : \alpha \in [0,1]\}$$

and $E(\mathbf{X}) = \alpha_1\mathbf{x}_1 + (1-\alpha_1)\mathbf{x}_2$ gives $E(f(\mathbf{X})) \geq f(E(\mathbf{X}))$ ■

Exercise III.7.6 Suppose $C_1, C_2 \subset R^k$ are convex. Prove that $C_1 \cap C_2$ is convex.

Exercise III.7.7 Suppose $C \subset R^k$ is convex and let $C_* = \mathbf{a} + BC = \{\mathbf{y} = \mathbf{a} + B\mathbf{x} : \mathbf{x} \in C\}$. Prove that C_* is convex.

Exercise III.7.8 If C is a linear subspace of R^k, then C is convex.

Proposition III.7.5 *(Supporting Hyperplane Theorem)* If $C \subset R^k$ is convex and $\mathbf{x}_0 \in R^k$ is not an interior point of C (there isn't a ball $B_r(\mathbf{x}_0) \subset C$ with $r > 0$), then there exists $\mathbf{c} \in R^k \backslash \{\mathbf{0}\}$ such that $\mathbf{c}'\mathbf{x} \geq \mathbf{c}'\mathbf{x}_0$ for every $\mathbf{x} \in C$.
Proof: See a text on convex analysis.

- for a set $A \subset R^k$ it is always possible to find a set of the form $\{\mathbf{x} \in R^k : \mathbf{a} + B\mathbf{x} = \mathbf{0}\}$ for some $\mathbf{a} \in R^l, B \in R^{l \times k}$ for some $l \leq k$ s.t. $A \subset \{\mathbf{x} \in R^k : \mathbf{a} + B\mathbf{x} = \mathbf{0}\}$

- e.g., take $\mathbf{a} = \mathbf{0} \in R^k, B = 0 \in R^{1 \times k}$ so $\{\mathbf{x} : \mathbf{a} + B\mathbf{x} = \mathbf{0}\} = R^k$

- the set $\{\mathbf{x} \in R^k : \mathbf{a} + B\mathbf{x} = \mathbf{0}\}$ is called an *affine subset* of R^k and it has a dimension (point has dimension 0, line has dimension 1, ..., hyperplane has dimension $k - 1, R^k$ has dimension k)

Definition III.7.3 If $A \subset R^k$ the *affine dimension of* A is the smallest dimension of an affine set containing A.

Proposition III.7.7 If $C \subset R^k$ is convex, $P_{\mathbf{X}}(C) = 1$ and $E(\mathbf{X}) \in R^k$, then $E(\mathbf{X}) \in C$.

Proof: (Induction on the affine dimension of C.)

If the affine dim of C is 0, then $C = \{\mathbf{x}\}$ and $E(\mathbf{X}) = \mathbf{x} \in C$ and the result holds.

Assume wlog that $E(\mathbf{X}) = \mathbf{0}$, else put $\mathbf{Y} = \mathbf{X} - E(\mathbf{X})$, $C_* = C - E(\mathbf{X})$ is convex (Exercise III.7.7) and note

$$P_{\mathbf{Y}}(C_*) = P(\mathbf{Y} \in C_*) = P(\mathbf{X} \in C) = P_{\mathbf{X}}(C) = 1$$

and $E(\mathbf{X}) \in C$ iff $E(\mathbf{Y}) = \mathbf{0} \in C_*$.

Now assume the result holds for affine dim $C < k$.

Suppose $\mathbf{0} \notin C$, then the SHT gives $\mathbf{c} \in R^k \backslash \{\mathbf{0}\}$ s.t. $\mathbf{c}'\mathbf{x} \geq \mathbf{c}'\mathbf{0} = 0$ for every $\mathbf{x} \in C$. This implies $P(\mathbf{c}'\mathbf{X} \geq 0) = 1$ (so $\mathbf{c}'\mathbf{X}$ is a nonnegative r.v.) and since $E(\mathbf{c}'\mathbf{X}) = \mathbf{c}'E(\mathbf{X}) = 0$ then $P(\mathbf{c}'\mathbf{X} = 0) = 1$. Therefore, $P(\mathbf{X} \in \{\mathbf{x} : \mathbf{c}'\mathbf{x} = 0\} \cap C) = 1$ and $\{\mathbf{x} : \mathbf{c}'\mathbf{x} = 0\} \cap C$ is a convex set (Exercises III.7.8 and III.7.6) having affine dimension no greater than $k - 1$. So by the inductive hypothesis $\mathbf{0} \in \{\mathbf{x} : \mathbf{c}'\mathbf{x} = 0\} \cap C$ which implies $\mathbf{0} \in C$ which is a contradiction. This implies $E(\mathbf{X}) = \mathbf{0} \in C$. ■

Proposition III.7.8 *(Jensen's Inequality)* If $C \subset R^k$ is convex, $P_{\mathbf{X}}(C) = 1, E(\mathbf{X}) \in R^k$, and $f : C \to R^1$ is convex, then

$$E(f(\mathbf{X})) \geq f(E(\mathbf{X})).$$

Equality is obtained iff $f(\mathbf{x}) \overset{wp1}{=} a + \mathbf{b}'\mathbf{x}$ for constants $a, \mathbf{b}$.

Proof: (Induction on the affine dimension of C.)

If affine dim C is 0, then $C = \{\mathbf{x}\}$ and $E(f(\mathbf{X})) = f(\mathbf{x}) = f(E(\mathbf{X}))$ and $f(\mathbf{x}) \overset{wp1}{=} f(\mathbf{x}) + \mathbf{0}'\mathbf{x}$ so the result holds.

Now assume the result holds for affine dim $C < k$. Let

$$S = \{(\mathbf{x}, y) : \mathbf{x} \in C, y \geq f(\mathbf{x})\},$$

note that $S \subset R^{k+1}$ is convex (**Exercise III.7.9)** and $(E(\mathbf{X}), f(E(\mathbf{X})))$ is a boundary point of S (not an interior point). Then by SHT there exists $\mathbf{c} \in R^{k+1} \backslash \{\mathbf{0}\}$ s.t. for every $\mathbf{z} \in S$

$$\mathbf{c}'\mathbf{z} = \sum_{i=1}^{k} c_i z_i + c_{k+1} z_{k+1} \geq \mathbf{c}' \begin{pmatrix} E(\mathbf{X}) \\ f(E(\mathbf{X})) \end{pmatrix} = \sum_{i=1}^{k} c_i E(X_i) + c_{k+1} f(E(\mathbf{X})).$$

If $c_{k+1} < 0$, then the inequality can be violated by taking z_{k+1} large so $c_{k+1} \geq 0$.

Case 1: $c_{k+1} > 0$

Let

$$Y = \sum_{i=1}^{k} c_i(X_i - E(X_i)) + c_{k+1}(f(\mathbf{X}) - f(E(\mathbf{X}))$$

and note that $P(Y \geq 0) = 1$ so $0 \leq E(Y) = c_{k+1}(E(f(\mathbf{X})) - f(E(\mathbf{X}))$ which implies $E(f(\mathbf{X})) \geq f(E(\mathbf{X}))$. Also $E(f(\mathbf{X})) = f(E(\mathbf{X})$ iff $E(Y) = 0$ which occurs iff $P(Y = 0) = 1$ and so

$$\begin{aligned} f(\mathbf{X}) &= f(E(\mathbf{X})) - \sum_{i=1}^{k} \frac{c_i}{c_{k+1}}(X_i - E(X_i)) \\ &= \left(f(E(\mathbf{X})) + \sum_{i=1}^{k} \frac{c_i}{c_{k+1}} E(X_i) \right) + \sum_{i=1}^{k} \left(-\frac{c_i}{c_{k+1}} \right) X_i \end{aligned}$$

which is of the required form.

Case 2: $c_{k+1} = 0$

Then $Y = \sum_{i=1}^{k} c_i(X_i - E(X_i))$ and since $P(Y \geq 0) = 1$ with $E(Y) = 0$, this implies $P(Y = 0) = 1$ which in turn implies

$$P(\mathbf{X} \in \{\mathbf{x} : \mathbf{c}'\mathbf{x} = \mathbf{c}'E(\mathbf{X})\} \cap C) = 1$$

and $\{\mathbf{x} : \mathbf{c}'\mathbf{x} = \mathbf{c}'E(\mathbf{X})\} \cap C$ is a convex set of affine dim $< k$ and so by the inductive hypothesis the result holds. ■

- $f : C \to R^1$ is concave and $P_{\mathbf{X}}(C) = 1, E(\mathbf{X}) \in R^k$ then the concave version of Jensen says $E(f(\mathbf{X})) \leq f(E(\mathbf{X}))$

Definition III.7.4 Suppose P, Q are probability measures on $(\Omega, \mathcal{A})$ with probability (density) functions p and q respectively. The *Kullback-Liebler distance* between P and Q is defined to be

$$KL(P \,||\, Q) = E_P\left(-\log\frac{q}{p}\right) = -\int_\Omega p(\omega)\log\frac{q(\omega)}{p(\omega)}\,\nu(d\omega)$$

when $E_P\left(-\log q/p\right)$ exists, where ν is counting (discrete case) or volume measure (abs. cont. case).

- $KL(P \,||\, Q)$ serves as a distance measure between probability measures P and Q

Proposition III.7.9 When $E_P\left(-\log q/p\right)$ exists then $KL(P \,||\, Q) \geq 0$ with equality iff $P = Q$.

Proof: Since $-\log x$ is convex on $(0, \infty)$ (Exercise III.7.5(vi)), applying Jensen gives

$$\begin{aligned} KL(P \,||\, Q) &\geq -\log\left(E_P\left(\frac{q}{p}\right)\right) = -\log\left(\int_\Omega p(\omega)\frac{q(\omega)}{p(\omega)}\,\nu(d\omega)\right) \\ &= -\log\left(\int_\Omega q(\omega)\,\nu(d\omega)\right) = -\log 1 = 0. \end{aligned}$$

Equality holds iff there exist a, b such that for every ω,

$$-\log \frac{q(\omega)}{p(\omega)} \overset{wp1}{=} a + b\frac{q(\omega)}{p(\omega)}. \qquad (*)$$

Then (*) holds when $p \overset{wp1}{=} q$, and so $P = Q$, with $a = b = 0$. Otherwise. (*) implies $a = -b$ since $KL(P \,||\, Q) = 0$ implies $0 = a + b$ by taking the expectation of both sides of (*) wrt P. This implies

$$-\log \frac{q(\omega)}{p(\omega)} \overset{wp1}{=} a\left(1 - \frac{q(\omega)}{p(\omega)}\right).$$

Now $-\log x$ and $a(1-x)$ agree at $x = 1$ and at most at one other point (draw the graphs). Let $A = \{\omega : q(\omega) = p(\omega)\}$. If $P(A) = 1$ then $P = Q$. If $P(A) < 1$, then on A^c we have $q(\omega) = rp(\omega)$ for some real number r. This implies $Q(A) = P(A), Q(A^c) = rP(A^c) = rQ(A^c)$ which implies $r = 1$ and $p \overset{wp1}{=} q$. ■

Exercise III.7.10 Suppose P is the $N(\mu_1, \sigma_1^2)$ probability measure and Q is the $N(\mu_2, \sigma_2^2)$ probability measure. Compute $KL(P \,||\, Q)$.

Exercise III.7.11 Does $KL(P \,||\, Q) = KL(Q \,||\, P)$?