CSC 665: Concentration of measure

Chicheng Zhang

August 29, 2019

1 Concentration of measure

Concentration of measure, informally, states the following:

> Given a set of independently and identically distributed (iid) random variables, their empirical mean concentrates around the true mean with overwhelming probability.

One important examples is Hoeffding's Inequality, where the distribution of each random variable is supported on an interval:

Theorem 1 (Hoeffding's Inequality)**.** *Suppose that $Z_1, \ldots, Z_n$'s are iid random variables such that $a \le Z_i \le b$ for all i. Denote by $\bar{Z} := \frac{1}{n}\sum_{i=1}^n Z_i$, and $\mu = \mathbb{E}X$. Then,*

$$\mathbb{P}(|\bar{Z} - \mu| > \epsilon) \le 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}. \tag{1}$$

In other words, with probability $1-\delta$,

$$|\bar{Z} - \mu| \le (b-a) \cdot \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}. \tag{2}$$

Why is Hoeffding's Inequality relevant in machine learning theory? Consider the binary classification setup: suppose examples (x, y)'s are drawn from a distribution D. In addition, we are (magically) given a classifier $h : \mathcal{X} \to \{-1, +1\}$. We would like to know the performance of h, which is measured by its *generalization error*, i.e.

$$\mathrm{err}(h, D) \triangleq \mathbb{P}(h(x) \neq y).$$

But we only have access to the training examples $S = (x_i, y_i)_{i=1}^m$ drawn iid from D.[1] How can we measure the performance of h? We can use the *training error* of h as a proxy, denoted as

$$\mathrm{err}(h, S) \triangleq \frac{1}{m}\sum_{i=1}^m \mathbf{1}(h(x_i) \neq y_i).$$

Now, applying Hoeffding's inequality with $Z_i = \mathbf{1}(h(x_i) \neq y_i)$, $a = 0$, $b = 1$, we get that with probability $1 - \delta$,

$$|\,\mathrm{err}(h, S) - \mathrm{err}(h, D)| \le \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

This show that with high probability, the generalization error of h will be concentrated around the empirical error of h.

[1] It is important that h should be independent of S here, otherwise h might well "overfit" to S.

1.1 Chernoff bound

Note that we can apply Chebyshev's Inequality to get a bound on $\mathbb{P}(|\bar{Z} - \mu| \geq \epsilon)$. Indeed, taking $X = \bar{Z}$, $\mu = \mathbb{E}\bar{Z}$, $\mathrm{Var}(\bar{Z}) = \frac{1}{n}\mathrm{Var}(Z_1) \leq \frac{(b-a)^2)}{n}$, we have

$$\mathbb{P}(|\bar{Z} - \mu| > \epsilon) \leq \frac{(b-a)^2)}{n\epsilon^2}.$$

If we set ϵ such that right hand side to be δ, then we get $\epsilon = (b-a)\sqrt{\frac{\frac{1}{\delta}}{n}}$; that is,

$$\mathbb{P}(|\bar{Z} - \mu| > (b-a)\sqrt{\frac{\frac{1}{\delta}}{n}}) \leq \delta.$$

In other words, with probability $1 - \delta$,

$$|\bar{Z} - \mu| \leq (b-a)\sqrt{\frac{\frac{1}{\delta}}{n}}. \tag{3}$$

Now compare Equation (2) with Equation (3), with constants ignored. We can immediately see that, when δ is small, Hoeffding's Inequality implies stonger concentration of the empirical mean to the true mean - indeed, the dependency of δ is $\ln\frac{1}{\delta}$ in Hoeffding's Inequality, which can be much smaller than $\frac{1}{\delta}$.

How can we get such a stronger result? Note that applying Chebyshev's Inequality only uses the second moment of $\bar{Z}$. The proof utilizes a new tool called the *moment generating function*, which (implicitly) uses all moments of $\bar{Z}$.

Definition 1. *ϕ_X, the moment generating function of a random variable X, is defined as $\phi_X(t) \triangleq \mathbb{E}[e^{tX}]$. ψ_X, the cumulant generating function of X, is defined as $\psi_X(t) \triangleq \ln\phi_X(t) = \ln\mathbb{E}[e^{tX}]$.*

Lemma 1 (Chernoff Bound)**.** *Suppose $Z_1, \ldots, Z_n$ has a common cumulant generating function ψ_Z. Then,*

$$\mathbb{P}(\bar{Z} - \mu \geq \epsilon) \leq \exp\left\{-n\left(\sup_{t\geq 0} t(\mu+\epsilon) - \psi_Z(t)\right)\right\} = \exp\left\{-n\left(\sup_{t\in\mathbb{R}} t(\mu+\epsilon) - \psi_Z(t)\right)\right\}, \tag{4}$$

$$\mathbb{P}(\bar{Z} - \mu \leq -\epsilon) \leq \exp\left\{-n\left(\sup_{t\leq 0} t(\mu-\epsilon) - \psi_Z(t)\right)\right\} = \exp\left\{-n\left(\sup_{t\in\mathbb{R}} t(\mu-\epsilon) - \psi_Z(t)\right)\right\}. \tag{5}$$

Proof. First, observe that for any $t \geq 0$, event $\left\{\bar{Z} - \mu \geq \epsilon\right\}$ is the same as $\left\{\sum_{i=1}^n Z_i \geq n(\mu+\epsilon)\right\}$, which is contained in $\left\{\sum_{i=1}^n tZ_i \geq tn(\mu+\epsilon)\right\}$. By Markov's Inequality,

$$\mathbb{P}(\bar{Z} - \mu \geq \epsilon) \leq e^{-nt(\mu+\epsilon)}\mathbb{E}e^{\sum_{i=1}^n tZ_i}$$

Observe that

$$\mathbb{E}e^{\sum_{i=1}^n tZ_i} = \mathbb{E}\prod_{i=1}^n e^{tZ_i} = \prod_{i=1}^n \mathbb{E}e^{tZ_i} = (\phi_Z(t))^n = e^{n\psi_Z(t)}.$$

where the second equality follows from the independence of Z_i's, and the third equality uses the definition of ϕ_Z, and the last equality uses the definition of ψ_Z.

Therefore,

$$\mathbb{P}(\bar{Z} - \mu \geq \epsilon) \leq e^{-nt(\mu+\epsilon)+n\psi_Z(t)} = e^{-n(t(\mu+\epsilon)-\psi_Z(t))}.$$

As the above inequality holds for any $t \geq 0$, the inequality of Equation (4) is concluded by observing that

$$\min_{t\geq 0} \exp\{-n(t(\mu+\epsilon)-\psi_Z(t))\} = \exp\left\{-n\left(\max_{t\geq 0} t(\mu+\epsilon)-\psi_Z(t)\right)\right\}.$$

For the equality of (4), we first note that by Jensen's Inequality,

$$\phi_Z(t) = \mathbb{E}[e^{tZ}] \geq e^{t\mathbb{E}Z} = e^{t\mu}.$$

This implies that for all $t < 0$, $t(\mu+\epsilon)-\psi_Z(t) \leq t\epsilon \leq 0(\mu+\epsilon)-\psi(0)$. Therefore,

$$\max_{t\geq 0} t(\mu+\epsilon)-\psi_Z(t) = \max_{t\in\mathbb{R}} t(\mu+\epsilon)-\psi_Z(t).$$

Equation (5) follows from the exact same reasoning, and is left as an exercise. □

2 Proof of Hoeffding's Inequality

Chernoff bound (Lemma 1) gives an generic tool to bound the tail probability of the mean of a set of iid random variables: it reduces the problem to establishing properties on the moment generating function of Z. The only information we have about Z is that it has range $[a, b]$ and has mean μ. What can we say about ϕ_Z and ψ_Z?

Lemma 2. *For a random variable Z such that $Z \in [a, b]$ and $\mathbb{E}Z = \mu$, we have*

$$\phi_Z(t) \leq e^{\mu t + \frac{(b-a)^2}{8}t^2},$$

consequently, $\psi_Z(t) \leq \mu t + \frac{(b-a)^2}{8}t^2$.

Proof. First, suppose $b - a = 0$. In this case, $Z = \mu$ with probability 1, therefore the lemma statement trivially holds.

Now suppose $b - a = 1$. (We will defer the case with general settings of $b - a$ to the end of the proof.)

The trick is to write Z as a convex combination of a and b: specifically, $Z = (Z-a)\cdot b + (b-Z)\cdot a$. Note that the coefficients $(Z - a)$ and $(b - Z)$ are both nonnegative and sum to 1. Now let's look at ϕ_Z.

$$\begin{aligned}
\phi_Z(t) &= \mathbb{E}[\exp\{(Z-a)\cdot tb + (b-Z)\cdot ta\}] \\
&\leq \mathbb{E}[(Z-a)\cdot e^{tb} + (b-Z)e^{ta}] \\
&= (\mu-a)e^{tb} + (b-\mu)e^{ta}
\end{aligned}$$

Taking log on both sides, and subtracting μt on both sides, we get,

$$\psi_Z(t) - \mu t \leq \ln\Big((\mu-a)e^{tb} + (b-\mu)e^{ta}\Big).$$

Hence,

$$\psi_Z(t) - \mu_t \leq \ln\Big((\mu-a)e^{t(b-\mu)} + (b-\mu)e^{t(a-\mu)}\Big). \tag{6}$$

Now, let $p = \mu - a$, therefore, $1 - p = b - \mu$. This implies that the right hand side of Equation 6 equals $\ln\big(pe^{t-tp} + (1-p)e^{-tp}\big) =: f(t)$. Using Lemma 3 (given below), we conclude that

$$\psi_Z(t) - \mu t \leq \frac{1}{8}t^2, \tag{7}$$

which gives the lemma statement.

Now consider the case of general $b-a$. For random variable Z that takes value between a and b, $\frac{Z}{b-a}$ takes values between range $a' = \frac{a}{b-a}$ and $b' = \frac{b}{b-a}$, and has mean $\mu' = \frac{\mu}{b-a}$. Note that $b' - a' = 1$. Using Equation 7, we have that for any s,

$$e^{s\frac{Z}{b-a}} \leq \exp\left\{\mu' s + \frac{1}{8}s^2\right\},$$

Let $t = \frac{s}{b-a}$, we conclude that

$$e^{tZ} \leq \exp\left\{\mu t + \frac{(b-a)^2}{8}t^2\right\}.$$

□

Lemma 3. *Suppose $f(t) = \ln\bigl(pe^t + 1 - p\bigr) - tp$ for some $p \in [0,1]$. Then $f(t) \leq \frac{1}{8}t^2$ for all $t \in \mathbb{R}$.*

Proof. We have the following properties of f:

1. $f(0) = 0$,
2. $f'(t) = \frac{pe^t}{pe^t+1-p} - p$, and $f'(0) = 0$,
3. $f''(t) = \frac{pe^t\cdot(1-p)}{(pe^t+1-p)^2}$, and by Arithmetic Mean-Geometric Mean inequality on the numerator, $f''(t) \leq \frac{1}{4}$ for all t in $\mathbb{R}$.

Therefore, by Taylor's Theorem, for all $t \in \mathbb{R}$, there exists ξ between 0 and t, such that

$$f(t) = f(0) + f'(0)\cdot t + \frac{f''(\xi)}{2}t^2 = \frac{f''(\xi)}{2}t^2.$$

As $f''(\xi) \leq \frac{1}{4}$, we get the lemma. □

The cumulant generating function bound (Lemma 2) and Chernoff bound (Lemma 1) allows us to conclude Hoeffding's Inequality.

Proof of Theorem 1. We first show that

$$\mathbb{P}(\bar{Z} - \mu > \epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}. \tag{8}$$

Now, applying Lemma 2,

$$\begin{aligned}
\sup_{t\in\mathbb{R}} t(\mu+\epsilon) - \psi_Z(t) &\geq \sup_{t\in\mathbb{R}} (\mu+\epsilon)t - (\mu t + \frac{(b-a)^2}{8}t^2) \\
&\geq \sup_{t\in\mathbb{R}} \epsilon t - \frac{(b-a)^2}{8}t^2 \\
&= \frac{2\epsilon^2}{(b-a)^2}.
\end{aligned}$$

Plugging into Equation (4) of Chernoff bound, we get Equation (8). Symmetrically, we have

$$\mathbb{P}(\bar{Z} - \mu < -\epsilon) \leq e^{-\frac{2n\epsilon^2}{(b-a)^2}}. \tag{9}$$

Equation (1) follows from union bound, along with the fact that $\bigl\{|\bar{Z} - \mu|\bigr\}$ is the union of $\bigl\{\bar{Z} - \mu > \epsilon\bigr\}$ and $\bigl\{\bar{Z} - \mu < \epsilon\bigr\}$.

Equation (2) follows directly from Equation (1), with the setting of $\epsilon = \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}$. □

3 Chernoff bound for binomial distribution

In binary classification setup, recall that the Z_i's are drawn iid from the Bernoulli distribution of mean $p = \mathrm{err}(h, D)$. Applying Hoeffding's inequality already gives us good concentration results of $\bar{Z}$ to p. But in fact we can say more for the speiclal Bernoulli case. Formally we have the following.

Theorem 2 (Binomial Chernoff bound)**.** *Suppose $Z_1, \ldots, Z_m$ are drawn iid from the Bernoulli distribution of mean p. Then,*

$$\mathbb{P}(\bar{Z} - \mu \geq \epsilon) \leq \exp\big\{-n \,\mathrm{kl}(p + \epsilon, p)\big\},$$

$$\mathbb{P}(\bar{Z} - \mu \leq -\epsilon) \leq \exp\big\{-n \,\mathrm{kl}(p - \epsilon, p)\big\},$$

where $\mathrm{kl}(p, q) = p \ln \frac{p}{q} + (1 - p) \ln \frac{1-p}{1-q}$.

Proof. Note that

$$\psi_Z(t) = \ln \mathbb{E} e^{tZ} = \ln\big(pe^t + (1 - p)\big),$$

□