Advanced Topics in Data Sciences

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 10: Concentration of Measure Inequalities
Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-731 (Spring 2016)
Outline

This lecture:

1. Cramér-Chernoff bound
2. Hoeffding bound
3. Herbst’s trick
4. Entropy function and its properties
5. Bounded differences inequality
Recommended Reading Materials


2. R. V. Handel, Probability in High Dimension. Lecture Notes, 2014 *(Section 3.3)*
Part I: Results and Examples
Concentration of Measure Phenomenon

Problem (a rough statement)

Given a random variable \( Y \), how “concentrated” is \( Y \) (e.g., around its mean)?

Concentration of Measure Inequalities

Suppose that we can find a deterministic value \( m \), such that

\[
\Pr \left( |Y - m| > t \right) \leq D(t)
\]

where \( D(t) \) decreases drastically to 0 in \( t \). We say that \( Y \) concentrates around \( m \).

Note: Typically \( m = \mathbb{E}[Y] \), and \( D(t) \) decreases exponentially: \( D(t) \sim e^{-tk} \) for some positive integer \( k \).

Example 1. In statistics, \( Y \) can be the estimation/prediction error.

Example 2. In optimization, \( Y \) can be the objective error \( f(x_k) - f(x^*) \), or the estimate of gradient \( \nabla f(x_k) \).

Example 3. In computer science, \( Y \) can be the outcomes of randomized algorithms.

Example 4. Many other applications in information theory, statistical physics, random matrices, statistical learning theory...
Concentration of Measure Phenomenon

Problem (a rough statement)

*Given a random variable \( Y \), how “concentrated” is \( Y \) (e.g., around its mean)?*

Concentration of Measure Inequalities

Suppose that we can find a deterministic value \( m \), such that

\[
\Pr(|Y - m| > t) \leq D(t)
\]

where \( D(t) \) decreases drastically to 0 in \( t \). We say that \( Y \) concentrates around \( m \).

**Note:** Typically \( m = \mathbb{E}[Y] \), and \( D(t) \) decreases exponentially: \( D(t) \sim e^{-t^k} \) for some positive integer \( k \).
Concentration of Measure Phenomenon

Problem (a rough statement)

*Given a random variable $Y$, how “concentrated” is $Y$ (e.g., around its mean)?*

Concentration of Measure Inequalities

Suppose that we can find a deterministic value $m$, such that

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

where $D(t)$ decreases drastically to 0 in $t$. We say that $Y$ concentrates around $m$.

**Note:** Typically $m = \mathbb{E}[Y]$, and $D(t)$ decreases exponentially: $D(t) \sim e^{-tk}$ for some positive integer $k$.

**Example**

1. In statistics, $Y$ can be the estimation/prediction error.
2. In optimization, $Y$ can be the objective error $f(x_k) - f(x^*)$, or the estimate of gradient $\nabla f(x_k)$.
3. In computer science, $Y$ can be the outcomes of randomized algorithms.
4. Many other applications in information theory, statistical physics, random matrices, statistical learning theory...
Example: Sums of Independent Random Variables

A simple example: $Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i$, where the $X_i$ are independent with mean $\mu$ and variance $\sigma^2$

- **Law of Large Numbers:** $\Pr(|Y_n - \mu| > \epsilon) \to 0$ as $n \to \infty$
- **Central Limit Theorem:** $\Pr \left(|Y_n - \mu| > \frac{\alpha}{\sqrt{n}}\right) \to 2\Phi \left(-\frac{\alpha}{\sigma}\right)$ as $n \to \infty$, where $\Phi$ is the standard normal CDF.
- **Large Deviations:** Under some technical assumptions, $\Pr(|Y_n - \mu| > \epsilon) \leq e^{-n \cdot c(\epsilon)}$
- **Moderate Deviations:** Decay rate of $\Pr(|Y_n - \mu| > \epsilon_n)$ when $\epsilon_n \to 0$ sufficiently slowly so that $\epsilon_n \sqrt{n} \to \infty$

In many applications, we want the bounds to be non-asymptotic.
Concentration of measure has many manifestations; we will only cover one today:

**A General Principle of Concentration of Measure: Functional Inequalities**

If $X_1, \ldots, X_n$ are independent random variables, then any function $f(x_1, \cdots, x_n)$ that is “not too sensitive” to any of the coordinates will concentrate around its mean:

$$P\left(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| > t\right) \leq e^{-t^2/c(f)},$$

where $c(f)$ depends on the sensitivity in its coordinates.

**Note:** *No assumptions* on the $X_i$ besides independence! (which can be relaxed)
In This Lecture

Definition (Bounded Difference Functions)

A function \( f : \mathcal{X}^n \rightarrow \mathbb{R} \) has the bounded differences property if for some positive \( c_1, \ldots, c_n \),

\[
\sup_{x_1, \ldots, x_n, x'_i \in \mathcal{X}} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x'_i, \ldots, x_n)| \leq c_i.
\]
### In This Lecture

<table>
<thead>
<tr>
<th>Definition (Bounded Difference Functions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A function $f : \mathcal{X}^n \to \mathbb{R}$ has the bounded differences property if for some positive $c_1, \ldots, c_n$,</td>
</tr>
<tr>
<td>$\sup_{x_1, \ldots, x_n, x'_i \in \mathcal{X}}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theorem (Bounded Differences Inequality)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Let $X_1, \ldots, X_n$ be independent random variables, and let $f$ satisfy the bounded differences property with $c_i$'s. Then</td>
</tr>
<tr>
<td>$P\left(</td>
</tr>
</tbody>
</table>
In This Lecture

Definition (Bounded Difference Functions)

A function $f : \mathcal{X}^n \rightarrow \mathbb{R}$ has the bounded differences property if for some positive $c_1, \ldots, c_n$

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

Theorem (Bounded Differences Inequality)

Let $X_1, \ldots, X_n$ be independent random variables, and let $f$ satisfy the bounded differences property with $c_i$'s. Then

$$P\left( |f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| > t \right) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2} \right).$$

To prove this result, we need the following fundamental notions:

- Cramér-Chernoff bound
- Hoeffding bound
- Herbst’s trick
- Entropy function and its properties
Bounded Differences: Example

Example (Chromatic Number of a Random Graph)

Let $V = \{1, \cdots, n\}$, and let $G$ be a random graph such that each pair $i, j \in V$ is independently connected with probability $p$. Let

$$X_{ij} = \begin{cases} 1 & (i, j) \text{ are connected} \\ 0 & \text{otherwise.} \end{cases}$$

The chromatic number of $G$ is the minimum number of colors needed to color the vertices such that no two connected vertices have the same color. Writing

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

we find that $f$ satisfies the bounded difference property with $c_{ij} = 1$.

In the later lectures, we will see an application of the bounded differences inequality to statistical learning theory.
Markov’s Inequality

Let $Z$ be a nonnegative random variable. Then $\operatorname{Pr}(Z \geq t) \leq \frac{\mathbb{E}[Z]}{t}$.

Proof: $$\int_{0}^{\infty} f_Z(z) 1\{z \geq t\} \, dz \leq \int_{0}^{\infty} \frac{z}{t} f_Z(z) 1\{z \geq t\} \, dz \leq \int_{0}^{\infty} \frac{z}{t} f_Z(z) \, dz = \frac{\mathbb{E}[Z]}{t}$$
Markov’s Inequality Applied to Functions

Let \( \phi \) denote any\textit{ nondecreasing} and\textit{ nonnegative} function. Let \( Z \) be any random variable. Then Markov’s inequality gives

\[
\Pr(Z \geq t) \leq \Pr(\phi(Z) \geq \phi(t)) \leq \frac{\mathbb{E}[\phi(Z)]}{\phi(t)}.
\]
Markov’s Inequality Applied to Functions

Let $\phi$ denote any nondecreasing and nonnegative function. Let $Z$ be any random variable. Then Markov’s inequality gives

$$\Pr(Z \geq t) \leq \Pr(\phi(Z) \geq \phi(t)) \leq \frac{E[\phi(Z)]}{\phi(t)}.$$ 

**Chebyshev’s Inequality:** Choose $\phi(t) = t^2$, and replace $Z$ by $|Z - E[Z]|$. Then

$$\Pr\left(|Z - E[Z]| \geq t\right) \leq \frac{\text{Var}[Z]}{t^2}.$$ 

**Chernoff Bound:** Choose $\phi(t) = e^{\lambda t}$ where $\lambda \geq 0$. Then we have

$$\Pr(Z \geq t) \leq e^{-\lambda t}E[e^{\lambda Z}].$$
Cramér-Chernoff Inequality

**Definition (Log-moment-generating function)**

The log-moment-generating function $\psi_Z(\lambda)$ of a random variable $Z$ is defined as

$$
\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}], \quad \lambda \geq 0.
$$

Clearly the Chernoff bound can be written as $\Pr(Z \geq t) \leq e^{-\psi_Z(\lambda)}$.
Cramér-Chernoff Inequality

**Definition (Log-moment-generating function)**

The log-moment-generating function $\psi_Z(\lambda)$ of a random variable $Z$ is defined as

$$\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}], \quad \lambda \geq 0.$$ 

Clearly the Chernoff bound can be written as $\Pr(Z \geq t) \leq e^{-(\lambda t - \psi_Z(\lambda))}$.

**Definition (Cramér transform)**

The Cramér transform of $Z$ is defined as

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

Note that $\psi^*_Z(t) \geq \psi^*_Z(0) = 0$. 

---

Advanced Topics in Data Sciences | Prof. Volkan Cevher, volkan.cevher@epfl.ch
### Cramér-Chernoff Inequality

#### Definition (Log-moment-generating function)

The log-moment-generating function $\psi_Z(\lambda)$ of a random variable $Z$ is defined as

$$\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}], \quad \lambda \geq 0.$$  

Clearly the Chernoff bound can be written as $\Pr(Z \geq t) \leq e^{-\lambda t - \psi_Z(\lambda)}$.

#### Definition (Cramér transform)

The Cramér transform of $Z$ is defined as

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

Note that $\psi^*_Z(t) \geq \psi^*_Z(0) = 0$.

#### Theorem (Cramér-Chernoff Inequality)

*For any random variable $Z$, we have*

$$\Pr(Z \geq t) \leq \exp(-\psi^*_Z(t)).$$
Sums of Independent Random Variables Revisited

Let $Z = X_1 + \cdots + X_n$ where $\{X_i\}$ are independent and identically distributed (i.i.d.).

**Chebyshev’s Inequality on the Sum:** We have $\text{Var}[Z] = n\text{Var}[X]$, and hence Chebyshev’s inequality with $t = n\epsilon$ gives

$$\Pr\left(\frac{1}{n}|Z - \mathbb{E}[Z]| \geq \epsilon\right) \leq \frac{\text{Var}[X]}{n\epsilon^2}.$$
Sums of Independent Random Variables Revisited

Let \( Z = X_1 + \cdots + X_n \) where \( \{X_i\} \) are independent and identically distributed (i.i.d.).

**Chebyshev’s Inequality on the Sum:** We have \( \text{Var}[Z] = n \text{Var}[X] \), and hence Chebyshev’s inequality with \( t = n\epsilon \) gives

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

**Cramér-Chernoff Inequality on the Sum:** We have

\[
\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}] = \log \mathbb{E}\left[ e^{\lambda \sum_{i=1}^n X_i} \right] = \log \mathbb{E}\left[ \prod_{i=1}^n e^{\lambda X_i} \right]
\]

\[
= \log \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}] = \log \left( \mathbb{E}[e^{\lambda X}] \right)^n = n\psi_X(\lambda),
\]

where on the second line we used independence and then the identical distribution property. Then the Cramér-Chernoff Inequality with \( t = n\epsilon \) gives

\[
\Pr(Z \geq n\epsilon) \leq \exp\left( -n\psi_X^*(\epsilon) \right).
\]
The Cramér-Chernoff Method

Cramér-Chernoff Inequality

For any random variable $Z$, we have

$$\Pr(Z \geq t) \leq \exp(-\psi^*_Z(t)).$$

Observation:

1. Given a random variable $X$, let $Z = X - \mathbb{E}[X]$. If we can provide an lower bound on the Cramér transform of $Z$, then we obtain a one-sided concentration inequality:

$$\Pr(X - \mathbb{E}[X] \geq t) \leq \exp(-\psi^*_Z(t)) \leq \exp[-(\text{lower bound of } \psi^*_Z(t))].$$

2. Applying the same argument to $-Z = X - \mathbb{E}[X]$ gives the other side.
**The Cramér-Chernoff Method**

**Cramér-Chernoff Inequality**

For any random variable $Z$, we have

$$\Pr(Z \geq t) \leq \exp(-\psi^*_Z(t)).$$

**Observation:**

1. Given a random variable $X$, let $Z = X - \mathbb{E}[X]$. If we can provide an lower bound on the Cramér transform of $Z$, then we obtain a one-sided concentration inequality:

$$\Pr(X - \mathbb{E}[X] \geq t) \leq \exp(-\psi^*_Z(t)) \leq \exp\left[-(\text{lower bound of } \psi^*_Z(t))\right].$$

2. Applying the same argument to $-Z = X - \mathbb{E}[X]$ gives the other side.

**Example (Gaussian random variables concentrate)**

Let $X \sim \mathcal{N}(0, \sigma^2)$. Then $\psi_X(\lambda) = \frac{\lambda^2 \sigma^2}{2}$, and thus $\psi^*_X(t) = \frac{t^2}{2\sigma^2}$. Therefore,

$$\Pr(|X| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

That is, Gaussian random variables *concentrate around their mean* — increasingly so for small $\sigma^2$. 
Sub-Gaussian Random Variables

Notice that if $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$, then $\psi_X^*(t) \geq \frac{t^2}{2\sigma^2}$. This motivates the following.

**Definition (Sub-Gaussian Random Variables)**

A *centered* random variable $X$ is said to be *sub-Gaussian* with parameter $\sigma^2$ if $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$, $\forall \lambda > 0$. Denote the set of all such random variables by $G(\sigma^2)$. 
Sub-Gaussian Random Variables

Notice that if $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$, then $\psi^*_X(t) \geq \frac{t^2}{2\sigma^2}$. This motivates the following.

**Definition (Sub-Gaussian Random Variables)**

A *centered* random variable $X$ is said to be *sub-Gaussian* with parameter $\sigma^2$ if $\psi_X(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$, $\forall \lambda > 0$. Denote the set of all such random variables by $G(\sigma^2)$.

**Basic Properties of Sub-Gaussian Random Variables**

1. $\Pr(|X| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right)$ (*sub-Gaussian random variables concentrate*)

2. If $X_i \in G(\sigma_i^2)$ are independent, then $\sum_{i=1}^{n} a_i X_i \in G\left(\sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$. 
Bounded Random Variables are Sub-Gaussian

One of the most important examples of sub-Gaussian random variable is the bounded random variable.

**Theorem (Hoeffding’s Lemma)**

Let $Y$ be a random variable with $\mathbb{E}[Y] = 0$, taking values in a bounded interval $[a, b]$. Let $\psi_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y}]$. Then $\psi''_Y(\lambda) \leq \frac{(b-a)^2}{4}$ and $Y \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$.

We will see the proof later in the lecture.
Hoeffding’s Inequality

Applying sub-Gaussian concentration to the previous slide, we find that for \( Y \in [a, b] \),

\[
\Pr \left( |Y - \mathbb{E}[Y]| > t \right) \leq 2 \exp \left( - \frac{2t^2}{(b-a)^2} \right).
\]

Using a similar argument along with the fact that sums of sub-Gaussian variables are sub-Gaussian, we obtain the following.

**Theorem (Hoeffding’s Inequality)**

Let \( Z = X_1 + \cdots + X_n \), where the \( X_i \) are independent and supported on \([a_i, b_i]\). Then

\[
\Pr \left( \frac{1}{n} |Z - \mathbb{E}[Z]| > \epsilon \right) \leq 2 \exp \left( - \frac{2n\epsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).
\]
Concentration in Applications: PAC Learnability

Recall the following from the previous lecture.

**Proposition**

Assume that the hypothesis class $\mathcal{H}$ consists of a finite number of functions $f(h, \cdot)$ taking values 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:** Define $\xi_i(h) = f(h, x_i)$, and define $S_n(h) := (1/n) \sum_{1 \leq i \leq n} (\xi_i(h) - \mathbb{E} \xi_i(h))$ for every $h \in \mathcal{H}$. Notice that then

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

By the union bound and Hoeffding’s inequality (with $a = 0$ and $b = 1$), we have

$$\mathbb{P} \left( \sup_{h \in \mathcal{H}} |S_n(h)| \geq \epsilon \right) \leq \sum_{h \in \mathcal{H}} \mathbb{P} (|S_n(h)| \geq \epsilon) \leq |\mathcal{H}| \cdot 2 \exp \left( -2n\epsilon^2 \right),$$

which is upper bounded by $\delta$ provided that $n \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$.
Concentration in Applications: Network Tomography

The problem in the case of \( n \) packets and \( p \) leaf nodes:

- \( X_k^{(i)} = 1 \{ \text{packet } i \text{ arrives at node } k \} \) for \( i = 1, \cdots, n \) and \( k = 1, \cdots, p \)
- Goal: Given these \( n \) independent samples, reconstruct the tree structure.
Concentration in Applications: Network Tomography

The problem in the case of $n$ packets and $p$ leaf nodes:

- $X_{k}^{(i)} = 1\{\text{packet } i \text{ arrives at node } k\}$ for $i = 1, \cdots, n$ and $k = 1, \cdots, p$
- Goal: Given these $n$ independent samples, reconstruct the tree structure.

Outline of analysis (Ni, 2011):

- Show that the tree can be recovered from the values $q_{kl} = \Pr(\text{packet reaches } x_k \text{ and } x_l)$
- Show robustness, in that any $\hat{q}$ with $|\hat{q}_{kl} - q_{kl}| \leq \epsilon$ suffices
- Set $\hat{q}_{kl} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_{k}^{(i)} = 1 \cap X_{l}^{(i)} = 1\}$, and bound using Hoeffding’s inequality:
  \[ \Pr(|\hat{q}_{kl} - q_{kl}| > \epsilon) \leq 2 \exp(-2n\epsilon^2). \]
- Apply the union bound to conclude $\Pr(\text{error}) \leq \delta$ if $n \geq \frac{1}{2\epsilon^2} \log \frac{p^2}{\delta}$. 
**Theorem (Johnson-Lindenstrauss)**

Let $x_1, \cdots, x_p$ be a collection of points in $\mathbb{R}^d$, and let $A \in \mathbb{R}^{n \times d}$ be a random matrix with independent $N\left(0, \frac{1}{\sqrt{n}}\right)$ entries. For any $\epsilon, \delta \in (0, 1)$, we have with probability at least $1 - \delta$ that

$$(1 - \epsilon) \|x_i - x_j\|_2^2 \leq \|Ax_i - Ax_j\|_2^2 \leq (1 + \epsilon) \|x_i - x_j\|_2^2$$

for all $i, j$, provided that $n \geq \frac{4}{\epsilon^2 (1 - \epsilon)} \log \frac{p^2}{\delta}$. 

---

**Concentration in Applications: Random Linear Projections**
**Theorem (Johnson-Lindenstrauss)**

Let $x_1, \ldots, x_p$ be a collection of points in $\mathbb{R}^d$, and let $A \in \mathbb{R}^{n \times d}$ be a random matrix with independent $N\left(0, \frac{1}{\sqrt{n}}\right)$ entries. For any $\epsilon, \delta \in (0, 1)$, we have with probability at least $1 - \delta$ that

$$(1 - \epsilon)\|x_i - x_j\|_2^2 \leq \|Ax_i - Ax_j\|_2^2 \leq (1 - \epsilon)\|x_i - x_j\|_2^2$$

for all $i, j$, provided that $n \geq \frac{4}{\epsilon^2(1-\epsilon)} \log \frac{p^2}{\delta}$.

The idea:

1. Show that $\mathbb{E}[\|Au\|_2^2] = \|u\|_2^2$ for any $u$
2. Use squared-Gaussian concentration (not covered in this lecture) to show that, for any $u$, $\Pr\left(\|Au\|_2^2 - \|u\|_2^2 > (1 + \epsilon)\|u\|_2^2\right) \leq 2 \exp \left(\frac{n}{4} \epsilon^2(1 - \epsilon)\right)$
3. Apply the union bound to conclude that the analogous event holding for some $u$ of the form $u = x_i - x_j$ is at most $p^2 \exp \left(\frac{n}{4} \epsilon^2(1 - \epsilon)\right)$. 
Other Examples of Concentration Inequalities

There are an extensive range of concentration inequalities in the literature; here are just two more examples to get a flavor for them (Boucheron et al., 2013).

**Theorem (Lipschitz Function of Gaussian RVs)**

Let $X_1, \ldots, X_n$ be independent Gaussian $N(0, 1)$ random variables, and let $f$ be $L$-Lipschitz (i.e., $|f(x) - f(x')| \leq L\|x - x'\|_2$ for any $x, x'$). Then

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

**Theorem (Separately Convex Lipschitz Function of Bounded RVs)**

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

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

We have considered probabilities of the form

$$P\left(\left|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]\right| > t\right)$$

In summary, there are several features of the random variables $X_i$ that tend to permit strong concentration guarantees:

- Boundedness
- Sub-Gaussian
- Moments $\mathbb{E}[|X|^c]$ (not covered here; see, e.g., Bernstein’s inequalities)
- ...

...and there are several properties of the function $f$ that tend to permit strong concentration guarantees:

- Bounded differences
- Lipschitz continuous
- ...

Many of the concentration results for sums of independent RVs have counterparts in sums of random matrices, but this is an ongoing area of research (Tropp, 2015).
Part II: Proofs
Bounded Random Variables are Sub-Gaussian

Theorem (Hoeffding’s Lemma)

Let $Y$ be a random variable with $\mathbb{E}[Y] = 0$, taking values in a bounded interval $[a, b]$. Let $\psi_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y}]$. Then $\psi''_Y(\lambda) \leq \frac{(b-a)^2}{4}$ and $Y \in \mathcal{G} \left( \frac{(b-a)^2}{4} \right)$. 
Bounded Random Variables are Sub-Gaussian

Theorem (Hoeffding’s Lemma)

Let $Y$ be a random variable with $\mathbb{E}[Y] = 0$, taking values in a bounded interval $[a, b]$. Let $\psi_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y}]$. Then $\psi''_Y(\lambda) \leq \frac{(b-a)^2}{4}$ and $Y \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$.

Outline of proof:

1. Prove that $\text{Var}[Z] \leq \frac{(b-a)^2}{4}$ for any $Z$ bounded on $[a, b]$.
2. Show $\psi_Y(0) = 0$, $\psi'_Y(0) = 0$, and $\psi''_Y(\lambda) = \text{Var}[Z]$, where $Z$ is a random variable with PDF $f_Z(z) = e^{-\psi_Y(\lambda)} e^{\lambda z} f_Y(z)$; hence $\psi''_Y(\lambda) \leq \frac{(b-a)^2}{4}$ by Step 1.
3. Taylor expand $\psi_Y(\lambda) = \psi_Y(0) + \lambda \psi'_Y(0) + \frac{\lambda^2}{2} \psi''_Y(\theta)$ (for some $\theta \in [0, \lambda]$) and substitute Step 2 to upper bound this by $\frac{\lambda^2}{2} \cdot \frac{(b-a)^2}{4}$.
Entropy of a Random Variable

Definition (Entropy)

Let $Z$ be a nonnegative random variable. The entropy of $Z$ is defined as

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

Rough intuition: A measure of variation that is scale-independent: $\text{Ent}[cZ] = \text{Ent}[Z]$

- Always non-negative by Jensen’s inequality; zero if and only if $Z$ is deterministic

Note: Not to be confused with Shannon entropy $H(Z) = \mathbb{E}[-\log f_Z(Z)]$. The two are related but not equivalent (in fact, $\text{Ent}(\cdot)$ is more related to the relative entropy).
Entropy of a Random Variable

Definition (Entropy)

Let $Z$ be a nonnegative random variable. The entropy of $Z$ is defined as

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

Rough intuition: A measure of variation that is scale-independent: $\text{Ent}[cZ] = \text{Ent}[Z]$
- Always non-negative by Jensen’s inequality; zero if and only if $Z$ is deterministic

Note: Not to be confused with Shannon entropy $H(Z) = \mathbb{E}[\log f_Z(Z)]$. The two are related but not equivalent (in fact, $\text{Ent}(\cdot)$ is more related to the relative entropy).

Definition (Conditional Versions of $\text{Ent}$ and $\mathbb{E}$)

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

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

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

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

**Theorem (Bounded Differences Inequality)**

Let $X_1, \ldots, X_n$ 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

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

**Theorem (Bounded Differences Inequality)**

Let $X_1, \ldots, X_n$ 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

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

Outline of proof ($Z = f(X_1, \cdots, X_n)$):

1. Show that $\frac{\text{Ent}^{(i)}(e^{\lambda Z})}{\mathbb{E}^{(i)}[e^{\lambda Z}]} \leq \frac{\lambda^2}{2} \cdot \frac{c_i^2}{4}$ (**Hoeffding-type Bound**)

2. Use $\text{Ent} \left[ f(X_1, \ldots, X_n) \right] \leq \mathbb{E} \left[ \sum_{i=1}^n \text{Ent}^{(i)}(f(X_1, \ldots, X_n)) \right]$ (**Subadditivity of Entropy**) to deduce that $\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\lambda^2}{2} \cdot \frac{1}{4} \sum_{i=1}^n c_i^2$.

3. Deduce that $Z - \mathbb{E}[Z]$ is sub-Gaussian with $\sigma^2 = \frac{1}{4} \sum_{i=1}^n c_i^2$ (**Herbst’s Trick**).
Herbst’s Trick

Theorem (Herbst’s Trick)

Suppose $Z$ is such that, for some $\sigma^2 > 0$, we have

$$\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\lambda^2 \sigma^2}{2}, \quad \forall \lambda \geq 0. \quad (1)$$

Then $Z - \mathbb{E}Z \in \mathcal{G}(\sigma^2)$; that is,

$$\psi_0(\lambda) := \psi(Z - \mathbb{E}Z)(\lambda) = \log \mathbb{E}e^{\lambda(Z - \mathbb{E}Z)} \leq \frac{\lambda^2 \sigma^2}{2}, \quad \forall \lambda \geq 0.$$
Herbst’s Trick

**Theorem (Herbst’s Trick)**

Suppose $Z$ is such that, for some $\sigma^2 > 0$, we have

$$
\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\lambda^2 \sigma^2}{2}, \quad \forall \lambda \geq 0. \tag{1}
$$

Then $Z - \mathbb{E}Z \in \mathcal{G}(\sigma^2)$; that is,

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

Outline of proof:

1. Write log-MGF of $Z - \mathbb{E}[Z]$ as $\psi_0(\lambda) = \log \mathbb{E}[e^{\lambda Z}] - \lambda \mathbb{E}[Z]$

2. Prove $\frac{d}{d\lambda} \frac{\psi_0(\lambda)}{\lambda} = \frac{\text{Ent}(e^{\lambda Z})}{\lambda^2 \mathbb{E}[e^{\lambda Z}]}$.

3. Integrate both sides of Step 2 from 0 to $\lambda$, and apply (1) to obtain $\frac{\psi_0(\lambda)}{\lambda} \leq \frac{\lambda \sigma^2}{2}$. 
Sub-Additivity of the Entropy

Theorem (Sub-Additivity of the Entropy)

For independent $X_1, \ldots, X_n$,

$$
\operatorname{Ent} (f(X_1, \ldots, X_n)) \leq \mathbb{E} \left[ \sum_{i=1}^{n} \operatorname{Ent}^{(i)} (f(X_1, \ldots, X_n)) \right].
$$

Outline of proof:

1. Show $\operatorname{Ent}(Z) = \sum_{i=1}^{n} \mathbb{E}[Z U_i]$ where $U_i = \log \frac{\mathbb{E}[Z|X_1, \ldots, X_i]}{\mathbb{E}[Z|X_1, \ldots, X_{i-1}]}$
2. Show $\mathbb{E}[e^{U_i} | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n] = 1$
3. Use variational formula to deduce $\mathbb{E}[Z U_i] \leq \mathbb{E}[\operatorname{Ent}^{(i)} (Z)]$, then average both sides
Sub-Additivity of the Entropy

Theorem (Sub-Additivity of the Entropy)

For independent $X_1, \cdots, X_n$,

$$\text{Ent}(f(X_1, \ldots, X_n)) \leq \mathbb{E} \left[ \sum_{i=1}^{n} \text{Ent}^{(i)}(f(X_1, \ldots, X_n)) \right].$$

Outline of proof:
1. Show $\text{Ent}(Z) = \sum_{i=1}^{n} \mathbb{E}[ZU_i]$ where $U_i = \log \frac{\mathbb{E}[Z|X_1,\cdots,X_i]}{\mathbb{E}[Z|X_1,\cdots,X_{i-1}]}$
2. Show $\mathbb{E}[e^{U_i} | X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n] = 1$
3. Use variational formula to deduce $\mathbb{E}[ZU_i] \leq \mathbb{E}[\text{Ent}^{(i)}(Z)]$, then average both sides

Theorem (Variational Formula for Entropy)

$$\text{Ent}(Z) = \sup_{X : \mathbb{E}[e^X]=1} \mathbb{E}[ZX].$$

Outline of proof:
1. Use Jensen’s inequality to show $\text{Ent}(Z) - \mathbb{E}[ZX] \geq 0$ whenever $\mathbb{E}[e^X] = 1$
2. Show that equality holds when $X = \log \frac{Z}{\mathbb{E}[Z]}$
Sub-Additivity of the Variance

As a side-note, the variance satisfies a similar property.

**Theorem (Efron-Stein Inequality – Sub-Additivity of the Entropy)**

*For independent $X_1, \cdots, X_n$,*

$$\text{Var}[f(X_1, \ldots, X_n)] \leq \mathbb{E}\left[ \sum_{i=1}^{n} \text{Var}^{(i)} f(X_1, \ldots, X_n) \right].$$

When $f(X_1, \cdots, X_n) = \sum_{i=1}^{n} X_i$, this becomes $\text{Var}\left[ \sum_{i=1}^{n} X_i \right] \leq \sum_{i=1}^{n} \text{Var}[X_i]$, which in fact holds with equality.

The above (Efron-Stein) inequality can be used to obtain useful concentration results in some settings, but the entropy is more useful for our purposes.
Bounded Differences Inequality

**Theorem (Bounded Differences Inequality)**

Let $X_1, \ldots, X_n$ 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

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

Outline of proof ($Z = f(X_1, \cdots, X_n)$):

1. Show that $rac{\text{Ent}^{(i)}(e^{\lambda Z})}{\mathbb{E}^{(i)}[e^{\lambda Z}]} \leq \frac{\lambda^2}{2} \cdot \frac{c_i^2}{4}$ (Hoeffding-type Bound)

2. Use $\text{Ent}(Z) \leq \mathbb{E} \left[ \sum_{i=1}^n \text{Ent}^{(i)}(Z) \right]$ (Subadditivity of Entropy) to deduce that

$$\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\lambda^2}{2} \cdot \frac{1}{4} \sum_{i=1}^n c_i^2.$$ 

3. Deduce that $Z - \mathbb{E}[Z]$ is sub-Gaussian with $\sigma^2 = \frac{1}{4} \sum_{i=1}^n c_i^2$ (Herbst's Trick)
References


