Concentration of Measure with Applications in Information Theory, Communications and Coding

Maxim Raginsky
UIUC
Urbana, IL 61801, USA
maxim@illinois.edu

Igal Sason
Technion
Haifa 32000, Israel
sason@ee.technion.ac.il

Presenters: Part 1 by I. Sason, and Part 2 by M. Raginsky

IEEE 2015 International Symposium on Information Theory (ISIT 2015)
Tutorial T-AM-2
Part 1 of 2

Hong-Kong
June 14, 2015.

Concentration of Measures

- In many situations, the statistical properties of random fluctuations of functions of many random variables are of interest.

Concentration of Measures

- In many situations, the statistical properties of random fluctuations of functions of many random variables are of interest.
- Roughly speaking, the concentration of measure phenomenon can be stated in the following simple way: "A random variable that depends in a smooth way on many independent random variables (but not too much on any of them) is essentially constant" (Talagrand, 1996).

Concentration of Measures

- In many situations, the statistical properties of random fluctuations of functions of many random variables are of interest.
- Roughly speaking, the concentration of measure phenomenon can be stated in the following simple way: "A random variable that depends in a smooth way on many independent random variables (but not too much on any of them) is essentially constant" (Talagrand, 1996).
- Concentration of measure has far-reaching consequences in
  - Pure and applied probability,
  - High-dimensional statistics,
  - Functional analysis,
  - Computer science,
  - Machine learning,
  - Statistical physics,
  - Complex graphs and networks,
  - Information theory, communication and coding theory.

Concentration Inequalities

- Let $X_1, \ldots, X_n$ be (possibly dependent) random variables taking values in some set $\mathcal{X}$.
- Let $f \colon \mathcal{X}^n \to \mathbb{R}$, and

$$Z = f(X_1, \ldots, X_n).$$

Question: How large are typical deviations of Z from its expected value ($\mathbb{E}[Z]$) or median ?

We seek upper bounds for

$$\mathbb{P}\{Z \geq \mathbb{E}[Z] + t\}, \quad \mathbb{P}\{Z \leq \mathbb{E}[Z] - t\}$$

for $t > 0$.

Various Approaches for Proving Concentration Inequalities

1. The martingale approach: Hoeffding (1963), Azuma (1967), Milman and Schechtman (1986), Shamir and Spencer (1987) and McDiarmid (1989, 1998), Sipser and Spielman (1996), Richardson and Urbanke (2001), and other contributions.
2. Talagrand's inequalities for product measures: Talagrand (1996).
3. The entropy method and logarithmic Sobolev inequalities: Ledoux (1996), Massart (1998), Lugusi et al. (1999, 2001), etc.
4. Transportation-cost inequalities: Ahlswede, Gács and Körner (1976), Marton (1986, 1996, 1997), Dembo (1997), Villani (2003, 2008), etc.
5. Stein's method (a.k.a. the method of exchangeable pairs): Chatterjee (2007), Chatterjee and Dey (2010), Goldstein et al. (2011, 2014), etc.

Main Focus of this Tutorial

This tutorial is mainly focused on the following issues:

1. The foundations of concentration of measure inequalities.
2. The interplay between concentration of measures and Shannon theory.
3. Applications to information theory, communications and coding.

Definition - Conditional Expectation with Respect to a σ-Algebra

Let

- $(\Omega, \mathscr{F}, P)$ be a probability space.
- $X \colon \Omega \to \mathbb{R}^n$ be a random vector defined on that probability space.
- $\mathscr{G} \subseteq \mathscr{F}$ be a sub-σ algebra of $\mathscr{F}$.

Then, a conditional expectation of X given $\mathscr{G}$, denoted by $\mathbb{E}[X|\mathscr{G}]$, is a $\mathscr{G}$-measurable function ($\Omega \to \mathbb{R}^n$) which satisfies:

$$\mathbb{E}[\mathbb{E}[X|\mathscr{G}]\, 1_G(X)] = \mathbb{E}[X\, 1_G(X)], \quad \forall\, G \in \mathscr{G}$$

where 1_G is the indicator function of the subset $G \subseteq \Omega$ (i.e., $1_G(x) = 1$ if $x \in G$, and $1_G(x) = 0$ if $x \in \Omega \setminus G$).

Existence and Uniqueness of the Conditional Expectation

- The conditional expectation exists and is unique up to a set of probability 0 (by the Radon-Nikodym theorem).
- The following relation holds between the conditional expectation and the Radon-Nikodym derivative:

$$\mathbb{E}[X|\mathscr{G}] = \frac{\mathrm{d}P|_{\mathscr{G}}}{\mathrm{d}P}$$

where $P|_{\mathscr{G}}$ is the restriction of P to events in the sub σ-algebra $\mathscr{G}$.

The tower property for conditional expectations

Let X be a random variable with finite mean, and let $\mathscr{G}_1 \subseteq \mathscr{G}_2$ be two sub-algebras. Then, with probability 1,

$$\mathbb{E}[\mathbb{E}[X|\mathscr{G}_2] \,|\, \mathscr{G}_1] = \mathbb{E}[X|\mathscr{G}_1].$$

That is, conditioning first on $\mathscr{G}_2$ and then on $\mathscr{G}_1$ is equivalent to conditioning just on the smaller sub σ-algebra $\mathscr{G}_1$.

Concentration via the Martingale Approach

Definition - [Discrete-Time Martingales]

- Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.
- Let $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots$ be a filtration, i.e., a sequence of sub σ-algebras of $\mathscr{F}$.

A sequence $X_0, X_1, \ldots$ of RVs is a martingale (w.r.t. this filtration) if

1. $X_i \colon \Omega \to \mathbb{R}$ is $\mathscr{F}_i$-measurable for every i, i.e.,
$$\{\omega \in \Omega \colon X_i(\omega) \leq a\} \in \mathscr{F}_i \quad \forall\, i \in \{0, 1, \ldots\},\ a \in \mathbb{R}.$$
2. $\mathbb{E}[|X_i|] < \infty$ for all i.
3. $X_i = \mathbb{E}[X_{i+1} | \mathscr{F}_i]$ almost surely for all $i \in \{0, 1, \ldots\}$.

Concentration via the Martingale Approach

Definition - [Discrete-Time Martingales]

- Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.
- Let $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \ldots$ be a filtration, i.e., a sequence of sub σ-algebras of $\mathscr{F}$.

A sequence $X_0, X_1, \ldots$ of RVs is a martingale (w.r.t. this filtration) if

1. $X_i \colon \Omega \to \mathbb{R}$ is $\mathscr{F}_i$-measurable for every i, i.e.,
$$\{\omega \in \Omega \colon X_i(\omega) \leq a\} \in \mathscr{F}_i \quad \forall\, i \in \{0, 1, \ldots\},\ a \in \mathbb{R}.$$
2. $\mathbb{E}[|X_i|] < \infty$ for all i.
3. $X_i = \mathbb{E}[X_{i+1} | \mathscr{F}_i]$ almost surely for all $i \in \{0, 1, \ldots\}$.

A Simple Example: Random walk

$X_n = \sum_{i=0}^{n} U_i$ where $\mathbb{P}(U_i = +1) = \mathbb{P}(U_i = -1) = \frac{1}{2}$ are i.i.d. RVs.

Martingales

Fact 1

Since $\{\mathscr{F}_i\}_{i=0}^n$ is a filtration, it follows that

$$X_j = \mathbb{E}[X_i|\mathscr{F}_j], \quad \forall\, i > j.$$

Also

$$\mathbb{E}[X_i] = \mathbb{E}\big[\mathbb{E}[X_i|\mathscr{F}_{i-1}]\big] = \mathbb{E}[X_{i-1}]$$

so, the expected value of a martingale sequence is constant.

Martingales

Fact 1

Since $\{\mathscr{F}_i\}_{i=0}^n$ is a filtration, it follows that

$$X_j = \mathbb{E}[X_i|\mathscr{F}_j], \quad \forall\, i > j.$$

Also

$$\mathbb{E}[X_i] = \mathbb{E}\big[\mathbb{E}[X_i|\mathscr{F}_{i-1}]\big] = \mathbb{E}[X_{i-1}]$$

so, the expected value of a martingale sequence is constant.

Fact 2

Given a RV $X \in \mathbb{L}^1(\Omega, \mathscr{F}, \mathbb{P})$ and a filtration $\{\mathscr{F}_i\}$, let

$$X_i = \mathbb{E}[X|\mathscr{F}_i] \quad i \in \{0, 1, \ldots\}.$$

Then, the sequence $\{X_i\}$ forms a martingale.

Martingales (Cont.)

Fact 3

Choose $\mathscr{F}_0 = \{\Omega, \emptyset\}$ and $\mathscr{F}_n = \mathscr{F}$, then Fact 2 gives a martingale sequence with the following first and last terms:

$$X_0 = \mathbb{E}[X|\mathscr{F}_0] = \mathbb{E}[X] \quad (\mathscr{F}_0 \text{ provides no information about } X).$$
$$X_n = \mathbb{E}[X|\mathscr{F}_n] = X \ \text{ a.s.} \quad (\mathscr{F}_n \text{ provides full information about } X).$$

Interpretation

1. At the beginning, we don't know anything about X, so the first guess is that X is equal to its expected value.
2. As time evolves, we get more and more information on X (due to the conditioning on sub σ-algebras that form a filtration).
3. At the end, we are provided with full information about X.

Super and Sub Martingales

To define sub- and super-martingales, we keep the first two conditions of the martingale, and

- $\mathbb{E}[X_i|\mathscr{F}_{i-1}] \geq X_{i-1}$ holds a.s. for sub-martingales.
- $\mathbb{E}[X_i|\mathscr{F}_{i-1}] \leq X_{i-1}$ holds a.s. for super-martingales.

Generalization:

From the tower property for conditional expectations, for sub-martingales

$$\mathbb{E}[X_i|\mathscr{F}_j] \geq X_j, \quad \forall\, i > j$$

and, for super-martingales,

$$\mathbb{E}[X_i|\mathscr{F}_j] \leq X_j, \quad \forall\, i > j.$$

Setting

- Let $X_1, \ldots, X_n$ be n independent RVs, getting values in the set $\mathcal{X}$.
- Let $f \colon \mathcal{X}^n \to \mathbb{R}$ be an arbitrary function.
- Let $\overline{X} = (X_1, \ldots, X_n)$, and $Z = f(\overline{X})$.
- Let

$$\overline{X}^{(k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n), \quad \forall\, k \in \{1, \ldots, n\}.$$

Lemma

Let

$$\xi_k = \mathbb{E}[Z \,|\, X_1, \ldots, X_k] - \mathbb{E}[Z \,|\, X_1, \ldots, X_{k-1}], \quad \forall\, k \in \{1, \ldots, n\}.$$

Then,

$$\text{Var}(Z) = \sum_{k=1}^{n} \mathbb{E}[\xi_k^2]. \tag{1}$$

Proof

- By the smoothing theorem

$$\mathbb{E}[\xi_l \,|\, X_1, \ldots, X_k] = 0, \quad \forall\, l > k.$$

so $\{\xi_k\}_{k=1}^n$ is a martingale-difference sequence w.r.t. the natural filtration $\{\mathscr{F}_k\}$ where

$$\mathscr{F}_k = \sigma(X_1, \ldots, X_k), \quad \forall\, k \in \{0, 1, \ldots, n\}.$$

- From Doob's martingale decomposition

$$Z - \mathbb{E}[Z] = \sum_{k=1}^{n} \xi_k.$$

Martingale Representation: the Variance

- Consequently, we have

$$\mathrm{Var}(Z) = \mathbb{E}\left[\left(\sum_{k=1}^{n} \xi_k\right)^2\right] = \sum_{k=1}^{n} \mathbb{E}[\xi_k^2] + 2\sum_{l>k} \mathbb{E}[\xi_l\,\xi_k].$$

- For $l > k$

$$\begin{aligned} \mathbb{E}[\xi_l \xi_k] &= \mathbb{E}\big[\,\mathbb{E}[\xi_l \xi_k \,|\, X_1, \ldots, X_k]\big] \\ &= \mathbb{E}\big[\xi_k\,\mathbb{E}[\xi_l \,|\, X_1, \ldots, X_k]\big] \\ &= \mathbb{E}[\xi_k \cdot 0] = 0 \end{aligned}$$

thus proving the claim.

First Efron-Stein-Steele Inequality

$$\mathsf{Var}(Z) \le \mathbb{E}\left[\sum_{k=1}^{n} \mathsf{Var}\big(Z \,|\, \overline{X}^{(k)}\big)\right]. \tag{2}$$

Proof

- Since $X_1, \ldots, X_n$ are independent, then

$$\begin{aligned}\xi_k &= \mathbb{E}[Z \,|\, X_1, \ldots, X_k] - \mathbb{E}[Z \,|\, X_1, \ldots, X_{k-1}] \\ &= \mathbb{E}[Z \,|\, X_1, \ldots, X_k] - \mathbb{E}[\mathbb{E}[Z \,|\, \overline{X}^{(k)}] \,|\, X_1, \ldots, X_k] \\ &= \mathbb{E}[Z - \mathbb{E}[Z \,|\, \overline{X}^{(k)}] \,|\, X_1, \ldots, X_k].\end{aligned}$$

Hence, from Jensen's inequality,

$$\xi_k^2 \le \mathbb{E}\big[(Z - \mathbb{E}[Z \,|\, \overline{X}^{(k)}])^2 \,|\, X_1, \ldots, X_k\big], \quad \forall\, k \in \{1, \ldots, n\}.$$

First Efron-Stein-Steele Inequality (Proof - Cont.)

- For all $k \in \{1, \ldots, n\}$, due to the independence of $X_1, \ldots, X_n$,

$$\begin{aligned}\mathbb{E}[\xi_k^2] &\le \mathbb{E}\Big[\mathbb{E}\big[(Z - \mathbb{E}[Z \,|\, \overline{X}^{(k)}])^2 \,|\, X_1, \ldots, X_k\big]\Big] \\ &= \mathbb{E}[(Z - \mathbb{E}[Z \,|\, \overline{X}^{(k)}])^2] \\ &= \mathbb{E}_{\overline{X}^{(k)}}\big[\mathbb{E}_{X_k}(Z - \mathbb{E}[Z \,|\, \overline{X}^{(k)}])^2\big] \\ &= \mathbb{E}_{\overline{X}^{(k)}}\big[\mathsf{Var}(Z \,|\, \overline{X}^{(k)})\big] \\ &= \mathbb{E}\big[\mathsf{Var}(Z \,|\, \overline{X}^{(k)})\big].\end{aligned}$$

- Finally, from (1) and the last inequality, we have

$$\mathsf{Var}(Z) = \sum_{k=1}^{n} \mathbb{E}[\xi_k^2] \le \mathbb{E}\left[\sum_{k=1}^{n} \mathsf{Var}\big(Z \,|\, \overline{X}^{(k)}\big)\right].$$

Second Efron-Stein-Steele Inequality

Let $X_1', \ldots, X_n'$ be independent copies of $X_1, \ldots, X_n$, and let

$$Z_k' = f(X_1, \ldots, X_{k-1}, X_k', X_{k+1}, \ldots, X_n), \quad \forall\, k \in \{1, \ldots, n\}.$$

Then,

$$\text{Var}(Z) \leq \frac{1}{2} \sum_{k=1}^{n} \mathbb{E}[(Z - Z_k')^2]. \tag{3}$$

Proof

- If X, Y are independent copies (i.i.d.), then $\text{Var}(X) = \frac{1}{2}\, \mathbb{E}[(X-Y)^2]$.
- Due to the independence of $X_1, \ldots, X_n$, given $\overline{X}^{(k)}$, the RVs Z, Z_k' are independent copies. Hence, $\text{Var}(Z|\overline{X}^{(k)}) = \frac{1}{2}\, \mathbb{E}[(Z - Z_k')^2]$.
- Now, it follows from the first Efron-Stein-Steele inequality.

Sum of independent RVs

If $Z = \sum_{k=1}^{n} X_k$ then the second Efron-Stein-Steele inequality is satisfied with equality.

Conclusion

Sums of independent RVs are the least concentrated of all functions.

Third Efron-Stein-Steele Inequality

Let

$$Z_k = f_k(X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n), \quad \forall\, k \in \{1, \ldots, n\}$$

for arbitrary functions $f_k \colon \mathcal{X}^{n-1} \to \mathbb{R}$ for all $k \in \{1, \ldots, n\}$. Then,

$$\text{Var}(Z) \le \mathbb{E}\left[\sum_{k=1}^{n} (Z - Z_k)^2\right]. \tag{4}$$

Proof

- $\text{Var}(X) \le \mathbb{E}[(X-a)^2]$ for all $a \in \mathbb{R}$ with equality if $a = \mathbb{E}[X]$.
- Hence, $\text{Var}(Z \,|\, \overline{X}^{(k)}) \le \mathbb{E}[(Z - Z_k)^2 \,|\, \overline{X}^{(k)}]$ for all k.
- From the first Efron-Stein-Steele inequality in (2), we have

$$\text{Var}(Z) \le \mathbb{E}\left[\sum_{k=1}^{n} \mathbb{E}\left[(Z - Z_k)^2 \,|\, \overline{X}^{(k)}\right]\right] = \sum_{k=1}^{n} \mathbb{E}[(Z - Z_k)^2].$$

In statistics, kernel density estimation is a non-parametric way to estimate the probability density function of a random variable.

Example: Kernel Density Estimation (KDE)

- Let $X_1, \ldots, X_n$ be i.i.d. RVs drawn from an unknown *pdf* ϕ.
- Let K be a kernel, i.e., an arbitrary non-negative real-valued integrable function over $\mathbb{R}$ satisfying the following requirements:
  1. $\displaystyle\int_{-\infty}^{\infty} K(u)\,\mathrm{d}u = 1,$
  2. $K(u) = K(-u)$ for all $u \in \mathbb{R}$.
  3. $\lim_{h \to 0} \frac{1}{h} K\left(\frac{x-u}{h}\right) = \delta(x-u)$ where δ is the Dirac function at zero.

  These requirements are satisfied, e.g., by a Gaussian $\mathcal{N}(0,1)$ *pdf*.
- Let $h > 0$ be fixed (h is called a smoothing parameter).
- The KDE is given by

$$\phi_n(x) = \frac{1}{nh} \sum_{k=1}^{n} K\left(\frac{x - X_k}{h}\right), \quad \forall\, x \in \mathbb{R}.$$

KDE Convergence

$$\begin{aligned}\mathbb{E}[\phi_n(x)] &= \frac{1}{nh}\sum_{k=1}^{n}\int_{-\infty}^{\infty} K\left(\frac{x-u}{h}\right)\phi(u)\,\mathrm{d}u \\ &= \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{x-u}{h}\right)\phi(u)\,\mathrm{d}u\end{aligned}$$

$$\Rightarrow \lim_{h\to 0}\mathbb{E}[\phi_n(x)] = \int_{-\infty}^{\infty}\delta(x-u)\,\phi(u)\,\mathrm{d}u = \phi(x), \quad \forall\, x \in \mathbb{R}.$$

L_1 Error of the KDE

$$Z_n = f(X_1,\ldots,X_n) \triangleq \int_{-\infty}^{\infty} |\phi_n(x)-\phi(x)|\,\mathrm{d}x.$$

In the following, we invoke the second Efron-Stein-Steele inequality (3) to provide a rigorous simple bound for the variance of Z_n.

L_1 Error of the KDE (Cont.)

$$\begin{aligned}
&\left|f(x_1,\dots,x_{k-1},x_k,x_{k+1},\dots,x_n) - f(x_1,\dots,x_{k-1},x'_k,x_{k+1},\dots,x_n)\right| \\
&\le \frac{1}{nh}\int_{-\infty}^{\infty}\left|K\left(\frac{x-x_k}{h}\right) - K\left(\frac{x-x'_k}{h}\right)\right| \, \mathrm{d}x \\
&\le \frac{1}{nh}\left[\int_{-\infty}^{\infty}K\left(\frac{x-x_k}{h}\right)\mathrm{d}x + \int_{-\infty}^{\infty}K\left(\frac{x-x'_k}{h}\right)\mathrm{d}x\right] \\
&= \frac{2}{n}.
\end{aligned}$$

From the second Efron-Stein-Steele inequality in (3), it follows that (Devroye, 1991)

$$\mathrm{Var}(Z_n) \le \frac{1}{2}\sum_{k=1}^{n}\left(\frac{2}{n}\right)^2 = \frac{2}{n}.$$

L_1 Error of the KDE (Cont.)

It is known that (Devroye - 1991)

$$\lim_{n\to\infty} \sqrt{n}\,\mathbb{E}Z_n = +\infty. \tag{5}$$

By Chebyshev's inequality, it follows that for every $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{Z_n}{\mathbb{E}[Z_n]} - 1\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(Z_n)}{\varepsilon^2\left(\mathbb{E}[Z_n]\right)^2} \leq \frac{2}{n\varepsilon^2\left(\mathbb{E}[Z_n]\right)^2}. \tag{6}$$

Hence, from (5) and (6),

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{Z_n}{\mathbb{E}[Z_n]} - 1\right| \geq \varepsilon\right) = 0, \quad \forall\, \varepsilon > 0$$

so $\frac{Z_n}{\mathbb{E}[Z_n]} \to 1$ in probability.

Weakly Self-Bounding Functions

A function $f\colon \mathcal{X}^n \to [0,\infty)$ is called a weakly (a,b) self-bounding function if there exist n functions $f_k\colon \mathcal{X}^{n-1} \to \mathbb{R}$, for $k \in \{1,\ldots,n\}$, such that

$$\sum_{k=1}^{n} \bigl(f(\overline{x}) - f_k(\overline{x}^{(k)})\bigr)^2 \le a\, f(\overline{x}) + b, \quad \forall\, \overline{x} \in \mathcal{X}^n.$$

Proposition

If f is a weakly (a,b) self-bounding function, then

$$\mathrm{Var}\bigl(f(\overline{X})\bigr) \le a\, \mathbb{E}\bigl[f(\overline{X})\bigr] + b. \tag{7}$$

Proof

Take $Z = f(\overline{X})$, and $Z_k = f_k(\overline{X}^{(k)})$ for $k \in \{1,\ldots,n\}$ in the third Efron-Stein-Steele inequality (4).

Weakly Self-Bounding Functions (Cont.)

Corollary

If

$$0 \leq f(\overline{x}) - f_k(\overline{x}^{(k)}) \leq 1, \quad \forall\, \overline{x} \in \mathcal{X}^n,\ k \in \{1, \ldots, n\}$$

and

$$\sum_{k=1}^{n} \big(f(\overline{x}) - f_k(\overline{x}^{(k)})\big) \leq f(\overline{x}), \quad \forall\, \overline{x} \in \mathcal{X}^n$$

then,

$$\mathit{Var}\big(f(\overline{X})\big) \leq \mathbb{E}\big[f(\overline{X})\big].$$

Proof

The last proposition is satisfied with $a = 1, b = 0$.

Before we proceed to an information-theoretic example, we introduce:

Han's Inequality for the Shannon Entropy

Let $X_1, \ldots, X_n$ be (possibly dependent) RVs, and let

$$\mathbf{X} = (X_1, \ldots, X_n), \quad \mathbf{X}^{(k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)$$

Then,

$$\sum_{k=1}^{n} \bigl(H(\mathbf{X}) - H(\mathbf{X}^{(k)})\bigr) \leq H(\mathbf{X}). \tag{8}$$

Proof

$$\begin{aligned} \sum_{k=1}^{n} \bigl(H(\mathbf{X}) - H(\mathbf{X}^{(k)})\bigr) &= \sum_{k=1}^{n} H(X_k \,|\, \mathbf{X}^{(k)}) \\ &\leq \sum_{k=1}^{n} H(X_k \,|\, X_1, \ldots, X_{k-1}) \\ &= H(\mathbf{X}). \end{aligned}$$

Han's Inequality (Equivalent Form)

$$\frac{1}{n}\sum_{k=1}^{n} H(\mathbf{X}^{(k)}) \leq H(\mathbf{X}) \leq \frac{1}{n-1}\sum_{k=1}^{n} H(\mathbf{X}^{(k)}).$$

Proof

The LHS is trivial since

$$H(\mathbf{X}^{(k)}) \leq H(\mathbf{X}), \quad \forall\, k \in \{1, \ldots, n\}.$$

The RHS is an equivalent form of (8).

Definition

A finite set of points in the plane is said to be in convex position if the interior of the polygon whose vertices are these points forms a convex set.

Combinatorial Entropies

Consider the following problem:

- Let $X_1, \ldots, X_n$ be independent points in $\mathbb{R}^2$ (not necessarily i.i.d.).
- Let N be the number of subsets of $\{X_1, \ldots, X_n\}$ that are in convex position.

Based on the last corollary for weakly self-bounding functions, we prove that

$$\text{Var}[\log_2 N] \leq \mathbb{E}[\log_2 N].$$

Combinatorial Entropies (Cont.)

The proof goes as follows:

- Let $\overline{x} = (x_1, \dots, x_n)$ be an arbitrary vector of points in the plane.
- Let N be the number of subsets of $\{x_k\}_{k=1}^n$ in convex position (N is a deterministic function of $\overline{x}$).
- Draw uniformly at random one of these N subsets. By assumption, it is in convex position.
- For such a subset, let $\{Y_k\}_{k=1}^n$ be an n-length binary sequence where $Y_k = 1$ if x_k is one of the vertices in the chosen subset.
- Let $f \colon \mathcal{X}^n \to \mathbb{R}$ and $f_k \colon \mathcal{X}^{n-1} \to \mathbb{R}$ be defined to satisfy

$$f(\overline{x}) = H(Y_1, \dots, Y_n),$$

$$f_k(\overline{x}^{(k)}) \geq H(Y_1, \dots, Y_{k-1}, Y_{k+1}, \dots, Y_n), \quad \forall\, k \in \{1, \dots, n\}.$$

Combinatorial Entropies (Cont.)

- Since the subset is chosen uniformly at random among N possible subsets, the random vector $(Y_1, \ldots, Y_n)$ has a uniform distribution on a size-N subset of $\{0,1\}^n$. Hence,

$$f(\overline{x}) = H(Y_1, \ldots, Y_n) = \log_2 N.$$

- Checking the first condition of the corollary, for all $k \in \{1, \ldots, n\}$,

$$\begin{aligned} & f(\overline{x}) - f_k(\overline{x}^{(k)}) \\ & \leq H(Y_1, \ldots, Y_n) - H(Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n) \\ & = H(Y_k \,|\, Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n) \end{aligned}$$

$$\Rightarrow 0 \leq f(\overline{x}) - f_k(\overline{x}^{(k)}) \leq H(Y_k) \leq 1, \quad \forall \, \overline{x}, \; k \in \{1, \ldots, n\}$$

since Y_k is a binary RV.

Combinatorial Entropies (Cont.)

- Checking the second condition of the corollary, from Han's inequality

$$
\begin{aligned}
&\sum_{k=1}^{n} \bigl(f(\overline{x}) - f(\overline{x}^{(k)})\bigr) \\
&= \sum_{k=1}^{n} \bigl(H(Y_1, \ldots, Y_n) - H(Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n)\bigr) \\
&\leq H(Y_1, \ldots, Y_n) \\
&= f(\overline{x}).
\end{aligned}
$$

- Hence, it follows from the corollary that

$$\mathsf{Var}\bigl(f(\overline{X})\bigr) \leq \mathbb{E}\bigl[f(\overline{X})\bigr].$$

- Recall that $f(\overline{x}) = \log_2 N$, $N = N(\overline{x})$. This completes the proof.

Han's Inequality for the Relative Entropy

- Let $Q = Q_1 \otimes \ldots \otimes Q_n$ be a product probability distribution on $\mathcal{X}^n$, and let P be an arbitrary probability distribution defined on $\mathcal{X}^n$.
- For all $k \in \{1, \ldots, n\}$ and $\overline{x}^{(k)} \in \mathcal{X}^{n-1}$, let $P^{(k)}$ and $Q^{(k)}$ be the marginals

$$P^{(k)}(\overline{x}^{(k)}) = \mathbb{E}\big[P(x_1, \ldots, x_{k-1}, X_k, x_{k+1}, \ldots, x_n)\big],$$

$$Q^{(k)}(\overline{x}^{(k)}) = \mathbb{E}\big[Q(x_1, \ldots, x_{k-1}, X_k, x_{k+1}, \ldots, x_n)\big] = \prod_{j \neq k} Q_j(x_j).$$

Then,

$$D(P\|Q) \geq \frac{1}{n-1} \sum_{k=1}^{n} D(P^{(k)}\|Q^{(k)}) \tag{9}$$

or, equivalently,

$$D(P\|Q) \leq \sum_{k=1}^{n} \Big(D(P\|Q) - D(P^{(k)}\|Q^{(k)})\Big). \tag{10}$$

Proof of Han's Inequality for the Relative Entropy

- Let $\overline{X} \sim P$, and $\overline{X}^{(k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_n)$.
- From Han's inequality for the Shannon entropy, we have

$$H(\overline{X}) \leq \frac{1}{n-1} \sum_{k=1}^{n} H\big(\overline{X}^{(k)}\big)$$

which is equivalent to

$$\sum_{\overline{x} \in \mathcal{X}^n} P(\overline{x}) \log P(\overline{x}) \geq \frac{1}{n-1} \sum_{k=1}^{n} \sum_{\overline{x}^k \in \mathcal{X}^{n-1}} P^{(k)}(\overline{x}^{(k)}) \, \log P^{(k)}(\overline{x}^{(k)}).$$

- To complete the proof of (9), it suffices to prove that

$$\sum_{\overline{x} \in \mathcal{X}^n} P(\overline{x}) \log Q(\overline{x}) = \frac{1}{n-1} \sum_{k=1}^{n} \sum_{\overline{x}^k \in \mathcal{X}^{n-1}} P^{(k)}(\overline{x}^{(k)}) \, \log Q^{(k)}(\overline{x}^{(k)}).$$

Proof of Han's Inequality for the Relative Entropy (Cont.)

- Since Q is a product measure, then $\log Q(\overline{x}) = \log Q^{(k)}(\overline{x}^k) + \log Q(x_k)$ for all $k \in \{1, \ldots, n\}$.

$$\Rightarrow \sum_{\overline{x} \in \mathcal{X}^n} P(\overline{x}) \, \log Q(\overline{x})$$

$$= \frac{1}{n} \sum_{k=1}^{n} \sum_{\overline{x} \in \mathcal{X}^n} P(\overline{x}) \Bigl(\log Q^{(k)}(\overline{x}^k) + \log Q(x_k) \Bigr)$$

$$= \frac{1}{n} \sum_{k=1}^{n} \sum_{\overline{x}^k \in \mathcal{X}^{n-1}} \sum_{x_k \in \mathcal{X}} P(\overline{x}) \, \log Q^{(k)}(\overline{x}^k) + \frac{1}{n} \sum_{\overline{x} \in \mathcal{X}^n} P(\overline{x}) \, \log Q(\overline{x})$$

$$= \frac{1}{n} \sum_{k=1}^{n} \sum_{\overline{x}^k \in \mathcal{X}^{n-1}} P^{(k)}(\overline{x}^{(k)}) \, \log Q^{(k)}(\overline{x}^k) + \frac{1}{n} \sum_{\overline{x} \in \mathcal{X}^n} P(\overline{x}) \, \log Q(\overline{x})$$

Proof of Han's Inequality for the Relative Entropy (Cont.)

- Rearranging terms in the last equality gives

  $$\sum_{\overline{x} \in \mathcal{X}^n} P(\overline{x}) \log Q(\overline{x}) = \frac{1}{n-1} \sum_{k=1}^{n} \sum_{\overline{x}^k \in \mathcal{X}^{n-1}} P^{(k)}(\overline{x}^{(k)}) \, \log Q^{(k)}(\overline{x}^{(k)})$$

  as required. This therefore completes the proof of (9).
- Eq. (10) is an equivalent form of (9) (as it follows by rearranging terms). This completes the proof of (10).

Basic Concentration Inequalities with the Martingale Approach

We are interested in sharp bounds on the *deviation probabilities*

$$\mathbb{P}\left(|U - \mathbb{E}U| \geq r\right), \quad \forall\, r > 0$$

where U is a real-valued RV that may be a function of a large number n of independent or weakly dependent RVs $X_1, \ldots, X_n$.

In a nutshell, the martingale approach in establishing concentration has two basic ingredients:

1. The martingale decomposition.
2. The Chernoff bounding technique.

The Martingale Decomposition (Doob)

- We first construct a suitable filtration $\{\mathscr{F}_i\}_{i=0}^n$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ that carries U, where $\mathscr{F}_0 = \{\emptyset, \Omega\}$ is the trivial σ-algebra, and $\mathscr{F}_n = \mathscr{F}$.
- Then, we decompose the difference $U - \mathbb{E}U$ as

$$\begin{aligned} U - \mathbb{E}[U] &= \mathbb{E}[U|\mathscr{F}_n] - \mathbb{E}[U|\mathscr{F}_0] \\ &= \sum_{k=1}^{n} \left(\mathbb{E}[U|\mathscr{F}_k] - \mathbb{E}[U|\mathscr{F}_{k-1}]\right). \end{aligned} \tag{11}$$

- The idea is to choose the σ-algebras $\{\mathscr{F}_k\}$ in such a way that the differences $\xi_k = \mathbb{E}[U|\mathscr{F}_k] - \mathbb{E}[U|\mathscr{F}_{k-1}]$ are bounded in some sense, e.g., almost surely (a.s.).
- The following equality holds $U - \mathbb{E}[U] = \sum_{k=1}^n \xi_k$ where, under the above assumption, $U - \mathbb{E}[U]$ is expressed as a sum of a bounded martingale-difference sequence.

The Chernoff Bounding Technique

- Let

$$\psi(\lambda) \triangleq \ln \mathbb{E}\Big[\exp\big(\lambda(U - \mathbb{E}[U])\big)\Big], \quad \forall\, \lambda \geq 0$$

designate the logarithmic moment-generating function of U.

- Using Markov's inequality, it follows that for non-negative λ

$$\begin{aligned}
&\mathbb{P}(U - \mathbb{E}[U] \geq t) \\
&= \mathbb{P}\big[\exp\big(\lambda\big(U - \mathbb{E}[U]\big) \geq \exp(\lambda t)\big)\big] \\
&\leq \exp(-\lambda t)\, \mathbb{E}\big[\exp\big(\lambda\big(U - \mathbb{E}[U]\big)\big)\big] \\
&\leq \exp\left(-\big(\lambda t - \psi(\lambda)\big)\right), \quad \forall\, \lambda \geq 0.
\end{aligned}$$

- Let

$$\psi^\star(t) = \sup_{\lambda \geq 0} \big(\lambda t - \psi(\lambda)\big)$$

be the Legendre dual of the logarithmic moment-generating function.

The Chernoff Bounding Technique (Cont.)

- An optimization over $\lambda \geq 0$ gives

$$\mathbb{P}(U - \mathbb{E}[U] \geq t) \leq \exp\bigl(-\psi^{\star}(t)\bigr), \quad \forall\, t \geq 0.$$

  Hence, the problem of bounding the deviation probability is reduced to the analysis of the logarithmic moment-generating function.

- Sanity check:

$$U \sim \mathcal{N}(m, \sigma^2), \quad \psi(\lambda) = \frac{\lambda^2\,\sigma^2}{2}, \quad \psi^{\star}(t) = \frac{t^2}{2\sigma^2}$$

  so, we get

$$\mathbb{P}(U - \mathbb{E}[U] \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \forall\, t \geq 0$$

  which is asymptotically tight as $t \gg 1$ since

$$\mathbb{P}(U - \mathbb{E}[U] \geq t) = Q\left(\frac{t}{\sigma}\right).$$

The Chernoff Bounding Technique (Cont.)

- To bound the probability of the lower tail, $\mathbb{P}(U - \mathbb{E}[U] \leq -t)$, for $t \geq 0$, we follow the same steps, but with $-U$ instead of U.
- The success of the whole enterprise hinges on our ability to obtain tight upper bounds on $\psi(\lambda)$.
- Exploiting the martingale decomposition (11), we write

$$\psi(\lambda) = \ln \mathbb{E}\left[\prod_{k=1}^{n} \exp(\lambda \xi_k)\right]$$

  which allows focusing on the behavior of individual terms $\exp(t\xi_k)$ (as it is clarified later).
- The logarithmic moment-generating function plays a key role in the theory of large deviations, which can be thought of as an asymptotic analysis of the concentration of measure phenomenon.

Hoeffding's Lemma

Let U be a real-valued random variable, such that $U \in [a, b]$ a.s. for some finite $a < b$. Then,

$$\mathbb{E}\left[\exp\left(t(U - \mathbb{E}U)\right)\right] \le \exp\left(\frac{t^2(b-a)^2}{8}\right), \quad \forall\, t \in \mathbb{R}.$$

Proof of Hoeffding's Lemma

- Define the function

$$H_p(\lambda) \triangleq \ln\left(pe^{\lambda(1-p)} + (1-p)e^{-\lambda p}\right)$$

where $p \in [0,1]$ and $\lambda \in \mathbb{R}$.

- Let $\xi = U - \mathbb{E}U$, where $\xi \in [a - \mathbb{E}U, b - \mathbb{E}U]$.
- Using the convexity of the exponential function, we have

$$\begin{aligned}\exp(t\xi) &= \exp\left(\frac{U-a}{b-a} \cdot t(b - \mathbb{E}U) + \frac{b-U}{b-a} \cdot t(a - \mathbb{E}U)\right) \\ &\leq \left(\frac{U-a}{b-a}\right)\exp\left(t(b-\mathbb{E}U)\right) + \left(\frac{b-U}{b-a}\right)\exp\left(t(a-\mathbb{E}U)\right).\end{aligned}$$

- Taking expectations of both sides, we get $\mathbb{E}[\exp(t\xi)] \leq \exp\left(H_p(\lambda)\right)$ where $p = \frac{\mathbb{E}U-a}{b-a}$ and $\lambda = t(b-a)$.

Proof (cont.)

- Consequently, it suffices to prove that for every $\lambda \in \mathbb{R}$

$$H_p(\lambda) \leq \frac{\lambda^2}{8}, \quad \forall \, p \in [0,1].$$

- We have $H_p(0) = H_p'(0) = 0$, and

$$H_p''(\lambda) = \frac{1}{4} \, \frac{pe^{\lambda} \cdot (1-p)}{\left(\frac{pe^{\lambda}+(1-p)}{2}\right)^2} \leq \frac{1}{4}, \quad \forall \, \lambda \in \mathbb{R}, \; p \in [0,1]$$

  where the last inequality holds since the geometric mean is less than or equal to the arithmetic mean.

- Hence, the upper bound follows from Taylor's series expansion of on $H_p(\lambda)$ (around zero) with a remainder of order 2.

Concentration via the Martingale Approach

Theorem - [Azuma-Hoeffding Inequality]

Let $\{X_k, \mathscr{F}_k\}_{k=0}^n$ be a real-valued martingale sequence. Suppose that there exist nonnegative reals $d_1, \dots, d_n$, such that $|X_k - X_{k-1}| \leq d_k$ a.s. for all $k \in \{1, \dots, n\}$. Then,

$$\mathbb{P}(|X_n - X_0| \geq r) \leq 2 \exp\left(-\frac{r^2}{2\sum_{k=1}^n d_k^2}\right), \quad \forall\, r > 0. \tag{12}$$

Ingredients of the Proof

1. Martingale decomposition
2. Chernoff's bound
3. Hoeffding's lemma

Proof of the Azuma-Hoeffding Inequality

- Let $\xi_k \triangleq X_k - X_{k-1}$ for $k \in \{1, \dots, n\}$ denote the differences of the martingale sequence. By hypothesis,
  - $|\xi_k| \leq d_k, \quad \forall\, k \in \{1, \dots, n\}$
  - $\mathbb{E}[\xi_k \,|\, \mathscr{F}_{k-1}] = 0$ a.s.
- By the martingale decomposition, $X_n - X_0 = \sum_{k=1}^n \xi_k$.
- Invoking Chernoff's bound gives

$$\mathbb{P}(X_n - X_0 \geq r) \leq \exp(-tr)\, \mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right], \quad \forall\, t \geq 0.$$

- By the tower property for conditional expectations,

$$\mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right] = \mathbb{E}\left[\exp\left(t \sum_{k=1}^{n-1} \xi_k\right) \mathbb{E}\big[\exp(t\xi_n) \,|\, \mathscr{F}_{n-1}\big]\right]$$

which holds since $\exp\big(t \sum_{k=1}^{n-1} \xi_k\big)$ is $\mathscr{F}_{n-1}$-measurable.

Proof (Cont.)

- We now apply Hoeffding's lemma with the conditioning on $\mathscr{F}_{n-1}$. Since $\mathbb{E}[\xi_n | \mathscr{F}_{n-1}] = 0$ and $\xi_n \in [-d_n, d_n]$ a.s., then

$$\mathbb{E}\big[\exp(t\xi_n) \,|\, \mathscr{F}_{n-1}\big] \leq \exp\left(\frac{t^2\, d_n^2}{2}\right).$$

- Continuing recursively in a similar manner gives

$$\mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right)\right] \leq \prod_{k=1}^{n} \exp\left(\frac{t^2\, d_k^2}{2}\right) = \exp\left(\frac{t^2}{2} \sum_{k=1}^{n} d_k^2\right).$$

- Substituting this bound into the Chernoff bound gives

$$\mathbb{P}(X_n - X_0 \geq r) \leq \exp\left(-tr + \frac{t^2}{2} \sum_{k=1}^{n} d_k^2\right), \quad \forall\, t \geq 0.$$

Proof (Cont.)

- Optimization of the bound w.r.t. $t \geq 0$ gives $t = r \left(\sum_{k=1}^n d_k^2\right)^{-1}$.
- The substitution of the optimized t in the bound gives

$$\mathbb{P}(X_n - X_0 \geq r) \leq \exp\left(-\frac{r^2}{2\sum_{k=1}^n d_k^2}\right), \; \forall \, r > 0.$$

- Since $\{X_k, \mathscr{F}_k\}$ is a martingale with bounded differences, so is $\{-X_k, \mathscr{F}_k\}$ (with the same bounds on its differences).
- This implies that the same bound is also valid for the probability $\mathbb{P}(X_n - X_0 \leq -r)$.
- Using these bounds, and the equality

$$\mathbb{P}(|X_n - X_0| \geq r) = \mathbb{P}(X_n - X_0 \geq r) + \mathbb{P}(X_n - X_0 \leq -r), \; \forall \, r > 0$$

completes the proof.

Example

- Let $\{Y_i\}_{i=0}^{\infty}$ be i.i.d. RVs where, for a fixed $d > 0$, $\mathbb{P}(Y_i = \pm d) = \frac{1}{2}$.
- Let $X_k = \sum_{i=0}^{k} Y_i$ for $k \in \{0, 1, \ldots, \}$, and define the natural filtration $\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \ldots$ where

$$\mathscr{F}_k = \sigma(Y_0, \ldots, Y_k)\,, \quad \forall\, k \in \{0, 1, \ldots, \}$$

  is the σ-algebra generated by $Y_0, \ldots, Y_k$.
- Note that $\{X_k, \mathscr{F}_k\}_{k=0}^{\infty}$ is a martingale sequence, and

$$|X_k - X_{k-1}| = |Y_k| = d, \quad \forall\, k \in \mathbb{N}.$$

- From the Azuma–Hoeffding inequality

$$\mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) \leq 2\exp\left(-\frac{\alpha^2}{2d^2}\right), \quad \forall\, \alpha \geq 0,\ n \in \mathbb{N}.$$

Example (Cont.)

- From the Central Limit Theorem (CLT)

$$\frac{1}{\sqrt{n}}(X_n - X_0) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} Y_k \Rightarrow \mathcal{N}(0, d^2).$$

$$\Rightarrow \lim_{n \to \infty} \mathbb{P}(|X_n - X_0| \geq \alpha\sqrt{n}) = 2\,Q\left(\frac{\alpha}{d}\right), \quad \forall \alpha \geq 0.$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t, \quad \forall\, x \in \mathbb{R}$$

is the complementary standard Gaussian CDF (a.k.a. the Q-function).

- The Q-function satisfies the exponential bounds:

$$\frac{1}{\sqrt{2\pi}} \frac{x}{1+x^2} \cdot \exp\left(-\frac{x^2}{2}\right) < Q(x) < \frac{1}{\sqrt{2\pi}\,x} \cdot \exp\left(-\frac{x^2}{2}\right), \ \ \forall\, x > 0.$$

Example (Cont.)

- The Azuma–Hoeffding inequality in the last example is exponentially tight.

Nevertheless, the Azuma-Hoeffding inequality is not tight in general.

$$r > \sum_{i=1}^{n} d_i \;\Rightarrow\; \mathbb{P}(|X_n - X_0| \geq r) = 0.$$

Theorem (Th. 2 – McDiarmid '89)

Let $\{X_k, \mathscr{F}_k\}_{k=0}^{\infty}$ *be a discrete-parameter real-valued martingale. Assume that, for some constants* $d, \sigma > 0$, *the following two requirements hold a.s.*

$$|X_k - X_{k-1}| \leq d,$$

$$\mathit{Var}(X_k | \mathscr{F}_{k-1}) = \mathbb{E}\big[(X_k - X_{k-1})^2 \,|\, \mathscr{F}_{k-1}\big] \leq \sigma^2$$

for every $k \in \{1, \ldots, n\}$. *Then, for every* $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n\, D\left(\frac{\delta + \gamma}{1 + \gamma} \Big|\Big| \frac{\gamma}{1 + \gamma}\right)\right)$$

where $\gamma \triangleq \frac{\sigma^2}{d^2}$, $\delta \triangleq \frac{\alpha}{d}$, *and* $D(p||q) \triangleq p \ln\left(\frac{p}{q}\right) + (1-p) \ln\left(\frac{1-p}{1-q}\right)$.

Corollary

Under the conditions of Theorem 2, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n f(\delta)\right)$$

where

$$f(\delta) = \begin{cases} \ln(2)\Big[1 - h_2\left(\frac{1-\delta}{2}\right)\Big], & 0 \leq \delta \leq 1 \\ +\infty, & \delta > 1 \end{cases}$$

and $h_2(x) \triangleq -x \log_2(x) - (1-x)\log_2(1-x)$ for $0 \leq x \leq 1$.

Corollary

Under the conditions of Theorem 2, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n f(\delta)\right)$$

where

$$f(\delta) = \begin{cases} \ln(2)\left[1 - h_2\left(\frac{1-\delta}{2}\right)\right], & 0 \leq \delta \leq 1 \\ +\infty, & \delta > 1 \end{cases}$$

and $h_2(x) \triangleq -x \log_2(x) - (1-x) \log_2(1-x)$ for $0 \leq x \leq 1$.

Proof

Set $\gamma = 1$ in Theorem 2.

Corollary

Under the conditions of Theorem 2, for every $\alpha \geq 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) \leq 2 \exp\left(-n f(\delta)\right)$$

where

$$f(\delta) = \begin{cases} \ln(2)\Big[1 - h_2\left(\frac{1-\delta}{2}\right)\Big], & 0 \leq \delta \leq 1 \\ +\infty, & \delta > 1 \end{cases}$$

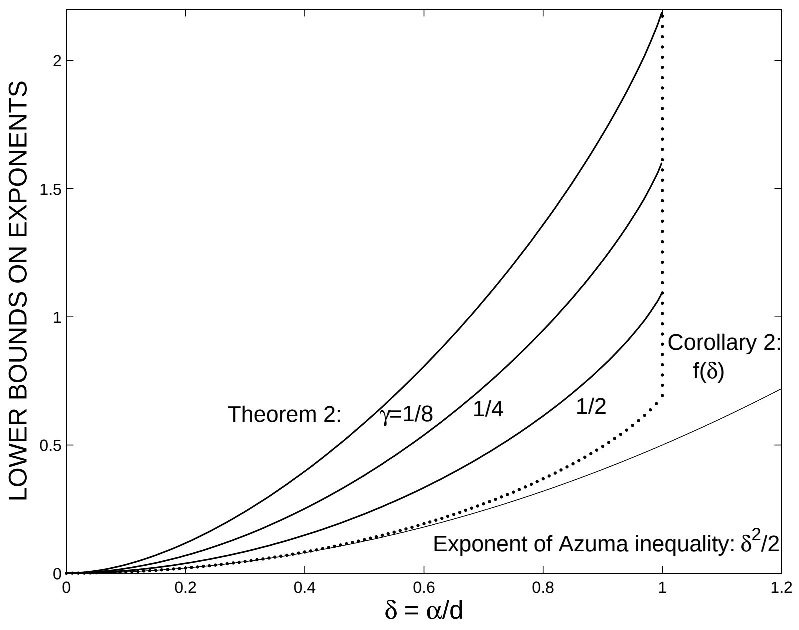
and $h_2(x) \triangleq -x \log_2(x) - (1-x) \log_2(1-x)$ for $0 \leq x \leq 1$.

Proof

Set $\gamma = 1$ in Theorem 2.

Observation (first set $\gamma = 1$, and then use Pinsker's Inequality)

Theorem 2 ⇒ Corollary ⇒ Azuma-Hoeffding inequality.


LOWER BOUNDS ON EXPONENTS
$\delta = \alpha/d$
0
0.2
0.4
0.6
0.8
1
1.2
0
0.5
1
1.5
2
Theorem 2:
γ=1/8
1/4
1/2
Corollary 2:
f(δ)
Exponent of Azuma inequality: $\delta^2/2$

Proof Technique for Theorem 2

- For an arbitrary $\alpha > 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) = \mathbb{P}(X_n - X_0 \geq \alpha n) + \mathbb{P}(X_n - X_0 \leq -\alpha n).$$

Proof Technique for Theorem 2

- For an arbitrary $\alpha > 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) = \mathbb{P}(X_n - X_0 \geq \alpha n) + \mathbb{P}(X_n - X_0 \leq -\alpha n).$$

- Define $X_n - X_0 = \sum_{k=1}^{n} \xi_k$ where $\xi_k \triangleq X_k - X_{k-1}$.

Proof Technique for Theorem 2

- For an arbitrary $\alpha > 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) = \mathbb{P}(X_n - X_0 \geq \alpha n) + \mathbb{P}(X_n - X_0 \leq -\alpha n).$$

- Define $X_n - X_0 = \sum_{k=1}^{n} \xi_k$ where $\xi_k \triangleq X_k - X_{k-1}$.
- $|\xi_k| \leq d$, $\mathbb{E}\big[\xi_k \,|\, \mathscr{F}_{k-1}\big] = 0$, $\text{Var}(\xi_k \,|\, \mathscr{F}_{k-1}) \leq \gamma d^2$. The RV ξ_k is $\mathscr{F}_k$-measurable.

Proof Technique for Theorem 2

- For an arbitrary $\alpha > 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) = \mathbb{P}(X_n - X_0 \geq \alpha n) + \mathbb{P}(X_n - X_0 \leq -\alpha n).$$

- Define $X_n - X_0 = \sum_{k=1}^{n} \xi_k$ where $\xi_k \triangleq X_k - X_{k-1}$.
- $|\xi_k| \leq d$, $\mathbb{E}\big[\xi_k \,|\, \mathscr{F}_{k-1}\big] = 0$, $\text{Var}(\xi_k \,|\, \mathscr{F}_{k-1}) \leq \gamma d^2$. The RV ξ_k is $\mathscr{F}_k$-measurable.
- From Chernoff's bound, for every $\alpha \geq 0$,

$$\mathbb{P}\left(X_n - X_0 \geq n\alpha\right) \leq e^{-n\alpha t}\, \mathbb{E}\left[\exp\left(t \sum_{k=1}^{n} \xi_k\right)\right], \quad \forall\, t \geq 0.$$

Proof Technique for Theorem 2

- For an arbitrary $\alpha > 0$,

$$\mathbb{P}(|X_n - X_0| \geq \alpha n) = \mathbb{P}(X_n - X_0 \geq \alpha n) + \mathbb{P}(X_n - X_0 \leq -\alpha n).$$

- Define $X_n - X_0 = \sum_{k=1}^n \xi_k$ where $\xi_k \triangleq X_k - X_{k-1}$.
- $|\xi_k| \leq d$, $\mathbb{E}\bigl[\xi_k \,|\, \mathscr{F}_{k-1}\bigr] = 0$, $\text{Var}(\xi_k \,|\, \mathscr{F}_{k-1}) \leq \gamma d^2$. The RV ξ_k is $\mathscr{F}_k$-measurable.
- From Chernoff's bound, for every $\alpha \geq 0$,

$$\mathbb{P}\left(X_n - X_0 \geq n\alpha\right) \leq e^{-n\alpha t}\, \mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right], \quad \forall\, t \geq 0.$$

- Since $\{\mathscr{F}_i\}$ forms a filtration, then for every $t \geq 0$

$$\mathbb{E}\left[\exp\left(t \sum_{k=1}^n \xi_k\right)\right] = \mathbb{E}\left[\exp\left(t \sum_{k=1}^{n-1} \xi_k\right) \mathbb{E}\bigl[\exp(t\xi_n) \,|\, \mathscr{F}_{n-1}\bigr]\right].$$

Proof Technique for Theorem 2 (Cont.)

- For proving Theorem 2, the use of Bennett's inequality gives

$$\mathbb{E}\left[\exp(t\xi_k) \,|\, \mathscr{F}_{k-1}\right] \leq \frac{\gamma \exp(td) + \exp(-\gamma t d)}{1+\gamma}.$$

Proof Technique for Theorem 2 (Cont.)

- For proving Theorem 2, the use of Bennett's inequality gives

$$\mathbb{E}\left[\exp(t\xi_k) \,|\, \mathscr{F}_{k-1}\right] \leq \frac{\gamma \exp(td) + \exp(-\gamma td)}{1+\gamma}.$$

- In both cases, one gets exponential bounds with the free parameter t.

Proof Technique for Theorem 2 (Cont.)

- For proving Theorem 2, the use of Bennett's inequality gives

$$\mathbb{E}\left[\exp(t\xi_k) \,|\, \mathscr{F}_{k-1}\right] \leq \frac{\gamma \exp(td) + \exp(-\gamma td)}{1+\gamma}.$$

- In both cases, one gets exponential bounds with the free parameter t.
- Optimization over $t \geq 0$ & union bound to get two-sided inequalities.

A prominent application of the martingale approach is a powerful inequality due to McDiarmid, a.k.a. the *bounded-difference inequality*.

McDiarmid's Inequality

Setting:

- Let $\mathcal{X}$ be a set.
- Let $f\colon \mathcal{X}^n \to \mathbb{R}$ satisfy the *bounded difference assumption*

$$\sup_{x_1,\ldots,x_n,x_i'\in\mathcal{X}} \Big| f(x_1,\ldots,x_{i-1},x_i,x_{i+1}\ldots,x_n) \\ - f(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)\Big| \le d_i \qquad (13)$$

  for all $i \in \{1,\ldots,n\}$, where $d_1,\ldots,d_n$ are arbitrary nonnegative constants.

- This is equivalent to saying that, for every given i, the variation of the function f with respect to its ith coordinate is upper bounded by d_i.

McDiarmid's Inequality (Cont.)

Theorem (McDiarmid's inequality)

Let $\{X_k\}_{k=1}^n$ *be independent (though not necessarily i.i.d.) random variables taking values in a set* $\mathcal{X}$*. Consider a random variable* $U = f(X^n)$ *where* $f\colon \mathcal{X}^n \to \mathbb{R}$ *is a measurable function that satisfies the bounded difference assumption* (13)*, and* $X^n \triangleq (X_1, \dots, X_n)$*. Then, for every* $r \geq 0$,

$$\mathbb{P}\left(\left|U - \mathbb{E}U\right| \geq r\right) \leq 2 \exp\left(-\frac{2r^2}{\sum_{k=1}^n d_k^2}\right). \tag{14}$$

McDiarmid's Inequality (Cont.)

Remark

1. *The Azuma–Hoeffding inequality can be used for a derivation of a concentration inequality in the considered setting.*
2. *The proof of McDiarmid's Inequality, however, provides an improvement by a factor of 4 in the exponent of the bound.*

Example: The Azuma–Hoeffding and McDiarmid's Inequality

- Let $g \colon \{1, \ldots, n\} \to \{1, \ldots, n\}$ be chosen uniformly at random from all n^n such possible functions.
- Let $L(g)$ denote the number of values $y \in \{1, \ldots, n\}$ for which the equation $g(x) = y$ has no solution (i.e., $g(x) \neq y$ for every $x \in \{1, \ldots, n\}$).

Example: The Azuma–Hoeffding and McDiarmid's Inequality

- Let $g \colon \{1, \ldots, n\} \to \{1, \ldots, n\}$ be chosen uniformly at random from all n^n such possible functions.
- Let $L(g)$ denote the number of values $y \in \{1, \ldots, n\}$ for which the equation $g(x) = y$ has no solution (i.e., $g(x) \neq y$ for every $x \in \{1, \ldots, n\}$).
- By the linearity of the expectation, we have

$$\mathbb{E}[L(g)] = n \left(1 - \frac{1}{n}\right)^n .$$

Consequently, for every $n \in \mathbb{N}$,

$$\frac{n-1}{e} < \mathbb{E}[L(g)] < \frac{n}{e}. \qquad (15)$$

Example: The Azuma–Hoeffding and McDiarmid's Inequality

- Let $g\colon \{1, \ldots, n\} \to \{1, \ldots, n\}$ be chosen uniformly at random from all n^n such possible functions.
- Let $L(g)$ denote the number of values $y \in \{1, \ldots, n\}$ for which the equation $g(x) = y$ has no solution (i.e., $g(x) \neq y$ for every $x \in \{1, \ldots, n\}$).
- By the linearity of the expectation, we have

$$\mathbb{E}[L(g)] = n \left(1 - \frac{1}{n}\right)^n.$$

Consequently, for every $n \in \mathbb{N}$,

$$\frac{n-1}{e} < \mathbb{E}[L(g)] < \frac{n}{e}. \tag{15}$$

- We wish to derive a concentration inequality for $L(g)$ around its expected value.

Example (Cont.): Invoking the Azuma-Hoeffding Inequality

- Let us construct a martingale sequence $\{X_k, \mathscr{F}_k\}_{k=0}^n$ by

$$X_k = \mathbb{E}[L(g) \,|\, \mathscr{F}_k], \quad \forall\, k \in \{0, \ldots, n\}$$

  with the natural filtration

$$\mathscr{F}_k = \sigma\bigl(g(1), \ldots, g(k)\bigr), \quad \forall\, k \in \{1, \ldots, n\}$$

  which denotes the σ-algebra that is generated by the first k values of the random function g.
- $\mathscr{F}_0 = \{\emptyset, \{1, \ldots, n\}\}$ is the minimal σ-algebra that only includes the empty set and the probability space.
- By construction, $X_0 = \mathbb{E}[L(g)]$ and $X_n = L(g)$.
- Since a modification of one value of g cannot change $L(g)$ by more than 1, it follows that $|X_k - X_{k-1}| \leq 1$ for every $k \in \{1, \ldots, n\}$.

Example (Cont.): Invoking the Azuma-Hoeffding Inequality

From the Azuma-Hoeffding inequality and the bounds on $\mathbb{E}[L(g)]$

$$\mathbb{P}\left(\left|L(g) - \frac{n}{e}\right| > \alpha\sqrt{n} + 1\right) \leq 2\exp\left(-\frac{\alpha^2}{2}\right), \quad \forall\, \alpha > 0. \tag{16}$$

Example (Cont.): Invoking McDiarmid's Inequality

This concentration result can be improved as follows:

- Let $f \colon \{1, \ldots, n\}^n \to \{1, \ldots, n\}$ be defined by $L(g) \triangleq f\bigl(g(1), \ldots, g(n)\bigr)$ so, the function f maps the n-length vector $(g(1), \ldots, g(n))$ to the number of elements $y \in \{1, \ldots, n\}$ where $g(x) \neq y$ for every $x \in \{1, \ldots, n\}$.
- Since by assumption $g(1), \ldots, g(n)$ are independent random variables, the variation of f with respect to each of its arguments (while all the other $n-1$ arguments of f are kept fixed) is no more than 1.
- Consequently, from McDiarmid's inequality,

$$\mathbb{P}\left(\left|L(g) - \frac{n}{e}\right| > \alpha\sqrt{n} + 1\right) \leq 2\exp\bigl(-2\alpha^2\bigr), \quad \forall\, \alpha > 0, \tag{17}$$

which implies that the exponent of the concentration inequality is improved by a factor of 4.

Concentration Phenomena for Codes Defined on Graphs

Motivation & Background

- The performance analysis of a particular code is difficult, especially for codes of large block lengths.

Concentration Phenomena for Codes Defined on Graphs

Motivation & Background

- The performance analysis of a particular code is difficult, especially for codes of large block lengths.
- Does the performance concentrate around the average performance of the ensemble ?

Concentration Phenomena for Codes Defined on Graphs

Motivation & Background

- The performance analysis of a particular code is difficult, especially for codes of large block lengths.
- Does the performance concentrate around the average performance of the ensemble ?
- The existence of such a concentration validates the use of the density evolution technique as an analytical tool to assess performance of long enough codes (e.g., LDPC codes) and to assess their asymptotic gap to capacity.

Concentration Phenomena for Codes Defined on Graphs

Motivation & Background

- The performance analysis of a particular code is difficult, especially for codes of large block lengths.
- Does the performance concentrate around the average performance of the ensemble ?
- The existence of such a concentration validates the use of the density evolution technique as an analytical tool to assess performance of long enough codes (e.g., LDPC codes) and to assess their asymptotic gap to capacity.
- The current concentration results for codes defined on graphs, which mainly rely on the Azuma-Hoeffding inequality, are weak since in practice concentration is observed at much shorter block lengths.

Bit-Flipping Decoding Algorithms for Expander Codes

Expansion Properties of Random Regular Bipartite Graphs

Theorem (Sipser and Spielman, 1996)

Let $\mathcal{G}$ be a bipartite graph that is chosen uniformly at random from the ensemble of bipartite graphs with n vertices on the left, a left degree l, and a right degree r. Let $\alpha \in (0,1)$ and $\delta > 0$ be fixed numbers. Then, with probability at least $1 - \exp(-\delta n)$, all sets of αn vertices on the left side of $\mathcal{G}$ are connected to at least

$$n \left[\frac{l\big(1-(1-\alpha)^r\big)}{r} - \sqrt{2l\alpha\left(h(\alpha)+\delta\right)} \right] \qquad (18)$$

vertices (neighbors) on the right side of $\mathcal{G}$, where h is the binary entropy function to base e (i.e., $h(x) = -x\ln(x) - (1-x)\ln(1-x)$, $x \in [0,1]$).

The proof starts by looking at the expected number of neighbors, and then exposing one neighbor at a time to bound the probability that the number of neighbors deviates significantly from this mean.

Proof

- Let $\mathcal{V}$ denote a given set of $n\alpha$ vertices on the left side of the selected bipartite graph $\mathcal{G}$. The set $\mathcal{V}$ has $n\alpha l$ outgoing edges in $\mathcal{G}$.
- Let $X(\mathcal{G})$ be a RV which denotes the number of neighbors of $\mathcal{V}$ on the right side of $\mathcal{G}$.
- Let $\mathbb{E}[X(\mathcal{G})]$ be the expected value of neighbors of $\mathcal{V}$ where all the bipartite graphs are chosen uniformly at random from the ensemble.
- This expected number is equal to $\mathbb{E}[X(\mathcal{G})] = \frac{nl\left(1-(1-\alpha)^r\right)}{r}$ since, for each of the $\frac{nl}{r}$ vertices on the right side of $\mathcal{G}$, the probability that it has at least one edge in the subset of $n\alpha$ chosen vertices on the left side of $\mathcal{G}$ is $1-(1-\alpha)^r$.

Proof (Cont.)

- Let us form a martingale sequence to estimate, via the Azuma–Hoeffding inequality, the probability that the actual number of neighbors deviates by a certain amount from its expected value.
- The set of $n\alpha$ vertices in $\mathcal{V}$ has $n\alpha l$ outgoing edges.
- Let us reveal the destination of each of these edges one at a time. More precisely, let S_i be the random variable denoting the vertex on the right side of $\mathcal{G}$ which the i-th edge is connected to, where $i \in \{1, \ldots, n\alpha l\}$.
- Let us define, for $i \in \{0, \ldots, n\alpha l\}$,

$$X_i = \mathbb{E}[X(\mathcal{G})|S_1, \ldots, S_{i-1}].$$

Note that this forms a martingale sequence where $X_0 = \mathbb{E}[X(\mathcal{G})]$ and $X_{n\alpha l} = X(\mathcal{G})$.

Proof (Cont.)

- For every $i \in \{1, \ldots, n\alpha l\}$, we have $|X_i - X_{i-1}| \leq 1$ since every time only one connected vertex on the right side of $\mathcal{G}$ is revealed, so the number of neighbors of the chosen set $\mathcal{V}$ cannot change by more than 1 at every single time.
- From the one-sided Azuma–Hoeffding inequality, it follows that

$$\mathbb{P}\Big(\mathbb{E}[X(\mathcal{G})] - X(\mathcal{G}) \geq \lambda\sqrt{l\alpha n}\Big) \leq \exp\left(-\frac{\lambda^2}{2}\right), \quad \forall\, \lambda > 0. \qquad (19)$$

- Since there are $\binom{n}{n\alpha}$ choices for the set $\mathcal{V}$, the event that there exists a set of size $n\alpha$ with fewer than $\mathbb{E}[X(\mathcal{G})] - \lambda\sqrt{l\alpha n}$ neighbors occurs with probability at most $\binom{n}{n\alpha} \exp\big(-\frac{\lambda^2}{2}\big)$, by the union bound.
- Since $\binom{n}{n\alpha} \leq e^{nh(\alpha)}$, we get the upper bound $\exp\left(nh(\alpha) - \frac{\lambda^2}{2}\right)$.
- Choosing $\lambda = \sqrt{2n\big(h(\alpha) + \delta\big)}$ in (19) gives the bound in (18).

Performance under Message-Passing Decoding

Theorem I - **[Concentration of performance under iterative message-passing decoding (Richardson and Urbanke, 2001)]**

Let $\mathcal{C}$, a code chosen uniformly at random from the ensemble LDPC(n, λ, ρ), be used for transmission over a memoryless binary-input output-symmetric (MBIOS) channel. Assume that the decoder performs l iterations of message-passing decoding, and let $P_{\mathsf{b}}(\mathcal{C}, l)$ denote the resulting bit error probability. Then, for every $\delta > 0$, there exists an $\alpha > 0$ where $\alpha = \alpha(\lambda, \rho, \delta, l)$ (*independent of the block length* n) such that

$$\mathbb{P}\left(|P_{\mathsf{b}}(\mathcal{C}, l) - \mathbb{E}_{\mathsf{LDPC}(n,\lambda,\rho)}[P_{\mathsf{b}}(\mathcal{C}, l)]| \geq \delta\right) \leq e^{-\alpha n}$$

Proof

The proof applies Azuma's inequality to a martingale sequence with bounded differences (IEEE Trans. on IT, Feb. 2001).

Conditional Entropy of LDPC code ensembles

Theorem II - **[Concentration of Conditional Entropy of LDPC code ensembles (Méasson et al. 2008)]**

Let $\mathcal{C}$ be chosen uniformly at random from the ensemble $\text{LDPC}(n, \lambda, \rho)$. Assume that the transmission of the code $\mathcal{C}$ takes place over an MBIOS channel. Let $H(\mathbf{X}|\mathbf{Y})$ designate the conditional entropy of the transmitted codeword $\mathbf{X}$ given the received sequence $\mathbf{Y}$ from the channel. Then, for any $\xi > 0$,

$$\mathbb{P}\left(|H(\mathbf{X}|\mathbf{Y}) - \mathbb{E}_{\text{LDPC}(n,\lambda,\rho)}[H(\mathbf{X}|\mathbf{Y})]| \geq \sqrt{n}\,\xi\right) \leq 2\exp(-B\xi^2)$$

where $B \triangleq \frac{1}{2(d_{\mathsf{c}}^{\max}+1)^2(1-R_{\mathsf{d}})}$, $d_{\mathsf{c}}^{\max}$ is the maximal check-node degree, and R_{d} is the design rate of the ensemble.

Concentration of Martingales in Coding Theory

Selected Papers Applying the Martingale Approach for LDPC Code Ensembles

- M. Sipser and D. A. Spielman, ”Expander codes,” *IEEE Trans. on Information Theory*, vol. 42, no. 6, pp. 1710-1722, November 1996.
- M.G. Luby, M. Mitzenmacher, M.A. Shokrollahi and D.A. Spielman, ”Efficient erasure correcting codes”, *IEEE Trans. on Information Theory*, vol. 47, no. 2, pp. 569-584, February 2001.
- T. Richardson and R. Urbanke, ”The capacity of low-density parity check codes under message-passing decoding.” *IEEE Trans. on Information Theory*, vol. 47, pp. 599–618, February 2001.
- A. Kavcic, X. Ma and M. Mitzenmacher, ”Binary intersymbol interference channels: Gallager bounds, density evolution, and code performance bounds,” *IEEE Trans. on Information Theory*, vol. 49, no. 7, pp. 1636-1652, July 2003.

(Cont.)

- A. Montanari, "Tight bounds for LDPC and LDGM codes under MAP decoding," *IEEE Trans. on Information Theory*, vol. 51, no. 9, pp. 3247–3261, September 2005.
- C. Méasson, A. Montanari and R. Urbanke, "Maxwell construction: The hidden bridge between iterative and maximum a-posteriori decoding," *IEEE Trans. on Information Theory*, vol. 54, pp. 5277–5307, December 2008.
- L.R. Varshney, "Performance of LDPC codes under faulty iterative decoding," *IEEE Trans. on Information Theory*, vol. 57, no. 7, pp. 4427–4444, July 2011.
- R. Eshel and I. Sason, "On concentration of measures for LDPC code ensembles," *Proc. of ISIT 2011*, pp. 1273–1277, August 2011.
- M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Information Theory, Communications and Coding*, *FnT in Comm. and IT, 2nd Edition*, pp. 1–250, NOW Publishers, September 2014.

Talagrand's Inequality for Product Spaces

We consider in the following a powerful concentration technique for product spaces, which was introduced by Michel Talagrand in his Landmark paper:

M. Talagrand, "Concentration of measure and isoperimetric inequalities in product spaces," *Publications Mathématique de l'Institute des Hautes Études Scientifiques*, vol. 81, no. 1, pp. 73–205, 1995.

Distances

Let $\Omega_1, \ldots, \Omega_n$ be n sample spaces, and let $\overline{x}, \overline{y} \in \prod_{k=1}^n \Omega_k$.

The Hamming distance between $\overline{x}$ and $\overline{y}$ is the sum of the number of coordinates where $\overline{x}$ and $\overline{y}$ are different, i.e., $d_{\mathsf{H}}(\overline{x}, \overline{y}) = \sum_{k=1}^n 1\{x_k \neq y_k\}$ where $1\{\cdot\}$ denotes the indicator function.

Generalization & normalization: Let $\overline{a} \in \mathbb{R}_+^n$ with $\|\overline{a}\|_2 = \frac{1}{n}$, i.e.,

$$\overline{a} = (a_1, \ldots, a_n), \quad \text{s.t. } a_k \geq 0, \; \forall\, k \in \{1, \ldots, n\}, \; \sum_{k=1} a_k^2 = 1.$$

Then, define

$$d_{\overline{a}}(\overline{x}, \overline{y}) = \sum_{k=1}^n a_k \, 1\{x_k \neq y_k\} = \sum_{k \colon x_k \neq y_k} a_k.$$

Special case: Let $\overline{a} = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$, then $d_{\mathsf{H}}(\overline{x}, \overline{y}) = \sqrt{n}\, d_{\overline{a}}(\overline{x}, \overline{y})$.

Distances (Cont.)

Let $\mathcal{A} \subseteq \prod_{k=1}^{n} \Omega_k$. Define

$$d_{\overline{a}}(\overline{x}, \mathcal{A}) = \min_{\overline{y} \in \mathcal{A}} d_{\overline{a}}(\overline{x}, \overline{y})$$

and define

$$d(\overline{x}, \mathcal{A}) = \max \left\{ d_{\overline{a}}(\overline{x}, \mathcal{A}) \colon \overline{a} \in \mathbb{R}_{+}^{n}, \ \|\overline{a}\|_2 = 1 \right\}$$

where the maximum always exists.

Distances (Cont.)

Let $\mathcal{A} \subseteq \prod_{k=1}^{n} \Omega_k$ and $\overline{x} \in \prod_{k=1}^{n} \Omega_k$.

Define

$$U_{\mathcal{A}}(\overline{x}) = \Big\{ \overline{s} \in \{0,1\}^n : \exists\, y \in \mathcal{A} \text{ s.t. } s_k = 0 \Rightarrow x_k = y_k \Big\}$$

$$= \Big\{ \overline{s} \in \{0,1\}^n : \exists\, y \in \mathcal{A} \text{ s.t. } s_k \geq 1\{x_k \neq y_k\}, \ \forall\, k \in \{1, \dots, n\} \Big\}.$$

Note that $U_{\mathcal{A}}(\overline{x})$ is a finite set whose cardinality is at most 2^n.

Fact 1

For all $\overline{a} \in \mathbb{R}_+^n$ with $\|\overline{a}\|_2 = 1$

$$d_{\overline{a}}(\overline{x}, \mathcal{A}) = \min_{\overline{t} \in U_{\mathcal{A}}(\overline{x})} (\overline{t}, \overline{a})$$

where $(\overline{t}, \overline{a}) = \sum_{k=1}^{n} t_k a_k$ is the scalar product.

Distances (Cont.)

Define

$$V_{\mathcal{A}}(\overline{x}) = \text{Convex hull } U_{\mathcal{A}}(\overline{x}).$$

Fact 2

$$d(\overline{x}, \mathcal{A}) = d_{\mathsf{E}}(\overline{0}, V_{\mathcal{A}}(\overline{x}))$$

where $d_{\mathsf{E}}(\overline{0}, V_{\mathcal{A}}(\overline{x}))$ denotes the Euclidean distance of the convex set $V_{\mathcal{A}}(\overline{x})$ to the origin.

Example

Let $\overline{x} = (a, b)$, $\mathcal{A} = \Big\{(a, a),\ (b, b)\Big\}$ with $a \neq b$. Then, by definition, $U_{\mathcal{A}}(\overline{x}) = \Big\{(1, 0),\ (0, 1),\ (1, 1)\Big\}$, and $V_{\mathcal{A}}(\overline{x})$ is the triangle in the plane whose vertices are the points $(1, 0)$, $(0, 1)$, $(1, 1)$.
The Euclidean distance of $V_{\mathcal{A}}(\overline{x})$ to the origin is $\frac{1}{\sqrt{2}}$, so $d(\overline{x}, \mathcal{A}) = \frac{1}{\sqrt{2}}$.

Talagrand's Theorem

Let $\mathcal{A}$ be a measurable, non-empty subset of $\Omega = \prod_{k=1}^{n} \Omega_k$. Then, for every product probability measure P on Ω

$$\mathbb{E}\left[\exp\left(\frac{d(\overline{X}, \mathcal{A})^2}{4}\right)\right] \leq \frac{1}{P(\mathcal{A})} \tag{20}$$

where $\overline{X} \sim P$. Furthermore,

$$\mathbb{P}\big(d(\overline{X}, \mathcal{A}) \geq r\big) \leq \frac{1}{P(\mathcal{A})} \cdot \exp\left(-\frac{r^2}{4}\right), \quad \forall\, r \geq 0. \tag{21}$$

Proof of Talagrand's Theorem

The original proof of (20) is by induction on the number of coordinates together with geometric arguments involving projections and sections to lower dimensions.

Inequality (21) follows easily from (20) and Markov's inequality:

$$\begin{aligned}
&\mathbb{P}\big(d(\overline{X}, \mathcal{A}) \geq r\big) \\
&= \mathbb{P}\left(\exp\left(\frac{d(\overline{X}, \mathcal{A})^2}{4}\right) \geq \exp\left(\frac{r^2}{4}\right)\right) \\
&\leq \exp\left(-\frac{r^2}{4}\right)\ \mathbb{E}\left[\exp\left(\frac{d(\overline{X}, \mathcal{A})^2}{4}\right)\right] \\
&\leq \frac{1}{P(\mathcal{A})} \cdot \exp\left(-\frac{r^2}{4}\right).
\end{aligned}$$

Talagrand's Inequality for Lipschitz Functions

- Consider a function $F\colon \Omega \to \mathbb{R}$ on the product space $\Omega = \prod_{k=1}^{n} \Omega_k$.
- Assume that, for every $\overline{x} \in \Omega$, there exists $\overline{a} = \overline{a}(\overline{x}) \in \mathbb{R}_+^n$ with $\|\overline{a}\|_2 = 1$ such that

  $$F(\overline{x}) \leq F(\overline{y}) + \sigma\, d_{\overline{a}}(\overline{x}, \overline{y}), \quad \forall\, \overline{y} \in \Omega$$

  for some $\sigma > 0$.
- Due to the symmetry property where $d_{\overline{a}}(\overline{x}, \overline{y}) = d_{\overline{a}}(\overline{y}, \overline{x})$, we have

  $$\big|F(\overline{x}) - F(\overline{y})\big| \leq \sigma\, d_{\overline{a}}(\overline{x}, \overline{y}), \quad \forall\, \overline{x}, \overline{y} \in \Omega$$

  where $\overline{a}$ is as above.

This function is called a Lipschitz function.

Talagrand's Inequality for Lipschitz Functions (Cont.)

- Let $\mathcal{A} \subseteq \Omega$ be defined as

$$\mathcal{A} = \left\{ \overline{y} \in \Omega \colon \; F(\overline{y}) \leq m \right\}$$

where $m \in \mathbb{R}$ is arbitrary. For brevity, we will write $\mathcal{A} = \{F \leq m\}$.

- From the Lipschitz property, for every $\overline{x} \in \Omega$ there exists $\overline{a} \in \mathbb{R}_+^n$ with $\|\overline{a}\|_2 = 1$ such that

$$\begin{aligned} & F(\overline{x}) \leq m + \sigma \, d_{\overline{a}}(\overline{x}, \overline{y}), \quad \forall \, y \in \mathcal{A}. \\ \Rightarrow \; & F(\overline{x}) \leq m + \sigma \, \min_{\overline{y} \in \mathcal{A}} d_{\overline{a}}(\overline{x}, \overline{y}) \\ & \qquad = m + \sigma \, d_{\overline{a}}(\overline{x}, \mathcal{A}). \end{aligned}$$

Recall that, in general, $\overline{a}$ <u>depends</u> on $\overline{x}$.

Talagrand's Inequality for Lipschitz Functions (Cont.)

- By taking a supremum over $\overline{a}$, we get
$$\begin{aligned} F(\overline{x}) &\leq m + \sigma \sup\Big\{ d_{\overline{a}}(\overline{x}, \mathcal{A}) \colon \overline{a} \in \mathbb{R}_+^n, \, \|\overline{a}\|_2 = 1 \Big\} \\ &= m + \sigma \, d(\overline{x}, \{F \leq m\}), \quad \forall \, m \in \mathbb{R}, \; \overline{x} \in \Omega. \end{aligned}$$
- Let $r \geq 0$ be arbitrary. If $F(\overline{x}) \geq m + r$, then $d\big(\overline{x}, \, \{F \leq m\}\big) \geq \frac{r}{\sigma}$, so
$$\mathbb{P}(F(\overline{X}) \geq m + r) \leq \mathbb{P}\Big(d(\overline{X}, \{F \leq m\}) \geq \frac{r}{\sigma}\Big), \quad \forall \, r \geq 0, \, m \in \mathbb{R}.$$

Combining it with Talagrand's inequality, with $\mathcal{A} = \{F \leq m\}$, gives:

Theorem: Talagrand's Inequality for Lipschitz Functions

For a Lipschitz Function $F \colon \Omega \to \mathbb{R}$, and $\overline{X} \sim P$ for a product measure P,
$$\mathbb{P}\big(F(\overline{X}) \leq m\big) \, \mathbb{P}\big(F(\overline{X}) \geq m + r\big) \leq \exp\left(-\frac{r^2}{4\sigma^2}\right), \quad \forall \, r \geq 0, \; m \in \mathbb{R}.$$

From Talagrand's Inequality to Concentration Around the Median

- Define the median of F by

$$m_F \triangleq \inf\Big\{m\colon\ \mathbb{P}\big(F(\overline{X}) \leq m\big) \geq \frac{1}{2}\Big\}, \quad \overline{X} \sim P.$$

- Due to the σ-additivity of a probability measure

$$\mathbb{P}\big(F(\overline{X}) \leq m_F\big) \geq \frac{1}{2}, \quad \mathbb{P}\big(F(\overline{X}) \geq m_F\big) \geq \frac{1}{2}.$$

- Picking $m = m_F$ and $m = m_F - r$, for $r \geq 0$, give for a Lipschitz function with constant $\sigma > 0$

$$\mathbb{P}\big(F(\overline{X}) \geq m_F + r\big) \leq 2\, e^{-\frac{r^2}{4\sigma^2}}$$

$$\mathbb{P}\big(F(\overline{X}) \leq m_F - r\big) \leq 2\, e^{-\frac{r^2}{4\sigma^2}}$$

$$\Rightarrow \mathbb{P}\big(|F(\overline{X}) - m_F| \geq r\big) \leq 4\, e^{-\frac{r^2}{4\sigma^2}}, \quad \forall\, r \geq 0.$$

Bound on the Distance Between the Median and the Expected Value

For a Lipschitz Function with constant $\sigma > 0$, and $\overline{X} \sim P$ where P is a product measure

$$\begin{aligned}
&\left|m_F - \mathbb{E}[F(\overline{X})]\right| \\
&\leq \mathbb{E}\big[\left|m_F - F(\overline{X})\right|\big] \\
&= \int_0^\infty \mathbb{P}\big(\left|F(\overline{X}) - m_F\right| \geq r\big)\, \mathsf{d}r \\
&\leq \int_0^\infty 4\, e^{-\frac{r^2}{4\sigma^2}}\, \mathsf{d}r \\
&= 4\sqrt{\pi}\,\sigma.
\end{aligned}$$

Talagrand's Inequality to Concentration Around the Expected Value

$$\mathbb{P}\big(|F(\overline{X}) - \mathbb{E}[F(\overline{X})]| \geq r\big) \leq 4\, e^{-\frac{(r-4\sqrt{\pi}\,\sigma)^2}{4\sigma^2}}, \quad \forall\, r \geq 4\sqrt{\pi}\,\sigma.$$

Example: Longest Increasing Subsequence

- Let $X_1, \ldots, X_n$ be i.i.d. and uniformly distributed on $[0, 1]$.
- Let $L_n = L(X_1, \ldots, X_n)$ be the longest increasing subsequence of $X_1, \ldots, X_n$ (i.e., it is the largest integer p where there exist indices $1 \leq i_1 < \ldots < i_p \leq n$ such that $X_{i_1} < \ldots < X_{i_p}$).
- It is possible to show that for some constants c_1, c_2

$$c_1\sqrt{n} < L_n < c_2\sqrt{n} \quad \text{almost surely.}$$

In the following, we consider the concentration of L_n.

A Concentration Inequality by Invoking McDiarmid's inequality

Since a change of one value in the sequence $x_1, \ldots, x_n$ may modify L_n by at most 1, McDiarmid's inequality yields that

$$\mathbb{P}\big(|L_n - \mathbb{E}[L_n]| \geq \alpha\sqrt{n}\big) \leq 2\exp(-2\alpha^2), \quad \forall\, \alpha > 0.$$

This is not particularly useful since $\mathbb{E}[L_n]$ is itself of the order of $\sqrt{n}$.

A Concentration Inequality by Invoking Talagrand's inequality

The function $L(\overline{x}) = L(x_1, \ldots, x_n)$ is not Lipschitz, though it possesses the following property:

Lemma

Let $\overline{x}, \overline{y} \in [0,1]^n$. If for some $m, r \geq 0$

$$L(\overline{x}) \geq m + r, \quad L(\overline{y}) \leq m$$

then there exists $\overline{a} = \overline{a}(\overline{x}) = (a_1, \ldots, a_n)$ with $\overline{a} \in \mathbb{R}_+^n$ and $\|\overline{a}\|_2 = 1$ s.t.

$$d_{\overline{a}}(\overline{x}, \overline{y}) \geq \frac{r}{\sqrt{m+r}}.$$

$$\Rightarrow \mathbb{P}(L(\overline{X}) \geq m + r) \leq \mathbb{P}\left(d(\overline{X}, \{L \leq m\}) \geq \frac{r}{\sqrt{m+r}}\right).$$

A Concentration Inequality by Invoking Talagrand's inequality (Cont.)

From Talagrand's theorem and the last inequality, it follows that

$$\begin{aligned}
&\mathbb{P}\big(L(\overline{X}) \leq m\big)\,\mathbb{P}\big(L(\overline{X}) \geq m+r\big) \\
&\leq \mathbb{P}\big(L(\overline{X}) \leq m\big)\,\mathbb{P}\left(d(\overline{X}, \{L \leq m\}) \geq \frac{r}{\sqrt{m+r}}\right) \\
&\leq \exp\left(-\frac{r^2}{4(m+r)}\right), \quad \forall\, m, r \geq 0. \qquad (22)
\end{aligned}$$

Note that if $m < 0$, the bound is trivial (since $\mathbb{P}\big(L(\overline{X}) \leq m\big) = 0$).

A Concentration Inequality by Invoking Talagrand's inequality (Cont.)

From Talagrand's theorem and the last inequality, it follows that

$$\begin{aligned}
&\mathbb{P}\big(L(\overline{X}) \leq m\big)\, \mathbb{P}\big(L(\overline{X}) \geq m + r\big) \\
&\leq \mathbb{P}\big(L(\overline{X}) \leq m\big)\, \mathbb{P}\left(d(\overline{X}, \{L \leq m\}) \geq \frac{r}{\sqrt{m+r}}\right) \\
&\leq \exp\left(-\frac{r^2}{4(m+r)}\right), \quad \forall\, m, r \geq 0. \qquad (22)
\end{aligned}$$

Note that if $m < 0$, the bound is trivial (since $\mathbb{P}\big(L(\overline{X}) \leq m\big) = 0$).
Let m_n be the median of L_n, then $\mathbb{P}(L_n \geq m_n) \geq \frac{1}{2}$, $\mathbb{P}(L_n \leq m_n) \geq \frac{1}{2}$.

A Concentration Inequality by Invoking Talagrand's inequality (Cont.)

From Talagrand's theorem and the last inequality, it follows that

$$\begin{aligned}
&\mathbb{P}\big(L(\overline{X}) \leq m\big)\,\mathbb{P}\big(L(\overline{X}) \geq m + r\big) \\
&\leq \mathbb{P}\big(L(\overline{X}) \leq m\big)\,\mathbb{P}\left(d(\overline{X}, \{L \leq m\}) \geq \frac{r}{\sqrt{m+r}}\right) \\
&\leq \exp\left(-\frac{r^2}{4(m+r)}\right), \quad \forall\, m, r \geq 0. \qquad (22)
\end{aligned}$$

Note that if $m < 0$, the bound is trivial (since $\mathbb{P}\big(L(\overline{X}) \leq m\big) = 0$).

Let m_n be the median of L_n, then $\mathbb{P}(L_n \geq m_n) \geq \frac{1}{2}$, $\mathbb{P}(L_n \leq m_n) \geq \frac{1}{2}$. Substituting $m = m_n$ and $m = m_n - r$ $(r \geq 0)$ in (22) gives, respectively,

$$\left.\begin{aligned}
&\mathbb{P}\big(L_n \geq m_n + r\big) \leq 2\,e^{-\frac{r^2}{4(m_n+r)}}, \\
&\mathbb{P}\big(L_n \leq m_n\big) \leq 2\,e^{-\frac{r^2}{4m_n}}.
\end{aligned}\right. \quad \forall\, r > 0.$$

A Concentration Inequality by Invoking Talagrand's inequality (Cont.)

$$\Rightarrow \; \mathbb{P}\big(|L_n - m_n| \geq r\big) \leq 4\, e^{-\frac{r^2}{4(m_n+r)}}, \quad \forall\, r > 0 \tag{23}$$

or, by letting $r = \alpha\sqrt{m_n}$ for $\alpha > 0$, we get

$$\begin{aligned} \mathbb{P}\big(|L_n - m_n| \geq \alpha\sqrt{m_n}\big) &\leq 4\, e^{-\frac{\alpha^2\, m_n}{4(m_n+\alpha\sqrt{m_n})}} \\ &\leq 4\, e^{-\frac{\alpha^2}{4(1+\alpha)}}, \quad \forall\, \alpha > 0 \end{aligned} \tag{24}$$

where the last inequality holds since $L_n \geq 1 \Rightarrow m_n \geq 1$.

A Concentration Inequality by Invoking Talagrand's inequality (Cont.)

A concentration result for L_n around $\mathbb{E}[L_n]$ can be further obtained from (23) and a derivation of an upper bound on $|\mathbb{E}[L_n] - m_n|$.

A Concentration Inequality by Invoking Talagrand's inequality (Cont.)

A concentration result for L_n around $\mathbb{E}[L_n]$ can be further obtained from (23) and a derivation of an upper bound on $|\mathbb{E}[L_n] - m_n|$. From (23),

$$\begin{aligned} |\mathbb{E}[L_n] - m_n| &\leq \mathbb{E}\big[|L_n - m_n|\big] \\ &= \int_0^\infty \mathbb{P}\big(|L_n - m_n| \geq r\big)\, \mathrm{d}r \\ &\leq \int_0^\infty 4\, e^{-\frac{r^2}{4(m_n+r)}}\, \mathrm{d}r \\ &\leq 4 \int_0^{m_n} e^{-\frac{r^2}{8m_n}}\, \mathrm{d}r + 4 \int_0^{m_n} e^{-\frac{r}{8}}\, \mathrm{d}r \\ &= 4\sqrt{2\pi m_n} + 32. \end{aligned}$$

Conclusion

- The expected value and median scale like $\sqrt{n}$ almost surely.
- Talagrand's inequality gives a concentration result for deviations that scale like $\sqrt[4]{n}$.
- McDiarmid's inequality gives a concentration result for deviations that scale like $\sqrt{n}$.

Hence, Talagrand's inequality provides a stronger concentration result than McDiarmid's inequality.

Further Analysis

A much stronger result determining the precise asymptotic distribution of L_n has been obtained by Baik, Deift and Johansson (1999) using deep analytical tools.

Orthogonal Frequency Division Multiplexing (OFDM)

- The OFDM modulation converts a high-rate data stream into a number of low-rate steams that are transmitted over parallel narrow-band channels.
- One of the problems of OFDM is that the peak amplitude of the signal can be significantly higher than the average amplitude.

  ⇒ Sensitivity to non-linear devices in the communication path (e.g., digital-to-analog converters, mixers and high-power amplifiers).

  ⇒ An increase in the symbol error rate and also a reduction in the power efficiency as compared to single-carrier systems.

OFDM (Cont.)

- Given an n-length codeword $\{X_i\}_{i=0}^{n-1}$, a single OFDM baseband symbol is described by

$$s(t; X_0, \ldots, X_{n-1}) = \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} X_i \exp\left(\frac{j\, 2\pi i t}{T}\right), \quad 0 \leq t \leq T.$$

- Assume that $X_0, \ldots, X_{n-1}$ are i.i.d. complex RVs with $|X_i| = 1$. Since the sub-carriers are orthonormal over $[0, T]$, then a.s. the power of the signal s over this interval is 1.
- The CF of the signal s, composed of n sub-carriers, is defined as

$$\mathrm{CF}_n(s) \triangleq \max_{0 \leq t \leq T} |s(t)|.$$

- The CF scales with high probability like $\sqrt{\log(n)}$ for large n.

Assumption

Consider the case where $\{X_j\}_{j=0}^{n-1}$ are independent complex-valued random variables with magnitude 1, attaining the M points of an M-ary PSK constellation with equal probability.

A concentration inequality (Litsyn and Wunder, 2006)

For every $c \geq 2.5$

$$\mathbb{P}\left(\left|\mathsf{CF}_n(s) - \sqrt{\log(n)}\right| < \frac{c \log\log(n)}{\sqrt{\log(n)}}\right) = 1 - O\left(\frac{1}{\bigl(\log(n)\bigr)^4}\right).$$

Concentration Inequalities

In the following, we invoke McDiarmid and Talagrand's inequalities for obtaining concentration results for the crest factor of OFDM signals.

Application of McDiarmid's inequality

McDiarmid's inequality implies that

$$\mathbb{P}(|\mathsf{CF}_n(s) - \mathbb{E}[\mathsf{CF}_n(s)]| \geq \alpha) \leq 2\exp\left(-\frac{\alpha^2}{2}\right), \quad \forall\, \alpha \geq 0.$$

The exponent improves by factor 4 in comparison to Azuma's inequality.

Note

The same kind of concentration result can be applied to QAM-modulated OFDM signals, since the independent RVs $\{X_j\}$ are bounded.

Establishing Concentration via Talagrand's Inequality

Let $x_0, y_0, \ldots, x_{n-1}, y_{n-1}$ have all absolute value of 1.

$$\begin{aligned}
&\max_{0 \leq t \leq T} \big| s(t; x_0, \ldots, x_{n-1}) \big| - \max_{0 \leq t \leq T} \big| s(t; y_0, \ldots, y_{n-1}) \big| \\
&\leq \max_{0 \leq t \leq T} \big| s(t; x_0, \ldots, x_{n-1}) - s(t; y_0, \ldots, y_{n-1}) \big| \\
&\leq \frac{1}{\sqrt{n}} \, \left| \sum_{i=0}^{n-1} (x_i - y_i) \exp\left(\frac{j\, 2\pi i t}{T}\right) \right| \\
&\leq \frac{1}{\sqrt{n}} \, \sum_{i=0}^{n-1} |x_i - y_i| \\
&\leq \frac{2}{\sqrt{n}} \sum_{i=0}^{n-1} 1\{x_i \neq y_i\} = 2 d_{\overline{a}}(\overline{x}, \overline{y})
\end{aligned}$$

where $\overline{a} \triangleq \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$.

Establishing Concentration via Talagrand's Inequality (Cont.)

- Talagrand's inequality for Lipschitz functions with constant $\sigma = 2$ yields that

$$\mathbb{P}(|\mathsf{CF}_n(s) - m_n| \geq \alpha) \leq 4\, \exp\left(-\frac{\alpha^2}{16}\right), \quad \forall\, \alpha > 0, \quad \forall\, \alpha \geq 0 \quad (25)$$

  where m_n is the median of the crest factor for OFDM signals that are composed of n sub-carriers.

- This inequality demonstrates the concentration of this measure around its median.

Establishing Concentration via Talagrand's Inequality (Cont.)

Corollary

The median and expected value of the crest factor differ by at most a constant, independently of the number of sub-carriers n.

Proof: From Talagrand's inequality in (25)

$$\begin{aligned}
&|\mathbb{E}[\mathsf{CF}_n(s)] - m_n| \\
&\leq \mathbb{E}\,|\mathsf{CF}_n(s) - m_n| \\
&= \int_0^\infty \mathbb{P}(|\mathsf{CF}_n(s) - m_n| \geq \alpha)\, \mathsf{d}\alpha \\
&\leq \int_0^\infty 4\,\exp\left(-\frac{\alpha^2}{16}\right)\, \mathsf{d}\alpha \\
&= 8\sqrt{\pi}.
\end{aligned}$$

Summary

- The area of concentration of measure has seen enormous growth and tremendous activity since the early '90s.
- This part of the tutorial focused on establishing concentration results via
  - The martingale approach
  - Talagrand's convex-hull distance inequality for product measures.
- The introduction of these two approaches was exemplified by some potential applications in information theory, communications and modern coding techniques.

Summary (Cont.)

- The second part of this tutorial continues this path while introducing more approaches that have their roots in information theory: the entropy method, and transportation-cost inequalities.
- Some more material is available in our recent monograph: M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Information Theory, Communications and Coding*, *Foundations and Trends in Communications and Information Theory*, *Second Edition*, pp. 1–250, NOW Publishers, Delft, the Netherlands, September 2014.
- This tutorial also includes material that was not presented there.