lions@epfl

EE-731: ADVANCED TOPICS IN DATA SCIENCES
LABORATORY FOR INFORMATION AND INFERENCE SYSTEMS
SPRING 2016

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

INSTRUCTOR: VOLKAN CEVHER

SCRIBER: HARSHIT GUPTA

PROBABILITY IN HIGH DIMENSIONS

Index terms: Concentration inequalities, Bounded difference function, Cramer-Chernoff bound, Johnson-Lindenstrauss Theorem, Hoeffding's Lemma, Sub-Gaussian random variable, Entropy, Herbst's trick

1 Introduction

Measuring the concentration of a random variable around some value has many applications in statistics, information theory, optimization, etc. In probability theory, this concentration of measures can be quantified as :

Definition 1. *Given a random variable Y and a constant m, Y is said to be concentrated around m if,*

$$P(|Y - m| > t) \leq D(t)$$

where, $D(t)$ decreases drastically to 0 in t.

These inequalities are called concentration of measure inequalities [1] [2].

Example 1.1. A simple example can be considered when $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, where X_i are independent random variables with mean μ and variance σ^2, then Y_n tends to concentrate around μ as $n \to \infty$. The concentration in this case can be quantified by following results:

- Law of Large Numbers : $P(|Y - m| > \epsilon) \to 0$ as $n \to \infty$.
- Central Limit Theorem : $P(|Y - m| > \frac{\alpha}{\sqrt{n}}) \to 2\Phi\left(-\frac{\alpha}{\sigma}\right)$ as $n \to \infty$, where Φ is the standard normal CDF.
- Large Deviations : Under some technical assumptions, $P(|Y - m| > \epsilon) \leq \exp(-n.c(\epsilon))$

It is quite interesting to note that these concentration properties on the average of the independent random variables hold for the more generalized scenario where a *not too sensitive* function of independent random variables concentrates on its expectation, formally:

Proposition 1.1. *If $x_1, \cdots, x_n$ are independent random variables then any function $f(x_1, \cdots, x_n)$ that is, not too sensitive to any of the co-ordinate will concentrate around its mean:*

$$P(|f(x_1, \cdots, x_n) - E[f(x_1, \cdots, x_n)]| > t) \leq e^{-t^2/c(t)}.$$

Here, $c(t)$ quantifies the sensitivity of the function to its variables.

2 Concentration Inequalities

Now we define several types of functions and the concentration inequalities:

2.1 Bounded Difference Function

Definition 2. *A function $f : \mathcal{X}^n \to \mathcal{R}$ is called a bounded difference function if for some non-negative $c_1, \cdots, c_n$,*

$$\sup_{\{x_1,\ldots,x_i,\ldots,x_n,x_i' \in \mathcal{X}\}} |f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i.$$

Example 2.1 (Chromatic number of a Random Graph). Let $V = \{1, \cdots, n\}$ and G be a random graph such that each pair $i, j \in V$ is independently connected with probability p. Let $X_{ij} = 1$ if (i, j) are connected and 0 otherwise. The **Chromatic number** of G is the number of colors needed to color the vertices such that no two connected vertices have the same color. Defining:

$$\text{Chromatic number} = f(X_{11}, \cdots, X_{ij}, \cdots, X_{nn})$$

it can be shown that f is a bounded difference function with $C_{ij} = 1$.

Theorem 2.1 (Bounded difference inequality)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}$ satisfy the bounded difference property with c_i's and let $X_1, \cdots, X_n$ be independent random variables. Then:*

$$P(|f(x_1, \cdots, x_n) - E[f(x_1, \cdots, x_n)]| > t) \leq 2 \exp\left(\frac{-2t^2}{\sum_{i=1}^N c_i^2}\right).$$

Proof. To prove this result we need to exploit some common probability bounds. As these bounds are essential to prove the inequality, therefore, we discuss them in details with examples, meanwhile, keeping the flow of the proof:

- **Cramer Chernoff Bound**

  - **Markov's Inequality** We start by stating Markov's inequality and its proof:

    Theorem 2.2 (**Markov's Inequality**)**.** *Let Z be a non-negative random variable then, $P(Z \geq t) \leq \frac{\mathbb{E}[Z]}{t}$.*

    Proof. This can be proven by showing that:

    $$\int_t^\infty P_Z(z)\, dz \leq \int_t^\infty \frac{z}{t} P_Z(z)\, dz \leq \int_0^\infty \frac{z}{t} P_Z(z)\, dz = \frac{\mathbb{E}[Z]}{t}.$$

    ∎

  - **Chebyshev's Inequality** We extend this result to any non-decreasing and non-negative function φ of the non-negative random varibale Z then,

    $$P(Z \geq t) \leq P(\varphi(Z) \geq \varphi(t)) \leq \frac{\mathbb{E}[\varphi(Z)]}{\varphi(t)}.$$

    On choosing $\varphi(Z) = Z^2$, and substituting $|Z - \mathbb{E}[Z]|$ into Z, we get the Chebyshev's inequality:

    $$P(|Z - \mathbb{E}[Z]| \geq t) \leq \frac{Var[Z]}{t^2}.$$

  - **Chernoff Bound** Similarly, by choosing $\varphi(Z) = e^{\lambda Z}$ where, $\lambda \geq 0$ and taking $\psi_Z(\lambda) = \log^{\lambda \mathbb{Z}}$, we get the Chernoff bound:

    $$P(Z \geq t) \leq \inf_{\lambda \geq 0} e^{-\lambda t} \mathbb{E}[e^{\lambda Z}] = \mathbb{E}\left[\exp\left(-\sup_{\lambda \geq 0}(\lambda t - \psi_Z(\lambda))\right)\right].$$

– **Cramer-Transform** To get the Cramer-Chernoff inequality from this Chernoff bound we first define Cramer Transform of the random variable Z,:

$$\psi_Z^*(t) = \sup_{\lambda \geq 0} \lambda t - \psi_Z(\lambda)$$

Thus, by putting the Cramer transform of Z in the Chernoff bound we derive:

$$P(Z \geq t) \leq \exp\left(-\psi_Z^*(t)\right)$$

Example 2.2. To illustrate the power of these bounds we apply them to the case when $Z = X_1 + \cdots + X_n$ where X_i are i.i.d. 's (independent and identical distributions). On applying Chebyshev's inequality to the sum and acknowledging that $Var[Z] = nVar[X]$, and putting $t = n\epsilon$, we obtain,

$$P\left(\frac{1}{n}|Z - \mathbb{E}[Z]| \geq \epsilon\right) \leq \frac{Var[X]}{n\epsilon^2}.$$

Whereas, by first deriving that:

$$\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}] = \log \mathbb{E}[e^{\lambda \sum_{i=1}^n X_i}] = \log \mathbb{E}[\Pi_{i=1}^n e^{\lambda X_i}] = \log \Pi_{i=1}^n \mathbb{E}[e^{\lambda X_i}] = \log(\mathbb{E}[e^{\lambda X}])^n = n\psi_X(\lambda)$$

using i.i.d. property of the $X_i s$ and putting this in the Cramer-Chernoff inequality on the sum we get:

$$P(Z \geq n\epsilon) \leq \exp\left(-n\psi_X^*(\epsilon)\right).$$

We also use the Cramer-Chernoff inequality on any centered random variable $Y = X - \mathbb{E}[X]$ and its negative $Y^- = -Y$:

$$P(Y \geq t) \leq \exp\left(-\psi_Y^*(t)\right), P(-Y \geq t) \leq \exp\left(-\psi_{-Y(t)}^*\right),$$

adding the two:

$$P(|Y| \geq t) \leq \exp\left(-\psi_Y^*(t)\right) + \exp\left(-\psi_{Y-(t)}^*\right).$$

In case the probability distribution of Y is symmetric around its mean. Then $\psi_{-Y}^*(t) = \psi_Y^*(t)$, and thus,

$$P(|Y| \geq t) \leq 2\exp\left(-\psi_{Y(t)}^*\right).$$

For example, for a random variable $Y = \mathcal{N}(0, \sigma^2)$, $\psi_Y(\lambda) = \frac{\lambda^2\sigma^2}{2}$ and $\psi_Y^*(t) = \frac{t^2}{2\sigma^2}$ and hence,

$$P(|Y| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}},$$

which means that the Gaussian random variables concentrate around their mean as $\sigma \to 0$.

- **Sub-Gaussian Random Variables** : To use the discussed inequalities to prove the concentration of measure for the bounded difference functions we describe the larger family of sub-Gaussian random variables to which they belong:

Definition 3 (Sub-Gaussian Random Variable). *If a centered random variable Z is such that $\psi_Z(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$, for all non-negative λ then it is called a sub-Gaussian random variable with parameter σ^2. The family of all such random variables with the parameter σ^2 is denoted by $\mathcal{G}(\sigma^2)$.*

Interestingly, when $\psi_Z(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$, then $\psi_Z^*(t) \geq \frac{t^2}{2\sigma^2}$ and $\psi_{-Z}^*(t) \geq \frac{t^2}{2\sigma^2}$, which brings us to the properties of the sub-Gaussian random variables:

– If $Z \in \mathcal{G}(\sigma^2)$ then, $P(|Z| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$ i.e. Sub-Gaussian random variables concentrate around their mean.

- If $Z_i \in \mathcal{G}(\sigma^2)$ are independent, then $\sum_{i=1}^n a_i X_i \in \mathcal{G}(\sum_{i=1}^n a_i^2 \sigma^2)$ i.e. linear combination of sub-Gaussian random variables is sub-Gaussian as well.

- **Hoeffding's Lemma** The result then can be used to state Hoeffding's Lemma:

  Lemma 2.3 (Hoeffding's Lemma)**.** *Let Z be a random variable with $\mathbb{E}[Z] = 0$, and bounded in $[a, b]$ then, $\psi_Z(\lambda) \leq \frac{\lambda^2}{2} \cdot \frac{(b-a)^2}{4}$ and thus, $Z \in \mathcal{G}((b-a)^2/4)$.*

  The details of the lemma are discussed later.

- Applying property 1 of the sub-Gaussian random variables on $Z = Y - \mathbb{E}[Y]$, we get,

$$P(|Y - \mathbb{E}[Y]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{(b-a)^2}\right),$$

  We then use the property 2 of the sub-Gaussian random variables to prove that $Z = Z_1 + \cdots + Z_n$ (where $Z_i = Y_i - \mathbb{E}[Y_i]$ and is bounded in $[a_i, b_i]$) is sub-Gaussian with parameter $\sum_{i=1}^n (b_i - a_i)^2/4$. Then, substituting $t = n\epsilon$ and using the result obtained we get :

$$P\left(\frac{1}{n}|Z - \mathbb{E}[Z]| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2n\epsilon^2}{\frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2}\right).$$

■

We state 2 more examples of concentration inequalities :

Theorem 2.4 (Lipschitz Function of Gaussian RVs)**.** *Let $X_1, \cdots, X_n$ be independent random variables with distribution $\mathcal{N}(0, 1)$ and let f be an $L-$Lipschitz i.e. ($|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq L\|\boldsymbol{x} - \boldsymbol{x}'\|_2$ for any $\boldsymbol{x}, \boldsymbol{x}'$). Then,*

$$P\left(|f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)]|\right) \leq 2 \exp\left(-\frac{t^2}{2L^2}\right).$$

Theorem 2.5. *Let $X_1, \cdots, X_n$ be independent random variables bounded in $[0, 1]$, and let $f : [0, 1]^n \to \mathbb{R}$ be $1-$Lipschitz and separately convex (i.e. convex in any given co-ordinate when the other ones are fixed). Then,*

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

3 Examples

We now discuss some applications of these concentration inequalities:

Example 3.1. **PAC Learnability**:
As discussed in the previous lecture, a hypothesis class $\mathcal{H}$ has the uniform convergence property, if there exists a function $n_{\mathcal{H}}(\epsilon, \delta)$, such that for every $\epsilon, \delta \in [0, 1]$ and any probability distribution $\mathbb{P}$, if $n \geq n_{\mathcal{H}}(\epsilon, \delta)$, we have

$$\sup_{h \in \mathcal{H}} |\hat{F}_n(h) - F(h)| \leq \epsilon,$$

with probability at least $1 - \delta$, where, $\hat{F}_n(h)$ is the empirical risk and $F(h)$ is the risk.
Given that the hypothesis class $\mathcal{H}$ includes a finite number of functions $f(h, \cdot)$ bounded in [0,1]. Then, $\mathcal{H}$ satisfies the uniform convergence property with

$$n_{\mathcal{H}}(\epsilon, \delta) = \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}.$$

Proof. Let $\xi(h) = f(h, x_i)$, and $S_n(h) = \frac{1}{n}\sum_{i=1}^{n}(\xi_i(h) - \mathbb{E}\xi_i(h))$. Then,

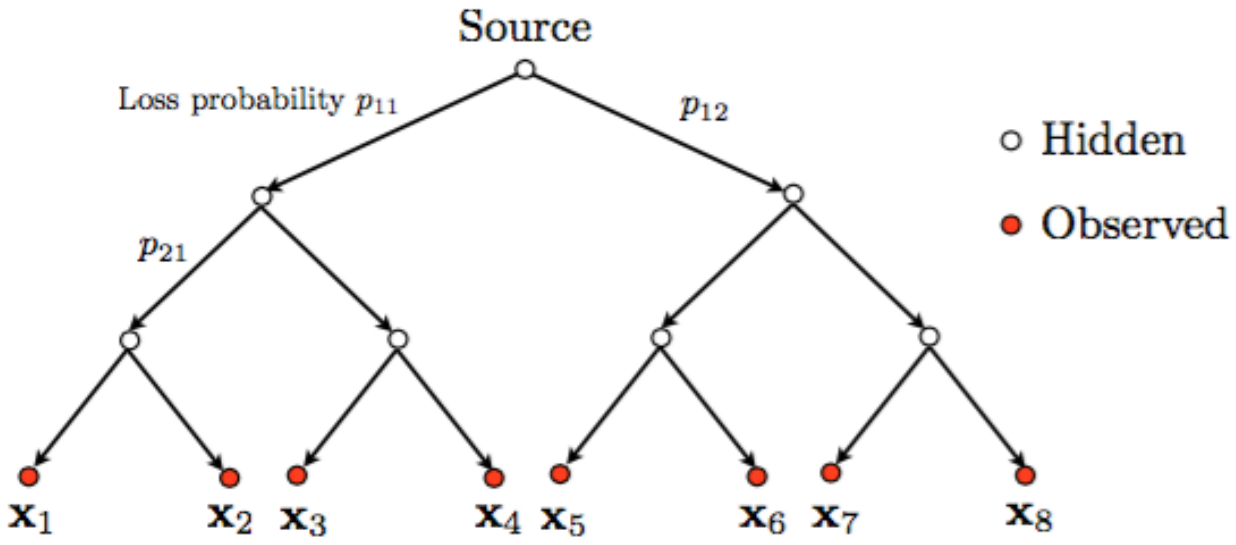
$$\sup_{h\in\mathcal{H}} |S_n(h)| = \sup_{h\in\mathcal{H}} |\hat{F}_n(h) - F(h)|.$$

As $f(h, x_i)$ are bounded in [0,1], therefore, they satisfy the bounding difference inequality such that,:

$$\sup_{h\in\mathcal{H}} \mathbb{P}(|S_n(h)| \geq \epsilon) \leq \sum_{h\in\mathcal{H}} \mathbb{P}(|S_n(h)| \geq \epsilon) \leq |\mathcal{H}|.2e^{-2n\epsilon^2} \leq \delta$$

for $n \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$. Here, $|\mathcal{H}|$ is the cardinality of the set $\mathcal{H}$. ■

Example 3.2. **Network Tomography**
The problem as shown in Fig 1. is to reconstruct the tree structure given n packets and p leaf nodes and the information $X_k^{(i)} = \mathbf{1}$ {Packet i arrives at the node k} for the n independent samples:
Solution: In [3], it has been shown that the tree structure can be recovered from the information q_{kl} = P(packet reaches



x_k and x_l). We approximate q_{kl} by the ensemble average $\hat{q}_{kl} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_k^{(i)} = 1 \cap X_l^{(i)} = 1\}$. But the approximation is feasible only if the system is robust i.e. $|\hat{q}_{kl} - q_{kl}| \leq \epsilon$ for any pair k, l. This can be achieved with probability $1 - \delta$ if $n \geq \frac{1}{2\epsilon^2}\log\frac{p^2}{\delta}$.

Proof. Using Hoeffding's inequality it is easy to see that:

$$P(|\hat{q}_{kl} - q_{kl}| \geq \epsilon) \leq 2\exp\left(-2n\epsilon^2\right)$$

To get the upper bound of this for any k and l,

$$\sup_{k,l\in n\times n} P(|\hat{q}_{kl} - q_{kl}| \geq \epsilon) \leq \sum_{k,l\in n\times n} P(|\hat{q}_{kl} - q_{kl}| \geq \epsilon) \leq 2.\frac{p^2}{2}\exp(-2n\epsilon^2) \leq \delta$$

Thus, the P(error)$\leq \delta$ if $n \geq \frac{1}{2\epsilon^2}\log\frac{p^2}{\delta}$.

■

Example 3.3. **Random Linear Projections**:
This is also called as Johnson-Lindenstrauss Theorem [4],

Theorem 3.1 (Johnson-Lindenstrauss Theorem). *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p$ be a collection of points in $\mathbb{R}^d$, and $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a random Gaussian matrix with $\mathcal{N}(0, \frac{1}{\sqrt{N}})$ distribution, then, for any pair $\epsilon, \delta \in (0,1)$ if $n \geq \frac{4}{\epsilon^2(1-\epsilon)} \log \frac{p^2}{\delta}$, with probability $1-\delta$ following is satisfied:*

$$(1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \leq \|\mathbf{A}\boldsymbol{x}_i - \mathbf{A}\boldsymbol{x}_j\|_2^2 \leq (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \tag{1}$$

Proof. Lets focus on $\mathbb{E}[\|\mathbf{A}\boldsymbol{u}\|_2^2]$, we observe that:

$$\mathbb{E}[\|\mathbf{A}\boldsymbol{u}\|_2^2] = \boldsymbol{u}^{\boldsymbol{T}}\mathbb{E}[\mathbf{A}^{\mathbf{T}}\mathbf{A}]\boldsymbol{u} = \boldsymbol{u}^{\boldsymbol{T}}\mathbf{I}\boldsymbol{u} = \|\boldsymbol{u}\|_2^2.$$

Also note that $Z_j = [\mathbf{A}\boldsymbol{x}]_j/\|\boldsymbol{x}\|$ is distributed as $\mathcal{N}(0,1)$ and Z_j are independent. Thus, $\|\mathbf{A}\boldsymbol{x}\|^2/\|\boldsymbol{x}\| = \sum_{i=1}^n Z_j^2 = \chi^2$ (where χ^2 is the chi-squared distribution with n degrees of freedom). We then use the squared-Gaussian concentration (Chapter 2, [5]) to show that for any $\boldsymbol{u}$,

$$P\left(\|\mathbf{A}\boldsymbol{u}\|_2^2 \geq (1+\epsilon)\|\boldsymbol{u}\|_2^2\right) \leq \exp\left(-\frac{n}{4}\epsilon\left(1-\epsilon^2\right)\right)$$

$$P\left(\|\mathbf{A}\boldsymbol{u}\|_2^2 \leq (1-\epsilon)\|\boldsymbol{u}\|_2^2\right) \leq \exp\left(-\frac{n}{4}\epsilon\left(1-\epsilon^2\right)\right)$$

implying that :

$$P\left(|\|\mathbf{A}\boldsymbol{u}\|_2^2 - \|\boldsymbol{u}\|_2^2| \geq \epsilon\|\boldsymbol{u}\|_2^2\right) \leq 2\exp\left(-\frac{n}{4}\epsilon\left(1-\epsilon^2\right)\right)$$

To get the upper bound of the LHS:

$$\sup_{k\in Q} P\left(|\|\mathbf{A}\boldsymbol{u}_k\|_2^2 - \|\boldsymbol{u}_k\|_2^2| \geq \epsilon\|\boldsymbol{u}_k\|_2^2\right) \leq \sum_{\boldsymbol{u}\in Q} P\left(|\|\mathbf{A}\boldsymbol{u}_i\|_2^2 - \|\boldsymbol{u}_i\|_2^2| \geq \epsilon\|\boldsymbol{u}_i\|_2^2\right) \leq 2|Q|\exp\left(-\frac{n}{4}\epsilon\left(1-\epsilon^2\right)\right)$$

On substituting $\boldsymbol{u}_i = \boldsymbol{x}_p - \boldsymbol{x}_q$, the cardinality of the set $|Q| = p^2/2$, we obtain:

$$P\left(\left|\|\mathbf{A}(\boldsymbol{x}_i - \boldsymbol{x}_j)\|_2^2 - \|(x_i - \boldsymbol{x}_j)\|_2^2\right| \geq \epsilon\|(\boldsymbol{x}_i - \boldsymbol{x}_j)\|_2^2\right) \leq p^2\exp\left(-\frac{n}{4}\epsilon\left(1-\epsilon^2\right)\right) \tag{2}$$

Assuming, $\delta = p^2\exp\left(-\frac{n}{4}\epsilon\left(1-\epsilon^2\right)\right)$, we obtain $n \geq \frac{4}{\epsilon^2(1-\epsilon)}\log\frac{p^2}{\delta}$ such that for this n the inequality in (2)1 is not true with probability $1-\delta$ i.e.

$$(1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \leq \|\mathbf{A}\boldsymbol{x}_i - \mathbf{A}\boldsymbol{x}_j\|_2^2 \leq (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$$

is true with probability $1-\delta$. ∎

4 Proofs

- Details of Hoeffding's Lemma:

  Lemma 4.1 (Hoeffding's Lemma)**.** *Let Z be a random variable with $\mathbb{E}[Z] = 0$, and bounded in $[a,b]$ then, $\psi_Z(\lambda) \leq \frac{\lambda^2}{2}\cdot\frac{(b-a)^2}{4}$ and thus, $Z \in \mathcal{G}((b-a)^2/4)$.*

  Proof. First note that if $a \neq b$ then, $a < 0$ and $b > 0$, since, $\mathbb{E}[Y] = 0$. Now, since $e^{\lambda y}$ is a convex function for a given λ then, using Jensen's inequality:

  $$e^{\lambda y} \leq \frac{b-y}{b-a}e^{\lambda a} + \frac{y-a}{b-a}e^{\lambda b}.$$

This can be simplified as :

$$\begin{aligned}
\mathbb{E}[e^{\lambda Y}] &\leq \frac{b-\mathbb{E}[Y]}{b-a}e^{\lambda a} + \frac{\mathbb{E}[Y]-a}{b-a}e^{\lambda b} \\
&= \frac{b}{b-a}e^{\lambda a} + \frac{-a}{b-a}e^{\lambda b} \\
&= \left(1-\frac{-a}{b-a}\right)e^{-\left(\lambda(b-a)\frac{-a}{b-a}\right)} + \left(\frac{-a}{b-a}\right)e^{\lambda(b-a)\left(1-\frac{-a}{b-a}\right)} \\
&= (1-p)e^{-hp} + pe^{h(1-p)} \\
&= e^{-hp}\left(1-p+pe^{h}\right) \\
&= e^{-hp+\ln\left(1-p+pe^{h}\right)} \\
&= e^{L(h)}
\end{aligned}$$

where, $h = \lambda(b-a)$, $p = \frac{-a}{b-a}$ and $L(h) = -hp+\ln(1-p+pe^h)$. Note that, $p > 0$ and $1-p+pe^h = p(\frac{1}{p}-1+e^h) = p(\frac{-b}{a}+e^h) > 0$ and thus, is a feasible argument for ln. By Taylor's theorem, there exists $h' \in [0,h]$ such that,

$$L(h) = L(0) + hL'(0) + \frac{1}{2}h^2L''(h')$$

It is easy to check that $L(0) = L'(0) = 0$ and

$$L''(h') = \frac{pe^{h'}}{1-p+pe^{h'}}\left(1-\frac{pe^{h'}}{1-p+pe^{h'}}\right) = t(1-t),$$

where $t = \frac{pe^{h'}}{1-p+pe^{h'}} > 0$. Note that, $L''(h')$ reaches its maximum $1/4$ at $t = 1/2$ thus,

$$L(h) \leq \frac{1}{2}h^2L''(h') \leq \frac{1}{2}h^2.\frac{1}{4} = \frac{1}{8}\lambda^2(b-a)^2.$$

This implies that:

$$\mathbb{E}[e^{\lambda Y}] \leq \exp\frac{1}{8}(b-a)^2$$

and that Y is sub-Gaussian $\mathcal{G}(\frac{(b-a)^2}{4})$. ■

- In literature, the bounded difference inequality is also proved using the entropy method [5]. We here outline the proof and the tools that are needed for the proof.

  Definition 4 (Entropy)**.** *: Let Z be a non-negative random variable. The entropy of Z is defined as:*

  $$Ent(Z) = \mathbb{E}[Z\log Z] - (\mathbb{E}[Z])\log(\mathbb{E}[Z]).$$

  Properties of Entropy:

  - It is a scale independent measure of variation i.e. $\mathbb{E}[cZ] = \mathbb{E}[Z]$ for a constant $c > 0$.
  - It is always non-negative and attains 0 when Z is deterministic (can be proven using Jensen's inequality).

  Note that, the entropy discussed here is different from the Shannon entropy $H(Z) = \mathbb{E}[-\log\mathbb{P}_Z(Z)]$. The two are related but not equivalent(in fact, Ent(.) is more related to the relative entropy).

  Definition 5. *Let $\{X_i\}_{i=1}^n$ be independent random variables and $f \geq 0$ be any function, and let*

  $$Ent^{(i)}(f(x_1,\cdots,x_n)) := Ent[f(x_1,\cdots,x_{i-1},X_i,x_{i+1},\cdots,x_n)].$$

  That is, $Ent^{(i)}f$ is the entropy of f with respect to the variable X_i only. Similarly,

  $$\mathbb{E}^{(i)}[f(x_1,\cdots,x_n)] := \mathbb{E}[f(x_1,\cdots,x_{i-1},X_i,x_{i+1},\cdots,x_n)].$$

Theorem 4.2 (Bounded Difference Inequality)**.** *: Let $X_1, \cdots, X_n$ be independent random variables, and let f satisfy the bounded differences property for some $\{c_i\}_{i=1}^n$. Set $\sigma^2 = \frac{1}{4}\sum_{i=1}^n c_i^2$. Then,*

$$P\left(|f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)]| > t\right) \leq 2e^{-\frac{t^2}{2\sigma^2}}.$$

Proof's outline : Taking $Z = f(X_1, \cdots, X_n)$, the proof contains three sub-parts:

- Showing that
$$\frac{Ent^{(i)}(e^{\lambda Z})}{\mathbb{E}^{(i)}[e^{\lambda Z}]} \leq \frac{\lambda^2}{2}.\frac{c_i^2}{4} \textbf{(Hoeffding type bound)}.$$
One approach to proving this is using Logarithmic Sobolev Inequalities [5] whereas, a direct approach can be found in section 2.3 in [[2]].

- In the next step we use the property of (**subadditivity of entropy**)[discussed later]
$$Ent(Z) \leq \mathbb{E}\left[\sum_{i=1}^n Ent^{(i)}(Z)\right]$$
to obtain:
$$\frac{Ent(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\lambda^2}{2}.\frac{\sum_{i=1}^n c_i^2}{4}.$$

- Then, we use (**Herbst's Trick**)[discussed later] to prove that, $Z - \mathbb{E}[Z] \in \mathcal{G}\left(\sigma^2 = \frac{1}{4}\sum_{i=1}^n c_i^2\right)$ which then can be used with the concentration inequality obtained for sub-Gaussian variables to prove the main result.

- We outline the proof of sub-additivity of entropy and Herbst's trick:

  - **Proof's outline for sub-additivity of entropy**:
    - First showing that $Ent(Z) = \sum_{i=1}^n \mathbb{E}\left[ZU_i\right]$ where $U_i = \log \frac{\mathbb{E}[Z|X_1,\cdots,X_i]}{\mathbb{E}[Z|X_1,\cdots,X_{i-1}]}$.
    - Then use the variational formula[discussed below] to deduce that $\mathbb{E}[ZU_i] \leq \mathbb{E}[Ent^{(i)}(Z)]$. After that we average both sides and reach the result.
      The variational formula for Entropy can be written as:
      $$Ent(Z) = \sup_{X:\mathbb{E}[e^X]=1} \mathbb{E}[ZX]$$
      This can be shown by :
      1. Using Jensen's inequality to show that $Ent(Z) - \mathbb{E}[ZX] \geq 0$ when $\mathbb{E}[e^X] = 1$.
      2. Then showing that equality holds when $X = \log \frac{Z}{\mathbb{E}[Z]}$.

  It is interesting to note that a similar property holds for the Variance i.e. (**Sub-additivity of the Variance**)
  For independent $X_1, \cdots, X_n$,
  $$Var[f(X_1, \cdots, X_n)] \leq \mathbb{E}\left[\sum_{i=1}^n Var^{(i)} f(X_1, \cdots, X_n)\right]$$
  This property is also called as **Effron-Stein Inequality**.

– **Proof's outline of Herbst's Trick**:
For a given random variable Z if

$$\frac{Ent(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\lambda^2\sigma^2}{2}, \forall \lambda \geq 0.$$

Then, $Z - \mathbb{E}[Z] \in \mathcal{G}(\sigma^2)$:

$$\psi_0(\lambda) := \psi_{(Z-\mathbb{E}[Z])}(\lambda) = \log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \leq \frac{\lambda^2\sigma^2}{2}, \forall \lambda \geq 0.$$

* The log-Moment generating function of $Z - \mathbb{E}[Z]$ is $\psi_0(\lambda) = \log \mathbb{E}[e^{\lambda Z}] - \lambda \mathbb{E}[Z]$.
* It can be shown that $\frac{d}{d\lambda}\frac{\psi_0(\lambda)}{\lambda} = \frac{Ent(e^{\lambda Z})}{\lambda^2 \mathbb{E}[e^{\lambda Z}]}$.
* Integrating $\int_0^\lambda \frac{d}{d\lambda}\frac{\psi_0(\lambda)}{\lambda} \leq \int_0^\lambda \frac{\sigma^2}{2}$ gives $\psi_0(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$.

5 Review

In this lecture we discussed:

- Concentration inequalities and there implications.
- Probability bounds to prove the concentration inequalities.
- Examples where these inequalities can be used.
- Alternative proof of concentration inequalities using entropy method.

References

[1] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

[2] R. van Handel, "Probability in high dimension," DTIC Document, Tech. Rep., 2014.

[3] J. Ni and S. Tatikonda, "Network tomography based on additive metrics," *Information Theory, IEEE Transactions on*, vol. 57, no. 12, pp. 7798–7809, 2011.

[4] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of johnson and lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[5] S. Boucheron, G. Lugosi, and P. Massart, "Concentration inequalities using the entropy method," *Annals of Probability*, pp. 1583–1614, 2003.