THE AMERICAN STATISTICIAN
2018, VOL. 0, NO. 0, 1–4: Teacher's Corner
https://doi.org/10.1080/00031305.2017.1419145

Sharpening Jensen's Inequality

J. G. Liao and Arthur Berg

Division of Biostatistics and Bioinformatics, Penn State University College of Medicine

ABSTRACT
This article proposes a new sharpened version of Jensen's inequality. The proposed new bound is simple and insightful, is broadly applicable by imposing minimum assumptions, and provides fairly accurate results in spite of its simple form. Applications to the moment generating function, power mean inequalities, and Rao-Blackwell estimation are presented. This presentation can be incorporated in any calculus-based statistical course.

ARTICLE HISTORY
Received July 2017
Accepted December 2017

KEYWORDS
Jensen gap; Power mean inequality; Rao-Blackwell estimator; Taylor series

1. Introduction

Jensen's inequality is a fundamental inequality in mathematics and it underlies many important statistical proofs and concepts. Some standard applications include derivation of the arithmetic-geometric mean inequality, non-negativity of Kullback-Leibler divergence, and the convergence property of the expectation-maximization algorithm (Dempster, Laird, and Rubin 1977). Jensen's inequality is covered in all major mathematical statistics textbooks for advanced undergraduate/beginning graduate students such as Casella and Berger (2002, sect. 4.7) and Wasserman (2013, sect. 4.2) as a basic mathematical tool for statistics.

Let X be a random variable with finite expectation and let $\varphi(x)$ be a convex function, then Jensen's inequality (Jensen 1906) establishes

$$\mathbb{E}\left[\varphi\left(X\right)\right]-\varphi\left(\mathbb{E}\left[X\right]\right)\geq 0. \tag{1}$$

This inequality, however, is not sharp unless $\text{var}(X) = 0$ or $\varphi(x)$ is a linear function of x. Therefore, there is substantial room for advancement. This article proposes a new sharper bound for the Jensen gap $\mathbb{E}[\varphi(X)] - \varphi(\mathbb{E}[X])$. Some other improvements of Jensen's inequality have been developed recently; see, for example, Walker (2014), Abramovich and Persson (2016); Horvath Khan, and Pecaric (2014) and references cited therein. Our proposed bound, however, has the following advantages. First, it has a simple, easy to use, and insightful form in terms of the second derivative $\varphi''(x)$ and $\text{var}(X)$. At the same time, it gives fairly accurate results in the several examples below. Many previously published improvements are much more complicated in form, much more involved to use, and can even be more difficult to compute than $E[\varphi(X)]$ itself as discussed in Walker (2014). Second, our method requires only the existence of $\varphi''(x)$ and is therefore broadly applicable. In contrast, some other methods require $\varphi(x)$ to admit a power series representation with positive coefficients (Abramovich and Persson 2016; Dragomir 2014; Walker 2014) or require $\varphi(x)$ to be super-quadratic (Abramovich, Persson, and Samko 2014). Third, we provide both a lower bound and an upper bound in a single formula.

2. Main Result

Theorem 1. Let X be a one-dimensional random variable with mean μ, and $P(X \in (a, b)) = 1$, where $-\infty \leq a < b \leq \infty$. Let $\varphi(x)$ be a twice differentiable function on (a, b), and define the function

$$h(x;\nu)\triangleq\frac{\varphi\left(x\right)-\varphi\left(\nu\right)}{\left(x-\nu\right)^{2}}-\frac{\varphi'\left(\nu\right)}{x-\nu}.$$

Then

$$\inf_{x\in(a,b)}\{h(x;\mu)\}\text{var}(X)\leq E\left[\varphi\left(X\right)\right]-\varphi\left(E[X]\right)\leq\sup_{x\in(a,b)}\{h(x;\mu)\}\text{var}(X). \tag{2}$$

Proof. Let $F(x)$ be the cumulative distribution function of X. Applying Taylor's theorem to $\varphi(x)$ about μ with a mean-value form of the remainder gives

$$\varphi(x)=\varphi(\mu)+\varphi'(\mu)(x-\mu)+\frac{\varphi''(g(x))}{2}(x-\mu)^{2},$$

where $g(x)$ is between x and μ. Explicitly solving for $\varphi''(g(x))/2$ gives $\varphi''(g(x))/2 = h(x; \mu)$ as defined above. Therefore

$$\begin{aligned}\mathbb{E}\left[\varphi\left(X\right)\right]-\varphi\left(\mathbb{E}\left[X\right]\right)&=\int_{a}^{b}\{\varphi(x)-\varphi(\mu)\}\,dF(x)\\&=\int_{a}^{b}\left\{\varphi'(\mu)(x-\mu)+h(x;\mu)(x-\mu)^{2}\right\}dF(x)\\&=\int_{a}^{b}h(x;\mu)(x-\mu)^{2}\,dF(x),\end{aligned}$$

and the result follows because $\inf_{x\in(a,b)} h(x; \mu) \leq h(x; \mu) \leq \sup_{x\in(a,b)} h(x; \mu)$. □

CONTACT J. G. Liao jg_liao@hotmail.com Division of Biostatistics and Bioinformatics, Penn State University College of Medicine.

Theorem 1 also holds when $\inf h(x;\mu)$ is replaced by $\inf \varphi''(x)/2$ and $\sup h(x;\mu)$ replaced by $\sup \varphi''(x)/2$ since

$$\inf \frac{\varphi''(x)}{2} \le \inf h(x;\mu) \quad \text{and} \quad \sup \frac{\varphi''(x)}{2} \ge \sup h(x;\mu).$$

These less tight bounds are implied in the working paper Becker (2012). Our lower and upper bounds have the general form $J \cdot \text{var}(X)$, where J depends on φ. Similar forms of bounds are presented in Abramovich and Persson (2016); Dragomir (2014); Walker (2014), but our J in Theorem 1 is much simpler and applies to a wider class of φ.

Inequality (2) implies Jensen's inequality when $\varphi''(x) \ge 0$. Note also that Jensen's inequality is sharp when $\varphi(x)$ is linear, whereas inequality (2) is sharp when $\varphi(x)$ is a quadratic function of x.

In some applications the moments of X present in (2) are unknown, although a random sample $x_1, \dots, x_n$ from the underlying distribution F is available. A version of Theorem 1 suitable for this situation is given in the following corollary.

Corollary 1. Let $x_1, \dots, x_n$ be any n datapoints in $(-\infty, \infty)$, and let

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{\varphi_x} = \frac{1}{n}\sum_{i=1}^{n} \varphi(x_i), \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Then

$$\inf_{x\in[a,b]} h(x;\bar{x})S^2 \le \overline{\varphi_x} - \varphi(\bar{x}) \le \sup_{x\in[a,b]} h(x;\bar{x})S^2,$$

where $a = \min\{x_1, \dots, x_n\}$ and $b = \max\{x_1, \dots, x_n\}$.

Proof. Consider the discrete random variable X with probability distribution $P(X = x_i) = 1/n$, $i = 1, \dots, n$. We have $E[X] = \bar{x}$, $E[\varphi(X)] = \overline{\varphi_x}$, and $\text{var}(X) = S^2$. Then the corollary follows from application of Theorem 1. □

Lemma 1. If $\varphi'(x)$ is convex, then $h(x;\mu)$ is monotonically increasing in x, and if $\varphi'(x)$ is concave, then $h(x;\mu)$ is monotonically decreasing in x.

Proof. We prove that $h'(x;\mu) \ge 0$ when $\varphi'(x)$ is convex. The analogous result for concave $\varphi'(x)$ follows similarly. Note that

$$\frac{dh(x;\mu)}{dx} = \frac{\frac{\varphi'(x)+\varphi'(\mu)}{2} - \frac{\varphi(x)-\varphi(\mu)}{x-\mu}}{\frac{1}{2}(x-\mu)^2},$$

so it suffices to prove

$$\frac{\varphi'(x) + \varphi'(\mu)}{2} \ge \frac{\varphi(x) - \varphi(\mu)}{x - \mu}.$$

Without loss of generality we assume $x > \mu$. Convexity of $\varphi'(x)$ gives

$$\varphi'(y) \le \varphi'(\mu) + \frac{\varphi'(x) - \varphi'(\mu)}{x - \mu}(y - \mu)$$

for all $y \in (\mu, x)$. Therefore we have

$$\begin{aligned}\varphi(x) - \varphi(\mu) &= \int_{\mu}^{x} \varphi'(y)dy \\ &\le \int_{\mu}^{x} \left\{\varphi'(\mu) + \frac{\varphi'(x) - \varphi'(\mu)}{x-\mu}(y-\mu)\right\} dy \\ &= \frac{\varphi'(x) + \varphi'(\mu)}{2}(x - \mu)\end{aligned}$$

and the result follows. □

Lemma 1 makes Theorem 1 easy to use as the follow results hold:

$$\begin{cases} \inf h(x;\mu) = \lim_{x\to a} h(x;\mu) \\ \sup h(x;\mu) = \lim_{x\to b} h(x;\mu), \end{cases} \quad \text{when } \varphi'(x) \text{ is convex}$$

$$\begin{cases} \inf h(x;\mu) = \lim_{x\to b} h(x;\mu) \\ \sup h(x;\mu) = \lim_{x\to a} h(x;\mu), \end{cases} \quad \text{when } \varphi'(x) \text{ is concave.}$$

Note the limits of $h(x;\mu)$ can be either finite or infinite. The proof of Lemma 1 borrows ideas from Bennish (2003). Examples of functions $\varphi(x)$ for which φ' is convex include $\varphi(x) = \exp(x)$ and $\varphi(x) = x^p$ for $p \ge 2$ or $p \in (0, 1]$. Examples of functions $\varphi(x)$ for which φ' is concave include $\varphi(x) = -\log x$ and $\varphi(x) = x^p$ for $p < 0$ or $p \in [1, 2]$.

3. Examples

Example 1 (Moment Generating Function). For any random variable X supported on (a, b) with a finite variance, we can bound the moment generating function $E[e^{tX}]$ using Theorem 1 to get

$$\begin{aligned}\inf_{x\in(a,b)} \{h(x;\mu)\}\text{var}(X) &\le E[e^{tX}] - e^{tE[X]} \\ &\le \sup_{x\in(a,b)} \{h(x;\mu)\}\text{var}(X),\end{aligned}$$

where

$$h(x;\mu) = \frac{e^{tx} - e^{t\mu}}{(x-\mu)^2} - \frac{te^{t\mu}}{x-\mu}.$$

For $t > 0$ and $(a, b) = (-\infty, \infty)$, we have

$$\inf h(x;\mu) = \lim_{x\to-\infty} h(x;\mu) = 0$$
$$\text{and } \sup h(x;\mu) = \lim_{x\to\infty} h(x;\mu) = \infty.$$

So Theorem 1 provides no improvement over Jensen's inequality. However, on a finite domain such as a nonnegative random variable with $(a, b) = (0, \infty)$, a significant improvement in the lower bound is possible because

$$\inf h(x;\mu) = h(0;\mu) = \frac{1 - e^{t\mu} + t\mu e^{t\mu}}{\mu^2} > 0.$$

Similar results hold for $t < 0$. We apply this to an example from Walker (2014), where X is an exponential random variable with mean 1 and $\varphi(x) = e^{tx}$ with $t = 1/2$. Here the actual Jensen's gap is $\mathbb{E}[e^{tX}] - e^{t\,\mathbb{E}[X]} = 2 - \sqrt{e} \approx 0.351$. Since $\mathrm{var}(X) = 1$, we have

$$0.176 \approx h(0;\mu) \le \mathbb{E}[e^{tX}] - e^{t\,\mathbb{E}[X]} \le \lim_{x\to\infty} h(x;\mu) = \infty.$$

The less sharp lower bound using $\inf \varphi''(x)/2$ is 0.125. Utilizing elaborate approximations and numerical optimizations Walker (2014) yielded a more accurate lower bound of 0.271.

Example 2 (Arithmetic vs Geometric Mean). Let X be a positive random variable on interval (a, b) with mean μ. Note that $-\log(x)$ is a convex function with a concave first derivative. Applying Theorem 1 and Lemma 1 gives

$$\lim_{x\to b} h(x;\mu)\mathrm{var}(X) \le -E\{\log(X)\} + \log\mu \le \lim_{x\to a} h(x;\mu)\mathrm{var}(X),$$

where

$$h(x;\mu) = \frac{-\log x + \log\mu}{(x-\mu)^2} + \frac{1}{\mu(x-\mu)}.$$

Now consider a sample of n positive data points $x_1, \ldots, x_n$. Let $\bar{x}$ be the arithmetic mean and $\bar{x}_g = (x_1x_2\cdots x_n)^{\frac{1}{n}}$ be the geometric mean. Applying Corollary 1 gives

$$\exp\{S^2h(b;\bar{x})\} \le \frac{\bar{x}}{\bar{x}_g} \le \exp\{S^2h(a;\bar{x})\},$$

where a, b, S^2 are as defined in Corollary 1. To give some numerical results, we generated 100 random numbers from a uniform distribution on [10, 100]. For these 100 numbers, the arithmetic mean $\bar{x}$ is 54.830 and the geometric mean $\bar{x}_g$ is 47.509. The above inequality becomes

$$1.075 \le \frac{\bar{x}}{(x_1x_2\cdots x_n)^{\frac{1}{n}}} = 1.154 \le 1.331,$$

which are fairly tight bounds. Replacing $h(x_n;\bar{x})$ by $\varphi''(x_n)/2$ and $h(x_1;\bar{x})$ by $\varphi''(x_1)/2$ leads to a less accurate lower bound 1.0339 and upper bound 21.698.

Example 3 (Power Mean). Let X be a positive random variable on a positive interval (a, b) with mean μ. For any real number $s \neq 0$, define the power mean as

$$M_s(X) = (EX^s)^{1/s}$$

Jensen's inequality establishes that $M_s(X)$ is an increasing function of s. We now give a sharper inequality by applying Theorem 1. Let $r \neq 0$, $Y = x^r$, $\mu_y = EY$, $p = s/r$ and $\varphi(y) = y^p$. Note that $EX^s = E\{\varphi(Y)\}$. Applying Theorem 1 leads to

$$\inf h(y;\mu_y)\mathrm{var}(Y) \le E[X^s] - (EX^r)^p \le \sup h(y;\mu_y)\mathrm{var}(Y),$$

where

$$h(y;\mu_y) = \frac{y^p - \mu_y^p}{(y-\mu_y)^2} - \frac{p\mu_y^{p-1}}{y-\mu_y}.$$

To apply Lemma 1, note that $\varphi'(y)$ is convex for $p \ge 2$ or $p \in (0, 1]$ and is concave for $p < 0$ or $p \in [1, 2]$ as noted in Section 2.

Applying the above result to the case of $r = 1$ and $s = -1$, we have $Y = X$, $p = -1$. Therefore

$$\left((EX)^{-1} + \lim_{y\to a} h(y;\mu_y)\mathrm{var}(X)\right)^{-1} \le \left(EX^{-1}\right)^{-1} \le \left((EX)^{-1} + \lim_{y\to b} h(y;\mu_y)\mathrm{var}(X)\right)^{-1}.$$

For the same sequence $x_1, \ldots, x_n$ generated in Example 2, we have $\bar{x}_{\text{harmonic}} = 39.113$. Applying Corollary 1 leads to

$$25.337 \le \bar{x}_{\text{harmonic}} = 39.113 \le 48.905.$$

Note that the upper bound 48.905 is much smaller than the arithmetic mean $\bar{x} = 54.830$ by Jensen's inequality. Replacing $h(b;\bar{x})$ by $\varphi''(b)/2$ and $h(a;\bar{x})$ by $\varphi''(a)/2$ leads to a less accurate lower bound 0.8298 and 51.0839.

In a recent article, de Carvalho (2016) revisited Kolmogorov's formulation of a generalized mean as

$$E_\varphi(X) = \varphi^{-1}(E[\varphi(X)]), \tag{3}$$

where φ is a continuous monotone function with inverse φ^{-1}. Example 2 corresponds to $\varphi(x) = -\log(x)$ and Example 3 corresponds to $\varphi(x) = x^s$. We can also apply Theorem 1 to bound $\varphi^{-1}(E\varphi(X))$ for a more general function $\varphi(x)$.

Example 4 (Rao-Blackwell Estimator). The Rao-Blackwell theorem (Theorem 7.3.17 in Casella and Berger, 2002; Theorem 10.42 in Wasserman, 2013) is a basic result in statistical estimation. Let $\hat{\theta}$ be an estimator of θ, $L(\theta,\hat{\theta})$ be a loss function convex in $\hat{\theta}$, and T a sufficient statistic. Then the Rao-Blackwell estimator, $\hat{\theta}^* = E[\hat{\theta} \mid T]$, satisifies the following inequality in risk function

$$E[L(\theta,\hat{\theta})] \ge E[L(\theta,\hat{\theta}^*)]. \tag{4}$$

We can improve this inequality by applying Theorem 1 to $\varphi(\hat{\theta}) = L(\theta,\hat{\theta})$ with respect to the conditional distribution of $\hat{\theta}$ given T:

$$E[L(\theta,\hat{\theta}) \mid T] - L(\theta,\hat{\theta}^*) \ge \inf_{x\in(a,b)} h(x;\hat{\theta}^*)\mathrm{var}(\hat{\theta} \mid T),$$

where function h is defined as in Theorem 1 for $\varphi(\hat{\theta})$ and $P(\hat{\theta} \in (a, b) \mid T) = 1$. Further taking expectations over T gives

$$E[L(\theta,\hat{\theta})] - E[L(\theta,\hat{\theta}^*)] \ge E\left[\inf_{x\in(a,b)} h(x;\hat{\theta}^*)\mathrm{var}(\hat{\theta} \mid T)\right].$$

In particular for square-error loss, $L(\theta,\hat{\theta}) = (\hat{\theta}-\theta)^2$, we have

$$E[(\theta-\hat{\theta})^2] - E[(\theta-\hat{\theta}^*)^2] = E\left[\mathrm{var}(\hat{\theta} \mid T)\right].$$

Using the original Jensen's inequality only establishes the cruder inequality in Equation (4).

4. Improved Bounds by Partitioning

As discussed in Example 1 above, Theorem 1 does not improve on Jensen's inequality if $\inf h(x;\mu) = 0$. In such cases, we can often sharpen the bounds by partitioning the domain (a, b) following an approach used in Walker (2014). Let

$$a = x_0 < x_1 < \cdots < x_m = b,$$

$I_j = [x_{j-1}, x_j)$, $\eta_j = P(X \in I_j)$, and $\mu_j = E(X|X \in I_j)$. It follows from the law of total expectation that

$$E[\varphi(X)] = \sum_{j=1}^{m} \eta_j E[\varphi(X) \mid X \in I_j]$$
$$= \sum_{j=1}^{m} \eta_j \varphi(\mu_j) + \sum_{j=1}^{m} \eta_j \left(E[\varphi(X) \mid X \in I_j] - \varphi(\mu_j) \right).$$

Let Y be a discrete random variable with distribution $P(Y = \mu_j) = \eta_j, j = 1, 2, \ldots, m$. It is easy to see that $EY = EX$. It follows by Theorem 1 that

$$\sum_{j=1}^{m} \eta_j \varphi(\mu_j) = E[\varphi(Y)] \geq \varphi(EY) + \inf_{y \in [\mu_1, \mu_m]} h(y; \mu_y) \text{var}(Y).$$

We can also apply Theorem 1 to each $E[\varphi(X|X \in I_j)] - \varphi(\mu_j)$ term:

$$E[\varphi(X \mid X \in I_j)] - \varphi(\mu_j) \geq \inf_{x \in I_j} h(x; \mu_j) \text{var}(X \mid X \in I_j).$$

Combining the above two equations, we have

$$E[\varphi(X)] - \varphi(EX) \geq \inf_{y \in [\mu_1, \mu_m]} h(y; \mu_y) \text{var}(Y) + \sum_{j=1}^{m} \eta_j \inf_{x \in I_j} h(x; \mu_j) \text{var}(X \mid X \in I_j). \quad (5)$$

Replacing inf by sup in the right hand side gives the upper bound.

For a convex function ϕ, the Jensen gap on the left side of (5) is positive if any of the $m + 1$ terms on the right is positive. In particular, the Jensen gap is positive if there exists an interval $I \subset (a_j, b_j)$ that satisfies $\inf_{x \in I} \varphi''(x) > 0$, $P(X \in I) > 0$ and $\text{var}(X|X \in I) > 0$. Note that a finer partition does not necessarily lead to a sharper lower bound in (5). The focus of the partition should therefore be on isolating the part of interval (a, b) in which $\varphi''(x)$ is close to 0.

Consider the example $X \sim N(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma = 1$ and $\varphi(x) = e^x$. We divide $(-\infty, \infty)$ into three intervals with equal probabilities. This gives

I_j	η_j	$E[X\|X \in I_j]$	$\text{var}(X\|X \in I_j)$	$\inf_{x \in I_j} h(x; \mu_j)$	$\sup_{x \in I_j} h(x; \mu_j)$
$(-\infty, -.431)$	1/3	−1.091	0.280	0.000	0.212
$(-0.431, 0.431)$	1/3	0.000	0.060	0.435	0.580
$(0.431, \infty)$	1/3	1.091	0.280	1.209	∞

The actual Jensen gap is $e^{\mu + \frac{\sigma}{2}} - e^{\mu} = 0.649$. The lower bound from (5) is 0.409, which is a huge improvement over Jensen's bound of 0. The upper bound ∞, however, provides no improvement over Theorem 1.

To summarize, this article proposes a new sharpened version of Jensen's inequality. The proposed bound is simple and insightful, is broadly applicable by imposing minimum assumptions on $\varphi(x)$, and provides fairly accurate result in spite of its simple form.

From a teaching perspective, the article enhances student understanding by providing an explicit expression of the Jensen's gap in terms of $\varphi''(x)$ and $\text{var}(X)$, which makes it easy to see that a larger nonnegative $\varphi''(x)$ leads to a larger nonnegative Jensen's gap. In particular, Jensen's inequality is sharp when either $\varphi(x)$ is a linear function or $\text{var}(X) = 0$. Our results also strengthen familiar statistical concepts and applications as presented in Section 3. We have incorporated this material in our classroom teaching with only slightly increased technical level and lecture time.

Acknowledgements

The authors gratefully acknowledge the careful review and editing of this article by the editor Dr. Nicole Lazar, an anonymous associate editor, and two anonymous reviewers.

References

Abramovich, S., and Persson, L.-E. (2016), "Some New Estimates of the Jensen Gap," *Journal of Inequalities and Applications*, 2016, 39. [1,2]

Abramovich, S., Persson, L.-E., and Samko, N. (2014), "Some New Scales of Refined Jensen And Hardy Type Inequalities," *Mathematical Inequalities & Applications*, 17, 1105–1114. [1]

Becker, R. A. (2012), "The Variance Drain and Jensen's Inequality," Technical Report, CAEPR Working Paper No. 2012-004. Available at *https://ssrn.com/abstract=2027471*. [2]

Bennish, J. (2003), "A Proof of Jensen's Inequality," *Missouri Journal of Mathematical Sciences*, 15. [2]

Casella, G., and Berger, R. L. (2002), *Statistical Inference*, vol. 2, Pacific Grove, CA: Duxbury. [1]

de Carvalho, M. (2016), "Mean, What Do You Mean? *The American Statistician*, 70, 270–274. [3]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 1–38. [1]

Dragomir, S. (2014), "Jensen Integral Inequality for Power Series with Nonnegative Coefficients and Applications," *RGMIA Res. Rep. Collect*, 17, 42. [1,2]

Horvath, L., Khan, K. A., and Pecaric, J. (2014), "Refinement of Jensen's Inequality for Operator Convex Functions," *Advances in inequalities and applications*, 2014. [1]

Jensen, J. L. W. V. (1906), "Sur les Fonctions Convexes et les Inégalités entre les Valeurs Moyennes," *Acta Mathematica*, 30, 175–193. [1]

Walker, S. G. (2014), "On a Lower Bound for the Jensen Inequality," *SIAM Journal on Mathematical Analysis*, 46, 3151–3157. [1,2,3]

Wasserman, L. (2013), *All of Statistics: A Concise Course in Statistical Inference*, Springer Science & Business Media. [1]