

CSC 665: Concentration of measure (2)

Chicheng Zhang

September 3, 2019

1 Chernoff bound for Bernoulli distributions

In the binary classification setup, recall that the $Z_i = I(h(X_i) \neq Y_i)$'s are drawn iid from the Bernoulli distribution of mean $p = \text{err}(h, D)$. As seen in the last lecture, applying Hoeffding's inequality already gives us strong concentration results of \bar{Z} to p (with tail bound exponentially decreasing with sample size). But in fact we can say more for this special Bernoulli case. Formally we have the following.

Theorem 1 (Binomial Chernoff bound). *Suppose Z_1, \dots, Z_m are drawn iid from the Bernoulli distribution with mean p . Then,*

$$\begin{aligned}\mathbb{P}(\bar{Z} - \mu \geq \epsilon) &\leq \exp\{-n \text{kl}(p + \epsilon, p)\}, \\ \mathbb{P}(\bar{Z} - \mu \leq -\epsilon) &\leq \exp\{-n \text{kl}(p - \epsilon, p)\},\end{aligned}$$

where $\text{kl}(p, q) = p \ln \frac{p}{q} + (1 - p) \ln \frac{1-p}{1-q}$ is the binary relative entropy.

Before going into the proof of the theorem, let us see several important consequences of the theorem.

1. As we have already seen in the calibration homework, for any $q \in [0, 1]$,

$$\text{kl}(q, p) \geq 2(q - p)^2.$$

This implies that both $\mathbb{P}(\bar{Z} - \mu \geq \epsilon)$ and $\mathbb{P}(\bar{Z} - \mu \leq -\epsilon)$ are at most $e^{-2n\epsilon^2}$. Notice that this is exactly what Hoeffding's Inequality implies for Bernoulli random variables.

2. Another fact we proved in the calibration homework is that $\text{kl}(q, p) \geq \frac{(q-p)^2}{2 \max(p, q)}$. Fix $\mu \in [0, 1]$, and let $\epsilon = \mu p$. We get that

$$\begin{aligned}\text{kl}(p(1 + \mu), p) &\geq \frac{\mu^2 p^2}{2(1 + \mu)p} \geq \frac{\mu p^2}{4}, \\ \text{kl}(p(1 - \mu), p) &\geq \frac{\mu^2 p^2}{2(1 - \mu)p} \geq \frac{\mu p^2}{4}.\end{aligned}$$

This implies that both $\mathbb{P}(\bar{Z} \geq p(1 + \mu))$ and $\mathbb{P}(\bar{Z} \leq p(1 - \mu))$ are at most $e^{-\frac{n\mu p^2}{4}}$.¹ This is oftentimes called a relative (or multiplicative) Chernoff bound for Bernoulli random variables (as it considers the ratio between empirical frequency and true mean), and is much tighter than Hoeffding's Inequality when p is small.

¹The constants in the exponents are by no means tight; in fact bounds with better constants (1/3 for the upper tail and 1/2 for the lower tail) can be found in the literature. However, in learning theory the constants are of secondary importance; the asymptotic orders of the convergence rates are often quantities of interest.

Proof. Using the Chernoff bound for general random variables (see Lemma 1 from our last note), it suffices to show that for any q in $[0, 1]$, $\sup_{t \in \mathbb{R}} (tq - \psi_Z(t)) = \text{kl}(q, p)$, where ψ_Z is the common cumulant generating function of all Z_i 's.

First, let us compute ψ_Z . As Z_i 's take value 1 with probability p and take value 0 with probability $1 - p$, ψ_Z has a closed form:

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

Now let $F(t) = tq - \psi_Z(t)$. Our goal is to show that $\sup_{t \in \mathbb{R}} F(t) = \text{kl}(q, p)$. Taking derivative of F with respect to t , we get that

$$F'(t) = q - \frac{pe^t}{(1 - p) + pe^t}.$$

Setting $F'(t) = 0$, we get that $t^* = \ln \frac{q(1-p)}{p(1-q)}$ is the only critical point of F . It can be readily checked that $F'(t) > 0$ if $t < t^*$, and $F'(t) \leq 0$ if $t > t^*$. Hence, t^* is the unique maximum of F , and

$$\sup_{t \in \mathbb{R}} F(t) = F(t^*) = qt^* - \ln(pe^{t^*} + (1 - p)) = \text{kl}(q, p). \quad \square$$

2 McDiarmid's Inequality

So far, we have seen concentration inequalities for averages of iid random variables. In this section, we go one step further: we consider general functions of iid random variables. Denote by f the function of interest, which takes into input x_1, \dots, x_n and outputs a real number $f(x_1, \dots, x_n)$. As long as f is not too sensitive on all its inputs (formally defined below), a random evaluation on f , i.e. $f(X_1, \dots, X_n)$, will be close to its expectation $\mathbb{E}f(X_1, \dots, X_n)$.

Definition 1 (Sensitivity). *Suppose f is a function from V^n to \mathbb{R} . f is called c -sensitive, if for every i in $\{1, \dots, n\}$, every x_1, \dots, x_n, x_i in V ,*

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

This property is also called *bounded difference*: suppose f have an input x_1, \dots, x_n , and we replace the i -th input with an arbitrary value x'_i , then the output of f only changes by c . Intuitively, if c is smaller, then f is more well-behaved.

Theorem 2 (McDiarmid's Inequality). *Suppose f is a function from V^n to \mathbb{R} that is c -sensitive. In addition, suppose X_1, \dots, X_n are iid random variables that take values in V . Then,*

$$\mathbb{P}(|f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n)| \geq \epsilon) \leq 2e^{-\frac{2\epsilon^2}{nc^2}}$$

Observe that Hoeffding's Inequality is a special case of McDiarmid's Inequality: Suppose $\{X_i\}_{i=1}^n$ are iid random variables that take values in $V = [a, b]$, with mean μ . We let $f(x_1, \dots, x_n) = \frac{\sum_{i=1}^n x_i}{n}$ be the empirical mean function. Note that f is $\frac{b-a}{n}$ -sensitive, moreover, $\mathbb{E}f(x_1, \dots, x_n) = \mu$.

This implies that

$$\mathbb{P}(|\bar{X} - \mu| \geq \epsilon) \leq 2 \exp\left\{-\frac{2\epsilon^2}{n \cdot \frac{(b-a)^2}{n^2}}\right\} = 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Proof. We will still consider the moment generating function of $f(X_1, \dots, X_n)$. However, we cannot directly apply Chernoff bound this time, as Chernoff bound only applies to the mean of a set of iid random variables.

We have the following key claim.

Claim 1. *For all t in \mathbb{R} ,*

$$\mathbb{E} \exp\{t(f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n))\} \leq \exp\left\{n \frac{c^2 t^2}{8}\right\}.$$

To see how the claim concludes the proof, we note that for all $t > 0$,

$$\begin{aligned} \mathbb{P}(f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n) \geq \epsilon) &= \mathbb{P}(\exp\{t(f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n))\} \geq \exp\{t\epsilon\}) \\ &\leq \mathbb{E} \exp\{t(f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n))\} \exp\{-t\epsilon\} \\ &\leq \exp\left\{n \frac{c^2 t^2}{8} - t\epsilon\right\} \end{aligned}$$

where the first inequality is Markov's Inequality, the second inequality is from Claim 1. Now, pick $t = \frac{4\epsilon}{nc^2} > 0$, we get that $\mathbb{P}(f(X_1, \dots, X_n) - \mathbb{E}f(X_1, \dots, X_n) \geq \epsilon) \leq e^{-\frac{2n\epsilon^2}{c^2}}$. The theorem follows from establishing the lower tail bound similarly, along with union bound. \square

Proof of Claim 1. We first setup some useful notation. We denote by f_n the original function f of n variables, and denote by f_0 the constant $\mathbb{E}f(X_1, \dots, X_n)$.

In addition, denote by f_{n-1} the function of $(n-1)$ variables, such that

$$f_{n-1}(x_1, \dots, x_{n-1}) = \mathbb{E}[f_n(X_1, \dots, X_n) | X_1 = x_1, \dots, X_{n-1} = x_{n-1}].$$

In other words, $f_{n-1}(x_1, \dots, x_{n-1})$ is the expectation of the output of f , given that the first $(n-1)$ -th inputs observed are x_1, \dots, x_{n-1} . Suppose that every x_i has a probability density function p , f_{n-1} has the following explicit form:

$$f_{n-1}(x_1, \dots, x_{n-1}) = \int_V f_n(x_1, \dots, x_{n-1}, x_n) p(x_n) dx_n.$$

We have the following important properties of f_{n-1} :

1. $\mathbb{E}f_{n-1}(X_1, \dots, X_{n-1}) = \int_{V^n} f_n(x_1, \dots, x_{n-1}, x_n) p(x_1) \dots p(x_n) dx_1 \dots dx_n = f_0$.
2. It can be checked that f_{n-1} is also c -sensitive. For example, consider changing the first coordinate from x_1 to x'_1 :

$$\begin{aligned} &|f_{n-1}(x_1, x_2, \dots, x_{n-1}) - f_{n-1}(x'_1, x_2, \dots, x_{n-1})| \\ &= \int_V (f_n(x_1, \dots, x_{n-1}, x_n) - f_n(x'_1, \dots, x_{n-1}, x_n)) p(x_n) dx_n \\ &\leq \int_V |f_n(x_1, \dots, x_{n-1}, x_n) - f_n(x'_1, \dots, x_{n-1}, x_n)| p(x_n) dx_n \\ &\leq \int_V c p(x_n) dx_n = c. \end{aligned}$$

We will show that for all t in \mathbb{R} ,

$$\mathbb{E} \exp\{t(f_n(X_1, \dots, X_n) - f_0)\} \leq \mathbb{E} \exp\{t(f_{n-1}(X_1, \dots, X_{n-1}) - f_0)\} \cdot \exp\left\{\frac{c^2 t^2}{8}\right\}. \quad (1)$$

To see why this implies the claim, we note that we can apply the same inequality again on $f_{n-1}(X_1, \dots, X_{n-1})$ and define function f_{n-2} similarly as before, getting

$$\mathbb{E} \exp\{t(f_{n-1}(X_1, \dots, X_{n-1}))\} \leq \mathbb{E} \exp\{t(f_{n-2}(X_1, \dots, X_{n-2}))\} \exp\left\{\frac{c^2 t^2}{8}\right\},$$

Repeatedly applying Equation (1) (with appropriate definitions of functions f_{n-i} 's), we get

$$\begin{aligned}
\mathbb{E} \exp\{t(f_n(X_1, \dots, X_n) - f_0)\} &\leq \mathbb{E} \exp\{t(f_{n-1}(X_1, \dots, X_{n-1}) - f_0)\} \cdot \exp\left\{1 \cdot \frac{c^2 t^2}{8}\right\} \\
&\leq \mathbb{E} \exp\{t(f_{n-2}(X_1, \dots, X_{n-2}) - f_0)\} \cdot \exp\left\{2 \cdot \frac{c^2 t^2}{8}\right\} \\
&\leq \dots \\
&\leq \mathbb{E} \exp\{t(f_1(X_1) - f_0)\} \cdot \exp\left\{(n-1) \cdot \frac{c^2 t^2}{8}\right\} \\
&\leq \mathbb{E} \exp\{t(f_0 - f_0)\} \cdot \exp\left\{n \cdot \frac{c^2 t^2}{8}\right\} = \exp\left\{n \cdot \frac{c^2 t^2}{8}\right\},
\end{aligned}$$

where the i -th inequality is by Equation (1) on f_{n-i+1} and the fact that f_{n-i+1} is c -sensitive.

Back to the proof of Equation (1). We first write down the left hand side explicitly:

$$\begin{aligned}
&\mathbb{E} \exp\{t(f_n(X_1, \dots, X_n) - f_0)\} \\
&= \mathbb{E} \exp\left\{t(f_n(X_1, \dots, X_n) - f_{n-1}(X_1, \dots, X_{n-1})) + t(f_{n-1}(X_1, \dots, X_{n-1}) - f_0)\right\} \\
&= \int_V \dots \int_V \exp\left\{t(f_n(x_1, \dots, x_n) - f_{n-1}(x_1, \dots, x_{n-1})) + t(f_{n-1}(x_1, \dots, x_{n-1}) - f_0)\right\} p(x_1) \dots p(x_n) dx_1 \dots dx_n \\
&= \int_V \dots \int_V p(x_1) \dots p(x_{n-1}) dx_1 \dots dx_{n-1} \exp\left\{t(f_{n-1}(x_1, \dots, x_{n-1}) - f_0)\right\} g(x_1, \dots, x_{n-1}), \tag{2}
\end{aligned}$$

where $g(x_1, \dots, x_{n-1}) \triangleq \int_V \exp\left\{t(f_n(x_1, \dots, x_n) - f_{n-1}(x_1, \dots, x_{n-1}))\right\} p(x_n) dx_n$, and the last equality is by reducing the multiple integral to an iterated integral.

Suppose for the moment that x_1, \dots, x_{n-1} are fixed numbers, and only X_n is random. Consider a random variable $Z = f_n(x_1, \dots, x_{n-1}, X_n)$. Note that Z takes values from interval $[a, b]$, where $a = \min_{x_n \in V} f_n(x_1, \dots, x_{n-1}, x_n)$ and $b = \max_{x_n \in V} f_n(x_1, \dots, x_{n-1}, x_n)$. Observe that $b - a \leq c$ as f_n is c -sensitive. By Lemma 2 in the last note (mgf bound for Hoeffding's Inequality),

$$\mathbb{E} e^{t(Z - \mathbb{E}Z)} \leq e^{\frac{(b-a)^2 t^2}{8}}.$$

Written in integral form, the above is

$$\int_V \exp\left\{t(f_n(x_1, \dots, x_n) - f_{n-1}(x_1, \dots, x_{n-1}))\right\} p(x_n) dx_n \leq e^{\frac{(b-a)^2 t^2}{8}},$$

i.e. $g(x_1, \dots, x_{n-1}) \leq e^{\frac{(b-a)^2 t^2}{8}}$. Plugging this inequality into Equation (2), we get

$$\begin{aligned}
&\mathbb{E} \exp\{t(f_n(X_1, \dots, X_n) - f_0)\} \\
&\leq \int_V \dots \int_V p(x_1) \dots p(x_{n-1}) dx_1 \dots dx_{n-1} \exp\left\{t(f_{n-1}(x_1, \dots, x_{n-1}) - f_0)\right\} \cdot e^{\frac{(b-a)^2 t^2}{8}} \\
&= \mathbb{E} \exp\left\{t(f_{n-1}(X_1, \dots, X_{n-1}) - f_0)\right\} \cdot e^{\frac{(b-a)^2 t^2}{8}}.
\end{aligned}$$

This concludes the proof of Equation (1), and the proof of the claim. □

Remark. For readers that are familiar with conditional expectation notation, the notation in the proof of Equation (1) can be simplified a bit. Specifically,

$$\begin{aligned}
& \mathbb{E} \exp\{t(f_n(X_1, \dots, X_n) - f_0)\} \\
= & \mathbb{E} \exp\left\{t(f_n(X_1, \dots, X_n) - f_{n-1}(X_1, \dots, X_{n-1})) + t(f_{n-1}(X_1, \dots, X_{n-1}) - f_0)\right\} \\
= & \mathbb{E} \left[\exp\left\{t(f_{n-1}(X_1, \dots, X_{n-1}) - f_0)\right\} \cdot \mathbb{E} \left[\exp\left\{t(f_n(X_1, \dots, X_n) - f_{n-1}(X_1, \dots, X_{n-1}))\right\} \middle| X_1, \dots, X_{n-1} \right] \right] \\
\leq & \mathbb{E} \left[\exp\left\{t(f_{n-1}(X_1, \dots, X_{n-1}) - f_0)\right\} \cdot e^{\frac{(b-a)^2 t^2}{8}} \right] \\
= & \mathbb{E} \left[\exp\left\{t(f_{n-1}(X_1, \dots, X_{n-1}) - f_0)\right\} \right] \cdot e^{\frac{(b-a)^2 t^2}{8}}.
\end{aligned}$$