Concentration Inequalities 1: Hoeffding, Azuma and McDiarmid's

March 17, 2019

Concentration inequalities provide bounds on the probability with which a random variable Z deviates from its typical value (usually expectation, but we will see examples of median at some point later). Specifically we will consider the case when Y is itself a function of random variables $X_1, \ldots, X_n$. The most typical example, is when Z is the sum or average of the X's. It has several applications all over, doesn't really need more motivation, lets just proceed.

1 Markov Inequality and Cramer Chernoff Method

Our starting point is as expected Markov's inequality which states that for any non-negative random variable Y,

$$P(Y > \theta) = \mathbb{E}[\mathbb{1}\{Y > \theta\}] \leq \mathbb{E}\left[\mathbb{1}\{Y > \theta\}\frac{Y}{\theta}\right] \leq \frac{\mathbb{E}[Y]}{\theta}$$

With the Markov inequality we are ready to take a stab at proving more interesting bounds using whats typically referred to as Cramer-Chernoff method. To this end, note that for any $\lambda \geq 0$,

$$P(Z > t) = P(\exp(\lambda Z) > \exp(\lambda t)) \leq \frac{\mathbb{E}\left[e^{\lambda Z}\right]}{e^{\lambda t}}$$

The function

$$M(\lambda) = \mathbb{E}\left[e^{\lambda Z}\right] \tag{1}$$

is referred to as the moment generating function. It is one of those magical functions. Why? Write down the expansion of the exponential function and then you immediately notice that if we take the k'th derivative if $M(\lambda)$ at $\lambda = 0$ it is the k'th moment. Hence $M(\lambda)$ generates all the moments. Any case, I belabor, now also define

$$C(\lambda) = \log\left(M(\lambda)\right) = \log\left(\mathbb{E}\left[e^{\lambda Z}\right]\right) \tag{2}$$

C is referred to as cumulant generating function. Its fairly straightforward to check that C is a convex function. From the above inequality we have that:

$$P(Z > t) \leq M(\lambda)e^{-\lambda t} \leq e^{-\lambda t + C(\lambda)}$$

But the choice of $\lambda \geq 0$ is arbitrary and so let us take infimum over λ's above to get:

$$P(Z > t) \leq \inf_{\lambda \geq 0} e^{-\lambda t + C(\lambda)} = e^{-\sup_{\lambda \geq 0}\{\lambda t - C(\lambda)\}} = e^{-\sup_{\lambda}\{\lambda t - C(\lambda)\}} = e^{-C^*(t)}$$

Where in the above we dropped the $\lambda \geq 0$ because $t > 0$ and C is continuous and if Z is a zero mean random variable then $C(0) = 0$ is the minimum and so optimum is only achieved for positive λ and C^* is the Legendre transform of C. The key idea for almost all concentration inequality proof techniques is that if we upper bound C by some function g, then we get an upper bound on $P(Z > t) \leq e^{-\sup_{\lambda}\{\lambda t - g(\lambda)\}}$.

Proposition 1. *If a function $g : \mathbb{R} \mapsto \mathbb{R}$ is such that $\forall \lambda, g(\lambda) \geq C(\lambda)$, then*

$$P(Z > t) \leq e^{-\sup_{\lambda}\{\lambda t - g(\lambda)\}} = e^{-g^*(\lambda)}$$

where g^* is the Legendre transform of g.

2 Hoeffding Inequality

For this section we assume $Z = \sum_{i=1}^{n} X_i$ where each X_i's are independent identically drawn zero mean random variable bounded as $|X_i| \leq b$.

Theorem 1. *Assume $X_1, \ldots, X_n$'s are independent identically drawn zero mean random variable bounded as by b then:*

$$P\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{t^2}{8nb^2}\right)$$

Proof. Note that

$$\begin{aligned} M(\lambda) &= \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n} X_t\right)\right] \\ &= \prod_{t=1}^{n} \mathbb{E}\left[\exp\left(\lambda X_t\right)\right] \\ &= \prod_{t=1}^{n} \mathbb{E}\left[\exp\left(\lambda\left(X_t - \mathbb{E}\left[X_t'\right]\right)\right)\right] \end{aligned}$$

By Jensen's inequality:

$$\begin{aligned} &\leq \prod_{t=1}^{n} \mathbb{E}\left[\exp\left(\lambda\left(X_t - X_t'\right)\right)\right] \\ &= \prod_{t=1}^{n} \mathbb{E}\left[\frac{\exp\left(\lambda\left(X_t - X_t'\right)\right) + \exp\left(\lambda\left(X_t' - X_t\right)\right)}{2}\right] \end{aligned}$$

Since $(e^y + e^{-y})/2 \leq e^{y^2/2}$ we have:

$$\leq \prod_{t=1}^{n} \mathbb{E}\left[\exp\left(\frac{\lambda^2}{2}\left(X_t - X_t'\right)^2\right)\right]$$
$$\leq \prod_{t=1}^{n} \exp\left(2\lambda^2 b^2\right) = \exp\left(2n\lambda^2 b^2\right)$$

Hence we have that $C(\lambda) = \log(M(\lambda)) \leq 2n\lambda^2 b^2 =: g(\lambda)$. Hence from Proposition **??**, we can conclude that:

$$P(Z \geq t) \leq e^{-\sup_\lambda\{\lambda t - g(\lambda)\}} = e^{-\sup_\lambda\left\{\lambda t - 2n\lambda^2 b^2\right\}} = e^{-t^2/8nb^2}$$

□

3 Hoeffding Azuma Inequality

For this section we assume $Z = \sum_{i=1}^{n} X_i$ where each $(X_i)_{i=1}^{n}$ is a martingale difference sequence with each $|X_i| \leq b$ almost surely. A very minor edit to the previous proof yields the same inequality and is referred to as Hoeffding Azuma Inequality.

Theorem 2. *Assume $\{X_i\}$ be a martingale difference sequence w.r.t. filtration $\{\mathcal{F}_i\}$ such that each $|X_i| \leq b$ almost surely. Then we have that*

$$P\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-\frac{t^2}{8nb^2}\right)$$

Proof. Before we proceed let us first introduce the short hands $\mathbb{E}_{i-1}[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_{i-1}]$. Note the idea is to do the previous proof more carefully, peeling indices one by one:

$$\begin{aligned}
M(\lambda) &= \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n} X_t\right)\right] \\
&= \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} X_t\right) \cdot e^{\lambda X_n}\right] \\
&= \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} X_t\right) \cdot \mathbb{E}_{n-1}\left[e^{\lambda X_n}\right]\right] \\
&= \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} X_t\right) \cdot \mathbb{E}_{n-1}\left[e^{\lambda(X_n - \mathbb{E}_{n-1}[X_n'])}\right]\right]
\end{aligned}$$

where in the above, X_n' has same conditional distribution as X_n when conditioned on $\mathcal{F}_{n-1}$ and since we have a MDS, $\mathbb{E}_{n-1}[X_n'] = 0$. Now by Jensen's inequality:

$$\leq \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} X_t\right) \cdot \mathbb{E}_{n-1}\left[e^{\lambda(X_n - X_n')}\right]\right]$$

Since X_n and X'_n conditioned on $\mathcal{F}_{n-1}$ are identically distributed, $X_n - X'_n$ are conditionally symmetric and so:

$$= \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} X_t\right) \cdot \mathbb{E}_{n-1}\left[\frac{e^{\lambda(X_n - X'_n)} + e^{\lambda(X'_n - X_n)}}{2}\right]\right]$$

Since $(e^y + e^{-y})/2 \leq e^{y^2/2}$ we have:

$$\leq \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} X_t\right) \cdot \mathbb{E}_{n-1}\left[e^{\lambda^2 (X_n - X'_n)^2/2}\right]\right]$$
$$\leq \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} X_t\right)\right] \cdot \exp\left(2\lambda^2 b^2\right)$$

Now peeling similarly $n-1$ term up to 1'st index we conclude that:

$$M(\lambda) \leq \exp\left(2n\lambda^2 b^2\right)$$

Hence we have that $C(\lambda) = \log(M(\lambda)) \leq 2n\lambda^2 b^2 =: g(\lambda)$. Hence from Proposition **??**, we can conclude that:

$$P(Z \geq t) \leq e^{-\sup_\lambda\{\lambda t - g(\lambda)\}} = e^{-\sup_\lambda\left\{\lambda t - 2n\lambda^2 b^2\right\}} = e^{-t^2/8nb^2}$$

□

4 McDiarmid's Inequality: Martingale Method

Now we move from sum to more general functions of independent random variables. We will use Hoeffding Azuma as our key inequality to do this.

Corollary 1. *Let $f : \mathcal{X}^n \mapsto \mathbb{R}$ be any function such that for any $i \in [n]$ and any $x_1, \ldots, x_n \in \mathcal{X}$ and $x'_i \in \mathcal{X}$,*

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)| \leq b$$

Then, if $X_1, \ldots, X_n$ are iid $\mathcal{X}$ valued random variables, we have that

$$P\left(f(X_1, \ldots, X_n) \geq \mathbb{E}[f] + t\right) \leq \exp\left(-\frac{t^2}{8nb^2}\right)$$

Proof. This is one of those cool tricks. We define whats referred to as Doob's martingale. Define,

$$Z_i = \mathbb{E}_i[f(X_1, \ldots, X_n)] - \mathbb{E}_{i-1}[f(X_1, \ldots, X_n)]$$

Notice first that $\mathbb{E}_{i-1}[Z_t] = 0$ (i.e. Z_i's are a martingale difference sequence) and

$$\sum_{i=1}^n Z_i = \sum_{i=1}^n \left(\mathbb{E}_i[f(X_1, \ldots, X_n)] - \mathbb{E}_{i-1}[f(X_1, \ldots, X_n)]\right) = f(X_1, \ldots, X_n) - \mathbb{E}[f]$$

Hence,

$$P\left(f(X_1,\ldots,X_n) \geq \mathbb{E}\left[f\right] + t\right) = P(\sum_{i=1}^{n} Z_i > t)$$

Further note that

$$\begin{aligned}
|Z_i| &= \left|\mathbb{E}_i\left[f(X_1,\ldots,X_n)\right] - \mathbb{E}_{i-1}\left[f(X_1,\ldots,X_n)\right]\right| \\
&= \left|\mathbb{E}_{X_{i+1},\ldots,X_n}\left[f(X_1,\ldots,X_n)\right] - \mathbb{E}_{X_i,\ldots,X_n}\left[f(X_1,\ldots,X_n)\right]\right| \\
&\leq \mathbb{E}_{X_{i+1},\ldots,X_n}\left[\left|f(X_1,\ldots,X_n) - \mathbb{E}_{X_i}\left[f(X_1,\ldots,X_n)\right]\right|\right] \\
&\leq \mathbb{E}_{X_{i+1},\ldots,X_n}\left[\left|f(X_1,\ldots,X_i,\ldots,X_n) - \mathbb{E}_{X_i'}\left[f(X_1,\ldots,X_i',\ldots,X_n)\right]\right|\right] \\
&\leq \mathbb{E}_{X_i',X_{i+1},\ldots,X_n}\left[\left|f(X_1,\ldots,X_i,\ldots,X_n) - \left[f(X_1,\ldots,X_i',\ldots,X_n)\right|\right] \leq b\right.
\end{aligned}$$

Now we are ready to apply Hoeffding Azuma and hence we get,

$$P\left(f(X_1,\ldots,X_n) \geq \mathbb{E}\left[f\right] + t\right) \leq \exp\left(-\frac{t^2}{8nb^2}\right)$$

□