A simple algorithmic explanation for the concentration of measure phenomenon

Igor C. Oliveira

October 10, 2014

Abstract

We give an elementary algorithmic argument that sheds light on the concentration of measure phenomenon observed in the sum of independent random variables. Roughly speaking, we observe that without concentration of measure, it would be possible to predict the outcome of a fair coin toss with probability greater than 1/2. We use this idea to derive an alternative proof for a particular case of the Chernoff-Hoeffding bound.

1 Introduction.

The Chernoff-Hoeffding bound [1, 2] is one of the most useful inequalities in discrete mathematics and theoretical computer science. In its simpler form, it states that the sum $X = \sum_i X_i$ of n independent uniformly distributed $0/1$ random variables is sharply concentrated around its expected value. The textbook proof of this result proceeds by applying Markov's inequality to the moment generating function of X.

We provide a new proof of a very useful case of this inequality, namely, a strong bound for deviations of magnitude at least εn around the expected value $n/2$, for any fixed constant $\varepsilon > 0$. Although quite simple combinatorial proofs exist for this particular concentration bound, we believe that our proof may be of independent interest.

The Chernoff-Hoeffding bound is the simplest example of the concentration of measure phenomenon observed in probability theory. From a conceptual point of view, our main contribution is an algorithmic explanation for the existence of concentration of measure. We observe that without this phenomenon, it would be possible to predict the outcome of a fair coin toss with non-trivial success probability.

Our argument relies on a learning algorithm discovered about two decades ago by Littlestone and Warmuth [3], the Weighted Majority Algorithm. Consider a sequence of n boolean trials, and a learner that makes a prediction for each trial. The final goal is to make as few mistakes as possible. The learner receives help from a pool of k experts $E_1, \ldots, E_k$, with each expert predicting the outcome of these trials according to some rule. Littlestone and Warmuth proved that there exists a strategy for the learner to make predictions based on the opinions of the experts that will always be almost as good as the predictions of the best expert in the pool. The non-trivial aspect of their result is that the learner of course does not know a priori which expert will perform better in the corresponding sequence of trials.

Our proof of the concentration bound then proceed as follows. A simple way to formulate the argument is by contradiction. First, fix a sequence of trials, unknown to the learner. It follows from our assumption of weak concentration that any expert E_i that predicts at random will be correct on slightly more than $n/2$ trials with some small but non-negligible probability. However, with k experts predicting (independently) at random, there will be with high probability an expert E^* with a small advantage over the expected number $n/2$ of correct predictions. Although the learner does not know a priori which expert will be more successful, the Weighted Majority Algorithm guarantees that we can make almost the same number of correct predictions as the best expert in the pool. In other words, for every fixed but unknown sequence of trials, it is possible to have some non-trivial advantage over random guessing (under our initial assumption of weak concentration, and by taking k sufficiently large). Now also make the sequence of boolean trials random. It follows from the previous discussion that we can predict the outcome of a sequence of random coin tosses with non-trivial advantage. A simple averaging argument then implies that there exists a particular random coin in this sequence that admits a randomized strategy which predicts its output with non-trivial success probability. This contradiction completes the proof.

We believe that our contribution is conceptual, so we do not attempt to optimize or generalize the argument. The proof of the concentration inequality is presented in Section 3, while the following section provides a brief introduction to the Weighted Majority Algorithm.

2 The Weighted Majority Algorithm.

In this section we review the Weighted Majority Algorithm and some of its properties. Consider a pool of k experts $E_1, \dots, E_k$, and a sequence of n trials. Each expert makes 0/1 predictions at each trial. Each trial event has a binary label associated to it, and we denote a sequence of trials of size n by a vector $b = (b_1, \dots, b_n) \in \{0,1\}^n$ of correct labels. The algorithm has a parameter $0 < \beta < 1$ associated to it, and it makes predictions based on the guesses of the experts as follows.

Weighted Majority Algorithm $\mathcal{A}^k_\beta$

1) Initialize weight $w_i = 1$ for each expert E_i.
2) At a given trial:

- Expert E_i predicts $z_i \in \{0,1\}$.
- Let $q_0 = \sum_{i:z_i=0} w_i$, and $q_1 = \sum_{j:z_j=1} w_j$.

3) Predict the value $y \in \{0,1\}$ such that $q_y \geq q_{1-y}$.
4) The algorithm has the following *update rule*: given the outcome $a \in \{0,1\}$ of a trial (correct label), for each i such that $z_i \neq a$, set $w_i \leftarrow w_i \cdot \beta$.

Observe that the algorithm is quite intuitive: it decreases the weight of experts that make wrong predictions, and obtain the next prediction by taking into account the guesses of the pool of experts according to their current weights. For completeness, we include the short proof that this algorithm never performs much worse than the best expert in the pool.

Theorem 2.1. *For any sequence of n trials, if the best expert among $E_1, \dots, E_k$ makes m mistakes, then Weighted Majority Algorithm with parameter β makes at most*

$$M \leq \frac{\log k + m \log \frac{1}{\beta}}{\log \frac{2}{1+\beta}}$$

mistakes.

Proof. The proof relies on the following elegant potential/energy argument. Let W_i be the total sum of the weights associated to the experts after the application of the i-th update rule. Initially, $W_0 = k$. If the algorithm makes a mistake, it is easy to see that its total weight is reduced to at most

$$\frac{W}{2} + \frac{W}{2} \cdot \beta = W \cdot \left(\frac{1+\beta}{2} \right),$$

where W is the total weight before the application of the update rule. Therefore, if the execution of the Weighted Majority Algorithm makes M mistakes, it follows that the final total weight W_n satisfies

$$W_n \leq k \cdot \left(\frac{1+\beta}{2}\right)^M .$$

However, since the best expert makes m mistakes, we must have a final total weight $W_n \geq \beta^m$. The upper bound on M now follows easily from the inequalities

$$\beta^m \leq W_n \leq k \left(\frac{1+\beta}{2}\right)^M .$$

□

Observe that this theorem provides a worst-case guarantee on the number of mistakes. It turns out that in our proof of the Chernoff-Hoeffding bound we only need an average-case guarantee. Fortunately, by making a weighted random choice for each prediction, in proportion to the current weight associated to each guess, it is possible to obtain a better (average-case) dependence on m for the number of mistakes. The proof is essentially the same, but considers instead the expected number of mistakes of the learner.[1] The interested reader is referred to the original paper of Littelestone and Warmuth [3] for further details. From now on, we let $\mathcal{A}^k_\beta$ be the randomized weighted majority algorithm with parameter β.

Fact 2.2. *For any* $0 \leq \varepsilon \leq 1/2$*, we have* $\ln\left(\frac{1}{1-\varepsilon}\right) \leq \varepsilon + \varepsilon^2$.

Theorem 2.3. *Consider any fixed sequence* $b \in \{0,1\}^n$ *of trials. If the best expert among* $E_1, \ldots, E_k$ *makes* $\leq m$ *mistakes, then*

$$M = \mathbb{E}_{\mathcal{A}^k_\beta}[\#\ \textit{mistakes of}\ \mathcal{A}^k_\beta(b)] \leq \frac{m \ln 1/\beta + \ln k}{1-\beta},$$

where the expectation if taken over the internal randomness of the randomized weighted majority algorithm. In particular, by taking $\beta = 1 - \zeta$*, it follows from* Fact 2.2 *that*

$$M \leq m(1+\zeta) + \frac{\ln k}{\zeta}.$$

[1] Needless to say, this proof is also elementary, and does not rely on any concentration bound.

3 Proof of the Concentration Bound.

We will use Theorem 2.3 to give an algorithmic proof of the following concentration result. We state and prove only one side of the concentration inequality, since we are in the symmetric case.

Theorem 3.1. *For any $\varepsilon > 0$, there exists a constant $\delta = \delta(\varepsilon) > 0$ for which the following holds. For any positive integer n, let $X = X_1 + \ldots + X_n$ be the sum of n independent uniformly distributed $0/1$ random variables. Then*

$$\mathbb{P}\left[X \geq \frac{n}{2} + \varepsilon n\right] \leq \exp(-\delta n).$$

Proof. Let $\mathbb{P}[X \geq n/2 + t] = \gamma(n,t)$, for some function $\gamma : \mathbb{N}^2 \to \mathbb{R}$. We need to upper bound the value of γ for $t = \varepsilon n$.

Consider a sequence of random independent trials $B = (B_1, \ldots, B_n)$, where each B_i is uniformly distributed over $\{0,1\}$. First, it is clear that for any randomized strategy/algorithm $\mathcal{A}$,

$$\mathbb{E}_{B,\mathcal{A}}[\#\ \textit{mistakes of } \mathcal{A} \textit{ on } B] = \sum_i \mathbb{E}_{B,\mathcal{A}}[\#\ \textit{mistakes of } \mathcal{A} \textit{ on } B_i] = \frac{n}{2}. \quad (1)$$

Now consider k experts $E_1, \ldots, E_k$, where each E_i tries to predict B_j by tossing a fair coin, and let $\mathcal{A}_\beta^k$ be the weighted majority algorithm with parameter $\beta > 0$ obtained from these experts. The values of k and β will be fixed appropriately later. Clearly,

$$\mathbb{E}_{B,\mathcal{A}_\beta^k}[\#\ \textit{mistakes of } \mathcal{A}_\beta^k] = \mathbb{E}_B[\mathbb{E}_{\mathcal{A}_\beta^k}[\#\ \textit{mistakes of } \mathcal{A}_\beta^k(B_1, \ldots, B_n)]]. \quad (2)$$

Consider the value of the inner expectation for a fixed sequence of trials $b = (b_1, \ldots, b_n) \in \{0,1\}^n$. Let Bad be the event that *no* expert E_i has $M_b^i \leq n/2 - t$, where M_b^i is a random variable representing the number of mistakes of E_i on sequence b. Since b is fixed, M_b^i has distribution $\mathsf{Binomial}(n, 1/2)$. In other words,

$$\mathbb{P}_{E_i}[M_b^i \leq n/2 - t] = \mathbb{P}_{E_i}[M_b^i \geq n/2 + t] = \gamma(n,t).$$

Therefore, $\mathbb{P}_{\mathcal{A}_\beta^k}[\mathsf{Bad}] = (1-\gamma)^k$. It follows that $\mathbb{E}_{\mathcal{A}_\beta^k}[\#\ \textit{mistakes of } \mathcal{A}_\beta^k(B)]$

$$\begin{aligned}
&= \mathbb{E}[\ \ldots \mid \mathsf{Bad}] \cdot \mathbb{P}[\mathsf{Bad}] + \mathbb{E}[\ \ldots \mid \neg\mathsf{Bad}] \cdot \mathbb{P}[\neg\mathsf{Bad}] \\
&\leq n \cdot (1-\gamma)^k + \mathbb{E}_{\mathcal{A}_\beta^k}[\#\ \textit{mistakes of } \mathcal{A}_\beta^k(B) \mid M_B^i \leq n/2 - t \textit{ for some } i] \\
&\leq n(1-\gamma)^k + (n/2 - t)(1+\zeta) + \frac{\ln k}{\zeta}, \qquad (3)
\end{aligned}$$

where the last inequality follows by Theorem 2.3 with the parametrization $\beta = 1 - \zeta$. It follows from equations (1), (2), (3) and $t = \varepsilon n$ that for any integer $k > 0$ and $\zeta \leq 1/2$, we have:

$$\frac{n}{2} \leq n(1-\gamma)^k + \left(\frac{1}{2} - \varepsilon\right)(1+\zeta)n + \frac{\ln k}{\zeta}.$$

By taking $\zeta = \zeta(\varepsilon) > 0$ to be a sufficiently small constant, we get

$$\alpha - \frac{\ln k}{\zeta n} \leq (1-\gamma)^k,$$

for some constant $\alpha(\zeta) > 0$. Observe that for $k = \exp\left(\frac{\alpha\zeta n}{2}\right)$, we have $\frac{\ln k}{n\zeta} \leq \frac{\alpha}{2}$, which implies that

$$e^{\ln\frac{\alpha}{2}} = \frac{\alpha}{2} \leq \alpha - \frac{\ln k}{n\zeta} \leq (1-\gamma)^k \leq e^{-\gamma k}.$$

This shows that $\ln\frac{\alpha}{2} \leq -\gamma k$, or equivalently,

$$\gamma \leq \frac{1}{k}\ln\frac{2}{\alpha} = \ln\left(\frac{2}{\alpha}\right)\cdot\exp\left(-\frac{\alpha\zeta n}{2}\right) \leq e^{-\delta n},$$

for some constant $\delta = \delta(\alpha, \zeta) = \delta(\varepsilon) > 0$, which completes the proof of Theorem 3.1. □

References

[1] Herman Chernoff, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, Ann. Math. Statistics **23** (1952), 493–507. MR 0057518 (15,241c)

[2] Wassily Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc. **58** (1963), 13–30. MR 0144363 (26 #1908)

[3] Nick Littlestone and Manfred K. Warmuth, *The weighted majority algorithm*, Inform. and Comput. **108** (1994), no. 2, 212–261. MR 1265851 (95d:68118)

Department of Computer Science, Columbia University.
oliveira@cs.columbia.edu