Optional notes on general proof for Jensen's inequality

Kevin H. Huang*

October 30, 2020

Clarifications. This is not the most general proof as we only consider real-valued random variable X here. For the most general form with X taking any values in a general topological space and with expectation allowed to be conditional expectations on any sub-sigma-algebra, one may refer to Wikipedia (`https://en.wikipedia.org/wiki/Jensen%27s_inequality#General_inequality_in_a_probabilistic_setting`), which requires knowledge on topology and measure theory.

Credits: This is adapted from James Norris's Year 1 Probability Notes for the Mathematics course at Cambridge (`http://www.statslab.cam.ac.uk/~james/Lectures/p.pdf`) by filling in some details for readers with a less mathematical background. We also extend the argument to $\mathbb{R}^d$ which is much cleaner to remember if one is familiar with convexity and subdifferentials. I am grateful to Yudong Chen for his corrections.

1 1D case

We start with some definitions and lemmas (which turn out to be alternative definitions) to set things up. We will see that Jensen's is straightforward once these are established.

Definition. (Integrability) A real-valued random variable X is integrable if $\mathbb{E}(|X|) < \infty$.

Definition. (Convexity in 1D) A function $f : I \to \mathbb{R}$ defined on a convex set $I \subset \mathbb{R}$ is convex if, for any $x, y \in I$ and $t \in [0, 1]$,

$$f\big(tx + (1-t)y\big) \leq tf(x) + (1-t)y.$$

Lemma 1. (Useful property (in fact alternative definition) of convexity) For a convex function $f : I \to \mathbb{R}$ defined on $I \subset \mathbb{R}$, given any $x, m, y \in I$ with $x < m < y$, we have

$$\frac{f(m) - f(x)}{m - x} \leq \frac{f(y) - f(m)}{y - m}.$$

Remark. *Intuitively this says the "local gradient" of the function is increasing or the "local second derivative" is non-negative.*

Proof. $x < m < y$ means there exists $t \in (0, 1)$ such that $m = tx + (1-t)y$, so expressing t in terms of

*PhD student, Gatsby Unit, UCL

m and applying convexity,

$$\begin{aligned}
f(m) &\le \frac{y-m}{y-x}f(x) + \frac{m-x}{y-x}f(y) \\
\frac{y-m}{y-x}f(m) + \frac{m-x}{y-x}f(m) &\le \frac{y-m}{y-x}f(x) + \frac{m-x}{y-x}f(y) \qquad \text{since } \frac{y-m}{y-x} + \frac{m-x}{y-x} = 1 \\
\frac{y-m}{y-x}\big(f(m) - f(x)\big) &\le \frac{m-x}{y-x}\big(f(y) - f(m)\big)
\end{aligned}$$

Scaling both sides by $\frac{y-x}{(y-m)(m-x)}$ gives the result. □

Lemma 2. (Another useful property (in fact alternative definition) of convexity) For a convex function $f : I \to \mathbb{R}$ defined on **an open interval** $I \subset R$, given any $m \in I$, there exists $a, b \in \mathbb{R}$ such that

1. $am + b = f(m)$,
2. $az + b \le f(z)$ for any $z \in I$.

Remark. *Intuitively this says that a function has a "supporting hyperplane" at any point $m \in I$. In fact in our extension to the n-dimensional case, and in the most general case on an abstract topology, this idea is characterised by replacing a with any subgradient of f.*

Proof. By Lemma 1, for any $x, y \in I$ with $x < m < y$ (which exist because I is open), we have

$$\frac{f(m) - f(x)}{m - x} \le \frac{f(y) - f(m)}{y - m},$$

therefore

$$\sup_{x \in I, x \le m} \frac{f(m) - f(x)}{m - x} \le \inf_{y \in I, y \ge m} \frac{f(y) - f(m)}{y - m}.$$

So there exists $a \in \mathbb{R}$ such that

$$\sup_{x \in I, x \le m} \frac{f(m) - f(x)}{m - x} \le a \le \inf_{y \in I, y \ge m} \frac{f(y) - f(m)}{y - m},$$

i.e. there exists $a \in \mathbb{R}$ independent of the choice of x and y such that

$$\frac{f(m) - f(x)}{m - x} \le a \le \frac{f(y) - f(m)}{y - m},$$

for all $x, y \in I$ with $x \le m \le y$. Rearranging we get $a(z - m) + f(m) \le f(z)$ for any $z \in I$ *(for $z \le m$ set $z = x$ above, for $z > m$ set $z = y$ above)*. Defining $b = f(m) - am$ gives us statement 1 and 2 in the lemma. □

Remark. *We have established that the convexity definition $\Rightarrow$ statement in Lemma 1 $\Rightarrow$ statements in Lemma 2. The second statement in Lemma 2 in fact imply the inequality in definition of convexity. One may see this by taking $m = tx + (1-t)y$ for $t \in (0, 1)$, noting $b = f(m) - am$ to get $a(x - m) + f(m) \le f(x)$ and $a(y - m) + f(m) \le f(y)$, and finally taking a convex combination. For $t = 0$ and $t = 1$ it is trivial. Therefore, the statements in both lemmas are in fact alternative definitions of convexity.*

Theorem 3. (Jensen's inequality, 1D) For an integrable random variable X taking values in an open interval $I \subset \mathbb{R}$, and a convex function $f : I \to \mathbb{R}$, we have

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

Proof. Take $m = \mathbb{E}(X)$ in Lemma 2, we have

$$f(\mathbb{E}(X)) \overset{\text{by 1 of Lemma 2}}{=} a\mathbb{E}(X) + b \overset{\text{by linearity of } \mathbb{E}}{=} \mathbb{E}(aX + b) \overset{\text{by 2 of Lemma 2}}{\leq} \mathbb{E}(f(X)).$$

□

Remark. *This is pretty trivial from Lemma 2, which we have seen is an alternative definition for convexity. In the case of strict convexity, one may show that statement 2 in Lemma 2 becomes strict inequality for all $z \neq m$. The implication on Jensen's is that, for strictly convex function f, we have strict equality if and only if $X = \mathbb{E}(X)$ $\mathbb{P}$-almost surely, i.e. X is constant with probability 1.*

2 n-dimensional case

Theorem 4. (Jensen's Inequality) For an integrable random variable X taking values in an open interval $I \subset \mathbb{R}^d$, and a convex function $f : I \to \mathbb{R}$, we have

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

Proof. I is open so $\mathbb{E}(X)$ (in fact any point in I) is in the interior of I, which implies subdifferential of f on that point is nonempty. Let g be a subgradient of f at $\mathbb{E}(X)$ then

$$\begin{aligned}
f(\mathbb{E}(X)) &= g^\top \mathbb{E}(X) + f(\mathbb{E}(X)) - g^\top \mathbb{E}(X) \\
&= \mathbb{E}\Big(g^\top X + f(\mathbb{E}(X)) - g^\top \mathbb{E}(X)\Big) \\
&\leq \mathbb{E}\Big(g^\top X + f(X) - g^\top X\Big) \\
&\qquad \text{by definition of a subgradient at } \mathbb{E}(X) \text{ applied to the last two terms with respect to another point } X \\
&= \mathbb{E}(f(X))
\end{aligned}$$

□

Remark. *The similarity to the proof for 1D case is clear by matching a with g and b with $f(\mathbb{E}(X)) - g^\top \mathbb{E}(X)$.*