A Standard Concentration Result via the Entropy Method

Justin Le
Department of Electrical & Computer Engineering
University of Nevada, Las Vegas
lejustin.lv@gmail.com

December 12, 2016

Abstract

Functions of independent random variables arise in information theory, signal processing, statistics, and many other fields deeply relevant to the digital era. Often, we seek to bound the deviation of such a function from its expectation and to precisely capture the behavior of this deviation, especially at the tails of its distribution. In this article, we present a widely successful technique for obtaining such bounds: the *Entropy Method*. We show by example how it improves on a simpler bound obtained by a previous approach involving the Efron-Stein inequality, using a *bounded differences* problem to illustrate the effect.

1 Introduction

The goal of this article is to show that a technique known as the Entropy Method can be used to obtain a precise concentration result, namely a bound on the probability that a function of random variables will deviate from its expectation. As the variance is a quantity closely related to this deviation, we first demonstrate that a bound on variance can be obtained via the Efron-Stein inequality. We then show how this bound can be greatly improved by one of the most successful tools that have been developed thus far in obtaining concentration inequalities: the Entropy Method. We also justify this improvement qualitatively. Along the way, we provide the basic definitions and tools that each method requires. No attempt is made to be comprehensive, and many technical (e.g., measure-theoretic) aspects are neglected for the sake of simplicity. Likewise, we restrict our demonstration to a "bounded differences" problem, which we define in the sequel. We refer the reader to [BLM03] for a complete treatment.

1.1 Setting and notation

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

Let $\mathbb{E}_i$ denote the expectation conditioned on $(X_1, \ldots, X_i)$ and $\mathbb{E}^{(i)}$ the expectation conditioned on $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. 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 depart somewhat from other authors' conventions by dropping 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 $\mathbb{E}$. For example, it should be apparent that $\log ab = \log(ab)$. If b should not be included as an argument, the expression would appear as $b \log a$ instead. The same is true for $\mathbb{E}$. However, when combining log with $\mathbb{E}$, which will be necessary on occasion, we include parentheses where appropriate.

2 The variance of functions with bounded differences

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

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

is said to be a function with bounded differences. We consider this class of functions as it requires only a simple assumption but facilitates a key historical example of the Entropy Method's use.

As the Efron-Stein inequality is essentially a statement regarding the variance of a function, we must first define variance. Then, we give a theorem regarding general convex functions, which will be the starting point in proving the Efron-Stein inequality.

Definition 2.1 (Variance). *The variance of a random variable Z is*

$$\mathrm{Var}(Z) = \mathbb{E}(Z - \mathbb{E}Z)^2. \tag{2}$$

To arrive at a more useful form of variance for proving 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

$$\begin{aligned} \mathrm{Var}(Z) &= \mathbb{E} \left(\sum_{i=1}^n \Delta_i \right)^2 \\ &= \sum_{i=1}^n \mathbb{E}\Delta_i^2 + 2 \sum_{i<j} \mathbb{E}\Delta_i \Delta_j \\ &= \sum_{i=1}^n \mathbb{E}\Delta_i^2. \end{aligned} \tag{3}$$

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. 3 facilitates the proof of Thm. 2.2. Before pursuing this proof, however, we note that the variance can be interpreted as the composition of expectation with a convex function, namely the square function. We thus require a bound for such a function involving its expectation. Jensen's inequality applies here, although it holds for other convex functions as well.

Theorem 2.1 (Jensen's Inequality)**.** *If $f : \mathcal{X}^n \to \mathbb{R}$ is convex, then*

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

Proof. Please refer to [CT06]. □

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 2.2 (Efron-Stein Inequality)**.**

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

Proof. By Jensen's inequality 2.1,

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

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. 3. Noting that expectation distributes over sums and that $\mathbb{E}\mathbb{E}_i = \mathbb{E}$ in this case, the right-hand side matches that of the theorem. □

This rendition of Thm. 2.2, proven by [RT86], is an optimal one of the original proven in [ES81]. Although the statement concerns variance, we will later see that the improved bound obtained via the Entropy Method does not directly concern variance but a conceptually equivalent quantity, namely the probability that Z will deviate from its expectation by an amount t. The argument of this quantity, $Z - \mathbb{E}Z$, appears in the definition of variance.

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

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

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. Although these examples demonstrate the flexibility of the Efron-Stein inequality, the literature has also shown that the results often compare unfavorably to the optimal bounds obtained by more sophisticated methods [BLM13]. In the sequel, we present one such method for tightening Ineq. 4.

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. (The reasons for exponentiality are beyond the scope of this article.) Although an exponential 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 bound 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* $K > 0$*,*

$$\begin{aligned} \frac{\mathrm{Ent}(e^{\lambda Z})}{\mathbb{E}e^{\lambda Z}} &\leq \lambda^2 K \qquad \forall \lambda > 0 \\ &\Rightarrow \\ \log \mathbb{E}e^{\lambda Z} &\leq \lambda^2 K \qquad \forall \lambda > 0. \end{aligned} \tag{5}$$

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

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

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

$$\frac{\mathrm{Ent}(e^{\lambda Z})}{\mathbb{E}e^{\lambda Z}} = \lambda\psi'(\lambda) - \psi(\lambda). \tag{6}$$

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,

$$\frac{d}{d\lambda}\left(\frac{\psi(\lambda)}{\lambda}\right) \leq K$$
$$\frac{\psi(\lambda)}{\lambda} \leq \lambda K.$$

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, as is often the case beyond this article, we may select λ prudently to arrive at an interpretable bound. Notice that Herbst's argument does not provide a bound on the probability of Z deviating from its expectation, which we desire. To obtain such a result, we must take an additional step with Markov's inequality.

Theorem 3.4 (Markov's Inequality)**.** *For $Z \geq 0$,*

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}Z}{t}.$$

Furthermore, for a convex function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{P}(Z \geq t) \quad \leq \quad \mathbb{P}(\phi(Z) \geq \phi(t)) \quad \leq \quad \frac{\mathbb{E}\phi(Z)}{\phi(t)}.$$

Proof. Please refer to [Chu74]. □

The latter variant of Markov's inequality (often referred to as Chebyshev's inequality) will be the one employed in this article, although both have widespread applications in obtaining probabilistic bounds.

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, which essentially bounds the probability that Z deviates from its mean. The proof follows the Entropy Method outlined previously. The result improves on Thm 2.2, as it succeeds in capturing the precise exponential behavior of this probability.

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

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

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,

$$\psi(\lambda) \leq \frac{v\lambda^2}{2}. \tag{7}$$

By Markov's inequality, the definition of $\psi(\lambda)$, and Eq. 7, we finally obtain, for all $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq \frac{\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}}{e^{\lambda t}} = e^{\psi\lambda - \lambda t}. \tag{8}$$

From Eq. 7, and by choosing $\lambda = t/v$,

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq e^{v\lambda^2/2-\lambda t} = e^{-t^2/(2v)}. \tag{9}$$

Due to our choice of λ, the result can be interpreted as a Gaussian (i.e., square exponential) bound on the probability that Z will deviate from its expectation. In contrast to the loose bound, Ineq. 4, obtained via the Efron-Stein inequality, the exponential curvature of Ineq. 9 more accurately reflects the empirical behavior of a wide variety of functions encountered in practice. More improvements can be made, but they would require more involved arguments and tools. Here, we have given only the simplest example of the Entropy Method's use.

4 Conclusion

In this article, we have shown how the Entropy Method can be used to obtain a bound on the probability that a function of independent random variables will deviate from its expectation. This bound was shown to be exponential, which more accurately reflects the behavior of the deviation and is thus an improvement over a similar bound previously obtained via the Efron-Stein inequality. We introduced the tools needed for both methods and demonstrated both of them on a simple problem involving functions with bounded differences. However, we have only considered discrete random variables, i.e., the case in which X has finitely many elements. Many interesting generalizations and applications remain to be considered, and a single article could not do justice to them all. Still, for the interested reader, we suggest [BLM13] and its bibliography as a more thorough introduction to the topic than the present article.

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.