Concentration inequalities

Paul Delatte
delatte@usc.edu
University of Southern California

Last updated: 31 December, 2022

Quantifying how much a random variable deviates from some value is a problem of general interest. Often, the value is the mean or the median, in which case the problem is similar to bounding the tails of the random variable. Concentration inequalities are useful in themselves as sharp non-asymptotic results quantifying deviation, but they are also useful as a tool to prove asymptotic results for a larger class of random variables.

Example 1. A motivation for the first point emerges from the classical central limit theorem. Since a sum of independent random variables is asymptotically normal, we may expect it to have exponentially decaying tails in the limit. Unfortunately, even if the sum concentrates well around the mean, the approximation error to the normal decays too slowly for the exponential decay of the tails to hold. Under some regularity conditions, we will show that it is actually possible to derive exponential tail bounds for sums of independent random variables, using more direct methods than the CLT.

The fact that random variable expressible as sums of independent random variables concentrate well around their mean (with exponentially decaying tails under some mild conditions) is a phenomenon with deep and far-reaching repercussions (from geometry to statistics to physics to ...). The extension of this concentration phenomenon (and its quantification through exponential bounds) to random variables expressible as more general functions (non necessarily linear) of independent (or weakly dependent) random variables has been explored in the last decades. A general principle, known as the **concentration of measure phenomenon**, stands out:

> "If $X_1, \dots, X_n$ are independent (or weakly dependent) random variables, then the random variable $f(X_1, \dots, X_n)$ is "close" to its mean $\mathbb{E}(f(X_1, \dots, X_n))$ provided that the function $(x_1, \dots, x_n) \mapsto f(x_1, \dots, x_n)$ is not too "sensitive" to any of the coordinates x_i." (van Handel in APC550)

1 Basic bounds from moments

We start with basic tools. The tail probability $\mathbb{P}(X \geq t)$ of a random variable X can first be bounded by controlling the moments of X. Controlling more moments gives sharper bounds.

Proposition 2. *Let X be a positive random variable. Then*

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X \geq t)\, dt.$$

This generalizes to the following result.

Proposition 3 (**Layer Cake Representation**). *Let X be a real random variable and $p > 0$. Then*

$$\mathbb{E}\left(|X|^p\right) = p \int_0^{+\infty} t^{p-1} \mathbb{P}(|X| \geq t)\, dt.$$

Proof. This is an application of Fubini's theorem noting first that $|X|^p = p \int_0^{+\infty} t^{p-1} \mathbb{1}_{|X| \geq t}\, dt$. See L.4.4 in Kallenberg FMP p.85 or T.1.13. in Lieb A p.26. □

The case $p = 1$ for positive X implies Markov's inequality by noting that $\int_0^{\infty} \mathbb{P}(X \geq t)\, dt \geq \int_0^a \mathbb{P}(X \geq a)\, dt = a\mathbb{P}(X \geq a)$. Alternatively, $x = x\mathbb{1}_{x \geq a} + x\mathbb{1}_{x<a}$ for any $x \in \mathbb{R}$, so that $\mathbb{E}(X) = \mathbb{E}(X\mathbb{1}_{X \geq a}) + \mathbb{E}(X\mathbb{1}_{X<a}) \geq E(a\mathbb{1}_{X \geq a}) = a\mathbb{P}(X \geq a)$.

Proposition 4 (**Markov's Inequality**). *Let X be a positive random variable. Then for all $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Proof. Above for two proofs. See T.2.1 in Roch MDP p.18. □

From Markov's inequality, it follows that if ϕ is a strictly increasing positive-valued function, then for any real random variable X and all $t \in \mathbb{R}$,

$$\mathbb{P}(X \geq t) = \mathbb{P}(\phi(X) \geq \phi(t)) \leq \frac{\mathbb{E}(\phi(X))}{\phi(t)}.$$

Applying this inequality for $x \mapsto x^2$ to $|X - \mathbb{E}\, X|$, we get Chebyshev's inequality.

Proposition 5 (**Chebyshev's Inequality**). *Let X be a random variable with $\mathbb{E}(X^2) < \infty$. Then for all $t > 0$,*

$$\mathbb{P}(|X - \mathbb{E}\, X| \geq t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Proof. Above. See T.2.2 in Roch MDP p.19. □

More generally, applying Markov's inequality for $x \mapsto x^q$ for any $q > 0$ to $|X - \mathbb{E}\, X|$ where $\mathbb{E}(X^q) < \infty$, we get for all $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}\, X| \geq t) \leq \frac{\mathbb{E}\,|X - \mathbb{E}\, X|^q}{t^q}.$$

One can choose q so as to minimize the left-side value of the inequality.

A similar idea underlies the Cramér–Chernoff bounding method which yields very good bounds. Applying Markov's inequality to $x \mapsto e^{\lambda x}$ for any $\lambda \geq 0$ to any random variable X with moment generating function, we get for all $t \in \mathbb{R}$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}\, e^{\lambda X}}{e^{\lambda t}}.$$

One can then choose λ so as to minimize the left-side value fo the inequality.

Proposition 6 (**Chernoff's Bound**). *Let X be a random variable with moment generating function. Then for all $t \in \mathbb{R}$,*

$$\mathbb{P}(X \geq t) \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}\, e^{\lambda X}}{e^{\lambda t}}.$$

Proof. Above. See L.2.32. in Roch MDP p.47. □

2 Bounds for sums of independent random variables

The Cramér–Chernoff bounding method applies well for sums of independent real random variables, for the expected value of a product of independent random variables is the product of the expected values. Take $X_1, \dots, X_n$ independent real random variables and define $S_n = \sum_{i=1}^n X_i$, then we have for all $t \in \mathbb{R}$,

$$\mathbb{P}(S_n - \mathbb{E}\, S_n \geq t) \leq \inf_{\lambda \geq 0} e^{-\lambda t} \prod_{i=1}^n \mathbb{E}\left(e^{\lambda(X_i - \mathbb{E}\, X_i)}\right) \qquad (*)$$

The objective is then to bound the moment generating functions $\mathbb{E}\,(e^{\lambda X_i})$ of the random variables X_i. This works well, in particular, for sub-Gaussian random variables (and thus to bounded random variables) which have by definition bounded moment generating functions. For completeness, we first bound the moment generating function of bounded variables (and incidentally prove they are sub-Gaussian), then get a bound for the sum of independent bounded random variables, and finally a more general bound for sums of independent sub-Gaussian random variables (both known as Hoeffding's inequalities).

Lemma 7 (**Hoeffding's Lemma**). *Let X be a random variable with $\mathbb{E}\, X = 0$ and supported on a finite interval $[a, b]$. Then for all $t \in \mathbb{R}$,*

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

Proof. L.2.42. in Roch MDP p.55 or L.8.1. in Lugosi PTPR p.122. □

Proposition 8 (**Hoeffding's Inequality for Bounded Variables**). *Let $X_1, \dots, X_n$ be independent random variables such that X_i is supported on a finite interval $[a_i, b_i]$. Define $S_n = \sum_{i=1}^n X_i$. Then for all $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}\, S_n \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

and

$$\mathbb{P}(S_n - \mathbb{E}\, S_n \leq -t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Proof. Direct application of Chernoff's bounding method with Hoeffding's lemma (see above). See T.2.40 in Roch MDP p.54 or T.8.1. in Lugosi PTPR p.122. □

Proposition 9 (**Hoeffding's Inequality**). *Let $X_1, \dots, X_n$ be independent sub-Gaussian random variables with proxy variance σ_i^2. Define $S_n = \sum_{i=1}^n X_i$. Then for all $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}\, S_n \geq t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right)$$

and

$$\mathbb{P}(S_n - \mathbb{E}\, S_n \leq -t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right).$$

Proof. Direct application of Chernoff's bounding method with the definition of sub-Gaussianity (see above). See T.2.39. in Roch MDP p.53. □

Hoeffding's inequalities are general but do not necessarily leverage all available information on the variance of the X_i when it exists. "The inequalities can sometimes be improved by introducing second moments in the upper bound. Bennett's one-sided tail bound provides a particularly elegant illustration of the principle." (Pollard).

Without loss of generality, assume $\mathbb{E}(X_i) = 0$. The objective is again to bound $\mathbb{E}(e^{\lambda X_i})$ in $(*)$. Let $\psi(x) := \exp(x) - x - 1$. Note that $\psi(x) \leq x^2/2$ for all $x \leq 0$ and $\psi(\lambda x) \leq x^2\psi(\lambda)$ for all $\lambda \geq 0$ and all $x \in [0, 1]$. If we assume that the X_i's are bounded such that $X_i \leq 1$, one can prove that $\mathbb{E}(e^{\lambda X_i}) \leq \exp(\psi(\lambda)\mathbb{E}(X_i^2))$. Plugging it into $(*)$ and minimizing over λ, we get Bennett's inequality.

Proposition 10 (**Bennett's Inequality**). *Let $X_1, \ldots, X_n$ be independent real random variables with $\mathbb{E}(X_i^2) < \infty$. Denote $S_n = \sum_{i=1}^n X_i$ and $\sigma^2 = \sum_{i=1}^n \mathbb{E}(X_i^2)$. If $X_i \leq b$ a.s. for some positive real b, then for all $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}\,S_n \geq t) \leq \exp\left(-\frac{\sigma^2}{b^2} h\left(\frac{t}{\sigma^2}\right)\right)$$

where $h(u) = (1+u)\ln(1+u) - u$ for all $u \geq 0$.

Proof. Above. See T.17. in Pollard's Mini2 notes or T.2.9. in Massart CI p.35. □

By applying the inequality $h(u) \geq u^2/(2 + 2u/3)$ for all $u \geq 0$ (the right side term is the $(2, 1)$-Padé approximation of $h(u)$ at $u = 0$ – the inequality can be obtained by taking the second derivative of the difference), we obtain a weaker result.

Proposition 11 (**Bernstein's Inequality 1**). *Let $X_1, \ldots, X_n$ be independent real random variables with $\mathbb{E}(X_i^2) < \infty$. Denote $S_n = \sum_{i=1}^n X_i$ and $\sigma^2 = \sum_{i=1}^n \mathbb{E}(X_i^2)$. If $X_i \leq b$ a.s. for some positive real b, then for all $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}\,S_n \geq t) \leq \exp\left(-\frac{t^2}{2(\sigma^2 + bt/3)}\right).$$

Proof. Above. See 22. in Pollard's Mini2 notes or C.2.11. in Massart CI p.38. □

If $\sigma^2 < t$, then the upper bound behaves like e^{-t} instead of the stronger e^{-t^2} guaranteed by Hoeffding's inequality. Intuitively, this is a manifestation of the approximation of binomial $\mathscr{B}(n, \theta/n)$ by Poisson $\mathscr{P}(\theta)$ whose tail decreases as $e^{-\theta}$.

The finiteness of the second moment and the boundedness of X_i imply a condition on the growth of $\mathbb{E}(|X_i|^k)$ for $k \geq 2$, namely $\mathbb{P}(|X_i|^k) \leq \mathbb{E}(X_i^2)b^{k-2}$ for $k \geq 2$. A weaker assumption on this growth, without assuming boundedness of X_i, yields a useful bound on the moment generating function of X_i as well as a more general Bernstein's inequality.

Proposition 12 (**Bernstein's Inequality 2**). *Let $X_1, \ldots, X_n$ be independent real random variables with $\mathbb{E}(X_i^2) < \infty$. Denote $S_n = \sum_{i=1}^n X_i$ and $\sigma^2 = \sum_{i=1}^n \mathbb{E}(X_i^2)$. If there is some positive real B such that $\mathbb{E}(|X_i|^k) \leq \frac{1}{2}\mathbb{E}(X_i^2)B^{k-2}k!$ for all $k \geq 2$, then for all $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}\,S_n \geq t) \leq e^{-H_2(x,B,\sigma^2)} \leq e^{-H_1(x,B,\sigma^2)},$$

where

$$H_1(x, B, \sigma^2) = \frac{t^2/\sigma^2}{2(1 + tB/\sigma^2)},$$
$$H_2(x, B, \sigma^2) = \frac{t^2/\sigma^2}{1 + tB/\sigma^2 + \sqrt{1 + 2tB/\sigma^2}}.$$

Proof. See T.23. in Pollard's Mini2 notes. The moment inequality is used to bound the moment generating function: one can prove that for $\lambda \in [0, 1/B), \mathbb{E}(\psi(\lambda X_i)) \leq \frac{1}{2}\frac{\mathbb{E}(X_i^2)\lambda^2}{1-B}$, so that $\mathbb{E}(e^{\lambda X_i}) \leq 1 + \lambda\mathbb{E}((X_i)) + \frac{1}{2}\frac{\mathbb{E}(X_i^2)\lambda^2}{1-B\lambda} \leq \exp\left(\lambda\mathbb{E}(X_i) + \frac{1}{2}\frac{\mathbb{E}(X_i^2)\lambda^2}{1-B\lambda}\right)$ for $0 \leq \lambda B < 1$. We then proceed as usual by plugging the bound into $(*)$ and minimizing over λ (correctly and incorrectly, see the comment of Pollard on Bennett). □

3 Martingale-based methods

To get bounds for more general functions than sums (of independent random variables), several methods have been developed. One is based on martingale decompositions.

We recall that a **(discrete-time) martingale** is a sequence $((Y_n, \mathcal{F}_n))_{n\in\mathbb{N}^*}$ where the $\mathcal{F}_n$'s are σ-algebras such that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for all $n \geq 0$ and the Y_n are integrable, $\mathcal{F}_n$-measurable random variables such that $\mathbb{E}(Y_{n+1}|\mathcal{F}_n) = Y_n$ a.s.. (We alternatively say that $(Y_n)_{n\in\mathbb{N}^*}$ is a martingale with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}^*}$.) This is equivalent to consider the **martingale difference sequence** $((\Delta_n, \mathcal{F}_n))_{n\in\mathbb{N}}$ defined by the conditions that the $\Delta_{n+1} = Y_{n+1} - Y_n$'s are integrable, $\mathcal{F}_n$-measurable such that $\mathbb{E}(\Delta_{n+1}|\mathcal{F}_n) = 0$ a.s. In this case, $Y_n = Y_0 + \Delta_1 + \Delta_2 + \cdots + \Delta_n = Y_{n-1} + \Delta_n$, or equivalently, $Y_n - Y_0 = \sum_{i=1}^m \Delta_i$. For simplicity, we take $\mathcal{F}_0$ to be the trivial σ-algebra, so that $Y_0 = \mu = \mathbb{E}(Y_n)$ for all $n \geq 0$. Sums of independent, mean zero random variables are examples of martingales: if $X_1, X_2, \ldots$ is a sequence of independent random variables with $\mathbb{E}\,X_i = 0$ for all i, then the partial sums $S_n = \sum_{i=1}^n X_i$ form a martingale with respect to the natural filtration generated by the X_n's; indeed, $\mathbb{E}(S_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n + X_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n|\mathcal{F}_n) + \mathbb{E}(X_{n+1}|\mathcal{F}_n) = S_n + \mathbb{E}(X_{n+1}) = S_n$.

> "The importance of martingales in modern probability stems at least in part from the fact that most of the essential properties of sums of independent [...] random variables are inherited (with minor modification) by martingales." (Lalley)

The concentration inequalities from last section can be proved to hold for martingales. The results can be equivalently stated for $Y_n - Y_0$ or for $\sum_{i=1}^n \Delta_i$. The results can be obtained by "conditional analogs of the moment generating function technique [used previously], with just a few precautions to avoid problems with negligible sets." (Pollard) The precautions consist in resorting to the version of the moment generating function of Δ_i obtained from its regular conditional distribution (whose existence is guaranteed), when applying the Cramér–Chernoff bounding method. (Most authors gloss over these technicalities – see Pollard's Mini3 notes for details.)

Proposition 13 (**Azuma–Hoeffding Inequality**). *Let $(Z_n)_{n\in\mathbb{N}^*}$ be a martingale with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}^*}$. Suppose that there exist predictable processes $(A_n)_{n\in\mathbb{N}}$*

and $(B_n)_{n\in\mathbb{N}}$ (i.e., A_n, B_n are $\mathcal{F}_{n-1}$ measurable) and strictly positive constants $(c_n)_{n\in\mathbb{N}}$ such that for all $n \in \mathbb{N}$, a.s.,

$$A_n \leq Z_n - Z_{n-1} \leq B_n \quad \textit{and} \quad B_n - A_n \leq c_n.$$

Then for all $t > 0$,

$$\mathbb{P}\left(\sup_{0\leq i\leq n}(Z_i - Z_0) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=0}^{n} c_i^2}\right)$$

Proof. T.3.52 in Roch MDP p.117 or C.2.20 in Wainwright HDS p.36 or T.9.1. in Lugosi PTPR p.135 or C.3.9. in van Handel APC550 p.51. □

By taking the inequality for $(-Z_n)_{n\in\mathbb{N}^*}$ yields a bound in the other direction.

In the Azuma–Hoeffding inequality, the difference sequence $(\Delta_n)_{n\in\mathbb{N}}$ is not only pairwise uncorrelated $\mathbb{E}(\Delta_i\Delta_j) = 0$ for all $i \neq j$ but also mutually uncorrelated in the sense $\mathbb{E}(\Delta_{i_1}\dots\Delta_{i_k})$ for all $k \in \mathbb{N}$ and all $i_1, \dots, i_n \in \mathbb{N}$. This is part of the reason why the sum $Z_n - Z_0 = \sum_{i=1}^{n}$ is so well concentrated (see E.3.1. in Roch MDP for a different proof of the Azuma–Hoeffding inequality).

Proposition 14 (Doob Martingale). *Let $(\mathcal{F}_n)_{n\in\mathbb{N}^*}$ be a filtration and Y a random variable with $\mathbb{E}|Y| < +\infty$. For all $n \in \mathbb{N}^*$, define $Z_n = \mathbb{E}(Y|\mathcal{F}_n)$. Then $(Z_n)_{n\in\mathbb{N}^*}$ is a martingale with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}^*}$ known as the Doob martingale.*

Proof. $\mathbb{E}|Z_n| \leq \mathbb{E}|Y| < +\infty$ and $\mathbb{E}(Z_n|\mathcal{F}_{n-1}) = \mathbb{E}(Y|\mathcal{F}_{n-1}) = Z_{n-1}$. □

The Doob martingale can be used with the Azuma–Hoeffding inequality to obtain a first manifestation of the concentration of measure phenomenon. To see this, consider the Doob martingale with $Y = f(X_1, \dots, X_n)$, where $(X_1, \dots, X_n)$ are independent random variables with values in arbitrary spaces $\mathcal{X}_1, \dots, \mathcal{X}_n$ and $f\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ is measurable such that $\mathbb{E}|f(X_1, \dots, X_n)| < +\infty$, and filtration given by $\mathcal{F}_0$ the trivial σ-algebra and $\mathcal{F}_i = \sigma(X_1, \dots, X_i)$ for all $1 \leq i \leq n$. Note that $Z_n = \mathbb{E}(f(X_1, \dots, X_n)|\sigma(X_1, \dots, X_n)) = f(X_1, \dots, X_n)$ and $Z_0 = \mathbb{E}(f(X_1, \dots, X_n))$, so that

$$f(X_1, \dots, X_n) - \mathbb{E}(f(X_1, \dots, X_n)) = Z_n - Z_0.$$

It is then possible to relate the martingale difference $Z_i - Z_{i-1}$ to the discrete derivatives of the function $f\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ defined as

$$\begin{aligned} D_i f(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \coloneqq \sup_{y\in\mathcal{X}_i} & f(x_1, \dots, x_{i_1}, y, x_{i+1}, \dots, x_n) \\ & - \inf_{y'\in\mathcal{X}_i} f(x_1, \dots, x_{i_1}, y', x_{i+1}, \dots, x_n). \end{aligned}$$

Take $X' = (X'_1, \dots, X'_n)$ an independent copy of $X = (X_1, \dots, X_n)$, and let $X^{(i)} = (X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n)$. Then

$$\begin{aligned} Z_i - Z_{i-1} &= \mathbb{E}(f(X)|\mathcal{F}_i) - \mathbb{E}(f(X)|\mathcal{F}_{i-1}) \\ &= \mathbb{E}(f(X)|\mathcal{F}_i) - \mathbb{E}(f(X^{(i)})|\mathcal{F}_{i-1}) \\ &= \mathbb{E}(f(X)|\mathcal{F}_i) - \mathbb{E}(f(X^{(i)})|\mathcal{F}_i) \\ &= \mathbb{E}(f(X) - f(X^{(i)})|\mathcal{F}_i). \end{aligned}$$

Let us write $x = (x_1, \ldots, x_n)$ and $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. By definition of the supremum we have

$$\begin{aligned}
|f(X) - f(X^{(i)})| &\leq |\sup_{x, y_i}(f(x) - f(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n))| \\
&\leq \sup_{x_{-i}} |\sup_{x_i, y_i}(f(x) - f(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n))| \\
&= \|D_i f\|_\infty \\
&(\leq \sup_{x, y_i} |f(x) - f(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)|),
\end{aligned}$$

so that by Jensen's inequality we have

$$|Z_i - Z_{i-1}| \leq \|D_i f\|_\infty.$$

Then if $\|D_i f\|_\infty < +\infty$ for all i, we can apply the Azuma–Hoeffding inequality to get bounds on $Z_n - Z_0$. This is the content of McDiarmid's inequality. (A more careful analysis leads to an improvement by a factor of 4 in the exponent of the bound.)

Proposition 15 (**McDiarmid's Inequality**). *Let $X_1, \ldots, X_n$ be independent random variables with values in $\mathcal{X}_1, \ldots, \mathcal{X}_n$ and $f\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ a measurable such that $\|D_i f\|_\infty < +\infty$ for all $1 \leq i \leq n$. Then for all $t > 0$,*

$$\mathbb{P}(f(X_1, \ldots, X_n) - \mathbb{E}\,(f(X_1, \ldots, X_n)) \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n \|D_i f\|_\infty^2}\right).$$

Proof. Above (for a weaker bound). (For the stated bound,) see T.3.61. in Roch MDP p.125 or T.3.41. in van Handel APC550 p.52 or T.9.2. in Lugosi PTPR p.136 (the latter for a slightly weaker result, as explained below). □

We can again apply the inequality to $-f$ to get a tail bound in the other direction. Note that McDiarmid's inequality is often stated under a stronger condition on f, known as bounded differences: a function f is said to have **bounded differences** if there exist positive constants c_i such that

$$\sup_{x, y_i} |f(x) - f(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

As shown above, the property of bounded differences implies the finiteness of $\|D_i f\|_\infty$ for all i. If we introduce the weighted Hamming metric $d_c(x, y) = \sum_{i=1}^n c_i \mathbb{1}_{x_i \neq y_i}$, then it possible to show that the finiteness of $\|D_i f\|_\infty$ for all i is equivalent to f being 1-Lipschitz with respect to d_c (see L.4.5. in van Handel APC550 p.74).

4 Entropy methods

The sub-Gaussian property does not behave well under tensorization. The martingale-based method used "a sort of poor man's tensorization [of the sub-Gaussian] property for sums of martingale increments" (van Handel ACP550 p. 51). (Recall that the big idea is to use tensorization to reduce the problem to the 1-dimensional case.) To develop better bounds, the objective is to find a formulation of the sub-Gaussian property that

behaves well under tensorization. The idea is to use the sub-Gaussian characterization $\frac{d}{d\lambda}\lambda^{-1}\psi(\lambda) \lesssim 1$ where ψ is the log-moment generating function.

Definition 16 (ϕ-Entropy)**.** Let $\mathscr{X}^+$ the space of positive integrable random variables and $\phi\colon \mathbb{R}_+ \to \mathbb{R}$ a convex function. The functional $H_\phi\colon \mathscr{X}^+ \to \mathbb{R}_+ \cup \{+\infty\}$ defined such that for all $X \in \mathscr{X}^+$,

$$H_\phi(X) = \mathbb{E}(\phi(X)) - \phi(\mathbb{E}(X))$$

is said to be the ***φ*-entropy**.

By Jensen's inequality, the ϕ-entropy is indeed a positive functional and it is finite if and only if $\mathscr{X}^+$ is restricted to the space of positive integrable random variables X such that $\mathbb{E}(|\phi(X)|) < +\infty$. The entropy is a measure of variability: in particular, we have $H_\phi(X) = 0$ if and only if $X = \mathbb{E}(X)$ (a.s.). The ϕ-entropy with $\phi = x \mapsto x^2$ is nothing more than the variance. If we restrict attention to random variables $Y = e^{\lambda X}$, the ϕ-entropy with $\phi = x \mapsto -\ln(x)$ is the centered log-moment generating function. In what follows, we should focus only on the ϕ-entropy with $\phi = x \mapsto x\ln x$ that we should simply denote H and call entropy, that is, for any $X \in \mathscr{X}^+$,

$$H(X) = \mathbb{E}(X\ln X) - \mathbb{E}(X)\ln\mathbb{E}(X).$$

Proposition 17 (**Herbst Argument**)**.** *Let X be a random variable with cumulative moment function. If for all $\lambda \geq 0$,*

$$H(e^{\lambda X}) \leq \frac{\lambda^2\sigma^2}{2}\mathbb{E}(e^{\lambda X}),$$

then for all $\lambda \geq 0$,

$$\ln\mathbb{E}(e^{\lambda(X-\mathbb{E}X)}) \leq \frac{\lambda^2\sigma^2}{2}.$$

Proof. L.3.13. in van Handel APC550 p.56 or P.3.2. in Wainwright HDS p.60 or S.5.2. in BLM CI p.121. □

If the assumption in the Herbst argument is satisfied for all $\lambda \in \mathbb{R}$, then applying the result to $-X$ shows that X is sub-Gaussian. (It can be proved that, up to constant factor, the converse holds.) The assumed bound on the entropy can thus be interpreted as another characterization of sub-Gaussianity. The benefit then comes from the tensorization property (also known as the sub-additivity property) of the entropy (property shared with many other ϕ-entropy such as the variance).

Let $X_1, \ldots, X_n$ be independent random variables with values in $\mathcal{X}_1, \ldots, \mathcal{X}_n$ and $f\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}_+$ a measurable function. Define for each $1 \leq i \leq n$ the function H_i by

$$H_i(f(X_1, \ldots, X_n)) = H(f(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, x_n)).$$

Proposition 18 (**Entropy Tensorization**)**.** *Let $X_1, \ldots, X_n$ be independent random variables with values in $\mathcal{X}_1, \ldots, \mathcal{X}_n$ and $f\colon \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}_+$ a measurable function. Then*

$$H(f(X_1, \ldots, X_n)) \leq \mathbb{E}\left(\sum_{i=1}^n H_i(f(X_1, \ldots, X_n))\right).$$

Proof. T.3.14. in van Handel APC550 p.57 or T.4.10. in BLM CI p.94. □

Lemma 19. *Let $D^- f := f - \inf f$. Then*

$$H(e^f) \leq \operatorname{cov}(f, e^f) \leq \mathbb{E}\left(|D^- f|^2 e^f\right).$$

Proof. L.3.16. in van Handel APC550 p.58. □

For a function $f\colon \mathcal{X}_1 \times \dots \mathcal{X}_n \to \mathbb{R}$, define the one-sided discrete derivatives

$$D_i^- f(x) := f(x_1, \dots, x_n) - \inf_{y_i} f(x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n),$$
$$D_i^+ f(x) := \sup_{y_i} f(x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_n).$$

Proposition 20 ((Improved) Bounded Differences Inequality). *Let $X_1, \dots, X_n$ be independent random variables with values in $\mathcal{X}_1, \dots, \mathcal{X}_n$ and $f\colon \mathcal{X}_1 \times \dots \times \mathcal{X}_n \to \mathbb{R}$ a measurable function.*

1. If $\| \sum_{i=1}^n |D_i^- f|^2 \|_\infty < +\infty$, then for all $t \geq 0$,

$$\mathbb{P}(f(X_1, \dots, X_n) - \mathbb{E}\left(f(X_1, \dots, X_n)\right) \geq t) \leq \exp\left(-\frac{t^2}{4\| \sum_{i=1}^n |D_i^- f|^2 \|_\infty}\right).$$

2. If $\| \sum_{i=1}^n |D_i^+ f|^2 \|_\infty < +\infty$, then for all $t \geq 0$,

$$\mathbb{P}(f(X_1, \dots, X_n) - \mathbb{E}\left(f(X_1, \dots, X_n)\right) \leq -t) \leq \exp\left(-\frac{t^2}{4\| \sum_{i=1}^n |D_i^+ f|^2 \|_\infty}\right),$$

Proof. T.3.18. in van Handel APC550 p.59 or T.6.7. in BLM CI p.176. □

5 Metric perspective and transportation methods

In the entropy approach, we noticed the connection between bounding the (discrete) gradient of f and f being Lipschitz when generating concentration inequalities for $f(X_1, \dots, X_n)$ where $X_1, \dots, X_n$ are independent. Both properties encapsulate the idea that f should not be too sensitive to any of its coordinates. If the two properties are sometimes equivalent, it is not always the case, so that developing "a metric viewpoint that emphasizes the role of Lipschitz functions" (van Handel) proves useful. For $K \geq 0$, let us write $\operatorname{Lip}_K(E, d)$ the space of real-valued K-Lipschitz functions from a metric space (X, d). That is, $f\colon E \to \mathbb{R}$ is in $\operatorname{Lip}_K(E, d)$ if and only if for all $x, y \in E$,

$$|f(x) - f(y)| \leq K d(x, y).$$

Equivalently, f is in $\operatorname{Lip}_K(E, d)$ if and only if

$$\|f\|_{\text{Lip}} := \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)} \leq K.$$

Remark. "There is an entirely different approach to investigating Lipschitz concentration properties that played an important role in the historical development of this area: the isoperimetric method. [...] Mathematical phenomena relating the size of a set to the size of its boundary are generally referred to as “isoperimetric problems.” [...] Isoperimetric inequalities are equivalent to tail bounds for Lipschitz functions. However, unlike most

of our previous results [...], the deviation here is from the median rather than from the mean. It turns out that deviation inequalities from the median and the mean are equivalent, however, up to constants." (P.4.2. in van Handel APC550 p.79).

We already know from Gaussian concentration and McDiarmid's inequality that: if $X \sim N(0,1)$, then $f(X)$ is sub-Gaussian with proxy variance 1 for every $f \in L_1(\mathbb{R}^n, \|\cdot\|)$; if X is a random vector with independent entries, then $f(X)$ is sub-Gaussian with proxy variance $\|c\|^2/4$ for every $f \in \mathrm{Lip}_1(E_1 \times \dots E_n, d_c)$. Then the general question we ask is: for which probability measure μ on a metric space (E,d) does it hold that if $\mathbb{P}_X = \mu$, then $f(X)$ is sub-Gaussian for every $f \in \mathrm{Lip}_1(E,d)$?

To answer this question, we need some new objects. Given a metric space (E,d), recall first that $\mathcal{M}_1(E)$ is the space of Borel probability measures on E. Let us then define for $p \in [1, +\infty)$ the space

$$\mathscr{P}_p(E,d) = \left\{ \mu \in \mathcal{M}_1(E) : \int_E d(x_0,x)^p \, d\mu(x) < +\infty \right\}$$

where $x_0 \in E$ is arbitrary (the finiteness of the integral does not depend on the choice of x_0 so that the space itself does not depend on the choice of x_0).

Definition 21 (Relative Entropy)**.** Let μ and ν be probability measures on a measurable space $(E, \mathcal{E})$. The function $D(\cdot\|\cdot)$ defined by

$$D(\nu\|\mu) = \begin{cases} H\left(\frac{d\nu}{d\mu}\right) & \text{if } \nu \ll \mu, \\ +\infty & \text{otherwise} \end{cases}$$

is said to be the **relative entropy** of ν relative to ν.

If $\nu \ll \mu$, then

$$\begin{aligned} D(\nu\|\mu) &= \int_E \ln\left(\frac{d\nu}{d\mu}\right) \frac{d\nu}{d\mu}\, d\mu - \left(\int_E \frac{d\nu}{d\mu}\, d\mu\right) \ln\left(\int_E \frac{d\nu}{d\mu}\, d\mu\right) \\ &= \int_E \ln\left(\frac{d\nu}{d\mu}\right) \frac{d\nu}{d\mu}\, d\mu, \end{aligned}$$

since by definition $\int_E \frac{d\nu}{d\mu}\, d\mu = \nu(E) = 1$. Equivalently,

$$D(\nu\|\mu) = \int_E \ln\left(\frac{d\nu}{d\mu}\right) \frac{d\nu}{d\mu}\, d\mu = \int_E \ln\left(\frac{d\nu}{d\mu}\right) d\nu.$$

Lemma 22. *$D(\nu\|\mu) \geq 0$ and $D(\nu\|\mu) = 0$ if and only if $\mu = \nu$ a.e..*

Proof. L.5.3. in Rassoul-Agha and Seppalainen CLDIGB p.68. □

Relative entropy provides a notion of "distance" between probability measures, but it is not properly a distance since it is not symmetric and even the symmetric sum does not satisfy the triangle inequality. The relative entropy is part of a general classes of "statistical distances" known as divergences (hence the notation). In particular, the relative entropy is also called the **Kullback–Leibler divergence**. In information theory, it is more customary to denote the entropy by Ent and the relative entropy directly by H.

Lemma 23 (**Chain Rule for Relative Entropy**)**.** *Let μ and ν be probability measures on a measurable space $(E, \mathcal{E})$. Let $\mathcal{F}$ a sub-σ-algebra of $\mathcal{E}$ and $\mu_{\mathcal{F}}$ and $\nu_{\mathcal{F}}$ the restrictions of μ and ν to $\mathcal{F}$. Suppose there exist regular conditional probability measures of μ and ν given $\mathcal{F}$, that is, $\mu^x(\cdot) := \mu(\cdot|\mathcal{F})(x)$ and $\nu^x(\cdot) := \nu(\cdot|\mathcal{F})(x)$ for all $x \in E$. Then*

$$D(\nu\|\mu) = D(\nu_{\mathcal{F}}\|\mu_{\mathcal{F}}) + \int_E D(\nu^x\|\mu^x)\, d\mu(x).$$

Proof. T.D.13. in Dembo and Zeitouni LDTA p.357 or E.5.13. in Rassoul-Agha and Seppalainen CLDIGB p.72 or L.4.18. in van Handel APC550 p.86. □

Lemma 24 (**Gibbs Variational Principle (Donsker–Varadhan, 1975)**)**.** *Let (E, d) be a metric space. Then for all $\mu, \nu \in \mathcal{M}_1(E)$,*

$$D(\nu, \mu) = \sup_{f \in C_b(E)} \left(\mathbb{E}_\nu(f) - \ln \mathbb{E}_\mu(e^f) \right)$$

Proof. T.5.6. in Rassoul-Agha and Seppalainen CLDIGB p.70 or L. 6.2.13. in Dembo and Zeitouni LDTA p.264 or L.4.10. in van Handel APC550 p.77. □

See S.3.3 in Polyanskiy and Wu's notes on Information Theory p.37 for explanations why more generally variational characterizations of divergences are useful.

Definition 25 (Kantorovich–Rubinstein Distance)**.** Let (E, d) be a metric space. The function $W_1 \colon \mathscr{P}_1(E, d) \times \mathscr{P}_1(E, d) \to \mathbb{R}_+$ defined for all $\mu, \nu \in \mathscr{P}_1(E, d)$ by

$$\mathcal{K}(\mu, \nu) = \sup_{f \in \mathrm{Lip}_1(E,d)} \left| \int f\, d\mu - \int f\, d\nu \right|$$

is said to be the **Kantorovich–Rubinstein distance**.

The metric $\mathcal{K}$ thus allows for the comparison of probability measures (by comparing the expectation of Lipschitz functions under different distributions). This choice is not arbitrary but originates from optimal transport. This origin is not anecdotal but provides deep results which are then put to great use in the manipulation of $\mathcal{K}$ to generate concentration inequalities. As such, it makes sense to introduce the theory in more details.

5.1 Optimal transport in 3 minutes

The general objective in optimal transport is to move goods or matter from a certain distribution to another distribution with minimal cost (the distributions are normalized as probability measures to encapsulate the fact that nothing is lost during transport). This translates into the **(primal) Monge–Kantorovich problem** in which we try to find

$$C(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{E \times E} c(x, y)\, d\gamma(x, y).$$

In other words, the goal is to find the minimal transport cost by selecting an optimal (randomized) transport policy (which takes the form of a coupling, which can be further decomposed into μ and a probability kernel κ yielding the distribution at arrival conditional on initial measured masses). (This formulation of the problem encompasses the too restricting Monge problem in which the policy function takes the form of a mapping

$T\colon E \to E$ such that $T_{\#}\mu = \nu$.) The Kantorovich problem is nothing more than an (infinite dimensional) linear program: the objective is to minimize the linear functional $\gamma \mapsto \int c\, d\gamma$ under the affine constraints $\pi^1_{\#}\gamma = \mu$, $\pi^2_{\#}\gamma = \nu$, and $\gamma > 0$.

As a linear program, the problem admits a (Lagrangian) dual representation (and interpretation). The **dual Kantorovich problem** can be defined as trying to find

$$P(\mu,\nu) = \sup_{\substack{\phi\in L^1(\nu),\psi\in L^1(\mu):\\ \phi(y)-\psi(x)\le c(x,y)}} \left(\int_E \phi(y)\, d\nu(y) - \int_E \psi(x)\, d\mu(x)\right).$$

That is, the goal is to find the maximal transport profit ("price times quantity at selling minus price times quantity at buying") by selecting (under a competitive cost constraint) some optimal pricing functions (one for buying and the other for selling). It can be showed that the cost constraint allows us to express one optimal price function in terms of the other (see Villani OTON p.66), so that only one pricing function need to be considered in the problem.

If we then take $c = d$ in the (modified) dual problem, the cost constraint naturally translates into a Lipschitz condition and we get back the Kantorovich–Rubinstein distance $\mathcal{K}$. If in the primal problem, we take $c = d^p$ for $p \in [1, +\infty)$, then we get (under separability) a distance on $\mathscr{P}_p(E,d)$ whose p^{th} root is called the Wasserstein distance of order p and denoted W_p. In the case $p = 1$, it can be showed (under separability) that $\mathcal{K} = W_1$. In other words, $\mathcal{K}$ can be interpreted as the dual representation of W_1.

Proposition 26 (**Wasserstein Distance**). *Let (E,d) be a metric space and $p \in [1,+\infty)$. Let $\mathcal{E}$ be a σ-algebra on E such that d is $\mathcal{E}\times\mathcal{E}$ measurable. Define the function W_p for all probability measures μ,ν on $(E,\mathcal{E})$ by*

$$W_p(\mu,\nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int_{E\times E} d(x,y)^p\, d\gamma(x,y)\right)^{1/p},$$

*where $\Gamma(\mu,\nu)$ is the set of all probability measures on $(E\times E, \mathcal{E}\times\mathcal{E})$ with marginals μ and ν. If (E,d) is separable and $\mathcal{E} = \mathcal{B}(E,d)$, then W_p is a well-defined function (possibly infinite) on $\mathcal{M}_1(E,d)\times\mathcal{M}_1(E,d)$ whose restriction to $\mathscr{P}_p(E,d)\times\mathscr{P}_p(E,d)$ defines a metric called the **Wasserstein distance** of order p (and also denoted W_p).*

Proof. If E separable and $\mathcal{E} = \mathcal{B}(E)$, then $\mathcal{E}\times\mathcal{E} = \mathcal{B}(E)\times\mathcal{B}(E) = \mathcal{B}(E\times E)$ so that d is $\mathcal{E}\times\mathcal{E}$ measurable since continuous, hence W_p is well-defined on $\mathcal{M}_1(E,d)\times\mathcal{M}_1(E,d)$. See D.6.1. in Villani OTON p.105 for positivity, finiteness, symmetry, and separation of the restriction (Villani assumes completeness but the proof for these properties does not use it – see remark in Villani OTON p.120). The triangle inequality is usually proved by disintegration which requires inner regularity (which obtains for instance under completeness). For a proof under separability only, see Clement and Desch (2008). □

Equivalently, in a more probabilistic flavor,

$$W_p(\mu,\nu) = \left(\inf_{X\sim\mu, Y\sim\nu} \mathbb{E}\left(d(X,Y)^p\right)\right)^{1/p},$$

where the infimum is taken over all pairs (X,Y) of jointly distributed random variables with values in $(E\times E, \mathcal{E}\times\mathcal{E})$ such that $\mathbb{P}_X = \mu$ and $\mathbb{P}_Y = \nu$. We are now ready to make clear the connection between $\mathcal{K}$ and W_1.

Theorem 27 (**Kantorovich–Rubinstein Theorem**). *Let (E, d) be a separable metric space. Then for all $\mu, \nu \in \mathscr{P}_1(E, d)$,*

$$W_1(\mu, \nu) = \mathcal{K}(\mu, \nu).$$

If, in addition, μ and ν are inner regular (for example if (E, d) is complete), then there is a probability measure in $\Gamma(\mu, \nu)$ for which the infimum in the definition of W_1 is attained.

Proof. T.11.8.2. in Dudley RAP p.421 or T.4.13 in van Handel APC550 p.81 or T.5.10 in Villani OTON p.70. □

5.2 Transportation inequalities and tensorization

Theorem 28 (**Induction Lemma**). *For $i = 1, \dots, n$, let (E_i, d_i) be Polish a metric space, $\mu_i \in \mathcal{M}_1(E_i, d_i)$, and $w_i \colon E_i \times E_i \to \mathbb{R}_+$ a $\mathcal{B}(E_i, d_i) \times \mathcal{B}(E_i, d_i)$ measurable function. Let $\phi \colon \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function. If for every $i = 1, \dots, n$,*

$$\inf_{\gamma \in \Gamma(\mu_i, \nu)} \phi(\mathbb{E}_\gamma(w_i(X, Y))) \leq 2\sigma^2 D(\nu \| \mu_i)$$

for all $\nu \in \mathcal{M}_1(E_i, d_i)$, then

$$\inf_{\gamma \in \Gamma(\mu_1 \times \cdots \times \mu_n, \nu)} \sum_{i=1}^{n} \phi(\mathbb{E}_\gamma(w_i(X_i, Y_i))) \leq 2\sigma^2 D(\lambda \| \mu_1 \times \cdots \times \mu_n)$$

for all probability measures λ on $(E_1 \times \cdots \times E_n, \mathcal{B}(E_1, d_1) \times \cdots \times \mathcal{B}(E_n, d_n))$.

Proof. L.8.13. in BLM CI p.256 or T.4.15. in van Handel APC550 p.85. □

Since E_i are assumed separable, we have $\mathcal{B}(E_1, d_1) \times \cdots \times \mathcal{B}(E_n, d_n) = \mathcal{B}(E_1 \times \cdots \times E_n, d_{E_1 \times \cdots \times E_n})$ where $d_{E_1 \times \cdots \times E_n}$ is any metric inducing the product topology.

5.3 T_1 inequalities and Gaussian concentration

Theorem 29 (**Bobkov–Götze Theorem**). *Let (E, d) be a separable metric space and $\mu \in \mathscr{P}_1(E, d)$. If a random variable X has distribution μ, then $f(X)$ is sub-Gaussian with proxy variance σ^2 for all $f \in \mathrm{Lip}_1(E, d)$ if and only if*

$$W_1(\nu, \mu) \leq \sqrt{2\sigma^2 D(\nu \| \mu)}$$

for all $\nu \in \mathscr{P}_1(E, d)$

Proof. T.4.8. in van Handel APC550 p.76 or T.3.4.3. in Raginsky and Sason CMI p.119 or P.6.1. in Ledoux CMP p.119. □

The induction lemma allows us to generate a number of tensorization results for the W_1 distance by taking $w_i = d_i$. The only thing that needs to be taken care of is that in this case the left-hand side tensorized quantity is not a W_1 distance. An extra step using Cauchy–Schwarz needs to be performed. For this, given $c = (c_1, \dots, c_n)$ with $c_i > 0$, define $d_c(x, y) = \sum_{i=1}^n c_i d_i(x_i, y_i)$ which can be proved to be a metric on the space $E_1 \times \cdots \times E_n$ generating the product topology. To make clear which metric is used on E_i or $E_1 \times \cdots \times E_n$ to define W_1, we may write $W_1(\mu, \nu; d)$.

Corollary 30 (Tensorization of T_1 under d_c)**.** *For $i = 1, \dots, n$, let (E_i, d_i) be a Polish metric space and $\mu_i \in \mathcal{M}_1(E_i, d_i)$. Let $c = (c_1, \dots, c_n)$ with $c_i > 0$ and define $d_c(x, y) = \sum_{i=1}^n c_i d_i(x_i, y_i)$. If $\sum_{i=1}^n c_i^2 = 1$ and if for every $i = 1, \dots, n$,*

$$W_1(\mu_i, \nu; d_i) \leq \sqrt{2\sigma^2 D(\nu \| \mu_i)}$$

for all $\nu \in \mathcal{M}_1(E_i, d_i)$, then

$$W_1(\mu_1 \times \cdots \times \mu_n, \tau; d_c) \leq \sqrt{2\sigma^2 D(\tau \| \mu_1 \times \cdots \times \mu_n)},$$

for all $\tau \in \mathcal{M}_1(E_1 \times \cdots \times E_n, d_c)$.

Proof. Apply Cauchy–Schwarz to the result of the induction lemma for $\phi(x) = x^2$ and $w_i = d_i$. See C.4.16. in van Handel APC550 p.85. □

Remark. It is possible to extend the result by taking $w_i = d_i$ but assuming that the spaces E_i are Polish with respect to different metrics ρ_i. In this case, one must be careful about: the measurability of d_i; for which measures the hypothesis must be verified (namely, $\mathcal{M}_1(E_i, \rho_i)$, and not $\mathcal{M}_1(E_i, d_i)$); and for which measures the result hold (namely, $\mathcal{M}_1(E_1 \times \cdots \times E_n, \rho)$ where ρ is any metric generating the product topology induced by the ρ_i). This extension allows, for example, to take the metric $d_i(x_i, y_i) = \mathbb{1}_{x_i \neq y_i}$ even when E_i is neither finite nor countable (so that (E_i, d_i) is not separable), yielding another proof of McDiarmid's inequality. Indeed, d_i is $\mathcal{B}(E_i, \rho_i) \times \mathcal{B}(E_i, \rho_i)$ measurable (since (E_i, ρ_i) is separable). Moreover, $W_1(\mu_i, \nu; d_i) = \|\mu_i - \nu_i\|_{TV}$ for all $\nu \in \mathcal{M}_1(E_i, \rho_i)$, so that Pinsker's inequality yields $W_1(\mu_i, \nu; d_i) \leq \sqrt{2\sigma^2 D(\nu \| \mu_i)}$ for all $\nu \in \mathcal{M}_1(E_i, \rho_i)$. By applying the extension of last corollary and the Bobkov–Gotze theorem, we get back McDiarmid's inequality.

5.4 Talagrand's concentration inequality from conditional transportation

So far, no new concentration inequalities have been obtained with the transportation method. We now show that the procedure above can be extended to hold under a one-sided Lipschitz condition (mirroring the transition but not equivalent to the one-sided derivative condition in the entropy method) which will generate new concentration results. The result we derive is known as Talagrand's concentration inequality (which he initially derived "in an isoperimetric form in terms of a 'convex distance' to a set" (van Handel) and was latter re-derived using the transportation method by Marton).

Proposition 31 (**Marton's d_2 (Asymmetric) Distance**)**.** *Let (E_i, ρ_i) be a Polish metric space for $i = 1, \dots, n$. Let ρ be any metric generating the product topology induced by the ρ_i (and $\rho = \rho_1$ if $n = 1$). Let $c_i \colon E_1 \times \cdots \times E_n \to \mathbb{R}_+$ be any $\mathcal{B}(E_1 \times \cdots \times E_n, \rho)$-measurable function for $i = 1, \dots, n$. Define the function $d_2 \colon \mathcal{M}_1(E_1 \times \cdots \times E_n, \rho) \times \mathcal{M}_1(E_1 \times \cdots \times E_n, \rho) \to \mathbb{R}_+$ by*

$$d_2(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \sup_{\mathbb{E}_\gamma(\sum_{i=1}^n c_i(X)^2) \leq 1} \mathbb{E}_\gamma \left(\sum_{i=1}^n c_i(X) \mathbb{1}_{X_i \neq Y_i} \right).$$

Then

$$d_2(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \sum_{i=1}^{n} \mathbb{E}_\gamma \left((\gamma(X_i \neq Y_i | X))^2 \right) \right)^{1/2}.$$

If $n = 1$*, then*

$$d_2(\mu,\nu) = \left(\int_E \frac{(f-g)_+^2}{g} \, d\tau \right)^{1/2},$$

where τ *is any dominating measure for* μ *and* ν *(for instance,* $(\mu+\nu)/2$*),* $f = d\mu/d\tau$*, and* $g = d\nu/d\tau$*.*

Proof. L.4.26. in van Handel APC550 p.94 for the first equality. L.8.3. in BLM CI p.243 for the second equality. □

In the definition of d_2, the Polish assumption ensures that $((x_1,\dots,x_n),(y_1,\dots,y_n)) \mapsto \sum_{i=1}^n c_i(x_1,\dots,x_n)\mathbb{1}_{x_i \neq y_i}$ is $\mathcal{B}(E_1 \times \cdots \times E_n, \rho) \times \mathcal{B}(E_1 \times \cdots \times E_n, \rho)$-measurable. If $n = 1$, the function d_2 can be more broadly defined for any measurable space $(E, \mathcal{E})$ such that c_1 is $\mathcal{E}$-measurable and the diagonal of the product space $E \times E$ is measurable with respect to the product σ-algebra $\mathcal{E} \times \mathcal{E}$ (which holds for instance if E is separable).

Proposition 32. *Let* (E, ρ) *be a separable metric space. Then for all* $\mu, \nu \in \mathcal{M}_1(E, \rho)$*,*

$$d_2^2(\nu,\mu) + d_2^2(\mu,\nu) \leq 2D(\nu\|\mu).$$

Proof. L.8.4. in BLM CI p.244. □

It is possible to slightly extend the induction lemma of last section to hold in this case. Combined with last result, the extension gives the following transportation inequalities.

Theorem 33 (**Marton's Conditional Transportation Inequality**)**.** *For* $i = 1,\dots,n$*, let* (E_i, ρ_i) *be a Polish metric space and* $\mu_i \in \mathcal{M}_1(E_i, \rho_i)$*. Then*

$$d_2^2(\nu, \mu_1 \times \cdots \times \mu_n) + d_2^2(\mu_1 \times \cdots \times \mu_n, \nu) \leq 2D(\nu\|\mu_1 \times \cdots \times \mu_n)$$

for all $\nu \in \mathcal{M}_1(E_1 \times \cdots \times E_n, \rho)$ *where* ρ *is any metric generating the product topology induced by the* ρ_i*.*

Proof. P.4.27. in van Handel APC550 p.65 or L.8.13. in BLM CI p.256 for the extension of the induction lemma. The result then follows immediately from the 1-dimensional transportation inequality above. See T.4.24. in van Handel APC550 p.93 or T.8.5. in BLM CI p.245. □

Theorem 34 (**Talagrand's Concentration Inequality**)**.** *For* $i = 1,\dots,n$*, let* (E_i, ρ_i) *be a Polish metric space and* X_i *a random variable in* E_i*. Let* ρ *be any metric generating the product topology induced by the* ρ_i*. Let* $f : E_1 \times \cdots \times E_n \to \mathbb{R}$ *be a* $\mathcal{B}(E_1 \times \cdots \times E_n, \rho)$*-measurable function. Denote* $X = (X_1,\dots,X_n)$*. If* $X_1,\dots,X_n$ *are independent and there exist for* $i = 1,\dots,n$ $\mathcal{B}(E_1 \times \cdots \times E_n, \rho)$*-measurable functions* $c_i : E_1 \times \cdots \times E_n \to \mathbb{R}_+$ *such that for all* $x, y \in E_1 \times \cdots \times E_n$

$$f(x) - f(y) \leq \sum_{i=1}^{n} c_i(x)\mathbb{1}_{x_i \neq y_i},$$

then for all $\lambda \geq 0$*,*

$$\ln \mathbb{E}\left(e^{\lambda(f(X)-\mathbb{E}f(X))}\right) \leq \frac{1}{2}\lambda^2 \left\| \sum_{i=1}^n c_i^2 \right\|_\infty$$

and

$$\ln \mathbb{E}\left(e^{\lambda(-f(X)+\mathbb{E}f(X))}\right) \leq \frac{1}{2}\lambda^2 \mathbb{E}\left(\sum_{i=1}^n c_i(X)^2\right).$$

Proof. T.4.20. in van Handel APC550 p.91 or T.8.6. in BLM CI p.245. □

In particular, we get, under the hypothesis of last theorem, for all $t \geq 0$,

$$\mathbb{P}(f(X) - \mathbb{E}f(X) \geq t) \leq e^{-t^2/2\| \sum_{i=1}^n c_i^2 \|_\infty}$$

and

$$\mathbb{P}(f(X) - \mathbb{E}f(X) \leq -t) \leq e^{-t^2/2\mathbb{E}\left(\sum_{i=1}^n c_i(X)^2\right)}.$$

Since $\| \sum_{i=1}^n c_i^2 \|_\infty \geq \mathbb{E}\left(\sum_{i=1}^n c_i(X)^2\right)$, the lower bound is sharper than the upper bound (and conversely when the one-sided Lipschitz property is reversed). Sometimes, the better bound is not included in the statement of the theorem, which can then be reformulated as: if $X_1, \ldots, X_n$ are independent and f satisfies the one-sided Lipschitz property for c_i, then $f(X_1, \ldots, X_n)$ is sub-Gaussian with variance proxy $\| \sum_{i=1}^n c_i^2 \|_\infty$.

Corollary 35. *If* $X_1, \ldots, X_n$ *are independent with values in* $[0, 1]$*, then* $f(X_1, \ldots, X_n)$ *is sub-Gaussian with variance proxy* $\| \|\nabla f\|^2 \|$ *for every convex function* f*.*

Proof. C.4.23. in van Handel APC550 p.92. □

5.5 T_2 inequalities and dimension-free Gaussian concentration

Corollary 36 (Tensorization of T_2 under the 2-norm product metric)**.** *For* $i = 1, \ldots, n$*, let* (E_i, d_i) *be a Polish metric space and* $\mu_i \in \mathcal{M}_1(E_i, d_i)$*. Define* $d(x, y) = \left(\sum_{i=1}^n d_i(x_i, y_i)^2\right)^{1/2}$*. If* $\sum_{i=1}^n c_i^2 = 1$ *and if for every* $i = 1, \ldots, n$*,*

$$W_2(\mu_i, \nu; d_i) \leq \sqrt{2\sigma^2 D(\nu \| \mu_i)}$$

for all $\nu \in \mathcal{M}_1(E_i, d_i)$*, then*

$$W_2(\mu_1 \times \cdots \times \mu_n, \tau; d) \leq \sqrt{2\sigma^2 D(\tau \| \mu_1 \times \cdots \times \mu_n)},$$

for all $\tau \in \mathcal{M}_1(E_1 \times \cdots \times E_n, d_E)$*.*

Proof. Apply $x \mapsto x^2$ to the result of the induction lemma for $\phi(x) = x$ and $w_i = d_i^2$. See C.4.30. in van Handel APC550 p.101. □

Naturally, the result does not extend to other p (because T_p inequalities are expressed in terms of the squared root of the relative entropy, not its p-th root). This is sometimes summed up by saying that T_p exactly tensorizes under the p-norm product metric only for $p = 2$. (Recall that the exact tensorization of T_1 we previously got was under a particular weighted 1-norm product metric, and not the 1-norm product metric.) It is possible to prove a related result: for (E, d) Polish and $p \in [1, 2]$, if $\mu \in \mathscr{P}_p(E, d)$ satisfies $T_p(c)$

on (E, d), then the product measure $\mu^{\times n}$ satisfies $T_p(cn^{1-2/p})$ on $(E^n, d_{p,n})$ where $d_{p,n}$ is the p-norm product metric (see P.22.5. in Villani OTON p.586 or P.3.4.3. in Raginsky and Sason CMI p.115).

Theorem 37 (**Gozlan's Theorem**). *Let (E, d) be a Polish metric space and $\mu \in \mathcal{M}_1(E, d)$. Let $(X_n)_{n\in\mathbb{N}}$ be i.i.d. with distribution μ. Define $d_{2,n}(x, y) = (\sum_{i=1}^n d(x_i, y_i)^2)^{1/2}$. Then the following propositions are equivalent:*

(i) μ satisfies the $T_2(\sigma^2)$ inequality on (E, d), that is, for all $\nu \in \mathcal{M}_1(E, d)$,

$$W_2(\mu, \nu) \leq \sqrt{2\sigma^2 D(\nu\|\mu)};$$

(ii) $\mu^{\times n}$ satisfies the $T_2(\sigma^2)$ inequality on $(E^n, d_{2,n})$ for every $n \geq 1$, that is, for qll $n \geq 1$ and all $\tau \in \mathcal{M}_1(E^n, d_{2,n})$,

$$W_1(\mu^{\times n}, \tau) \leq \sqrt{2\sigma^2 D(\tau\|\mu^{\times n})};$$

3. there is a constant C such that

$$\mathbb{P}(f(X_1, \ldots, X_n) - \mathbb{E}(f(X_1, \ldots, X_n)) \geq t) \leq Ce^{-t^2/2\sigma^2}$$

for all $n \geq 1$, all $t \geq 0$, and all 1-Lipschitz function $f \in \mathrm{Lip}_1(E^n, d_{2,n})$.

Proof. T.4.31. in van Handel APC550 p.102. □

References

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Clement, P., and W. Desch (2008): "An elementary proof of the triangle inequality for the Wasserstein metric," *Proceedings of the American Mathematical Society*, 136(1), 333–339.

Dembo, A., and O. Zeitouni (2009): *Large deviations techniques and applications*, vol. 38. Springer.

Devroye, L., L. Györfi, and G. Lugosi (2013): *A probabilistic theory of pattern recognition*, vol. 31. Springer.

Dudley, R. M. (2004): *Real analysis and probability*. Cambridge University Press.

Kallenberg, O. (2021): *Foundations of modern probability*. Springer.

Ledoux, M., and M. Talagrand (2013): *Probability in Banach Spaces: isoperimetry and processes*. Springer.

Lieb, E. H., and M. Loss (2001): *Analysis*, vol. 14. American Mathematical Soc.

Pollard, D. (2016): "A few good inequalities," *Lecture notes at Yale*.

Raginsky, M., and I. Sason (2014): "Concentration of Measure Inequalities in Information Theory, Communications and Coding," *Foundations and Trends in Communications and Information Theory*.

Rassoul-Agha, F., and T. Seppäläinen (2015): *A course on large deviations with an introduction to Gibbs measures*, vol. 162. American Mathematical Soc.

Roch, S. (2020): *Modern discrete probability*. Manuscript.

van Handel, R. (2016): "Probability in high dimension," *Lecture notes for ACP550 at Princeton*.

Villani, C., et al. (2009): *Optimal transport: old and new*, vol. 338. Springer.

Wainwright, M. J. (2019): *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.