McDiarmid's inequality via the Entropy Method

Justin Le
December 12, 2016

1 Introduction

We first demonstrate the use of the Efron-Stein inequality for bounding the variance of a *function with bounded differences*. We then show that the Entropy Method improves on this result by providing Gaussian tail bounds. This material is adapted from [BLM03].

1.1 Setting and notation

In this article, we consider random variables $X_1, \dots, X_n$ that are independent. Define $X = (X_1, \dots, X_n)$ and $Z = f(X_1, \dots, X_n)$. These assumptions hold for the theorems presented herein unless stated otherwise.

Let $\mathbb{E}_i$ denote the expectation conditioned on $(X_1, \dots, X_i)$ and $\mathbb{E}^{(i)}$ the expectation conditioned on $(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$. Use of the subscript i and superscript (i) are defined similarly for the entropy $\mathrm{Ent}(\cdot)$. Let $\mathbb{P}(Z \geq t)$ denote the probability that a random variable $Z \in \mathbb{R}$ is greater than or equal to $t \in \mathbb{R}$. To avoid clutter in notation, we drop parentheses from most instances of log and the expectation $\mathbb{E}$ of a random variable. Any expression appearing after log is assumed to be included in the argument and likewise for $\mathbb{E}$.

2 The variance of functions with bounded differences

A function $f : \mathcal{X}^n \to \mathbb{R}$ that satisfies

$$\sup_{x_1, \dots, x_n, x_i'} \left| f(x_1, \dots, x_n) - f(x_1, \dots, x_i', \dots, x_n) \right| \leq c_i \qquad 1 \leq i \leq n \tag{1}$$

is said to be a function with bounded differences.

To arrive at an expression of the variance that would facilitate the proof of the Efron-Stein inequality, we may define the quantity

$$\Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z = \mathbb{E}_i (Z - \mathbb{E}^{(i)} Z).$$

Note that $\mathbb{E}_i \mathbb{E}^{(i)}(\cdot) = \mathbb{E}_{i-1}(\cdot)$. Furthermore, note that

$$Z - \mathbb{E}Z = \left(\sum_{i=1}^n \Delta_i \right)^2.$$

Then the variance can be equivalently expressed as

$$\begin{aligned} \mathrm{Var}(Z) &= \mathbb{E} \left(\sum_{i=1}^n \Delta_i \right)^2 \\ &= \sum_{i=1}^n \mathbb{E}\Delta_i^2 + 2 \sum_{i<j} \mathbb{E}\Delta_i \Delta_j \\ &= \sum_{i=1}^n \mathbb{E}\Delta_i^2. \end{aligned} \tag{2}$$

The final line follows from the fact that for $i < j$,

$$\mathbb{E}_i \Delta_i \Delta_j = \Delta_i \mathbb{E}_i \Delta_j = 0,$$

and therefore $\mathbb{E}\Delta_i\Delta_j = 0$. We will see that Eq. 2 facilitates the proof of Thm. 2.1.

Now we are well-equipped to give a concise proof of the first key result in this article. This result has found many applications, and it is therefore interesting far beyond the setting considered here. Its complete statement and proof involve substantially more detail than what has been given here, but we refer the reader to [RT86] for a more thorough treatment.

Theorem 2.1 (Efron-Stein Inequality)**.**

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}(Z - \mathbb{E}^{(i)} Z)^2.$$

Proof. By Jensen's inequality,

$$\Delta_i^2 \leq \mathbb{E}_i (Z - \mathbb{E}^{(i)} Z)^2.$$

Summing over i and then taking the expectation of both sides, the left-hand side becomes

$$\mathbb{E} \sum_{i=1}^{n} \Delta_i^2 = \mathbb{E} \sum_{i=1}^{n} \left[\mathbb{E}_i (Z - \mathbb{E}^{(i)} Z) \right]^2,$$

which, by the linearity of expectation, equals $\mathrm{Var}(Z)$ in the form given by Eq. 2. Noting that $\mathbb{E}\mathbb{E}_i = \mathbb{E}$ in this case, the right-hand side matches that of the theorem. □

This rendition of the theorem, proven in [RT86], is an optimal successor to the original proven in [ES81].

With the above foundations, a bound on variance can now be obtained for a function with bounded differences. To see this, we choose

$$Z_i = \frac{1}{2} \left(\sup_{X_1, \ldots, X_n} f(X_1, \ldots, X_n) + \inf_{x_i'} f(X_1, \ldots, x_i', \ldots, X_n) \right).$$

Then by the bounded differences property, we have for all $X_1, \ldots, X_n$

$$\begin{aligned} Z - Z_i &= f(X_1, \ldots, X_n) - \frac{1}{2} \sup_{X_1, \ldots, X_n} f(X_1, \ldots, X_n) - \frac{1}{2} \inf_{x_i'} f(X_1, \ldots, x_i', \ldots, X_n) \\ &\leq \frac{1}{2} \left(f(X_1, \ldots, X_n) - \inf_{x_i'} f(X_1, \ldots, x_i', \ldots, X_n) \right) \\ &\leq \frac{c_i}{2}. \end{aligned}$$

Squaring both sides and combining the result with the Efron-Stein inequality, we have

$$\mathrm{Var}(Z) \leq \frac{1}{4} \sum_{i=1}^{n} c_i^2. \tag{3}$$

The Efron-Stein inequality greatly simplifies the task of bounding variance in a number of other settings in which variance estimation would be difficult, such as kernel density estimation, first passage percolation, and in bounding the variance of the largest eigenvalue of a random matrix.

3 The Entropy Method

As we have seen for functions with bounded differences, the Efron-Stein inequality provides a simple path toward a bound on variance. However, it lacks the exponential curvature that these functions often have in their tail probabilities. Although an exponentially decaying bound can, in fact, be obtained via the Efron-Stein inequality (e.g., see [BLM13]), its precision is still sub-optimal when compared to the bound obtained via the Entropy Method. Specifically, the Entropy Method gives not only an exponential decay but a Gaussian one, i.e., one with a square factor in the exponent. For the problem of bounded differences, this result is known as McDiarmid's inequality and was first proven in [McD89].

Importantly, we further note that the Efron-Stein inequality does not depend on the distribution of the random variables, so it is natural to ask whether a precise exponential bound can be obtained that would also be distribution-independent. As we will see, the Entropy Method provides one gracefully.

3.1 Tools of the Entropy Method

Generally, the Entropy Method begins with an entropic inequality, such as the sub-additivity property in Ineq. 3.1 or else a sort of log-Sobolev inequality. It then proceeds with an application of Herbst's argument. We present these components here before demonstrating their use.

Theorem 3.1 (Sub-Additivity of Entropy)**.**

$$\mathbb{E}\phi(Z) - \phi(\mathbb{E}Z) \leq \mathbb{E}\sum_{i=1}^{n} \mathbb{E}^{(i)}\phi(Z) - \phi(\mathbb{E}^{(i)}Z).$$

Equivalently, by introducing the notation Ent$(\cdot)$*, we have*

$$\mathrm{Ent}(Z) \leq \mathbb{E}\sum_{i=1}^{n} \mathrm{Ent}^{(i)}(Z).$$

Proof. Please refer to [BLM13]. □

The following results will be useful in obtaining the square quantity that will eventually arise in the exponential bound we seek.

Theorem 3.2 (Hoeffding's Lemma)**.** *Let Z be distributed over $[a, b]$ such that $\psi_Z(\lambda) = \log \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}$. Then*

$$\psi_Z''(\lambda) \leq \frac{(b-a)^2}{4}.$$

Proof. Please refer to [BLM13]. □

To simplify the presentation of the following proof, we assume that $\mathbb{E}Z = 0$, but when used in the proof of McDiarmid's inequality, we simply replace instances of Z in Thm. 3.3 with the more general deviation $Z - \mathbb{E}Z$ that interests us in this article.

Theorem 3.3 (Herbst's Argument)**.** *For some $K > 0$,*

$$\begin{aligned} \frac{\mathrm{Ent}(e^{\lambda Z})}{\mathbb{E}e^{\lambda Z}} &\leq \lambda^2 K \qquad \forall \lambda > 0 \\ &\Rightarrow \\ \log \mathbb{E}e^{\lambda Z} &\leq \lambda^2 K \qquad \forall \lambda > 0. \end{aligned} \tag{4}$$

Proof. Recall that in stating the sub-additivity of entropy, the notation $\mathrm{Ent}(\phi(Z)) = \mathbb{E}\phi(Z) - \phi(\mathbb{E}Z)$ was introduced. Choosing $\phi(x) = x \log x$ in this definition, we find that

$$\begin{aligned}\frac{\mathrm{Ent}(e^{\lambda Z})}{\mathbb{E}e^{\lambda Z}} &= \frac{\mathbb{E}(e^{\lambda Z}\log e^{\lambda Z})}{\mathbb{E}e^{\lambda Z}} - \log \mathbb{E}e^{\lambda Z} \\ &= \frac{\lambda \mathbb{E}Ze^{\lambda Z}}{\mathbb{E}e^{\lambda Z}} - \log \mathbb{E}e^{\lambda Z}.\end{aligned}$$

To simplify the remaining steps, define $\psi(\lambda) = \log \mathbb{E}e^{\lambda Z}$. Then

$$\frac{\mathrm{Ent}(e^{\lambda Z})}{\mathbb{E}e^{\lambda Z}} = \lambda\psi'(\lambda) - \psi(\lambda). \tag{5}$$

From the antecedent of the theorem, we divide by λ^2:

$$\begin{aligned}\lambda\psi'(\lambda) - \psi(\lambda) &\leq \lambda^2 K \\ \frac{\psi'(\lambda)}{\lambda} - \frac{\psi(\lambda)}{\lambda^2} &\leq K.\end{aligned}$$

By the product rule of differentiation and then by integration,

$$\begin{aligned}\frac{d}{d\lambda}\left(\frac{\psi(\lambda)}{\lambda}\right) &\leq K \\ \frac{\psi(\lambda)}{\lambda} &\leq \lambda K.\end{aligned}$$

Substituting the definition of $\psi(\lambda)$ and multiplying both sides by λ, the statement of the theorem is recovered. □

We will later see that in proving McDiarmid's inequality we may select λ prudently to arrive at an interpretable bound.

3.2 Application: functions with bounded differences

With the given tools, we may now give an elementary proof of the main result in this article, McDiarmid's inequality. The proof follows the Entropy Method outlined previously. The result improves on Thm 2.1, as it captures Gaussian tail behavior.

Applying the bounded differences property in 1 to Hoeffding's Lemma 3.2 and conditioning Z on $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, we have $a - b \leq c$. Thus

$$\frac{\mathrm{Ent}^{(i)}e^{\lambda Z}}{\mathbb{E}^{(i)}e^{\lambda Z}} \leq \frac{(c_i\lambda)^2}{8}.$$

Applying this result to the sub-additivity of entropy, we have

$$\mathrm{Ent}(e^{\lambda(Z-\mathbb{Z})}) \leq \sum_{i=1}^{n} \mathbb{E}\frac{(c_i\lambda)^2}{8}e^{\lambda(Z-\mathbb{E}Z)}.$$

In order to apply Herbst's argument efficiently, may define the quantities

$$\begin{aligned}\psi(\lambda) &= \log \mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \\ v &= \sum_{i=1}^{n} c_i^2.\end{aligned}$$

Then, continuing from above,

$$\frac{\mathrm{Ent}(e^{\lambda(Z-\mathbb{E}Z)})}{\mathbb{E}e^{Z-\mathbb{E}Z}} \leq \frac{v\lambda^2}{2},$$

and applying Herbst's argument,

$$\psi(\lambda) \leq \frac{v\lambda^2}{2}. \tag{6}$$

By Markov's inequality, the definition of $\psi(\lambda)$, and Eq. 6, we finally obtain, for all $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq \frac{\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}}{e^{\lambda t}} = e^{\psi\lambda - \lambda t}.$$

From Eq. 6, and by choosing $\lambda = t/v$,

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq e^{v\lambda^2/2 - \lambda t} = e^{-t^2/(2v)}.$$

References

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. "Concentration Inequalities: A Nonasymptotic Theory of Independence". In: 3 (2013).

[2] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. "Concentration inequalities using the entropy method". In: *Ann. Probab.* 31.3 (July 2003), pp. 1583–1614. DOI: 10.1214/aop/1055425791. URL: http://dx.doi.org/10.1214/aop/1055425791.

[3] Kai Lai Chung. *A course in probability theory.* Probability and mathematical statistics : a series of monographs and textbooks. Index. San Diego, Cal., New York, Boston: Academic Press, 1974. ISBN: 0-12-174650-X. URL: http://opac.inria.fr/record=b1079080.

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).* Wiley-Interscience, 2006. ISBN: 0471241954.

[5] B. Efron and C. Stein. "The Jackknife Estimate of Variance". In: *Ann. Statist.* 9.3 (May 1981), pp. 586–596. DOI: 10.1214/aos/1176345462. URL: http://dx.doi.org/10.1214/aos/1176345462.

[6] Colin McDiarmid. "On the method of bounded differences". In: *Surveys in Combinatorics* (1989), pp. 148–188.

[7] WanSoo T. Rhee and Michel Talagrand. "Martingale inequalities and the Jackknife estimate of variance". In: *Statistics & Probability Letters* 4.1 (1986), pp. 5–6. ISSN: 0167-7152. DOI: http://dx.doi.org/10.1016/0167-7152(86)90029-5. URL: http://www.sciencedirect.com/science/article/pii/0167715286900295.