A Bennett Inequality for the Missing Mass

Bahman Yari Saeed Khanloo  bahman.khanloo@monash.edu

Abstract

Novel concentration inequalities are obtained for the missing mass, i.e. the total probability mass of the outcomes not observed in the sample. We derive distribution-free deviation bounds with sublinear exponents in deviation size for missing mass and improve the results of Berend and Kontorovich (2013) and Yari Saeed Khanloo and Haffari (2015) for small deviations which is the most important case in learning theory.

Our derivation is based on Bennett’s inequality (Bennett (1962)). Below, we provide and prove a suitable representation of this inequality.

Theorem. [Bennett] Let $Z_1, ..., Z_N$ be independent zero-mean random variables such that $|Z_i| \leq \alpha$ almost surely for all $i$. Then, Bennett’s inequality states for all $\epsilon > 0$ that:

$$P\left( \sum_{i=1}^{N} Z_i > \epsilon \right) \leq \exp \left( - \frac{V}{\alpha^2} h\left( \frac{\alpha \epsilon}{V} \right) \right),$$  

(1)

where $V = \sum_{i=1}^{N} \mathbb{E}[Z_i^2]$ and $h(u) = (1 + u) \ln(1 + u) - u$ for $u > 0$.

Now if we consider the sample average $\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i$, and let $\bar{\sigma}^2$ be the average sample variance of the $Z_i$, i.e. $\bar{\sigma}^2 := \frac{1}{n} \sum_{i=1}^{n} \text{VAR}[Z_i] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Z_i^2]$.

Using (1) with $n \cdot \epsilon$ in the role of $\epsilon$, we get

$$P(\bar{Z} > \epsilon) \leq \exp \left( - \frac{n \bar{\sigma}^2}{\alpha^2} h\left( \frac{\alpha \epsilon}{\bar{\sigma}^2} \right) \right),$$  

(2)

If $Z_i$s are not just independent but also identically distributed, then $\bar{\sigma}^2$ is equal to $\sigma^2$ i.e. the variance of $Z$.

Proof. Suppose without loss of generality that $Z_1, ..., Z_N$ are zero-mean random variables. Denoting $S_N = \sum_{i=1}^{N} Z_i$, using Chernoff’s exponential moment method one has

$$P(S_N - \mathbb{E}[S_N] \geq \epsilon) \leq e^{-\lambda \epsilon} \prod_{i=1}^{N} \mathbb{E}[e^{\lambda |Z_i - \mathbb{E}[Z_i]|}].$$  

(3)

Thus, we need to bound $\mathbb{E}[e^{\lambda Z_i}]$. Let us use the notation $\text{VAR}[Z_i] = \mathbb{E}[Z_i^2]$ to refer to individual variance and introduce

$$F_i = \sum_{r=2}^{\infty} \frac{\lambda^{r-2} \mathbb{E}[Z_i^r]}{r! \cdot \text{VAR}[Z_i]}.$$  

(4)
Using Taylor’s expansion for the function $e^{\lambda z}$, we have

$$
E[e^{\lambda Z}] = 1 + \lambda E[Z] + \sum_{r=2}^{\infty} \frac{\lambda^r E[Z]^r}{r!} 
= 1 + \lambda^2 \text{VAR}[Z] \leq e^{\lambda^2 \text{VAR}[Z]}.
$$

(5)

Therefore, we have the following upperbound on deviation probability

$$
\mathbb{P}(S_N - E[S_N] \geq \epsilon) \leq \exp\left(\frac{\lambda^\alpha - \lambda \alpha - 1}{\alpha^2 - \lambda \epsilon} \cdot V\right),
$$

(6)

which is minimized by setting

$$
\lambda^* = \frac{1}{2} \ln \left(1 + \frac{\epsilon}{V}\right).
$$

(7)

Replacing optimal $\lambda$ given by (7) in (6) yields (1) which concludes the proof. ■

1 Results

Consider the following functions

$$
\gamma_\epsilon = -W^{-1}\left(-\frac{\epsilon}{\gamma}\right),
$$

(8)

$$
c(\epsilon) = \frac{2 \ln 2 - 1}{\gamma_\epsilon}.
$$

(9)

Let $Y$ denote the missing mass, $n$ the sample size and $\epsilon$ the deviation size.

**Theorem 1.** For any $0 < \epsilon < 1$ and any $n \geq \lceil \gamma_\epsilon \rceil - 1$, we obtain the following upper deviation bound

$$
\mathbb{P}(Y - E[Y] \geq \epsilon) \leq e^{-c(\epsilon) \cdot n \epsilon}.
$$

(10)

**Theorem 2.** For any $0 < \epsilon < 1$ and any $n \geq \lceil \gamma_\epsilon \rceil - 1$, we obtain the following lower deviation bound

$$
\mathbb{P}(Y - E[Y] \leq -\epsilon) \leq e^{-c(\epsilon) \cdot n \epsilon}.
$$

(11)

2 Proof Sketch

We provide a brief outline of the proof below. The key idea of the proof is based on the thresholding technique introduced in Yari Saeed Khanloo and Haffari (2015). Here, we will explain the main steps and advise the reader to refer to the above paper for more details.

Choose $1 < \theta < n$ such that $f(\theta) = e^{-\theta} = \frac{\epsilon}{\gamma}$ and $\theta = f^{-1}(\frac{\epsilon}{\gamma}) = \ln(\frac{\epsilon}{\gamma})$ for any $0 < \epsilon < 1$ and $\epsilon \gamma < e^n \epsilon$ as generic domain for $\gamma$, we derive the upper
deviation bound for missing mass as follows

\[ P(Y - E[Y] \geq \epsilon) \leq (12) \]

\[ P(Y' - E[Y] \geq \epsilon) = (13) \]

\[ P(Y'' - E[Y] \geq (\frac{\gamma - 1}{\gamma})\epsilon) \leq (16) \]

\[ P(Y'' - E[Y'] + f(\theta) \geq \epsilon) = (15) \]

\[ P(Y'' - E[Y'] \geq (\frac{\gamma - 1}{\gamma})\epsilon) \leq (17) \]

\[ \exp \left( -\frac{V_{U'}}{\alpha_n^2} h(\frac{\alpha_n}{V_{U'}}) \right) \leq (18) \]

\[ \inf_{\gamma} \{ \exp \left( -c_0 \cdot (\frac{\gamma - 1}{\gamma}) n \epsilon \right) \} = e^{-c(\epsilon) \cdot n \epsilon}. (20) \]

Here, we have defined

\[ \epsilon' = (\frac{\gamma - 1}{\gamma}) \cdot \epsilon, \]

\[ c_0 = h(1) = 2 \ln 2 - 1, \]

\[ \gamma_\epsilon = -W_{-1} \left( -\frac{\epsilon}{\epsilon} \right), \]

\[ c(\epsilon) = \frac{c_0}{\gamma_\epsilon}. \] (21)

An upperbound for the variance proxy \( V \) can be derived as follows

\[ V_{U'} = \sum_{i: w_i \in U'} w_i^2 (1 - w_i)^n \left( 1 - (1 - w_i)^n \right) \]

\[ \leq \alpha \cdot \sum_{i: w_i \in U'} w_i (1 - w_i)^n \left( 1 - (1 - w_i)^n \right) \]

\[ \leq \alpha \cdot \sum_{i: w_i = \alpha; \sum_i w_i = 1} w_i (1 - w_i)^n \]

\[ \leq |I(\theta, n)| \cdot \left( \frac{\theta}{n} \right)^2 \cdot \left( 1 - \frac{\theta}{n} \right)^n \]

\[ \leq \frac{\theta}{n} \cdot e^{-\theta} < \frac{\theta}{n} \cdot \epsilon = \alpha \cdot \epsilon. (22) \]

To establish (20), consider \( c(\gamma, \epsilon) = \frac{c(\gamma - 1)}{\gamma \ln(\frac{\gamma}{\epsilon})} \) and let us examine the first and second derivative as follows

\[ \frac{\partial c(\gamma, \epsilon)}{\partial \gamma} = \frac{\epsilon (\ln \left( \frac{\gamma}{\epsilon} \right) - \gamma + 1)}{\gamma^2 \ln^2 \left( \frac{\gamma}{\epsilon} \right)}, (23) \]

\[ \frac{\partial^2 c(\gamma, \epsilon)}{\partial \gamma^2} = \frac{\epsilon}{\gamma^3 \ln^3 \left( \frac{\gamma}{\epsilon} \right)} \left[ -2 \ln^2 \left( \frac{\gamma}{\epsilon} \right) + (\gamma - 3) \ln \left( \frac{\gamma}{\epsilon} \right) + 2(\gamma - 1) \right]. (24) \]
Solving for the first derivative given by (23), we obtain
\[
\gamma(\epsilon) = -W_{-1}\left(-\frac{\epsilon}{e}\right).
\] (25)

Further, solving the second derivative given by (24) implies that the function \(c(\gamma, \epsilon)\) is concave with respect to \(\gamma\) for all \(\gamma \in \mathbb{R}\). Recall, moreover, that there are interrelated restrictions on \(\gamma\), \(\epsilon\) and \(n\) in derivation of (19) and (20) which are compactly expressed as
\[
\max\{e \cdot \epsilon, 1, \gamma(1)\} < \gamma < e^n, \quad n \geq \lceil \gamma \epsilon \rceil - 1.
\] (26)

The proof for lower deviation bound is identical to that of upper deviation.

References

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

Daniel Berend and Aryeh Kontorovich. On the concentration of the missing mass. *Electronic Communications in Probability*, 18:no. 3, 1–7, 2013.

Stephane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Bahman Yari Saeed Khanloo and Gholamreza Haffari. Novel bernstein-like concentration inequalities for the missing mass. *CoRR*, arXiv:1402.6262, 2015.