

Lecture 1 — Measure concentration

Lecturer: Sanjoy Dasgupta

Scribe: Nakul Verma, Aaron Arvey, and Paul Ruvolo

1.1 Concentration of measure: examples

We start with some examples of “concentration of measure”. This phenomenon is very useful in analyzing machine learning algorithms and can be used to bound things like error probabilities. The first and the most standard of concentration results is for averages. It states that “the average of bounded independent random variables is tightly concentrated around its expectation”.

1.1.1 Example: coin tosses

Suppose a coin of unknown bias p is tossed n times: $X_1, \dots, X_n \in \{0, 1\}$. Then the average of the X_i is tightly concentrated around p . Specifically

$$\mathbb{P}\left(\left|\frac{X_1 + \dots + X_n}{n} - p\right| \geq \epsilon\right) \leq 2e^{-2\epsilon^2 n}$$

Figure 1.1 shows the quick drop-off of the probability that the sample mean deviates from its expectation. So for a large enough n , we can estimate p quite accurately.

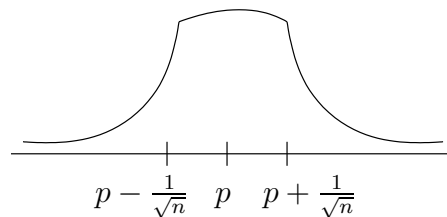


Figure 1.1. Shows the exponential decay in the probability of the sample mean deviating from its expectation (p) in the coin tossing experiment.

1.1.2 Example: random points in a d -dimensional box

Pick a point $X \in [-1, +1]^d$ uniformly at random. Then it can be shown that $\|X\|$ is tightly concentrated around $\sqrt{d/3}$. To see this we note that $X = (X_1, \dots, X_d)$, then

$$\mathbb{E}\|X\|^2 = \mathbb{E}[X_1^2 + \dots + X_d^2] = \sum_{i=1}^d \mathbb{E}X_i^2 = \sum_{i=1}^d \int_{-1}^{+1} \frac{1}{2}x^2 dx = \frac{d}{3}$$

where the second equality is due to the linearity of expectation. Now since each of the X_i 's are independent (and bounded), we can show the concentration

$$\mathbb{P}\left(\left|\|X\|^2 - \frac{d}{3}\right| \geq \epsilon d\right) \leq 2e^{-2\epsilon^2 d}.$$

This provides us with the counter-intuitive result that the volume of the high-dimensional cube tends to lie in its corners where the points have length approximately $\sqrt{d/3}$.

Note that above examples are special cases of *Hoeffding's Inequality*:

Lemma 1 (Hoeffding's inequality). *Suppose X_1, \dots, X_n are independent and bounded variables, such that $a_i \leq X_i \leq b_i$. Then,*

$$\mathbb{P} \left[\left| \frac{X_1 + \dots + X_n}{n} - \mathbb{E} \left(\frac{X_1 + \dots + X_n}{n} \right) \right| \geq \epsilon \right] \leq 2e^{-2\epsilon^2 n^2 / \sum_i (b_i - a_i)^2}. \quad (1.1)$$

We will soon prove a much more general version of this, which is introduced next.

1.1.3 Concentration of Lipschitz functions

Observing the Hoeffding bound, one might wonder whether such concentration applies only to averages of random variables. After all, what is so special about averages? It turns out that the relevant feature of the average that yields tight concentration is that it is “smooth”. In fact any “smooth” function of bounded independent random variables is tightly concentrated around its expectation. The notion of smoothness we will use is *Lipschitz*.

Definition 2. $f : \mathbb{R}^n \mapsto \mathbb{R}$ is λ -Lipschitz w.r.t. the l_p -metric if, for all x, y ,

$$|f(x) - f(y)| \leq \lambda \|x - y\|_p.$$

Example. For $x = (x_1, \dots, x_n)$, define the average: $a(x) = \frac{1}{n}(x_1 + \dots + x_n)$. Then $a(\cdot)$ is $(1/n)$ -Lipschitz with respect to l_1 metric, since for any x, x' ,

$$|a(x) - a(x')| = \left| \frac{1}{n} [(x_1 - x'_1) + \dots + (x_n - x'_n)] \right| \leq \frac{1}{n} (|x_1 - x'_1| + \dots + |x_n - x'_n|) = \frac{1}{n} \|x - x'\|_1.$$

It turns out that Hoeffding's bound holds for all Lipschitz (with respect to l_1) functions.

Lemma 3 (Concentration of Lipschitz functions wrt l_1 metric). *Suppose X_1, \dots, X_n are independent and bounded with $a_i \leq x_i \leq b_i$. Then, for any $f : \mathbb{R}^n \mapsto \mathbb{R}$ which is λ -Lipschitz w.r.t. l_1 -metric*

$$\mathbb{P} [f \geq \mathbb{E}f + \epsilon] \leq e^{-2\epsilon^2 / \lambda^2 \sum_i (b_i - a_i)^2}$$

Proof. See Section 1.5. □

Remark. Since $-f$ is also λ -Lipschitz, we can bound both above and below as

$$\mathbb{P} [|f - \mathbb{E}f| \geq \epsilon] \leq 2e^{-2\epsilon^2 / \lambda^2 \sum_i (b_i - a_i)^2}. \quad (1.2)$$

We now look at bounds for functions that are Lipschitz with respect to other metrics.

1.1.4 Concentration of Lipschitz functions w.r.t. l_2 metric

Let S^{d-1} denote the surface of the unit sphere in \mathbb{R}^d , and let μ be the uniform distribution over S^{d-1} . The following is known (we will prove it later in the course):

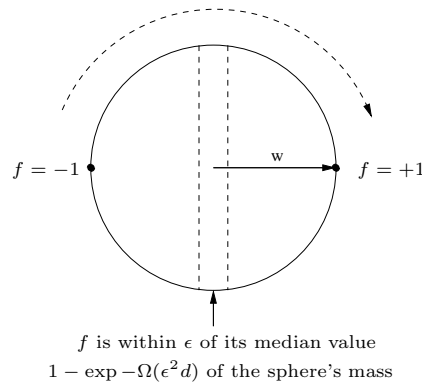


Figure 1.2. The function $f(x) = w \cdot x$ is -1 at one pole of the sphere, $+1$ at the other pole, and increases steadily from -1 to $+1$ as one moves from one pole to the other. Since f is 1-Lipschitz on S^{d-1} , most of the volume lies in a thin slice near the equator of the sphere (perpendicular to w).

Lemma 4. Let $f : S^{d-1} \mapsto \mathbb{R}$ be λ -Lipschitz w.r.t. l_2 -metric. Then,

$$\mu[f \geq \text{med}(f) + \epsilon] \leq 4e^{-\epsilon^2 d / 2\lambda^2} \quad (1.3)$$

where $\text{med}(f)$ is a median value of f .

One immediate consequence of (1.3) is most of the volume of the sphere lies in a thin slice around the equator (for all equators!). To see this, fix any unit vector $w \in S^{d-1}$. Then for $X \sim \mu$ (this notation means “ X drawn from distribution μ ”), $\mathbb{E}(w \cdot X) = 0$ and also $\text{med}(w \cdot X) = 0$. Moreover, the function $f(x) = w \cdot x$ is 1-Lipschitz wrt the l_2 norm: for all $x, y \in S^{d-1}$,

$$|f(x) - f(y)| = |w \cdot x - w \cdot y| = |w \cdot (x - y)| \leq \|w\|_2 \|x - y\|_2 = \|x - y\|_2$$

where the second-to-last inequality uses *Cauchy-Schwarz*. Thus by (1.3), f is tightly concentrated around its median, i.e.,

$$\mu[X : |w \cdot X| \geq \epsilon] \leq 4e^{-\epsilon^2 d / 2}.$$

See Figure 1.2. Moreover, since there is nothing special about this particular w ; the above bound is true for any equator!

1.1.5 Types of concentration

Types of concentration we’ll encounter in this course:

- Concentration of a product measure $X = (X_1, \dots, X_n)$ where X_i are independent and bounded, with respect to l_1 and Hamming metric.
- Concentration of a uniform measure over S^{d-1} , with respect to l_2 metric.
- Concentration of multivariate Gaussian measure, with respect to l_2 metric.

1.2 Probability review

1.2.1 Warm-up problem

Question. Let σ be a random permutation of $\{1 \dots n\}$. Let S be the number of fixed points of this permutation. What is the expected value and variance of S ?

Answer. Use n indicator random variables $X_i = \mathbf{1}(\sigma(i) = i)$, so that $S = \sum_i X_i$. By linearity of expectation, we can solve the first problem as follows:

$$\mathbb{E}S = \mathbb{E}(X_1 + \dots + X_n) = \sum_{i=1}^n \mathbb{E}X_i = \sum_{i=1}^n P(X_i = 1) = \sum_i \frac{1}{n} = 1.$$

For the second problem, we use $\text{var}(S) = \mathbb{E}(S^2) - (\mathbb{E}S)^2 = \mathbb{E}(S^2) - 1$, and

$$\begin{aligned} \mathbb{E}(S^2) &= \mathbb{E}(X_1 + \dots + X_n)^2 = \mathbb{E}\left(\sum_i X_i^2 + \sum_{i \neq j} X_i X_j\right) \\ &= \sum_i \mathbb{E}X_i^2 + \sum_{i \neq j} \mathbb{E}(X_i X_j) \quad (\text{linearity of expectation}) \\ &= \sum_i \frac{1}{n} + \sum_{i \neq j} \frac{1}{n(n-1)} = 2 \end{aligned}$$

Thus $\text{var}(S) = 1$.

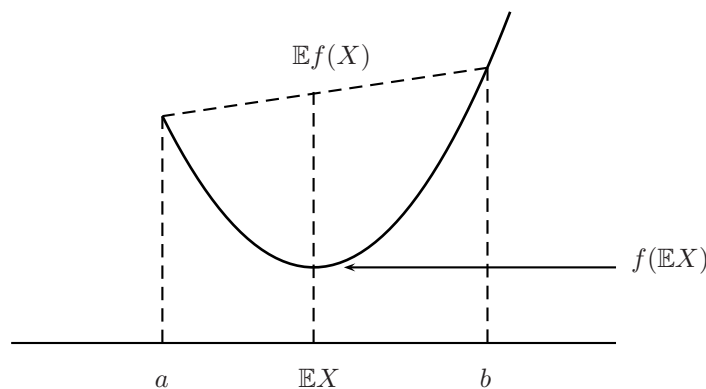
1.2.2 Some basics

Property 5 (Linearity of expectation). $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$ (holds even if X and Y are not independent).

Property 6. $\text{var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$.

Property 7 (Jensen's inequality). If f is a convex function, then $\mathbb{E}f(X) \geq f(\mathbb{E}X)$.

Here's a picture to help you remember this enormously useful property of convex functions:



Lemma 8. *If X_1, \dots, X_n are independent, then*

$$\text{var}(X_1 + \dots + X_n) = \text{var}(X_1) + \dots + \text{var}(X_n).$$

Proof. Let X_1, \dots, X_n be n independent random variables. Set $Y_i = X_i - \mathbb{E}X_i$. Thus Y_1, \dots, Y_n are independent with mean zero, and

$$\begin{aligned} \text{var}(X_1 + \dots + X_n) &= \mathbb{E}[(X_1 - \mathbb{E}X_1) + \dots + (X_n - \mathbb{E}X_n)]^2 \\ &= \mathbb{E}(Y_1 + \dots + Y_n)^2 \\ &= \mathbb{E}\left[\sum_i Y_i^2 + \sum_{i \neq j} Y_i Y_j\right] \\ &= \sum_i \mathbb{E}Y_i^2 + \sum_{i \neq j} \mathbb{E}Y_i \mathbb{E}Y_j \\ &= \sum_i \mathbb{E}Y_i^2 = \sum_i \mathbb{E}(X_i - \mathbb{E}X_i)^2 = \sum_i \text{var}(X_i) \end{aligned}$$

□

As an example of an *incorrect* application, had we mistakenly assumed that the X_i in the warmup problem were independent we would have found that the variance was $\frac{n-1}{n}$ instead of 1. Not too far off, since those X_i are approximately independent (for large n).

Lemma 9 (Markov's inequality).

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

Proof. Observe: $|X| \geq a \cdot 1(|X| \geq a)$; take expectations of both sides, using $\mathbb{E}[1(|X| \geq a)] = \mathbb{P}(|X| \geq a)$. □

Example. A simple application of Markov's inequality to the random variable S , which is always positive, is $\mathbb{P}(S \geq k) \leq 1/k$.

Lemma 10 (Chebyshev's inequality).

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) \leq \frac{\text{var}(X)}{a^2}.$$

Proof. Apply Markov's inequality to $(X - \mathbb{E}X)^2$:

$$\mathbb{P}(|X - \mathbb{E}X| \geq a) = \mathbb{P}((X - \mathbb{E}X)^2 \geq a^2) \leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{a^2} = \frac{\text{var}(X)}{a^2}$$

□

Example. Again, S is a strictly positive random variable, thus $\mathbb{P}(S \geq k) \leq 1/(k-1)^2$. Note that this is generally a better bound than that given by Markov's inequality.

1.2.3 Example: symmetric random walk

A symmetric random walk is a stochastic process on the line. One starts at the origin and at each time step moves either one unit to the left or one unit to the right, with equal probability. The move at time t is thus a random variable X_t , where

$$X_t = \begin{cases} +1 & \text{(right) with probability } 1/2 \\ -1 & \text{(left) with probability } 1/2 \end{cases}$$

Let $S_n = \sum_{i=1}^n X_i$ be the position after n steps of the random walk. What are the expected value and variance of S_n ?

The expected value of X_i is 0 since we are equally likely to obtain $+1$ and -1 , so

$$\mathbb{E}S_n = \mathbb{E} \sum_{i=1}^n X_i = \sum_{i=1}^n \mathbb{E}X_i = 0.$$

Similarly, since the X_i are independent, variance becomes linear as well. The variance of X_i is $\mathbb{E}X_i^2 = 1$, therefore

$$\text{var}(S_n) = \text{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{var}(X_i) = n.$$

The standard deviation of S_n is thus \sqrt{n} ; so we would expect that S_n is $\pm O(\sqrt{n})$. We can make this more precise by using Markov's and Chebyshev's inequalities.

$$\begin{aligned} \text{(Markov)} \quad \mathbb{P}(|S_n| \geq c\sqrt{n}) &\leq \frac{\mathbb{E}|S_n|}{c\sqrt{n}} \leq \frac{\sqrt{\mathbb{E}S_n^2}}{c\sqrt{n}} = \frac{\sqrt{\text{var}(S_n)}}{c\sqrt{n}} = \frac{1}{c} \\ \text{(Chebyshev)} \quad \mathbb{P}(|S_n| \geq c\sqrt{n}) &\leq \frac{\text{var}(S_n)}{(c\sqrt{n})^2} = \frac{1}{c^2} \end{aligned}$$

1.2.4 Moment-generating functions

The Chebyshev inequality is just the Markov inequality applied to X^2 ; this often yields a better bound, as in the case of the symmetric random walk. We could similarly apply Markov's inequality to X^4 , or X^6 , or even higher powers of X . For the symmetric random walk, the bounds would get better and better (they would look like $O(1/c^k)$ for increasing powers of k). The natural culmination of all this is to apply Markov's inequality to e^X (or, for a little flexibility, e^{tX} , where t is a constant we will optimize).

Lemma 11. (*Chernoff's Bounding Method*)

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}e^{tX}}{e^{tc}} \quad \text{for any } t > 0.$$

Proof. Again, we use Markov's inequality,

$$\mathbb{P}(X \geq c) = \mathbb{P}(e^{tX} \geq e^{tc}) \leq \frac{\mathbb{E}e^{tX}}{e^{tc}}.$$

□

Definition 12. *The moment generating function of random variable X is the function $\psi(t) = \mathbb{E}e^{tX}$.*

Example. If X is Gaussian with mean 0 and variance 1,

$$\psi(t) = \int e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = e^{t^2/2}.$$

In general, the value $\mathbb{E}e^{tX}$ may not always be defined. However, if $\mathbb{E}e^{t_0X}$ is defined for some $t_0 > 0$, then:

1. $\mathbb{E}e^{tX}$ is defined for all $|t| < t_0$.
2. All moments of X are finite and $\psi(t)$ has derivatives of all orders at $t = 0$, with

$$\mathbb{E}X^k = \left. \frac{\partial^k \psi}{\partial t^k} \right|_{t=0}.$$

3. $\{\psi(t), |t| \leq t_0\}$ uniquely determines the distribution of X .

1.3 Bounding $\mathbb{E}e^{tX}$

We can compute this expectation directly if we know the distribution of X (simply do an integral), but can we get bounds on it given just some coarse statistics of X ?

Lemma 13. *If $X \in [a, b]$ and X has mean 0, then $\mathbb{E}e^{tX} \leq e^{t^2(b-a)^2/8}$.*

Proof. As shown in Figure 1.3, e^{tx} is a convex function.

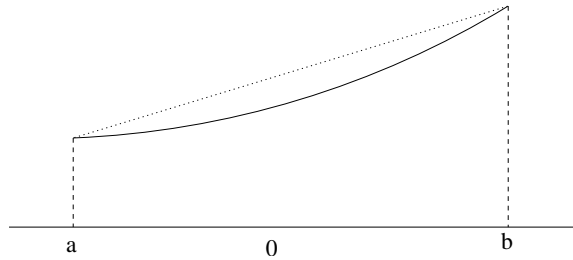


Figure 1.3. e^{tx} is a convex function.

If we write $x = \lambda a + (1 - \lambda)b$ (where $0 \leq \lambda \leq 1$), convexity tells us that

$$e^{tx} \leq \lambda e^{ta} + (1 - \lambda)e^{tb}.$$

Plugging in $\lambda = (b - x)/(b - a)$ then gives

$$e^{tx} \leq \frac{b - x}{b - a} e^{ta} + \frac{x - a}{b - a} e^{tb}$$

Take expectations of both sides, using linearity of expectation and the fact that $\mathbb{E}X = 0$.

$$\mathbb{E}e^{tX} \leq \frac{b - \mathbb{E}X}{b - a} e^{ta} + \frac{\mathbb{E}X - a}{b - a} e^{tb} = \frac{be^{ta} - ae^{tb}}{b - a} \leq e^{t^2(b-a)^2/8}$$

where the last step is just calculus. □

1.4 Hoeffding's Inequality

Theorem 14 (Hoeffding's inequality). *Let X_1, \dots, X_n be independent and bounded with $a_i \leq X_i \leq b_i$. Let $S_n = X_1 + \dots + X_n$. Then for any $\epsilon > 0$,*

$$\begin{aligned}\mathbb{P}(S_n - \mathbb{E}S_n \geq \epsilon) &\leq e^{-2\epsilon^2 / \sum_i (b_i - a_i)^2} \\ \mathbb{P}(S_n - \mathbb{E}S_n \leq -\epsilon) &\leq e^{-2\epsilon^2 / \sum_i (b_i - a_i)^2}\end{aligned}$$

Proof. We'll just do the upper bound (lower bound proof is very similar). Define $Y_i = X_i - \mathbb{E}X_i$; then $\{Y_i\}$ are independent, with mean zero and range $[a_i - \mathbb{E}X_i, b_i - \mathbb{E}X_i]$. For any $t > 0$,

$$\mathbb{P}(S_n - \mathbb{E}S_n \geq \epsilon) = \mathbb{P}(Y_1 + \dots + Y_n \geq \epsilon) = \mathbb{P}(e^{t(Y_1 + \dots + Y_n)} \geq e^{t\epsilon}) \leq \frac{\mathbb{E}e^{t(Y_1 + \dots + Y_n)}}{e^{t\epsilon}}$$

by Chernoff's bounding method. Exploiting the independence of the Y_i 's, and using our generic bound (Lemma 13) for each Y_i , we get

$$\begin{aligned}\mathbb{P}(S_n - \mathbb{E}S_n \geq \epsilon) &\leq \frac{\mathbb{E}e^{tY_1} \mathbb{E}e^{tY_2} \dots \mathbb{E}e^{tY_n}}{e^{t\epsilon}} \\ &\leq \frac{e^{t^2(b_1 - a_1)^2/8} e^{t^2(b_2 - a_2)^2/8} \dots e^{t^2(b_n - a_n)^2/8}}{e^{t\epsilon}} \\ &\leq e^{-2\epsilon^2 / \sum (b_i - a_i)^2}\end{aligned}$$

by choosing $t = 4\epsilon / (\sum (b_i - a_i)^2)$. □

Next: generalize to Lipschitz functions.

1.5 Concentration in metric spaces

1.5.1 Basic definitions

Definition 15. *A metric space (S, d) consists of a set S and a function $d : S \times S \rightarrow \mathbb{R}$ which satisfies three properties.*

1. $d(x, y) \geq 0$, with equality iff $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Example. $(\mathbb{R}^n, l_p\text{-distance})$ is a metric space for any $p \geq 1$.

Definition 16. $f : S \rightarrow \mathbb{R}$ is λ -Lipschitz if $f(x) - f(y) \leq \lambda d(x, y)$ for all $x, y \in S$.

Now suppose that μ is a probability measure on S , and that we want to bound

$$\mu\{f \geq \mathbb{E}f + \epsilon\} = \mathbb{P}_{X \sim \mu}(f(X) \geq \mathbb{E}f + \epsilon).$$

Once again, it would be natural to look at the moment-generating function

$$\mathbb{E}_\mu e^{tf} = \int e^{tf(x)} \mu(dx).$$

But we want a bound that holds for all Lipschitz functions, so we take the supremum of this quantity.

Definition 17. The Laplace functional of metric measure space (S, d, μ) is

$$\mathbb{L}_{(S,d,\mu)}(t) = \sup \mathbb{E}_\mu e^{tf}$$

where the supremum is taken over all 1-Lipschitz functions with mean 0.

1.5.2 Metric spaces of bounded diameter

We start with an analog of Lemma 13.

Lemma 18. If (S, d) has bounded diameter $D = \sup_{x,y \in S} d(x, y) < \infty$, then for any probability measure μ on S ,

$$\mathbb{L}_{(S,d,\mu)}(t) \leq e^{t^2 D^2 / 2}.$$

Proof. First some intuition. Pick any function $f : S \rightarrow \mathbb{R}$ which is 1-Lipschitz and has mean zero. Then certainly $f(x) \leq D$ for all x , and so $\mathbb{E} e^{tf} \leq e^{tD}$. The bound we seek is much tighter than this for small values of t (recall that in Hoeffding's proof we chose $t = O(\epsilon^2)$). To see why it is plausible, let's write out the Taylor expansion of e^{tf} and make an unjustifiable approximation:

$$\mathbb{E} e^{tf} = \mathbb{E} \left[1 + tf + \frac{t^2 f^2}{2} + \frac{t^3 f^3}{3!} + \dots \right] \approx 1 + t\mathbb{E}f + \frac{t^2 \mathbb{E}f^2}{2} \leq 1 + \frac{t^2 D^2}{2} \leq e^{t^2 D^2 / 2}.$$

We've exploited the fact that $\mathbb{E}f = 0$ to eliminate the first term of the series. However, notice that $e^{t^2 D^2 / 2}$ contains all the even powers of t , and so we really need to eliminate all the odd terms in the original Taylor series. When is $\mathbb{E}f^i = 0$ for odd i ? Answer: when the distribution of f is symmetric around zero. Since this might not be the case, we need to explicitly symmetrize f .

Now let's start the real proof. Take any 1-Lipschitz mean-0 function $f : S \rightarrow \mathbb{R}$. First note that by Jensen's inequality,

$$\mathbb{E}_\mu e^{-tf} \geq e^{-t\mathbb{E}_\mu f} = 1.$$

Let X, Y be two independent draws from distribution μ . Then:

$$\mathbb{E}_\mu e^{tf} \leq \mathbb{E}_\mu e^{tf} \mathbb{E}_\mu e^{-tf} = \mathbb{E}_{X \sim \mu} e^{tf(X)} \mathbb{E}_{Y \sim \mu} e^{-tf(Y)} = \mathbb{E}_{X, Y \sim \mu} e^{t(f(X) - f(Y))},$$

which is just what we wanted because $f(X) - f(Y)$ has a symmetric distribution. Thus its odd powers have zero mean:

$$\mathbb{E}_\mu e^{tf} \leq \mathbb{E} \left[\sum_{i=0}^{\infty} \frac{t^i}{i!} (f(X) - f(Y))^i \right] = \sum_{i=0}^{\infty} \frac{t^i}{i!} \mathbb{E} (f(X) - f(Y))^i = \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)!} \mathbb{E} (f(X) - f(Y))^{2i}.$$

Now we use the fact that $f(X) - f(Y) \leq D$, along with the inequality $(2i)! \geq i! \cdot 2^i$, to get

$$\mathbb{E}_\mu e^{tf} \leq \sum_{i=0}^{\infty} \frac{t^{2i} D^{2i}}{(2i)!} \leq \sum_{i=0}^{\infty} \left(\frac{t^2 D^2}{2} \right)^i \frac{1}{i!} = e^{t^2 D^2 / 2},$$

and we're done. \square

In fact, by being a little more careful and using the same technique as in Lemma 13, we can get a slightly better bound.

Lemma 19. Under the same conditions as Lemma 18, $\mathbb{L}_{(S,d,\mu)}(t) \leq e^{t^2 D^2 / 8}$.

We will apply this lemma to individual coordinates, as we did in Hoeffding's proof.

1.5.3 Product spaces

Lemma 20. If (S, d) and (T, δ) are metric spaces so is $(S \times T, d + \delta)$.

Example. $S = T = \mathbb{R}$ and $d(x, y) = |x - y| = \delta(x, y)$. In this case, the metric on the product space is l_1 distance.

Definition 21. If μ is a measure on S and ν is a measure on T , let $\mu \otimes \nu$ denote the product measure on $S \times T$, i.e., which satisfies $(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$ for all measurable $A \subset S, B \subset T$.

Lemma 22. If (S, d, μ) and (T, δ, ν) are metric measure spaces then

$$\mathbb{L}_{(S \times T, d + \delta, \mu \otimes \nu)}(t) \leq \mathbb{L}_{(S, d, \mu)}(t) \mathbb{L}_{(T, \delta, \nu)}(t).$$

Proof. Pick any 1-Lipschitz $f : S \times T \rightarrow \mathbb{R}$ which has mean zero. For any $y \in T$, define $\bar{f}(y) = \mathbb{E}_{X \sim \mu} f(X, y)$. Then \bar{f} has mean zero, over $Y \sim \nu$. Moreover, it is 1-Lipschitz on (T, δ) since for any $y, y' \in T$,

$$\bar{f}(y) - \bar{f}(y') = \mathbb{E}_{X \sim \mu} [f(X, y)] - \mathbb{E}_{X \sim \mu} [f(X, y')] = \mathbb{E}_{X \sim \mu} [f(X, y) - f(X, y')] \leq \delta(y, y')$$

(the last step uses the fact that f is 1-Lipschitz).

Now for any fixed y , the function $f(x, y) - \bar{f}(y)$ is 1-Lipschitz on (S, d) and has mean zero over $X \sim \mu$. Therefore,

$$\begin{aligned} \mathbb{E}_{\mu \otimes \nu} e^{tf} &= \mathbb{E}_{X \sim \mu} \mathbb{E}_{Y \sim \nu} \left[e^{t\bar{f}(Y)} e^{t(f(X, Y) - \bar{f}(Y))} \right] \\ &= \mathbb{E}_{Y \sim \nu} \left[e^{t\bar{f}(Y)} \mathbb{E}_{X \sim \mu} e^{t(f(X, Y) - \bar{f}(Y))} \right] \\ &\leq \mathbb{E}_{Y \sim \nu} \left[e^{t\bar{f}(Y)} \mathbb{L}_{(S, d, \mu)}(t) \right] \\ &\leq \mathbb{L}_{(S, d, \mu)}(t) \mathbb{L}_{(T, \delta, \nu)}(t). \end{aligned}$$

□

Theorem 23. Let $(S_1, d_1, \mu_1), \dots, (S_n, d_n, \mu_n)$ be metric measure spaces of bounded diameters $D_i < \infty$. Let $S = (S_1 \times S_2 \times \dots \times S_n, d_1 + d_2 + \dots + d_n)$ be the product space and $\mu = \mu_1 \otimes \mu_2 \otimes \dots \otimes \mu_n$ the product measure. Then for any 1-Lipschitz function $f : S \rightarrow \mathbb{R}$,

$$\mu \{f \geq \mathbb{E}f + \epsilon\} \leq e^{-2\epsilon^2 / \sum D_i^2}.$$

Proof. Combining Lemmas 19 and 22, we see that $\mathbb{L}_{(S, d, \mu)}(t) \leq e^{(t^2/8)(\sum_i D_i^2)}$. Now it is a simple matter of applying Chernoff's bounding method, using the fact that $f - \mathbb{E}f$ is 1-Lipschitz with mean zero:

$$\mu \{f - \mathbb{E}f \geq \epsilon\} = \mu \left\{ e^{t(f - \mathbb{E}f)} \geq e^{t\epsilon} \right\} \leq \frac{\mathbb{E}_{\mu} e^{t(f - \mathbb{E}f)}}{e^{t\epsilon}} \leq \frac{\mathbb{L}_{(S, d, \mu)}(t)}{e^{t\epsilon}}$$

and the rest is algebra. □

Example. Take $S_i = \mathbb{R}$ and $d_i(x, y) = |x - y|$. Then $S = \mathbb{R}^n$ and $d(x, y) = \|x - y\|_1$. This leads to the following corollary.

Corollary 24. Let X_1, \dots, X_n be independent and bounded with $a_i \leq X_i \leq b_i$. Then for any 1-Lipschitz function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ with respect to the l_1 metric,

$$\mathbb{P}(|f(X_1, \dots, X_n) - \mathbb{E}f| \geq \epsilon) \leq 2e^{-2\epsilon^2 / \sum (b_i - a_i)^2}.$$

Remark. Hoeffding's inequality is a special case of this corollary where $f(x_1, \dots, x_n) = x_1 + \dots + x_n$.