CSC 665: Concentration of measure (1)

Chicheng Zhang

August 30, 2019

1 Concentration of measure

Concentration of measure, (narrowly speaking) 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 example is Hoeffding's Inequality, where the distribution of each random variable is supported on an bounded 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, for any $\epsilon > 0$,*

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

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

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

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 is much smaller than $\frac{1}{\delta}$ for small δ.

How can Hoeffding's Inequality obtain a stronger result? Note that applying Chebyshev's Inequality only uses the second moment of $\bar{Z}$. In contrast, the proof of Hoeffding's Inequality utilizes a new tool called the *moment generating function*, which (implicitly) uses all moments of $\bar{Z}$; in addition, it takes advantage of the independence structure of all Z_i's in a clever way, as we will set next.

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 $\psi_X(t) \triangleq \ln\phi_X(t) = \ln\mathbb{E}[e^{tX}]$.* [2]

Lemma 1 (Chernoff Bound)**.** *Suppose $Z_1, \ldots, Z_n$ are iid, and have a common cumulant generating function ψ_Z. Then for any $\epsilon > 0$,*

$$\mathbb{P}(\bar{Z} - \mu \geq \epsilon) \leq \exp\left\{-n\left(\sup_{t\geq 0} t(\mu+\epsilon) - \psi_Z(t)\right)\right\} = \exp\left\{-n\left(\sup_{t\in\mathbb{R}} t(\mu+\epsilon) - \psi_Z(t)\right)\right\},^{3} \tag{4}$$

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

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 $\left\{\sum_{i=1}^n tZ_i \geq tn(\mu+\epsilon)\right\}$. Exponentiating both sides, the above event is $\left\{e^{\sum_{i=1}^n tZ_i} \geq e^{tn(\mu+\epsilon)}\right\}$.

Applying Markov's Inequality on the random variable $e^{\sum_{i=1}^n tZ_i}$, we get:

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

[2] If we write $\psi_X(t)$ as a infinite series $\sum_{n=0}^\infty a_n t^n$, then $n!a_n$ is called the n-th cumulant of X; specifically, it can be checked that the first cumulant is the mean of X and the second cumulant is the variance of X.

[3] The term $\sup_{t\in\mathbb{R}} t(\mu+\epsilon) - \psi_Z(t)$ is often written as $\psi_Z^\star(\mu+\epsilon)$; here for a function f, we denote by its Fenchel conjugate $f^\star(y) = \sup_{x\in\mathbb{R}}(xy - f(x))$. We will formally introduce this definition in future lectures.

Observe that the expectation of $e^{\sum_{i=1}^n tZ_i}$ has the following simple form:

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

where the first equality is simple algebraic manipulation, the second equality follows from the independence of Z_i's (this shows the power of exponentiation!), the third equality uses the definition of ϕ_Z, and the last equality uses the fact that $\psi_Z = \ln \phi_Z$.

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 $t < 0$, $t(\mu+\epsilon) - \psi_Z(t) \leq t\epsilon \leq 0 = 0(\mu+\epsilon) - \psi(0)$. Therefore,

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

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

2 Proof of Hoeffding's Inequality

Chernoff bound (Lemma 1) offers an generic tool to bound the tail probability of the mean of a set of iid random variables Z_i's: it reduces the problem to establishing properties on the moment generating function of each Z_i. In the condition of Hoeffding's Inequality, the only information we have about Z_i is that it has range $[a, b]$ and has mean μ. What can we say about ϕ_Z and ψ_Z? It turns out that we can say something quite nontrivial, as shown in the next lemma.

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,

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

For any t, consider $s = (b-a)t$ in the above inequality, we get

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

The lemma follows. □

Lemma 3. *Suppose $f(t) = \ln\big(pe^t + 1 - p\big) - tp$ 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

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

By Property 3 above, $f''(\xi) \leq \frac{1}{4}$, we get the lemma. □

The cumulant generating function bound on bounded random variable (Lemma 2) and Chernoff bound (Lemma 1) together allow 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,

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

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 Equations 8 and (9), along with union bound (event $\left\{|\bar{Z} - \mu| > \epsilon\right\}$ is the union of events $\left\{\bar{Z} - \mu > \epsilon\right\}$ and $\left\{\bar{Z} - \mu < \epsilon\right\}$).

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