Concentration of Measure

Muchang Bahng

Spring 2024

Contents

1 High-Dimensional Geometry **2**

2 Basic Concentration Inequalities **3**
2.1 Talagrand's Gaussian Inequality 4

3 Variance Bounds and Poincare Inequalities **7**
3.1 Markov Semigroups 15
3.2 Poincare Inequalities 18
3.2.1 The Gaussian Poincare Inequality 19
3.3 Variance Identities and Exponential Ergodicity 23

4 Subgaussian Concentration and log-Sobolev Inequalities **24**
4.1 Subgaussian Variables and Chernoff Bounds 24
4.2 The Martingale Method 28
4.3 The Entropy Method 32
4.4 Modified log-Sobolev Inequalities 34

5 Lipschitz Concentration and Transportation Inequalities **35**
5.1 Concentration in Metric Spaces 35

An informal statement of concentration of measure is the following: *If $X_1, \dots, X_n$ are independent 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 $f(x_1, \dots, x_n)$ is not too "sensitive" to any of the coordinates x_i.* Intuitively, say that we have a bunch of independent random variables X_i and sample from them, to get some values x_i. Calculating $f(x_1, \dots, x_n)$, we have sampled from $f(X_1, \dots, X_n)$. Since f depends smoothly w.r.t. its arguments, to drastically change f, we must drastically change all the arguments. This is not likely, since all the X_i's are independent.

1 High-Dimensional Geometry

Most of our intuition about probability in low-dimensional spaces breaks down in high-dimensional ones (on the order of perhaps 10 or 20). We start off with two geometric examples in high-dimensional space.

Example 1.1 (Uniform Measure on Sphere)

Let μ_n be the uniform probability distribution on the n-sphere $\mathbf{S}^n \subset \mathbb{R}^{n+1}$. That is, let us consider any measurable set $A \subset \mathbb{S}^n$ such that $\mu_n(A) \geq 1/2$. Then, if we let $d(x, A)$ be the geodesic distance between $x \in \mathbb{S}^n$ and A , we define the expanded set

$$A_t = \{x \in \mathbb{S}^n \mid d(x, A) < t\}$$

and it turns out that

$$\mu_n(A_t) \geq 1 - e^{-(n-1)t^2/2}$$

which states that given *any* length $t > 0$, no matter how small, A_t almost covers the whole space. Then, for large enough n, μ_n is highly concentrated around the equator.

Note that the bounds decay *exponentially* (or of greater order).

Example 1.2 (Uniform Measure on Cube)

Example 1.3 (High Dimensional Gaussian)

Given iid $X_1, \dots, X_n \sim \mathcal{N}(0, \sigma^2)$, then let $\mathbf{X}$ be the random n-vector of these random variables. Then, the random variable

$$||\mathbf{X}|| = \sqrt{X_1^2 + \dots, X_n^2}$$

has a distribution that is very concentrated around the expectation

$$\mathbb{E}[||\mathbf{X}||] = \sqrt{\frac{n}{3}}$$

Naturally, this concentration phenomenon extends to random variables.

Example 1.4 ()

Let us have iid random variables X_i with $\mathbb{P}(X_i = 1) = 1/2$ and $\mathbb{P}(X_i = -1) = 1/2$. Then, let's define $S_n = \sum_{i=1}^n X_i$. The strong law of large numbers tell us that

$$\frac{S_n}{n} \xrightarrow{a.s.} 0$$

while the central limit theorem tells us that

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

since $\mathbb{E}[X_i] = 0$ and $\text{Var}[X_i] = 1$. The CLT result shows us that the fluctuations (variance) of S_n of are order n. However, note that $|S_n|$ can take values as large as n, so the maximum value of S_n/n is of order 1. If we measure S_n using this scale, then $\frac{S_n}{n}$ is essentially 0. The actual bound looks like

$$\mathbb{P}\left(\frac{|S_n|}{n} \geq r\right) \leq 2e^{-nr^2/2}$$

2 Basic Concentration Inequalities

Lemma 2.1 (Markov's Inequality)

Given any random variable X, we have

$$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$$

Lemma 2.2 (Chebyshev's Inequality)

Given X with finite variance and expectation, we have

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \alpha) \leq \frac{\text{Var}[X]}{\alpha^2}$$

An inequality that we will use often in proofs is Jensen's inequality.

Lemma 2.3 (Jensen's Inequality)

Given a convex function $g : \mathbb{R} \to \mathbb{R}$ and random variable X, we have

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$$

Proof.

We will assume that f is differentiable for simplicity and let $\mathbb{E}[X] = \mu$. Define the linear function centered at μ to be $l(x) \coloneqq f(\mu) + f'(\mu)(x - \mu)$. Then, we know that $f(x) \geq l(x)$ for all x, so

$$\begin{aligned}
\mathbb{E}[f(X)] &\geq \mathbb{E}[l(X)] \\
&= \mathbb{E}[f(\mu) + f'(\mu)\,(X - \mu)] \\
&= \mathbb{E}[f(\mu)] + f'(\mu)(\mathbb{E}[X] - \mu) \\
&= \mathbb{E}[f(\mu)] \\
&= f(\mathbb{E}[X])
\end{aligned}$$

Definition 2.1 (Lipschitz Continuity)

A function $f : (X, d_X) \longrightarrow (Y, d_Y)$ is **Lipschitz continuous**, with Lipschitz constant A, if it satisfies

$$d_Y\big(f(\mathbf{x}), f(\mathbf{y})\big) \leq A\, d_X(\mathbf{x}, \mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in X$.

2.1 Talagrand's Gaussian Inequality

Lemma 2.4 (Gaussian Integration by Parts Formula)

For Gaussian random variables $x, x_1, \dots, x_n$ and a function F of moderate growth at infinity, we have

$$\mathbb{E}\big[x\, F(x_1, \dots, x_n)\big] = \sum_{i=1}^{n} \mathbb{E}[x\, x_i]\, \mathbb{E}\left[\frac{\partial F}{\partial x_i}(x_1, \dots, x_n)\right]$$

Theorem 2.1 (Talagrand's Gaussian Inequality)

Consider a Lipschitz function $F : \mathbb{R}^N \longrightarrow \mathbb{R}$ (with Lipschitz constant A). Let $x_1, \dots, x_N \sim \mathcal{N}(0, 1)$ be iid, and let $\mathbf{x} = (x_1, \dots, x_N)$. Then, for each $t > 0$, we have

$$\mathbb{P}\big(|F(\mathbf{x}) - \mathbb{E}F(\mathbf{x})| \geq t\big) \leq 2 \exp\left(-\frac{t^2}{4A^2}\right)$$

Proof.

For this proof, we assume that F is not only Lipschitz, but C^2. This is the case in most applications of this theorem, and if it is not the case, then we can regularize F by convolving with a smooth function to solve the problem. We begin with a parameter s and consider the function $G : \mathbb{R}^{2N} \longrightarrow \mathbb{R}$ defined

$$G(z_1, \dots, z_{2N}) = \exp\Big(s\big[F(z_1, \dots, z_N) - F(z_{N+1}, \dots, z_{2N})\big]\Big)$$

For clarity, we will denote variables of F with x_i and variables of G with z_i. Let $u_1, \dots, u_{2N} \sim \mathcal{N}(0, 1)$ be iid, and let $v_1, \dots, v_n \sim \mathcal{N}(0, 1)$ be iid, with $v_{N+1}, \dots, v_{2N}$ copies of the first N. For shorthand, we can denote the collection as $\mathbf{u}$ and $\mathbf{v}$. Then, we have

$$\mathbb{E}[u_i u_j] - \mathbb{E}[v_i v_j] = 0$$

except when $j = i + M$ or $i = j + M$, in which case we have

$$\mathbb{E}[u_i u_j] - \mathbb{E}[v_i v_j] = 0 - 1 = -1$$

since $v_i v_j = X^2$, where $X \sim \mathcal{N}(0, 1) = \chi^2_1$, a Chi-Squared distribution with 1 degree of freedom. We consider the transformed random variable

$$\mathbf{f}(t) \coloneqq \sqrt{t}\, \mathbf{u} + \sqrt{1 - t}\, \mathbf{v} \sim \mathcal{N}(0, 1) \text{ for all } t$$

that is essentially some smooth path from $\mathbf{f}(0) = \mathbf{u}$ and $\mathbf{f}(1) = \mathbf{v}$. Note that given some $t \in [0, 1]$, $\mathbf{f}(t)$ is some random vector, $G(\mathbf{f}(t))$ is some random variable, and $\mathbb{E}[G(\mathbf{f}(t))]$ is some number. We can define the function $\phi : [0, 1] \longrightarrow \mathbb{R}$ as

$$\begin{aligned} \phi(t) = \mathbb{E}[G(\mathbf{f}(t))] &= \int_{\mathbb{R}} x\; p_{G(f(t))}(x)\, dx \\ &= \int_{\mathbb{R}^{2N}} G(y)\; p_{f(t)}(y)\, dy \end{aligned}$$

where p_X is the PDF of the distribution X. Take the derivative with respect to t to get the first line, and we can simplify using Gaussian integration by parts

$$\begin{aligned}
&\phi'(t)\mathbb{E}\Big[\sum_{i=1}^{2N} \frac{d}{dt} f_i(t)\ \frac{\partial G}{\partial z_i}\big(\mathbf{f}(t)\big)\Big] \\
&= \sum_{i=1}^{2N} \mathbb{E}\Big[\frac{d}{dt} f_i(t)\ \frac{\partial G}{\partial z_i}\big(\mathbf{f}(t)\big)\Big] \\
&= \sum_{i=1}^{2N}\sum_{j=1}^{2N} \mathbb{E}\Big[\Big(\frac{\partial}{\partial t} f_i(t)\Big)\ f_i(t)\Big]\ \mathbb{E}\Big[\frac{\partial^2 G}{\partial z_i \partial z_j}\mathbf{f}(t)\Big]
\end{aligned}$$

But we can simplify

$$\begin{aligned}
\mathbb{E}\Big[\Big(\frac{\partial}{\partial t} f_i(t)\Big)\ f_i(t)\Big] &= \mathbb{E}\Big[\Big(\frac{1}{2\sqrt{t}}u_i - \frac{1}{2\sqrt{1-t}}v_i\Big)\big(\sqrt{t}u_j - \sqrt{1-t}\,v_j\big)\Big] \\
&= \frac{1}{2}\big(\mathbb{E}[u_i u_j] - \mathbb{E}[v_i v_j]\big) = \begin{cases} -1 & \text{if } j = i+M, i = j+M \\ 0 & \text{else} \end{cases}
\end{aligned}$$

So, we can simplify the above to

$$\phi'(t) = -\mathbb{E}\Big[\sum_{i=1}^{N} \frac{\partial^2 G}{\partial z_i\, \partial z_{i+M}}\big(\mathbf{f}(t)\big)\Big]$$

and computing the second derivative using the chain rule gives

$$\begin{aligned}
\frac{\partial G}{\partial z_i}(\mathbf{z}) &= \frac{\partial G}{\partial F}\frac{\partial F}{\partial x_i}(z_1, \dots, z_N) \\
&= s\ G(\mathbf{z})\ \frac{\partial F}{\partial x_i}(z_1, \dots, z_N) \\
\frac{\partial^2 G}{\partial z_i \partial z_{i+N}}(\mathbf{z}) &= -s^2\, G(\mathbf{z})\ \frac{\partial F}{\partial x_i}(z_1, \dots, z_N)\ \frac{\partial F}{\partial x_i}(z_{N+1}, \dots, z_{2N})
\end{aligned}$$

for all $\mathbf{z}$. So we have for all $t \in [0,1]$,

$$\begin{aligned}
\phi'(t) &= s^2\, \mathbb{E}\Big[\sum_{i=1}^{N} G(\mathbf{f}(t))\ \frac{\partial F}{\partial x_i}\big(f_1(t), \dots, f_N(t)\big)\ \frac{\partial F}{\partial x_i}\big(f_{N+1}(t), \dots, f_{2N}(t)\big)\Big] \\
&\le s^2 \mathbb{E}\Big[G(\mathbf{f}(t)) \sum_{i=1}^{N} \frac{\partial F}{\partial x_i}\big(f_1(t), \dots, f_N(t)\big)\ \frac{\partial F}{\partial x_i}\big(f_{N+1}(t), \dots, f_{2N}(t)\big)\Big] \\
&\le s^2 \mathbb{E}\big[G(\mathbf{f}(t))\, A^2\big] \\
&\le s^2 A^2 \mathbb{E}[G(\mathbf{f}(t))] = s^2 A^2 \phi(t)
\end{aligned}$$

Solving the inequality for ϕ gives

$$\begin{aligned}
\phi'(t)/\phi(t) \le s^2 A^2 \implies \int \phi'(t)/\phi(t)\, dt &\le \int s^2 A^2\, dt \\
\implies \log \phi(t) &\le s^2 A^2 t + C \\
\implies \phi(t) &\le e^{s^2 A^2 t} \le e^{s^2 A^2}
\end{aligned}$$

Recalling that $\mathbf{f}(1) = \mathbf{u}$, we have

$$\mathbb{E}[\exp\{s(F(u_1, \dots, u_N) - F(u_{N+1}, \dots, u_{2N}))\}] \le e^{s^2 A^2}$$

and by independence of the u_i's, the LHS equals $\mathbb{E}[e^{sF(u_1,\dots,u_N)}]\,\mathbb{E}[e^{-sF(u_{N+1},\dots,u_{2N})}]$ and by Jensen's inequality, we have $\mathbb{E}[e^{-sF(u_{N+1},\dots,u_{2N})}] \geq e^{-s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]}$. We can derive as follows:

$$\begin{aligned}
e^{s^2A^2} &\geq \mathbb{E}[e^{sF(u_1,\dots,u_N)}]\,\mathbb{E}[e^{-sF(u_{N+1},\dots,u_{2N})}] \\
&\geq \mathbb{E}[e^{sF(u_1,\dots,u_N)}]\,e^{-s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]} \\
&= \mathbb{E}[e^{sF(u_1,\dots,u_N)}]\,\mathbb{E}[e^{-s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]}] \\
&= \mathbb{E}[e^{sF(u_1,\dots,u_N)-s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]}] \\
&= \mathbb{E}[\exp\big(sF(u_1,\dots,u_N) - s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]\big)]
\end{aligned}$$

and by Markov's inequality, we get for a random vector of standard Gaussian random variables $\mathbf{x}$

$$\begin{aligned}
\mathbb{P}\big(F(\mathbf{x}) - \mathbb{E}[F(\mathbf{x})] \geq t\big) &= \mathbb{P}\big(e^{s(F(\mathbf{x})-\mathbb{E}[F(\mathbf{x})]} \geq e^{st}\big) \\
&\leq \frac{\mathbb{E}[e^{s(F(\mathbf{x})-\mathbb{E}[F(\mathbf{x})]}]}{e^{st}} \\
&\leq e^{s^2A^2-st} \\
&= e^{-t^2/4A^2} \text{ when } s = t/2A^2
\end{aligned}$$

3 Variance Bounds and Poincare Inequalities

Let us first describe this concentration phenomenon by investigating bounds on the variance

$$\text{Var}[f(x_1, \dots, x_n)] \coloneqq \mathbb{E}\big[\big(f(x_1, \dots, x_n) - \mathbb{E}[f(x_1, \dots, x_n)]\big)^2\big]$$

We can first bound

$$\text{Var}[f(X_1, \dots, X_n)] = \mathbb{E}\big[\big(f(X_1, \dots, X_n)\big)^2\big] - \mathbb{E}\big[f(X_1, \dots, X_n)\big]^2 \leq \mathbb{E}\big[\big(f(X_1, \dots, X_n)\big)^2\big]$$

and since adding a constant term to f doesn't affect the variance, we can utilize this to get our first variance bound.

Lemma 3.1 ()

Let $\mathbf{X}$ be a random variable or vector. Then,

$$\text{Var}[f(\mathbf{X})] \leq \mathbb{E}\big[\big(f(\mathbf{X}) - \inf f\big)^2\big] \text{ and } \text{Var}[f(\mathbf{X})] \leq \mathbb{E}\big[\big(\sup f - f(\mathbf{X})\big)^2\big]$$

and

$$\text{Var}[f(\mathbf{X})] \leq \frac{1}{4}(\sup f - \inf f)^2$$

Proof.

Since $\text{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2$ from above, we have

$$\text{Var}[f(\mathbf{X})] = \text{Var}[f(\mathbf{X}) - a] = \mathbb{E}[(f(\mathbf{X}) - a)^2] - \mathbb{E}[f(\mathbf{X}) - a]^2 \leq \mathbb{E}[(f(\mathbf{X}) - a)^2]$$

By letting $a = \inf f$, we get the first inequality. By letting $a = (\sup f + \inf f)/2$ be the "middle" of f, we have $|f(\mathbf{X}) - a| \leq (\sup f - \inf f)/2 \implies [f(\mathbf{X}) - a]^2 \leq (\sup f - \inf f)^2/4$, and so

$$\text{Var}[f(\mathbf{X})] \leq \mathbb{E}[(f(\mathbf{X}) - a)^2] \leq \frac{1}{4}(\sup f - \inf f)^2$$

which gives our third inequality. We can also see that

$$\text{Var}[f(\mathbf{X})] = \text{Var}[-f(\mathbf{X})] = \text{Var}[b - f(\mathbf{X})] \leq \mathbb{E}[(b - f(\mathbf{X}))^2]$$

to get our second.

This allows us to bound the random vector $f(\mathbf{X})$ if f itself is bounded, no matter what $\mathbf{X}$ is. But this generally turns out to be a very conservative bound, which is unsurprising since we assume so little about $\mathbf{X}$. For example, if we let $X_1, \dots, X_n$ be iid random variables taking values in $[-1, 1]$, and let $f(x_1, \dots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$. Then, f takes values in $[-1, 1]$, and by the previous lemma, we have

$$\text{Var}[f(X_1, \dots, X_n)] \leq \frac{1}{4}(1 - (-1))^2 = 1$$

which looks good, until we see that we can derive a better bound from direct computation (which becomes much better as n increases).

$$\text{Var}[f(X_1, \dots, X_n)] = \frac{1}{n^2}\sum_{i=1}^n \text{Var}[X_i] = \frac{1}{n}$$

However, this computation assumes independence of X_i's, which the previous lemma doesn't. This is the reason we're able to get a better bound, since if we took n copies of the same X, we would have

$$\text{Var}[f(X_1, \dots, X_n)] = \text{Var}[nX/n] = \text{Var}[X] = 1$$

Therefore, we will capitalize on the independence of these random variables in high dimensions to obtain better bounds. Now in the next result, we shall show that the variance of a high dimensional $f(X_1, \dots, X_n)$ can be bounded by the variances of each random variable. Those quantities, like the variance, that behave well in high dimensions is said to *tensorize*.

Consider independent random variables $X_1, \dots, X_n$ and a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. If we fix values $x_1, \dots, x_n$, then we can define for all $k = 1, \dots, n$ the function $g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n) : \mathbb{R} \to \mathbb{R}$ as

$$g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)(z) = f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n)$$

where

$$(g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n))'(z) = \frac{\partial}{\partial x_k} f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n)$$

and $g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)(X_k)$ is a random variable of X_k. Then, we can define

$$\begin{aligned}
\operatorname{Var}_k f(x_1, \dots, x_n) &= \operatorname{Var}_{X_k}[f(x_1, \dots, x_{k-1}, X_k, x_{k+1} \dots, x_n)] \\
&= \mathbb{E}_{X_k}\left[\left(f(x_1, \dots, x_{k-1}, X_k, x_{k+1}, \dots, x_n) - \mathbb{E}_{X_k}[f(x_1, \dots, x_{k-1}, X_k, x_{k+1}, \dots, x_n)]\right)^2\right] \\
&= \operatorname{Var}[g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)(X_k)] \\
&= \operatorname{Var}_{X_k}[g(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)]
\end{aligned}$$

which takes the variance of f with respect to X_k, keeping all other variables fixed. However, this value will change for different $x_1, \dots, x_n$'s, and so we can loosen the restriction that they are fixed. We can take

$$g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)(z) = f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n)$$

where $g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)(X_k)$ is a random variable of $X_1, \dots, X_n$. Now if we calculate its partial variance, we get

$$\begin{aligned}
\operatorname{Var}_k f(X_1, \dots, X_n) &= \operatorname{Var}_{X_k}[f(X_1, \dots, X_k, \dots, X_n)] \\
&= \operatorname{Var}[g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)(X_k)] \\
&= \operatorname{Var}_{X_k}[g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)]
\end{aligned}$$

which is now a random variable of all X_i's, $i \neq k$, that outputs the variance of f with respect to X_k. **But is it true that**

$$\mathbb{E}_{X_k}[f(X_1, \dots, X_n)] = \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n]?$$

Now, we can show a very useful property of variance: that the variance of some arbitrary function can be bounded by the expected sum of the partial variances.

Theorem 3.1 (Tensorization of Variance)

That is, $\operatorname{Var}_i f(\mathbf{x})$ is the variance of $f(X_1, \dots, X_n)$ w.r.t. the variable X_i only, the remaining variables kept fixed. Then, we have

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

Proof.

We try to mimic the fact that the variance of the sum of independent random variables is the sum of the variances. At first sight, the general function $f(x_1, \dots, x_n)$ need not look anything like a sum, but we can expand it as a telescoping sum of random variables. We will prove this using the *martingale method*, which constructs this random variable $f(X_1, \dots, X_n)$ as a sum of finer and finer increments

starting from the "coarse" constant function $\mathbb{E}[f(X_1, \dots, X_n)]$. We define the random variable

$$\Delta_k \coloneqq \mathbf{E}[f(X_1, \dots, X_n) \mid X_1, \dots X_k] - \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}]$$

Then, we can express

$$f(X_1, \dots, X_n) - \mathbf{E}[f(X_1, \dots, X_n)] = \sum_{k=1}^{n} \Delta_k$$

Note that $\mathbb{E}[\Delta_k \mid X_1, \dots, X_{k-1}] = 0$ (i.e. Δ_k's are martingale increments). In particular, even though the Δ_k's are not independent, if we have $l < k$, then

$$\begin{aligned}
\mathbb{E}[\Delta_k \Delta_l] &= \mathbb{E}[\mathbb{E}[\Delta_k \Delta_l \mid X_1, \dots, X_{k-1}]] \\
&= \mathbb{E}[\mathbb{E}[\Delta_k \mid X_1, \dots X_{k-1}] \, \mathbb{E}[\Delta_l \mid X_1, \dots X_{k-1}]] \\
&= \mathbb{E}[\mathbb{E}[\Delta_k \mid X_1, \dots X_{k-1}] \, \Delta_l] \\
&= \mathbb{E}[0 \cdot \Delta_l] = 0
\end{aligned}$$

and so, the variance can be expanded into terms that vanish.

$$\begin{aligned}
\text{Var}[f(X_1, \dots, X_n)] &= \mathbb{E}\big[\big(f(X_1, \dots, X_n) - \mathbf{E}[f(X_1, \dots, X_n)]\big)^2\big] \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{n} \Delta_k\right)^2\right] = \sum_{k=1}^{n} \mathbb{E}[\Delta_k^2]
\end{aligned}$$

Now it remains to show that $\mathbb{E}[\Delta_k^2] \leq \mathbb{E}[\text{Var}_k f(X_1, \dots, X_n)]$ for every k. Let us define

$$\tilde{\Delta}_k = f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n]$$

to be the approximation of $f(X_1, \dots, X_n)$ "one step" before the final increment. Then, we have

$$\Delta_k = \mathbb{E}[\tilde{\Delta}_k \mid X_1, \dots, X_k]$$

and as X_k and $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n$ are independent, we have

$$\text{Var}_k f(X_1, \dots, X_n) = \mathbb{E}[\tilde{\Delta}_k^2 \mid X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n]$$

and therefore using Jensen's inequality we can prove

$$\mathbb{E}[\Delta_k^2] = \mathbb{E}[\mathbb{E}[\tilde{\Delta}_k \mid X_1, \dots, X_k]^2] \leq \mathbb{E}[\tilde{\Delta}_k^2] = \mathbb{E}[\text{Var}_k f(X_1, \dots, X_n)]$$

What we want to eventually do is prove an inequality of the form where for any function $h : \mathbb{R} \to \mathbb{R}$ and some $X \sim \mu$,

$$\text{Var}_\mu[h] = \text{Var}[h(X)] \leq ||\mathcal{L}(h)||^2_{L^2(\mu)}$$

where $\mathcal{L}$ is an operator on h. This will allow us to bound

$$\text{Var}[g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)(X_k)] \leq ||\mathcal{L}(g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n))||^2$$

for all $x_1, \dots, x_n$, simply by taking $h = g(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)$. Since this works for all $x_1, \dots, x_n$, we can claim that this inequality holds for all $X_1(\omega), \dots, X_n(\omega)$ for all $\omega \in \Omega$. That is, we can loosen the fixed values into random variables.

$$\begin{aligned}
\text{Var}_k f(X_1, \dots, X_n) &= \text{Var}[g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)(X_k)] \\
&\leq ||\mathcal{L}(g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n))||^2_{L^2(\mu)}
\end{aligned}$$

Note that all terms are random variables of $X_1, \dots, X_n$, and so the same inequality holds for their expectations over the entire joint measure.

$$\mathbb{E}[\text{Var}_k f(X_1, \dots, X_n)] \leq \mathbb{E}\big[||\mathcal{L}(g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n))||^2_{L^2(\mu)}\big]$$

and so by tensorization (i.e. summing them up), we get

$$\text{Var}[f(X_1, \dots, X_n)] \leq \sum_{i=1}^n \mathbb{E}\big[\text{Var}_i f(X_1, \dots, X_n)\big] \leq \sum_{i=1}^n \mathbb{E}\big[||\mathcal{L}(g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n))||^2_{L^2(\mu)}\big]$$

Furthermore, this bound is sharp when f is linear. Let us demonstrate this by letting $f(x_1, \dots, x_n) = a_1x_1 + \dots + a_nx_n$. On the left hand side, we have

$$\text{Var}[f(X_1, \dots, X_n)] = \text{Var}\left[\sum_{i=1}^n a_iX_i\right] = \sum_{i=1}^n a_i^2 \,\text{Var}[X_i]$$

and on the right hand side, each component divides up to

$$\begin{aligned}\text{Var}_i f(x_1, \dots, x_n) &= \text{Var}[f(x_1, \dots, X_i, \dots, x_n)] \\ &= \text{Var}[a_1x_1 + \dots + a_iX_i + \dots a_nx_n] \\ &= \text{Var}[a_iX_i] \\ &= a_i^2 \,\text{Var}[X_i]\end{aligned}$$

Then? Note that since f is linear, the values of all $x_j, j \neq i$ have no effect on the variance of X_i, and so $\text{Var}_i f(X_1, \dots, X_n)$, which is originally a random variable of $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, is really just the constant (random variable) $a_i^2 \,\text{Var}[X_i]$. This is because no matter what values $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$ are realized, these values will only contribute to a translation of the random variable $f(X_1, \dots, X_n)$, and hence will not affect the variance w.r.t. X_i. So, the right hand side also becomes

$$\mathbb{E}\left[\sum_{i=1}^n \text{Var}_i f(X_1, \dots, X_n)\right] = \mathbb{E}\left[\sum_{i=1}^n a_i^2 \,\text{Var}[X_i]\right] = \sum_{i=1}^n a_i^2 \,\text{Var}[X_i]$$

which is the same as the LHS.

We can view the tensorization of the variance in itself as an expression of the concentration phenomenon. $\text{Var}_i f(\mathbf{x})$ quantifies the sensitivity of the function $f(\mathbf{x})$ of the coordinate x_i in a distribution-dependent manner. If this sensitivity w.r.t. each coordinate ($\mathbb{E}[\text{Var}_i f(X_1, \dots, X_n)]$) is small, then $f(X_1, \dots, X_n)$ is close to its mean. However, it might not be so straightforward to compute $\text{Var}_i f$, since it depends on both the function f and on the distribution of X_i. So, we can try combining this with a suitable bound on the component-wise variance.

Let us define the quantities:

$$D_i f(\mathbf{x}) \coloneqq \sup_z f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n) - \inf_z f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n)$$

and

$$D_i^- f(\mathbf{x}) \coloneqq f(x_1, \dots, x_n) - \inf_z f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n)$$

which quantifies the sensitivity of the function f to the coordinate x_i in a distribution-independent manner. Now we can introduce the following bounds.

Corollary 3.1 ()

We have

$$\text{Var}[f(X_1, \dots, X_n)] \leq \frac{1}{4}\mathbb{E}\left[\sum_{i=1}^n \big(D_i f(X_1, \dots, X_n)\big)^2\right]$$

Proof.

We start off with

$$\begin{aligned}\operatorname{Var}_i f(X_1, \dots, X_n) &= \operatorname{Var}[f(X_1, \dots, X_i, \dots, X_n)] \\ &\leq \frac{1}{4}\big(D_i f(X_1, \dots, X_n)\big)^2\end{aligned}$$

Since these a random variables follow this inequality (for all $\omega \in \Omega$), we can attach an expectation on them to get

$$\mathbb{E}[\operatorname{Var}_i f(X_1, \dots, X_n)] \leq \mathbb{E}\left[\frac{1}{4}\big(D_i f(X_1, \dots, X_n)\big)^2\right]$$

and substituting in the previous theorem gives

$$\begin{aligned}\operatorname{Var}[f(X_1, \dots, X_n)] &\leq \mathbb{E}\left[\sum_{i=1}^n \operatorname{Var}_i f(X_1, \dots, X_n)\right] \\ &= \sum_{i=1}^n \mathbb{E}\big[\operatorname{Var}_i f(X_1, \dots, X_n)\big] \\ &\leq \sum_{i=1}^n \mathbb{E}\left[\frac{1}{4}\big(D_i f(X_1, \dots, X_n)\big)^2\right] \\ &= \frac{1}{4}\mathbb{E}\left[\sum_{i=1}^n \big(D_i f(X_1, \dots, X_n)\big)^2\right]\end{aligned}$$

Example 3.1 (Random Matrices)

Exercise 3.1 (Banach-Valued Sums)

Let $X_1, X_2, \dots, X_N$ be independent random variables with values in a Banach space $(B, \|\cdot\|_B)$. Suppose these random variables are bounded in the sense that $\|X_i\|_B \leq C$ a.s. for every i. Show that

$$\operatorname{Var}\left(\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_B\right) \leq \frac{C^2}{n}$$

This is a simple vector-valued variant of the elementary fact that the variance of $\frac{1}{n}\sum_{k=1}^n X_k$ for real-valued random variables X_k is of order $\frac{1}{n}$.

Solution 3.1

We can tensorize the variance to get

$$\begin{aligned}\operatorname{Var}_k\left\|\frac{1}{n}\sum_{k=1}^n X_k\right\|_B &= \operatorname{Var}\left\|\frac{1}{n}X_k\right\|_B = \frac{1}{n^2}\operatorname{Var}\|X_k\|_B \\ &\leq \frac{1}{n^2}\left(\frac{1}{4}(C-(-C))^2\right) = \frac{C^2}{n^2}\end{aligned}$$

and so letting $f(X_1, \dots, X_n) = \left|\left|\frac{1}{n}\sum_{k=1}^n X_k\right|\right|_B$, we get

$$\begin{aligned}\text{Var}[f(X_1, \dots X_n)] &\leq \sum_{k=1}^n \mathbb{E}[\text{Var}_k\, f(X_1, \dots, X_n)] \\ &\leq \sum_{k=1}^n \frac{C^2}{n^2} = \frac{C^2}{n}\end{aligned}$$

Exercise 3.2 (Rademacher Processes)

Let $\epsilon_1, \dots, \epsilon_n$ be independent symmetric Bernoulli random variables $\mathbb{P}(\epsilon_i = \pm 1) = \frac{1}{2}$ (also called Rademacher variables), let $T \subset \mathbb{R}^n$. The following identity is completely trivial:

$$\sup_{t\in T} \text{Var}\left[\sum_{k=1}^n \epsilon_k t_k\right] = \sup_{t\in T}\sum_{k=1}^n t_k^2$$

Prove the following nontrivial fact:

$$\text{Var}\left[\sup_{t\in T}\sum_{k=1}^n \epsilon_k t_k\right] \leq 4\sup_{t\in T}\sum_{k=1}^n t_k^2$$

Solution 3.2

Let us consider a fixed $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$ and index $i \in [n]$. Then, consider the random variable formed by taking the value $f(\epsilon_1, \dots, \epsilon_n)$ and loosening ϵ_i to be an random variable. That is,

$$\begin{aligned}\mathbb{P}\Big[f(\epsilon_1, \dots, \epsilon_n) = \sup_{t\in T}\{\epsilon_1 t_1 + \dots + 1t_i + \dots + \epsilon_n t_n\}\Big] &= \frac{1}{2} \\ \mathbb{P}\Big[f(\epsilon_1, \dots, \epsilon_n) = \sup_{t\in T}\{\epsilon_1 t_1 + \dots - 1t_i + \dots + \epsilon_n t_n\}\Big] &= \frac{1}{2}\end{aligned}$$

Then, we compute

$$D_i^- f(\epsilon_1, \dots, \epsilon_n) = \inf_{\epsilon_i \in \{-1,1\}} \sup_{t\in T}\sum_{k=1}^n \epsilon_k t_k$$

and we can estimate

$$\begin{aligned}D_i^- f(\boldsymbol{\epsilon}) &= f(\epsilon_1, \dots, \epsilon_n) - D_i f(\epsilon_1, \dots, \epsilon_n) \\ &= \sup_{t\in T}\sum_{k=1}^n \epsilon_k t_k - \inf_{\epsilon_i \in \{-1,1\}} \sup_{t\in T}\sum_{k=1}^n \epsilon_k t_k \\ &\leq \sup_{t\in T} 2|t_i|\end{aligned}$$

We can finally bound

$$\begin{aligned}\text{Var}[f(\epsilon_1, \dots, \epsilon_n)] &\leq \mathbb{E}\left[\sum_{i=1}^n \left(D_i^- f(\boldsymbol{\epsilon})\right)^2\right] \\ &\leq 4\mathbb{E}\left[\sum_{i=1}^n \sup_{t\in T} t_i^2\right] \\ &= 4\sup_{t\in T}\sum_{i=1}^n t_i^2\end{aligned}$$

Exercise 3.3 (Bin Packing)

This is a classical application of bounded difference inequalities. Let $X_1, \dots, X_n$ i.i.d. random variables with values in $[0,1]$. Each X_i represents the size of a package to be shipped. The shipping containers are bins of size 1 (so each bin can hold a set packages whose sizes sum to at most 1). Let $B_n = f(X_1, \dots, X_n)$ be the minimal number of bins needed to store the packages. Note that computing B_n is a hard combinatorial optimization problem, but we can bound its mean and variance by easy arguments.

1. Show that $\text{Var}[B_n] \leq n/4$
2. Show that $\mathbb{E}[B_n] \geq n\mathbb{E}[X_1]$

Thus the fluctuations $\sim \sqrt{n}$ of B_n are much smaller than its magnitude $\sim n$.

Solution 3.3

Listed.

1. Given fixed sizes $X_1, \dots, X_n$ and some $i \in [n]$, we can see that a property of f is that

$$f(X_1, \dots, X_{i-1}, 0, X_{i+1}, \dots, X_n) + 1 = f(X_1, \dots, X_{i-1}, 1, X_{i+1}, \dots, X_n)$$

since for an extra package with size 1, you would for sure need one more bin. So the maximum difference of f based on the x_i value is the constant random variable

$$\begin{aligned} D_i f(X_1, \dots, X_n) &= \sup_{z \in [0,1]} f(X_1, \dots, z, \dots, X_n) - \inf_{z \in [0,1]} f(X_1, \dots, z, \dots, X_n) \\ &= f(X_1, \dots, 1, \dots, X_n) - f(X_1, \dots, 0, \dots, X_n) = 1 \end{aligned}$$

and so by the bounded difference inequalities,

$$\begin{aligned} \text{Var}[B_n] = \text{Var}[f(X_1, \dots, X_n)] &\leq \frac{1}{4}\mathbb{E}\left[\sum_{i=1}^n \big(D_i f(X_1, \dots, X_n)\big)^2\right] \\ &= \frac{1}{4}\sum_{i=1}^n \mathbb{E}\big[\big(D_i f(X_1, \dots, X_n)\big)^2\big] \\ &\leq \frac{n}{4} \end{aligned}$$

2. Given the sizes $X_1, \dots, X_n$, B_n must satisfy

$$B_n = f(X_1, \dots, X_n) \geq X_1 + \dots + X_n$$

since the total volume of bins B_n must exceed the total volume $X_1 + \dots + X_n$ of packages. So,

$$\mathbb{E}[B_n] \geq \mathbb{E}\left[\sum_{k=1}^n X_k\right] = n\mathbb{E}[X_1]$$

Exercise 3.4 (Order Statistics and Spacings)

Let $X_1, \dots, X_n$ be independent random variables, and denote by $X_{(1)} \geq \dots \geq X_{(n)}$ their decreasing rearrangement ($X_{(1)} = \max_i X_i$, $X_{(n)} = \min_i X_i$, etc.). Show that

$$\text{Var}[X_{(k)}] \leq k\,\mathbb{E}[(X_{(k)} - X_{(k+1)})^2] \text{ for } 1 \leq k \leq n/2$$

and that

$$\text{Var}[X_{(k)}] \leq (n-k+1)\,\mathbb{E}[(X_{(k-1)} - X_{(k)})^2] \text{ for } n/2 < k \leq n$$

Exercise 3.5 (Convex Poincare Inequality)

Let $X_1, \dots, X_n$ be independent random variables taking values in $[a, b]$. The bounded difference inequalities estimate the variance $\text{Var}[f(X_1, \dots, X_n)]$ in terms of *discrete* derivatives $D_i f$ or $D_i^- f$ of the function f. The goal of this problem is to show that if the function f is convex, then one can obtain a similar bound in terms of the ordinary notion of derivative $\nabla_i f(x) = \partial f(x)/\partial x_i$ in $\mathbb{R}^n$.

1. Show that if $g : \mathbb{R} \longrightarrow \mathbb{R}$ is convex, then
$$g(y) - g(x) \geq g'(x)\,(y - x) \text{ for all } x, y \in \mathbb{R}$$
2. Show using part (a) and the bounded difference inequalities that if $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then
$$\text{Var}[f(X_1, \dots, X_n)] \geq (b-a)^2 \mathbb{E}[||\nabla f(X_1, \dots, X_n)||^2]$$
3. Conclude that if f is convex and L-Lipschitz, i.e. $|f(x) - f(y)| \leq L||x - y||$ for all $x, y \in [a, b]^n$, then $\text{Var}[f(X_1, \dots, X_n)] \geq L^2(b-a)^2$.

Solution 3.4

Listed.

1. Assuming g is differentiable, let us choose any $x, y \in \mathbb{R}$ and define some $z = \lambda x + (1-\lambda)y$ in between. Then, pictorially, we would like to formally show that
$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(x)}{y - x}$$
and take the limit as $z \to x$ to get $f'(x)$ on the LHS. By definition, we have
$$f(z) = f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$$
Subtracting $f(x)$ and then dividing by $1 - \lambda > 0$ on both sides gives
$$\frac{f(z) - f(x)}{1 - \lambda} \leq f(y) - f(x)$$
Note that $z - x = \lambda x + (1 - \lambda y) - x = (1-\lambda)(y - x)$. So, dividing by $y - x > 0$ on both sides gives
$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(x)}{y - x}$$
and taking the limit on the LHS gives
$$f'(x) = \lim_{z \to x} \frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(x)}{y - x}$$
Since $y - x > 0$, we can multiply both on the same side to get
$$f(y) - f(x) \geq f'(x)\,(y - x)$$
If $y < x$, then the proof is the same, and the inequality sign ends up getting switched around twice, leading to the same conclusion.
2. Note that from the above result, we can multiply both sides by -1 to get that $g(x) - g(y) \leq g'(x)(x - y)$ for all $x, y \in \mathbb{R}$, and then swap the two variables to get $g(y) - g(x) \leq g'(y)(y - x)$. Let us consider fixed $x_1, \dots, x_n$ and some $i \in [n]$. Given $f : \mathbb{R}^n \to \mathbb{R}$, we define $f_i(\mathbf{x}) : \mathbb{R} \to \mathbb{R}$ by unfixing the ith variable. Then, given some $\alpha, \beta \in [a, b]$,
$$f_i(\mathbf{x})(\beta) - f_i(\mathbf{x})(\alpha) \leq g'(\beta)(\beta - \alpha)$$

or equivalently,

$$f(x_1, \dots, \beta, \dots, x_n) - f(x_1, \dots, \alpha, \dots, x_n) \leq \frac{\partial f}{\partial x_i}(x_1, \dots, \beta, \dots, x_n)\,(\beta - \alpha)$$

Now let $z^* \in [a, b]$ be the value s.t.

$$z^* = \arg\min_{z \in [a,b]} f(x_1, \dots, z, \dots, x_n)$$

Then,

$$D_i^- f(\mathbf{x}) = f(x_1, \dots, x_i, \dots x_n) - f(x_1, \dots, z^*, \dots, x_n) \leq \frac{\partial f}{\partial x_i}(x_1, \dots, x_i, \dots, x_n)\,(x_i - z^*)$$

and so

$$\left(D_i^- f(\mathbf{X})\right)^2 \leq \nabla_i f(\mathbf{x})^2\,(x_i - z^*)^2 \leq \nabla_i f(\mathbf{x})^2\,(b-a)^2$$

which gives from the bounded difference inequality

$$\begin{aligned}\text{Var}[f(X_1, \dots, X_n)] &\leq \mathbb{E}\left[\sum_{i=1}^n \left(D_i^- f(X_1, \dots, X_n)\right)^2\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^n \nabla_i f(\mathbf{x})^2\,(b-a)^2\right] \\ &= (b-a)^2 \mathbb{E}\left[\left|\left|\nabla f(\mathbf{X})\right|\right|^2\right]\end{aligned}$$

3. If f is L-lipschitz, then $||\nabla f(\mathbf{X})|| \leq L$, and so

$$\text{Var}[f(X_1, \dots, X_n)] \leq (b-a)^2 L^2$$

3.1 Markov Semigroups

Definition 3.1 (Markov Process)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(S, \mathcal{S})$ be a measurable space. A homogeneous Markov process $\{X_t\}_{t \geq 0}$ is a stochastic process that satisfies the **Markov property**: for every bounded measurable function f and $s, t \geq 0$, there exists a bounded measurable function $P_s f$ satisfying

$$\mathbb{E}[f(X_{t+s}) \mid \{X_r\}_{r \leq t}] = (P_s f)(X_t) = \mathbb{E}[f(X_{t+s}) \mid X_t]$$

Definition 3.2 (Stationary Measure)

A probability measure μ is called **stationary** or **invariant** if

$$\mathbb{E}_\mu[f] = \mathbb{E}_\mu[P_t f] \text{ i.e. } \int_S f\,d\mu = \int_S P_t f\,d\mu$$

for all $t \geq 0$ and bounded measurable f. By abusing notation, this is conventionally written

$$\mu(f) = \mu(P_t f)$$

To interpret this notion, suppose that $X_0 \sim \mu$. Then,

$$\mathbb{E}[f(X_t)] = \mathbb{E}[\mathbb{E}[f(X_t) \mid X_0]] = \mathbb{E}[P_t f(X_0)] = \mathbb{E}_\mu[P_t f]$$

and if μ is stationary, then we have $\mathbb{E}[f(X_t)] = \mathbb{E}_\mu[f]$. If $f = 1_A$ for some measurable $A \subset S$, then $\mathbb{E}[1_A(X_t)] = \mathbb{P}(X_t \in A)$, and

$$\mathbb{P}(X_t \in A) = \mathbb{E}_\mu[1_A] = \int_S 1_A \, d\mu = \int_A d\mu = \mu(A) = \mathbb{P}(X_0 \in A)$$

which means that the probability that for all $A \in \mathcal{S}$ and all $t \geq 0$, the probability of X_t realizing in A is equivalent to the initial probability of X_0 realizing in A. This means that the process remains distributed according to the stationary measure $X_t \sim \mu$ for every time t. In summary, stationary measures describe the equilibrium or steady-state behavior of the Markov process.

From now, given the state space $(S, \mathcal{S})$ we can put a measure μ on it to get a measure space $(S, \mathcal{S}, \mu)$. The Banach space of all μ-measurable functions $f : (S, \mathcal{S}, \mu) \to (\mathbb{R}, \mathcal{R})$ (i.e. for every Borel $B \in \mathcal{R}$, $f^{-1}(B) \in \mathcal{S}$) will be denoted $L^p(\mu)$, equipped with the norm

$$||f||_{L^p(\mu)} \coloneqq \mathbb{E}_\mu[f^p]^{1/p} = \left(\int_S |f|^p \, d\mu \right)^{1/p}$$

If $p = 2$, then we can define the inner product

$$\langle f, g \rangle_\mu \coloneqq \mathbb{E}_\mu[fg] = \int_S fg \, d\mu$$

Lemma 3.2 ()

Let μ be a stationary measure. Then, the following hold for all $p \geq 1$, $t, s \geq 1$, $\alpha, \beta \in \mathbb{R}$, and bounded measurable functions f, g.

1. Contraction:
$$||P_t f||_{L^p(\mu)} \leq ||f||_{L^p(\mu)} = \mathbb{E}_\mu[f^p]^{1/p}$$
2. Linearity:
$$P_t(\alpha f + \beta g) = \alpha P_t f + \beta P_t g$$
3. Semigroup Property:
$$P_{t+s} f = P_t P_s f$$
4. Conservativeness:
$$P_t 1 = 1$$

Lemma 3.3 ()

Let μ be a stationary measure. Then, $t \mapsto \text{Var}_\mu[P_t f]$ is a decreasing function of time for every function $f \in L^2(\mu)$.

Proof.

Note that

$$\begin{aligned} \text{Var}_\mu[P_t f] &= ||P_t f - \mu f||^2_{L^2(\mu)} = ||P_t(f - \mu f)||^2_{L^2(\mu)} = ||P_{t-s} P_s (f - \mu f)||^2_{L^2(\mu)} \\ &\leq ||P_s(f - \mu f)||^2_{L^2(\mu)} = ||P_s f - \mu f||^2_{L^2(\mu)} = \text{Var}_\mu(P_s f) \end{aligned}$$

We now define the analogous operator to the transition rate matrix in discrete time chains with a finite state space.

Definition 3.3 (Generator)

The generator $\mathscr{L}$ is defined as

$$\mathscr{L}f \coloneqq \lim_{t \downarrow 0} \frac{P_t f - f}{t}$$

for every $f \in L^2(\mu)$ for which the above limit exists in $L^2(\mu)$. The set of f for which $\mathscr{L}f$ is defined is called the domain $\text{Dom}(\mathscr{L})$ of the generator, and $\mathscr{L}$ defines a linear operator from $\text{Dom}(\mathscr{L}) \subset L^2(\mu)$ to $L^2(\mu)$.

We have defined the generator $\mathscr{L}$ from the Markov semigroup $\{P_t\}_{t \geq 0}$. Now, let's try to define the semigroup in terms of the generator $\mathscr{L}$. Given that we have some map $\mathscr{L}$), can we define some semigroup $\{P_t\}$ satisfying the definition? To do this, we must solve the differential equation:

$$\frac{d}{dt} P_t = \lim_{\delta \downarrow 0} \frac{P_{t+\delta} - P_t}{\delta} = \lim_{\delta \downarrow 0} \frac{P_t P_\delta - P_t}{\delta} = P_t \lim_{\delta \downarrow 0} \frac{P_\delta - I}{\delta} = P_t \mathscr{L}$$

For function P_t to satisfy this differential equation, we have the solution

$$P_t = e^{t\mathscr{L}}$$

which also implies that $\mathscr{L}$ and P_t must commute.

Definition 3.4 (Reversibility)

The Markov semigroup $\{P_t\}_{t \geq 0}$ with stationary measure μ is called **reversible** if

$$\langle f, P_t g \rangle_\mu = \langle P_t f, g \rangle_\mu$$

for every $f, g \in L^2(\mu)$. Equivalently, we can say that P_t is self-adjoint on $L^2(\mu)$, or since $P_t = e^{t\mathscr{L}}$, we have $\mathscr{L}$ is self-adjoint.

Definition 3.5 (Ergodicity)

The Markov semigroup $\{P_t\}_{t \geq 0}$ with stationary measure μ if called **ergodic** if

$$P_t f \to \mu f$$

in $L^2(\mu)$ as $t \to +\infty$ for every $f \in L^2(\mu)$. Note that $\mu f = \mu(f)$ is the constant function in $L^2(\mu)$.

Exercise 3.6 (Elementary Identities)

Let P_t be a Markov semigroup with generator $\mathscr{L}$ and stationary measure μ. Prove the following elementary facts.

1. Show that $\mu(\mathscr{L}f) = 0$ for every $f \in L^2(\mu)$
2. If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, then $P_t \phi(f) \geq \phi(P_t f)$ when $f, \phi(f) \in L^2(\mu)$
3. If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, then $\mathscr{L}\phi(f) \geq \phi'(f)\mathscr{L}f$ when $f, \phi(f) \in L^2(\mu)$
4. Let $f \in L^2(\mu)$. Show that the following process is a martingale.

$$M_t^f \coloneqq f(X_t) - \int_0^t \mathscr{L}f(X_s)\, ds$$

Solution 3.5

Listed.

1. This is simply a property of the generator. Not worrying about interchanging limits and integrals, we have

$$\begin{aligned}\mu(\mathscr{L}f) = \mathbb{E}_\mu[\mathscr{L}f] &= \int_S \lim_{t\downarrow 0} \frac{P_t f - P_0 f}{t}\, d\mu \\ &= \lim_{t\downarrow 0} \int_S \frac{P_t f - P_0 f}{t}\, d\mu \\ &= \lim_{t\downarrow 0} \frac{1}{t}\big(\mathbb{E}_\mu[P_t f] - \mathbb{E}_\mu[f]\big) = \lim_{t\downarrow 0} \frac{1}{t}\cdot 0 = 0\end{aligned}$$

2. By Jensen's inequality,

$$\begin{aligned}P_s\phi(f) &= \mathbb{E}[\phi(f)(X_{t+s}) \mid X_t] \\ &\geq \phi\bigg(\mathbb{E}[f(X_{t+s} \mid X_t]\big) = \phi(P_s f)\end{aligned}$$

3.2 Poincare Inequalities

Recall that a Poincare inequality for μ is, informally, of the form

$$\text{variance}(f) \leq \mathbb{E}_\mu[||\text{gradient}(f)||^2]$$

At first sight, such an inequality has nothing to do with Markov processes. However, the validity of a Poincare inequality for μ turns out to be related to the rate of convergence of an ergodic Markov process for which μ is the stationary distribution. That is, a measure μ satisfies a Poincare inequality for a certain notion of gradient if and only if an ergodic Markov semigroup associated to this gradient converges exponentially fast to μ.

Definition 3.6 (Dirichlet Form)

Given a Markov process with generator $\mathscr{L}$ and stationary measure μ, the corresponding Dirichlet form is defined as

$$\mathcal{E}(f,g) \coloneqq -\langle f, \mathscr{L}g\rangle_\mu$$

Theorem 3.2 (Poincare Inequality)

Let P_t be a reversible ergodic Markov semigroup with stationary measure μ. The following are equivalent given $c \geq 0$.

1. $\text{Var}_\mu(f) \leq c\mathcal{E}(f,f)$ for all f (Poincare Inequality)
2. $||P_t f - \mu f||_{L^2(\mu)} \leq e^{-t/c}||f - \mu f||_{L^2(\mu)}$
3. $\mathcal{E}(P_t f, P_t f) \leq e^{-2t/c}\mathcal{E}(f,f)$ for all f, t
4. For every f there exists $\kappa(f)$ s.t. $||P_t f - \mu f||_{L^2(\mu)} \leq \kappa(f)e^{-t/c}$
5. For every f there exists $\kappa(f)$ s.t. $\mathcal{E}(P_t f, P_t f) \leq \kappa(f)e^{-2t/c}$

We should view properties 2 through 5 as different notions of exponential convergence of the Markov semigroup P_t to the stationary measure μ. Properties 2 and 4 directly measure the rate of convergence of $P_t f$ to μf in $L^2(\mu)$, while properties 3 and 5 measure the rate of convergence of the "gradient" (now depicted as $\mathcal{E}$) of $P_t f$ to 0.

3.2.1 The Gaussian Poincare Inequality

Definition 3.7 (Ornstein-Uhlenbeck Process)

Given standard Brownian motion $(W_t)_{t\geq 0}$, the **Ornstein-Uhlenbeck process** is defined as

$$X_t = e^{-t}X_0 + e^{-t}W_{e^{2t}-1}$$

Lemma 3.4 (Gaussian Integration by Parts)

If $\xi \sim \mathcal{N}(0,1)$, then

$$\mathbb{E}[\xi f(\xi)] = \mathbb{E}[f'(\xi)]$$

Proof.

Assuming that f is smooth with compact support, we have by integration by parts

$$\begin{aligned}
\mathbb{E}[f'(\xi)] &= \int_{-\infty}^{\infty} f'(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx \\
&= \frac{e^{-x^2/2}}{\sqrt{2\pi}}f(x)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(x)\frac{d}{dx}\left(\frac{e^{-x^2/2}}{\sqrt{2\pi}}\right)dx \\
&= -\int_{-\infty}^{\infty} -xf(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx \\
&= \int_{-\infty}^{\infty} \big(xf(x)\big)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx = \mathbb{E}[\xi f(\xi)]
\end{aligned}$$

Theorem 3.3 ()

The Ornstein-Uhlenbeck Process $(X_t)_{t\geq 0}$

1. is a Markov process with semigroup

$$P_t f(x) = \mathbb{E}\big[f(e^{-t}x + \sqrt{1-e^{-2t}}\xi)\big] \text{ with } \xi \sim \mathcal{N}(0,1)$$

2. admits $\mu = \mathcal{N}(0,1)$ as its stationary measure
3. is ergodic
4. has generator and Dirichlet form given by

$$\mathscr{L}f(x) = -xf'(x) + f''(x), \qquad \mathcal{E}(f,g) = \langle f', g'\rangle_\mu$$

5. is reversible

Proof.

Let $s \geq t$.

1. By definition of X_t, we have $X_t = e^{-t}X_0 + e^{-t}W_{e^{2t-1}}$ and

$$X_s = e^{-s}X_0 + e^{-s}W_{e^{2s}-1} \implies X_0 = (X_s - e^{-s}W_{e^{2s}-1})e^s$$

Substituting in the equation for X_s gives

$$\begin{aligned}
X_t &= e^{-(t-s)}X_s + e^{-t}(W_{e^{2t}-1} - W_{e^{2s}-1}) \\
&= e^{-(t-s)}X_s + \sqrt{1-e^{-2(t-s)}}\xi
\end{aligned}$$

where $\xi = (W_{e^{2t}-1} - W_{e^{2s}-1})/\sqrt{e^{2t}-e^{2s}} \sim N(0,1)$ is independent of $\{X_r\}_{r\leq s}$. Therefore, we can write

$$\mathbb{E}[f(X_t) \mid \{X_r\}_{r\leq s}] = P_{t-s}f(X_s) = \mathbb{E}\big[f\big(e^{-(t-s)}X_s + \sqrt{1-e^{-2(t-s)}}\xi\big)\big]$$

which proves the Markov property and gives the semigroup.

2. We can clearly see that if $X_t \sim N(0,1)$, then $X_{t+s} = e^{-s}X_t + \sqrt{1-e^{-2s}}\xi$ is a sum of Gaussians, one with variance e^{-2s} and the other with variance $1-e^{-2s}$, and so their sum has variance 1.
3. We will take for granted that this is ergodic.
4. To compute the generator, we use the chain rule (and not worry about whether we take the derivative within the expectation integral) and then use Gaussian integration by parts to get

$$\begin{aligned}\frac{d}{dt}P_tf(x) &= \mathbb{E}\left[f'(e^{-t}x + \sqrt{1-e^{-2t}}\xi)\left(\frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\xi - e^{-t}x\right)\right] \\ &= \mathbb{E}\big[e^{-t}xf'(e^{-t}x + \sqrt{1-e^{-2t}}\xi) + e^{-2t}f''(e^{-t}x + \sqrt{1-e^{-2t}}\xi)\big]\end{aligned}$$

and therefore have

$$\frac{d}{dt}P_tf(x) = \left(-x\frac{d}{dx} + \frac{d^2}{dx^2}\right)P_tf(x)$$

The Dirichlet form can be simplified using the Gaussian integration by parts as

$$\begin{aligned}\mathcal{E}(f,g) &= -\langle f, \mathscr{L}g\rangle_\mu \\ &= \mathbb{E}[f(\xi)\big(xg'(\xi) - g''(\xi)\big)] \\ &= \mathbb{E}[\xi f(\xi)g'(\xi)] - \mathbb{E}[f(\xi)g''(\xi)] \\ &= \mathbb{E}[f'(\xi)g'(\xi) + f(\xi)g''(\xi)] - \mathbb{E}[f(\xi)g''(\xi)] \\ &= \mathbb{E}[f'(\xi)g'(\xi)]\end{aligned}$$

5. Since $\mathcal{E}(f,g) = \mathbb{E}[f'(\xi)g'(\xi)]$, it is symmetric and so $\mathscr{L}$ is self-adjoint.

From the previous theorem part 4, we can see that

$$\mathcal{E}(f,f) = \langle f', f'\rangle_\mu = ||f'||^2_{L^2(\mu)} = \mathbb{E}_\mu[f'^2]$$

which means that the Dirichlet form of an Ornstein-Uhlenbeck process is precisely the expected square gradient of function f! Therefore, with the Poincare inequality, we can bound the variance of f with the Dirichlet form, which is the expected square gradient of f.

Theorem 3.4 ()

Let $\mu = \mathcal{N}(0,1)$. Then,

$$\mathrm{Var}_\mu[f] \leq ||f'||^2_{L^2(\mu)}$$

Proof.

We have from the properties of the Ornstein-Uhlenbeck process that

$$\begin{aligned}\frac{d}{dx}P_tf(x) &= \frac{d}{dx}\mathbb{E}[f(e^{-t}x+\sqrt{1-e^{-2t}}\xi)] \\ &= \mathbb{E}\left[\frac{d}{dx}f(e^{-t}x+\sqrt{1-e^{-2t}}\xi)\right] \\ &= \mathbb{E}[f'(e^{-t}x+\sqrt{1-e^{-2t}}\xi)\ e^{-t}] \\ &= e^{-t}\mathbb{E}[f'(e^{-t}x+\sqrt{1-e^{-2t}}\xi)] \\ &= e^{-t}P_tf'(x)\end{aligned}$$

Thus

$$\mathcal{E}(P_tf,P_tf) = ||(P_tf)'||^2_{L^2(\mu)} = e^{-2t}||P_tf'||^2_{L^2(\mu)} \leq e^{-2t}||f'||^2_{L^2(\mu)} = e^{-2t}\mathcal{E}(f,f)$$

where the inequality follows from contraction.

By tensorization, we can prove the following.

Corollary 3.2 (Gaussian Poincare Inequality)

Let $X_1,\dots,X_n \sim N(0,1)$ be iid. Then,

$$\text{Var}[f(X_1,\dots,X_n)] \leq \mathbb{E}[||\nabla f(X_1,\dots,X_n)||^2]$$

Proof.

Computation.

$$\begin{aligned}\text{Var}[f(X_1,\dots,X_n)] &\leq \mathbb{E}\left[\sum_{i=1}^n \text{Var}_i f(X_1,\dots,X_n)\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^n \left|\left|\frac{d}{dx_i}f(X_1,\dots,X_n)\right|\right|^2\right] \\ &= \mathbb{E}[||\nabla f(X_1,\dots,X_n)||^2]\end{aligned}$$

So what have we done so far? If we have some distribution μ and want to prove an inequality that bounds $\text{Var}_\mu[f]$, then we should choose some (reversible ergodic) Markov process that has a stationary distribution μ. We can identify its semigroup, generator, and ultimately its Dirichlet form $\mathcal{E}(f,g)$, which will allow us to invoke the Poincare inequality to bound

$$\text{Var}_\mu[f] \leq c\mathcal{E}(f,f)$$

and since $\mu = N(0,1)$, we have shown above using both the properties of the generator of the Ornstein-Uhlenbeck process and Gaussian integration by parts that this Dirichlet form is precisely the norm of f'. This is clear since the Dirichlet form $\langle f, \mathscr{L}g\rangle_\mu$ only depends on $\mathscr{L}$ and μ. However, the Dirichlet form does not have to be this form.

1. If μ is some other distribution, we would not be able to reduce $\mathcal{E}(f,f)$ to the norm of its derivative, and so it make take on a different form.
2. If we choose a different Markov process, even with the same stationary measure $\mu = N(0,1)$, the generator may be different and so will the Dirichlet form.

Exercise 3.7 (Carre du Champ)

We have interpreted the Dirichlet form $\mathcal{E}(f,f)$ as a general notion of "expected square gradient" that arises in the study of Poincare inequalities. There is an analogous quantity $\Gamma(f,f)$ that plays the role of "square gradient" in this setting (without the expectation). In good probabilistic tradition, it is universally known by its French name carre du champ (literally, "square of the field"). The carre du champ is defined as

$$\Gamma(f,g) \coloneqq \frac{1}{2}\big[\mathscr{L}(fg) - f\mathscr{L}g - g\mathscr{L}f\big]$$

in terms of the generator $\mathscr{L}$ of a Markov process with stationary measure μ.

1. Show that $\mathcal{E}(f,f) = \int \Gamma(f,f)\,d\mu$ and that $\mathcal{E}(f,g) = \int \Gamma(f,g)\,d\mu$ if the Markov process is in addition reversible.
2. Show that $\Gamma(f,f) \geq 0$ so it can indeed by interpreted as a square.
3. Prove the Cauchy-Schwartz inequality $\Gamma(f,g)^2 \leq \Gamma(f,f)\,\Gamma(g,g)$
4. Compute the carre du champ of the Ornstein-Uhlenbeck process and confirm that it should indeed be interpreted as the appropriate notion of "square gradient."

Solution 3.6

Listed.

1. By stationarity, we have
$$\mu(\mathscr{L}f) = \int_S \mathscr{L}f\,d\mu = 0$$
for all $f \in L^2(\mu)$, which reduces the first term below to 0. So, we can reduce the carre du champ to
$$\begin{aligned}\int_S \Gamma(f,f)\,d\mu &= \frac{1}{2}\bigg(\int_S \mathscr{L}(f^2)\,d\mu - 2\int_S f\mathscr{L}f\,d\mu\bigg)\\ &= -\int_S f\mathscr{L}f\,d\mu = -\langle f, \mathscr{L}f\rangle_\mu = \mathcal{E}(f,f)\end{aligned}$$
Furthermore, assuming that P_t is reversible, we have
$$\mathcal{E}(f,g) = -\langle f, \mathscr{L}g\rangle_\mu = -\langle \mathscr{L}f, g\rangle_\mu = -\langle g, \mathscr{L}f\rangle_\mu = \mathcal{E}(g,f)$$
and so
$$\begin{aligned}\int \Gamma(f,g)\,d\mu &= \frac{1}{2}\bigg(\int \mathscr{L}(fg)\,d\mu - \int f\mathscr{L}g\,d\mu - \int g\mathscr{L}f\,d\mu\bigg)\\ &= \frac{1}{2}\big(-\langle f, \mathscr{L}g\rangle_\mu - \langle g, \mathscr{L}f\rangle_\mu\big)\\ &= -\langle f, \mathscr{L}g\rangle_\mu = \mathcal{E}(f,g)\end{aligned}$$
2. Since $\Gamma(f,f) = \frac{1}{2}\big(\mathscr{L}(f^2) - 2f\mathscr{L}f\big)$, the problem now reduces to proving that $\mathscr{L}(f^2) \geq 2f\mathscr{L}f$. By Jensen's inequality, we have $P_t(f^2) \geq (P_tf)^2$, and so
$$\begin{aligned}\mathscr{L}(f^2) &= \lim_{t\downarrow 0}\frac{P_t(f^2) - f^2}{t} \geq \lim_{t\downarrow 0}\frac{(P_tf)^2 - f^2}{t}\\ &= \frac{d}{dt}(P_tf)^2\bigg|_{t=0} = \bigg(2(P_tf)\cdot\frac{d}{dt}(P_tf)\bigg)\bigg|_{t=0} = 2f\mathscr{L}f\end{aligned}$$
3. We know that $\Gamma(f+tg, f+tg) \geq 0$ from above, and so if we expand out, we get
$$\begin{aligned}\Gamma(f+tg, f+tg) &= \frac{1}{2}\Big[\mathscr{L}\big((f+tg)^2\big) - 2(f+tg)\mathscr{L}(f+tg)\Big]\\ &= \Gamma(g,g)t^2 + 2\Gamma(f,g)t + \Gamma(f,f) \geq 0\end{aligned}$$

for all t. Since this quadratic is nonnegative, its discriminant must be ≤ 0, and so

$$\Delta = \big(2\Gamma(f,g)\big)^2 - 2\Gamma(g,g)\Gamma(f,f) \leq 0 \implies \Gamma(f,g)^2 \leq \Gamma(f,f)\Gamma(g,g)$$

4. The generator of the Ornstein-Uhlenbeck process is $\mathscr{L}f(x) = -xf'(x) + f''(x)$. Therefore,

$$\begin{aligned}\Gamma(f,g)(x) &= \frac{1}{2}\big[\mathscr{L}(fg)(x) - f(x)\mathscr{L}g(x) - g(x)\mathscr{L}f(x)\big] \\ &= \frac{1}{2}\Big[\big(-x(fg)'(x) + (fg)''(x)\big) - f(x)\big(-xg'(x) + g''(x)\big) - g(x)\big(-xf'(x) + f''(x)\big)\Big]\end{aligned}$$

which simplifies down to $f'(x)g'(x)$, and so $\Gamma(f,f) = [f'(x)]^2$ can be interpreted as the square gradient of f.

3.3 Variance Identities and Exponential Ergodicity

Now, let us develop some intuition on the connection between Markov semigroups, $\text{Var}_\mu[f]$ and the Dirichlet form $\mathcal{E}(f,f)$.

Lemma 3.5 ()

The following identity holds.

$$\frac{d}{dt}\text{Var}_\mu[P_t f] = -2\mathcal{E}(P_t f, P_t f)$$

Proof.

By stationarity, $\mu(P_t f) = \mu(f)$, and so

$$\begin{aligned}\frac{d}{dt}\text{Var}_\mu[P_t f] &= \frac{d}{dt}\big\{\mu((P_t f)^2) - \mu(P_t f)^2\big\} \\ &= \frac{d}{dt}\big\{\mu((P_t f)^2) - \mu(f)^2\big\} = \frac{d}{dt}\mu((P_t f)^2) \\ &= \frac{d}{dt}\int_S (P_t f)^2\, d\mu = \int_S \frac{d}{dt}(P_t f)^2\, d\mu = 2\int_S (P_t f)\,\frac{d}{dt}P_t f\, d\mu \\ &= 2\mathbb{E}_\mu[P_t f, \mathscr{L}(P_t f)] = 2\langle P_t f, \mathscr{L}P_t f\rangle_\mu = -2\mathcal{E}(P_t f, P_t f)\end{aligned}$$

Theorem 3.5 ()

$\mathcal{E}(f,f) \geq 0$ for every f.

Proof.

We know that $t \mapsto \text{Var}_\mu[P_t f]$ is a decreasing function of t (by contraction of P_t), so

$$\frac{d}{dt}\text{Var}_\mu[P_t f] = -2\mathcal{E}(P_t f, P_t f) \leq 0$$

Theorem 3.6 ()

Suppose that the Markov semigroup is ergodic. Then, we have for every f

$$\text{Var}_\mu[f] = 2\int_0^\infty \mathcal{E}(P_t f, P_t f)\, dt$$

4 Subgaussian Concentration and log-Sobolev Inequalities

4.1 Subgaussian Variables and Chernoff Bounds

We should first consider how one might go about proving that a random variable satisfies a Gaussian tail bound. Most tail bounds in probability theory are proved using some form of Markov's inequality.

Lemma 4.1 (Markov's Inequality)

Given a nonnegative random variable X, we have

$$\mathbb{P}(X > \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$$

which means that the probability that $X > \alpha$ goes down at least as fast as $1/\alpha$.

Markov's inequality is very conservative but very general, too. If we make further assumptions about the random variable X, we can often make stronger bounds. Chebyshev's inequality assumes a (possibly negative) random variable with finite variance and states that the probability will go down as $1/x^2$.

Theorem 4.1 (Chebyshev Inequality)

Given (possibly negative) random variable X, if $\mathbb{E}[X] = \mu < +\infty$ and $\text{Var}(X) = \sigma^2 < +\infty$, then for all $\alpha > 0$,

$$\mathbb{P}\big(|X - \mu| > k\sigma\big) \leq \frac{1}{k^2} \iff \mathbb{P}(|X - \mu| > \alpha) \leq \frac{\text{Var}[X]}{\alpha^2}$$

That is, the probability that X takes a value further than k standard deviations away from μ goes down by $1/k^2$. Therefore, if σ is small, then this bound will be small since there is more concentration in the mean.

Proof.

We apply Markov's inequality to the non-negative random variable $|X - \mu|$.

$$\mathbb{P}(|X - \mu| > \alpha) = \mathbb{P}(|X - \mu|^2 > \alpha^2) \leq \frac{\mathbb{E}(|X - \mu|^2)}{\alpha^2} = \frac{\text{Var}[X]}{\alpha^2}$$

since the numerator on the RHS is the definition of variance.

Using higher powers, we can obtain better and better bounds, but not exponential ones. To obtain these Gaussian tail bounds, we must use more sophisticated methods.

Lemma 4.2 (Chernoff Bound)

Define the log-moment generating function ψ of a random variable X and its Legendre dual ψ^* as

$$\psi_X(\lambda) := \log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] = \mathbb{E}[e^{\lambda X}] - \lambda \mathbb{E}[X] \qquad \psi_X^*(t) = \sup_{\lambda \geq 0}\{\lambda t - \psi_X(\lambda)\}$$

Then, the following is known as the **Chernoff bound**.

$$\mathbb{P}[X - \mathbb{E}[X] \geq t] \leq e^{-\psi_X^*(t)}$$

for all $t \geq 0$. We can lower bound it too with

$$\mathbb{P}[X - \mathbb{E}[X] \leq -t] \leq e^{-\psi_X^*(t)}$$

and union bounding them gives

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\psi_X^*(t)}$$

Proof.

We take some $\lambda \geq 0$ and given that the map $x \mapsto e^{\lambda x}$ is nondecreasing, we can exponentiate and then use Markov's inequality:

$$\mathbb{P}[X - \mathbb{E}[X] \geq t] = \mathbb{P}[e^{\lambda(X - \mathbb{E}[X])} \geq e^{\lambda t}] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] = e^{-(\lambda t - \psi_X(\lambda))} \leq e^{-\psi_X^*(t)}$$

as the left hand does not depend on the choice of λ, we have the additional flexibility of tuning λ to get potentially better bounds. We can also use Chernoff bound on the random variable $-X$ to bound

$$\begin{aligned}
\mathbb{P}(X - \mathbb{E}[X] \leq -t) &= \mathbb{P}(-X - \mathbb{E}[-X] \geq t) \\
&= \mathbb{P}(e^{\lambda(-X + \mathbb{E}[X])} \geq e^{\lambda t}] \\
&\leq e^{-\lambda t}\mathbb{E}[e^{\lambda(-X+\mathbb{E}[X])}] \\
&= e^{-(\lambda t - \psi_{-X}(\lambda))} \leq e^{-\psi_{-X}^*(t)}
\end{aligned}$$

There seems to be a minor problem in the fact that $-\psi_X^*$ and $-\psi_{-X}^*$ are different, and so provide different bounds for the upper and lower tail. But note that $\psi_X(\lambda) = \psi_{-X}(-\lambda)$, and so their maximum will coincide and $\psi_X^*(t) = \psi_{-X}^*(t)$, allowing us to get the union bound.

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\psi^*(t)}$$

To observe how the Chernoff bound can give rise to Gaussian tail bounds, let us first consider the case of an actual Gaussian random variable.

Example 4.1 ()

Let $X \sim N(\mu, \sigma^2)$. Then, $\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] = e^{\lambda^2\sigma^2/2}$, so

$$\psi(\lambda) = \frac{\lambda^2\sigma^2}{2}, \qquad \psi^*(t) = \sup_{\lambda \geq 0}\left\{\lambda t - \frac{\lambda^2\sigma^2}{2}\right\} = \frac{t^2}{2\sigma^2}$$

and by the Chernoff bound, we have $\mathbb{P}(X - \mathbb{E}[X] \geq t] \leq e^{-t^2/2\sigma^2}$.

Note that in order to get the tail bound, the fact that X is Gaussian was not actually important. It would suffice to assume that the log-MGF is bouded from above by a Gaussian.

Definition 4.1 (Subgaussian Random Variables)

A random variable is called σ^2-**subgaussian** if its log-MGF satisfies

$$\psi(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$$

for all $\lambda \in \mathbb{R}$. The constant σ^2 is called the **variance proxy**.

Remember that if $\psi(\lambda)$ is the log-MGF of a random variable X, then $\psi(-\lambda)$ is the log-MGF of the random variable $-X$. For a σ^2-subgaussian random variable X, we can therefore apply the Chernoff bound to both the upper and lower tails and union bound to obtain

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq 2e^{-t/2\sigma^2}$$

We have only worked with Gaussians, which are trivially subgaussian. A nontrivial results is that every bounded random variable is subgaussian.

Lemma 4.3 (Hoeffding's Lemma)

Let $a \leq X \leq b$ a.s. for some $a, b \in \mathbb{R}$. Then,

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

That is, X is $(b-a)^2/4$-subgaussian.

Proof.

We assume without loss of generality that $\mathbb{E}[X] = 0$. Then, we have $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$, and we can compute

$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}, \quad \psi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}\right)^2$$

and thus

$$\psi''(\lambda) = \int_\Omega X^2 \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}\, d\mathbb{P} - \left(\int_\Omega X \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}\, d\mathbb{P}\right)^2$$

can be interpreted as the variance of the random variable X under the twisted probability measure $d\mathbb{Q} = \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}\, d\mathbb{P}$. But $a \leq X \leq b$, so we can bound the variance by its infimum and suprememum $\psi''(\lambda) = \text{Var}_{\mathbb{Q}}[X] \leq (b-a)^2/4$, and the fundamental theorem of calculus yields

$$\psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(\rho)\, d\rho\, d\mu \leq \frac{\lambda^2 (b-a)^2}{8}$$

using $\psi(0) = 0$ and $\psi'(0)$.

Exercise 4.1 (Subgaussian Variables)

There are several different notions of random variables with a Gaussian tail that are all essentialy equivalent up to constants. The aim of this problem is to obtain some insight into these notions.

1. Show that if X is σ^2-subgaussian, then $\text{Var}[X] \leq \sigma^2$.

2. Show that for any increasing and differentiable function Φ,

$$\mathbb{E}[\Phi(|X|)] = \Phi(0) + \int_0^\infty \Phi'(t)\,\mathbb{P}(|X| \geq t)\,dt$$

In the following, we will assume for simplicity that $\mathbb{E}[X] = 0$. We now prove that the following three properties are equivalent for suitable constants σ, b, c: (1) X is σ^2-subgaussian; (2) $\mathbb{P}(|X| \geq t) \leq 2e^{-bt^2}$; and (3) $\mathbb{E}[e^{cX^2}] \leq 2$.

3. Show that if X is σ^2-subgaussian , then $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/2\sigma^2}$
4. Show that if $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/2\sigma^2}$, then $\mathbb{E}[e^{X^2/6\sigma^2}] \leq 2$.
5. Show that if $\mathbb{E}[e^{X^2/6\sigma^2}] \leq 2$, then X is $18\sigma^2$-subgaussian.

In addition, the subgaussian property of X is equivalent to the fact that the moments of X scale as is the case for the Gaussian distribution.

6. Show that if X is σ^2-subgaussian, then $\mathbb{E}[X^{2q}] \leq (4\sigma^2)^q q!$ for all $q \in \mathbb{N}$.
7. Show that if $\mathbb{E}[X^{2q}] \leq (4\sigma^2)^q q!$ for all $q \in \mathbb{N}$, then $\mathbb{E}[e^{X^2/8\sigma^2}] \leq 2$.

Solution 4.1

Listed.

1. We can expand out

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mathbb{E}X}] &= \mathbb{E}\left[1 + \lambda(X - \mathbb{E}X) + \frac{\lambda^2}{2}(X - \mathbb{E}X)^2 + \dots\right] \\ &= 1 + \frac{\lambda^2}{2}\operatorname{Var}[X] + o(\lambda^2) \\ &\leq e^{\lambda^2\sigma^2/2} = 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2)\end{aligned}$$

   which is true for all λ. Setting $\lambda = 0$, we get $\operatorname{Var}[X] \leq \sigma^2$.
2. Unfinished.
3. Since X is σ^2 subgaussian, its log-MGF satisfies $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}] \leq \frac{\lambda^2\sigma^2}{2} \implies -\psi(\lambda) \geq -\frac{\lambda^2\sigma^2}{2}$. Then, its Legendre dual is

$$\psi^*(t) = \sup_{\lambda \geq 0}\{\lambda t - \psi(\lambda)\} \geq \sup_{\lambda \geq 0}\{\lambda t - \frac{\lambda^2\sigma^2}{2}\} = \frac{t^2}{2\sigma^2}$$

   where we optimize the quadratic w.r.t. λ. Therefore, $-\psi^*(t) \leq -\frac{t^2}{2\sigma^2} \implies \mathbb{P}(X \geq t) \leq e^{-\psi^*(t)} \leq e^{-t^2/2\sigma^2}$.
4. By using the identity above with $\Phi(t) = e^{t^2/6\sigma^2}$, we have

$$\begin{aligned}\mathbb{E}[e^{X^2/6\sigma^2}] &= \mathbb{E}[e^{|X|^2/6\sigma^2}] \\ &= e^{0^2/6\sigma^2} + \int_0^\infty e^{t^2/6\sigma^2}\,\frac{t}{3\sigma^2}\mathbb{P}(|X| \geq t)\,dt \\ &\leq 1 + \frac{1}{3t^2}\int_0^\infty te^{t^2/6\sigma^2}\,2e^{-t^2/2\sigma^2}\,dt \\ &= 1 + \frac{2}{3\sigma^2}\int_0^\infty te^{-\frac{1}{3}\frac{t^2}{\sigma^2}}\,dt \\ &= 1 - \frac{1}{\sigma^2}\int_0^\infty \Big(-\frac{2}{3\sigma}t\Big)\,e^{-\frac{t^2}{3\sigma^2}}\,dt \\ &= 1 - e^{-\frac{t^2}{3\sigma^2}}\Big|_0^\infty \\ &= 1 - (0 - 1) = 2\end{aligned}$$

5. Unfinished.
6. We know $X^{2q} = |X|^{2q}$ for all $q \in \mathbb{N}$. By setting $\Phi(t) = t^{2q}$ from the identity above, we can get

$$\mathbb{E}[|X|^{2q}] = 0^{2q} + \int_0^\infty (2q)t^{2q-1}\mathbb{P}(|X| \geq t)\, dt$$

and from (3), we get the first line, where we can just keep doing integration by parts:

$$\begin{aligned}
\mathbb{E}[|X|^{2q}] &\leq \int_0^\infty (2q)t^{2q-1}e^{-t^2/2\sigma^2}\, dt \\
&= 2(4q\sigma^2)\int_0^\infty (2q-2)t^{2q-3}e^{-t^2/2\sigma^2}\, dt \\
&= 2(4q\sigma^2)(4(q-1)\sigma^2)\int_0^\infty (2q-4)t^{2q-5}e^{-t^2/2\sigma^2}\, dt \\
&= \dots \\
&= 2(4q\sigma^2)\dots(4\cdot 2\sigma^2)\int_0^\infty 2te^{-t^2/2\sigma^2}\, dt \\
&= \prod_{k=1}^{q}(4k\sigma^2) = (4\sigma^2)^q q!
\end{aligned}$$

7. We can expand and from the inequality above, we get

$$\begin{aligned}
\mathbb{E}[e^{X^2/8\sigma^2}] = \mathbb{E}\left[1 + \frac{X^2}{8\sigma^2} + \frac{1}{2}\left(\frac{X^2}{8\sigma^2}\right)^2 + \dots\right] \\
&= 1 + \sum_{q=1}^{\infty}\frac{1}{(8\sigma^2)^q q!}\mathbb{E}[X^{2q}] \\
&\leq 1 + \sum_{q=1}^{\infty}\frac{1}{(8\sigma^2)^q q!}(4\sigma^2)^q q! \\
&= 1 + \sum_{q=1}^{\infty}\frac{1}{2^q} = 2
\end{aligned}$$

Exercise 4.2 (Tightness of Hoeffding's Lemma)

Show that the bound on Hoeffding's lemma is the best possible by consider $\mathbb{P}(X = a) = \mathbb{P}(X = b) = \frac{1}{2}$.

Solution 4.2

From computing the expectation

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] = e^{\lambda(a-\frac{a+b}{2})}\mathbb{P}(X=a) + e^{\lambda(b-\frac{a+b}{2})}\mathbb{P}(X=b) = \frac{1}{2}e^{\lambda\frac{a-b}{2}} + \frac{1}{2}e^{\lambda\frac{b-a}{2}}$$

we know that this is always less than $\lambda^2(b-a)^2/8$ for all λ. But setting $\lambda = 0$ satisfies equality.

4.2 The Martingale Method

In this section, we will use the martingale method to derive useful results. Recall that in order to derive some property (like tensorization of variance) of $f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]$, we can expand it as a

telescoping sum of martingale differences

$$f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] = \sum_{k=1}^{n} \Delta_k$$

where

$$\Delta_k = \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k] - \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}]$$

and then deriving bounds on each difference. Note that these are martingale differences because given the filtration $\mathbb{F} = \{\mathcal{F}_k = \sigma(X_1, \dots, X_k)\}$, the stochastic process

$$Y_k = \sum_{i=1}^{k} \Delta_i = \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k] - \mathbb{E}[f(X_1, \dots, X_n)]$$

is a martingale.

Lemma 4.4 (Azuma)

Let $\mathbb{F} = \{\mathcal{F}_k\}_{k \leq n}$ be any filtration, and $\Delta_1, \dots, \Delta_n$ be random variables that satisfy the following properties for $k = 1, \dots, n$.

1. Martingale Difference Property: Δ_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}] = 0$
2. Conditional Subgaussian Property: $\mathbb{E}[e^{\lambda \Delta_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \sigma_k^2 / 2}$ a.s.

Then, the sum $\sum_{k=1}^{n} \Delta_k$ is subgaussian with variance proxy $\sum_{k=1}^{n} \sigma_k^2$.

Proof.

For any $1 \leq k \leq n$, we can compute

$$\mathbb{E}[e^{\lambda \sum_{i=1}^{k} \Delta_i}] = \mathbb{E}[e^{\lambda \sum_{i=1}^{k-1} \Delta_i} \mathbb{E}[e^{\lambda \Delta_k} \mid \mathcal{F}_{k-1}]] \leq e^{\lambda^2 \sigma_k^2 / 2} \, \mathbb{E}[e^{\lambda \sum_{i=1}^{k-1} \Delta_i}]$$

and by induction, this proof is finished. Note that $\mathbb{E}[e^{\lambda \Delta_k} \mid \mathcal{F}_{k-1}] \leq e^{\lambda^2 \sigma_k^2 / 2}$ can only hold if $\mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}] = 0$.

What this lemma basically says is that if we decompose a random variable into martingale differences, and each martingale difference is conditionally subgaussian, then their sum is also subgaussian. Now, if we just assume that each of these martingale differences are bounded, then we can use Hoeffding's lemma on each of them to make them subgaussian, and then use Azuma's lemma to show that their sum is subgaussian. This is exactly what we do here.

Theorem 4.2 (Azuma-Hoeffding Inequality)

Let $\mathbb{F} = \{\mathcal{F}_k\}_{k \leq n}$ be any filtration, and let Δ_k, A_k, B_k satisfy the following properties for $k = 1, \dots, n$.

1. Martingale Difference Property: Δ_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}] = 0$
2. Predictable bounds: A_k, B_k are $\mathcal{F}_{k-1}$-measurable and $A_k \leq \Delta_k \leq B_k$ a.s.

Then, $\sum_{k=1}^{n} \Delta_k$ is subgaussian with variance proxy $\frac{1}{4} \sum_{k=1}^{n} ||B_k - A_k||_\infty^2$. In particular, we obtain for every $t \geq 0$ the tail bound

$$\mathbb{P}\left(\sum_{k=1}^{n} \Delta_k \geq t \right) \leq \exp\left(- \frac{2t^2}{\sum_{k=1}^{n} ||B_k - A_k||_\infty^2} \right)$$

The Azuma-Hoeffding's inequality is often applied in the following setting. Let $X_1, \dots, X_n$ be independent random variables s.t. $a \leq X_i \leq b$ for all i (we can interpret a and b as simply constant random variables). Then, let $\Delta_k = (X_k - \mathbb{E}[X_k])/n$ be martingale differences, which we can show that Δ_k is clearly $\mathcal{F}_k$-measurable and that by independence of X_i's, $\mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}] = \mathbb{E}[\Delta_k] = 0$. Therefore, we can show that its

sum satisfies

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n}\{X_k - \mathbb{E}[X_k]\} \geq t\right) \leq e^{-2nt^2/(b-a)^2}$$

which is consistent with the central limit theorem.

Now we can return to the case of functions $f(X_1, \dots, X_n)$ of independent random variables. Recall that the discrete derivative is defined

$$D_k f(x) = \sup_z f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n) - \inf_z f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n)$$

Theorem 4.3 (McDiarmid)

For $X_1, \dots, X_n$ independent, $f(X_1, \dots, X_n)$ is subgaussian with variance proxy $\frac{1}{4}\sum_{k=1}^{n} ||D_k f||^2$. That is,

$$\mathbb{P}\big[f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq t\big] \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^{n} ||D_k f||_\infty^2}\right)$$

Proof.

We use the martingale method again to write

$$f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] = \sum_{k=1}^{n} \Delta_k$$

where

$$\Delta_k = \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k] - \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}]$$

What we want to do is set some upper and lower bound on $\mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k]$, which will set bounds on Δ_k. We can do this by bounding f by the infimum and supremum w.r.t. each element, getting

$$\begin{aligned}
&\mathbb{E}[\inf_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) \mid X_1, \dots, X_k] \\
&\quad \leq \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k] \\
&\quad\quad \leq \mathbb{E}[\sup_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) \mid X_1, \dots, X_k]
\end{aligned}$$

but by independence of X_k's, we have

$$\mathbb{E}[\inf_z f(X_1, \dots, z, \dots, X_n) \mid X_1, \dots, X_k] = \mathbb{E}[\inf_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) \mid X_1, \dots, X_{k-1}]$$

So, setting

$$\begin{aligned}
A_k &= \mathbb{E}[\inf_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) - f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}] \\
B_k &= \mathbb{E}[\sup_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) - f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}]
\end{aligned}$$

we have $A_k \leq \Delta_k \leq B_k$ for all k, and by Azuma-Hoeffding's inequality along with the fact that $||B_k - A_k|| \leq ||D_k f||_\infty$, we get

$$\mathbb{P}[f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^{n} ||B_k - A_k||_\infty^2}\right) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^{n} ||D_k f||_\infty^2}\right)$$

We should treat McDiarmid's inequality as a subgaussian form of the bounded difference inequality

$$\text{Var}[f(X_1, \dots, X_n)] \leq \frac{1}{4}\mathbb{E}\left[\sum_{k=1}^{n} \big(D_k f(X_1, \dots, X_n)\big)^2\right]$$

The bounded difference inequality says that the variance is controlled by the expectation of the square gradient of the function f. In contrast, McDiarmid's inequality asserts the stronger subgaussian inequality, but under the stronger condition that the variance proxy is controlled by a uniform upper bound on the square gradient rather than its expectation. This will be a recurring theme:

1. the expectation of the square gradient controls the variance
2. a uniform bound on the square gradient controls the subgaussian property

Note that McDiarmid's theorem is not satisfactory. The appropriate notion of a square gradient in both inequalities is the random variable $\sum_{k=1}^n |D_k f|^2$. To control the variance, we want to take its expectation $\mathbb{E}[\sum_{k=1}^n |D_k f|^2]$, and to control the upper bound of the square gradient, we simply want to take its supremum $||\sum_{k=1}^n |D_k f|^2||_\infty$. However, McDiarmid's inequality only yields control in terms of the larger quantity $\sum_{k=1}^n ||D_k f||_\infty^2$ (by triangle inequality), which gets worse in higher dimensions. Rather than taking the supremum of square gradient, we just take the supremum of each (squared) component and add them up, which may be much greater than the actual upper bound. Therefore, the martingale method is far too crude to capture this idea, and we will need new techniques for more refined bounds.

Exercise 4.3 (Bin Packing)

For the Bin packing problem previoulsly, show that the variance bound $\text{Var}[B_n] \leq n/4$ can be strengthened to a Gaussian tail bound

$$\mathbb{P}(|B_n - \mathbb{E}B_n| \geq t) \leq 2e^{-2t^2/n}$$

Solution 4.3

We can see that

$$D_k f(X_1, \dots, X_n) = f(X_1, \dots, X_{k-1}, 1, X_{k+1}, \dots, X_n) - f(X_1, \dots, X_{k-1}, 1, X_{k+1}, \dots, X_n) = 1$$

and by McDiarmid's inequality, we are done.

Exercise 4.4 (Rademacher Processes)

Exercise 4.5 (Sums in Hilbert Space)

Let $X_1, \dots, X_n$ be independent random variables with zero mean that map to a Hilbert space, and suppose that $||X_k|| \leq C$ a.s. for every k.

1. Show that for all $t \geq 0$,

$$\mathbb{P}\left[\left|\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right|\right| \geq \mathbb{E}\left|\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right|\right| + t\right] \leq e^{-nt^2/2C^2}$$

2. Show that

$$\mathbb{E}\left|\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right|\right| \leq Cn^{-1/2}$$

3. Conclude that for all $t \geq Cn^{-1/2}$,

$$\mathbb{P}\left[\left|\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right|\right| \geq t\right] \leq e^{-nt^2/8C^2}$$

4. Finally, argue that for all $t \geq 0$,

$$\mathbb{P}\left[\left|\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right|\right| \geq t\right] \leq e^{-nt^2/8C^2}$$

4.3 The Entropy Method

In order to develop more sophisticated concentration inequalities, let us introduce another term that is used to measure the deviation of a random variable.

Definition 4.2 (Entropy)

The **entropy** of a nonnegative random variable Z is defined

$$\text{Ent}[Z] \coloneqq \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z]$$

Lemma 4.5 (Herbst)

Suppose that random variable X satisfies

$$\text{Ent}[e^{\lambda X}] \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda X}] \text{ for all } \lambda \geq 0$$

Then, X is σ^2-subgaussian. That is,

$$\psi(\lambda) \coloneqq \log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \frac{\lambda^2 \sigma^2}{2} \text{ for all } \lambda \geq 0$$

Proof.

As $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}] - \lambda \mathbb{E}[X]$, we have

$$\frac{d}{d\lambda}\frac{\psi(\lambda)}{\lambda} = \frac{1}{\lambda}\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \frac{1}{\lambda^2}\log \mathbb{E}[e^{\lambda X}] = \frac{1}{\lambda^2}\frac{\text{Ent}[e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \leq \frac{\sigma^2}{2}$$

where the last inequality yields from the assumption. By the fundamental theorem of calculus, we have

$$\frac{\psi(\lambda)}{\lambda} = \lim_{\lambda \downarrow 0} \frac{\psi(\lambda)}{\lambda} + \int_0^\lambda \frac{1}{t^2}\frac{\text{Ent}[e^{tX}]}{\mathbb{E}[e^{tX}]}\, dt \leq \frac{\lambda \sigma^2}{2} \implies \psi(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$$

Exercise 4.6 ()

It turns out that the converse is true up to a constant: If X is $\frac{\sigma^2}{4}$-subgaussian, then

$$\text{Ent}[e^{\lambda X}] \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda X}]$$

Solution 4.4

We know that by Jensen's inequality and concavity of the logarithm,

$$\log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \geq \mathbb{E}[\lambda(X-\mathbb{E}X)] = 0 \implies \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \geq 1$$

Furthermore, note that given $Z = e^{\lambda X}/\mathbb{E}[e^{\lambda X}]$, we have

$$\begin{aligned}
\mathbb{E}[Z \log Z] &= \mathbb{E}\left[\frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]} \log\left(\frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}\right)\right] \\
&= \frac{1}{\mathbb{E}[e^{\lambda X}]}\mathbb{E}\left[e^{\lambda X}\left(\log e^{\lambda X} - \log \mathbb{E}[e^{\lambda X}]\right)\right] \\
&= \frac{1}{\mathbb{E}[e^{\lambda X}]}\mathbb{E}\left[e^{\lambda X}\lambda X - e^{\lambda X}\log \mathbb{E}[e^{\lambda X}]\right] \\
&= \frac{1}{\mathbb{E}[e^{\lambda X}]}\Big(\mathbb{E}[e^{\lambda X}\lambda X] - \mathbb{E}[e^{\lambda X}]\log \mathbb{E}[e^{\lambda X}]\Big) \\
&= \frac{\text{Ent}[e^{\lambda X}]}{\mathbb{E}[e^{[\lambda X}]}
\end{aligned}$$

Since this theorem assumes a bound on $\text{Ent}[e^{\lambda X}]$ rather than $\text{Ent}[X]$, we will mainly be working with the entropy of exponentials of a random variable.

It turns out that entropy behaves very similarly to variance and extends nicely into the subgaussian setting. Just like variance, we define the partial entropy of function $f(x_1, \dots, x_n)$ as

$$\text{Ent}_k\, f(x_1, \dots, x_n) \coloneqq \text{Ent}[f(x_1, \dots, x_{k-1}, X_k, x_{k+1}, \dots, x_n)]$$

That is, $\text{Ent}[f(X_1, \dots, X_n)]$ is the entropy of $f(X_1, \dots, X_n)$ with respect to the variable X_k only, the remaining variables kept fixed.

Theorem 4.4 (Tensorization of Entropy)

Given that $X_1, \dots, X_n$ are independent,

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

Recall that the basic method for deriving Poincare inequalities is that we have some bound on the variance of a single random variable

$$\text{Var}_\mu[g] \leq \mathbb{E}[|\nabla g|^2]$$

and by tensorization, we can take the multivariate function f and derive

$$\text{Var}_\mu[f] \leq \mathbb{E}[||\nabla g||^2]$$

In here, we derive modified log-Sobolev inequalities by bounding the entropy of the form

$$\text{Ent}_\mu[e^g] \leq \mathbb{E}[|\nabla g|^2\, e^g]$$

and then using tensorization to bound

$$\text{Ent}_\mu[e^{\lambda f}] \leq \mathbb{E}[||\nabla(\lambda f)||^2\, e^{\lambda f}]$$

Lemma 4.6 (Discrete Modified log-Sobolev)

Let $D^- f \coloneqq f - \inf f$. Then,

$$\text{Ent}[e^f] \leq \text{Cov}[f, e^f] \leq \mathbb{E}[|D^- f|^2 e^f]$$

Proof.

Note that $\log \mathbb{E}[e^f] \geq \mathbb{E}[f]$ by Jensen's inequality. Therefore,

$$\text{Ent}[e^f] = \mathbb{E}[fe^f] - \mathbb{E}[e^f] \log \mathbb{E}[e^f] \leq \mathbb{E}[fe^f] - \mathbb{E}[f]\mathbb{E}[e^f] = \text{Cov}[f, e^f]$$

To prove the second part, we have

$$\text{Cov}[f, e^f] = \mathbb{E}[(f - \mathbb{E}[f]))(e^f - \mathbb{E}[e^f])] \leq \mathbb{E}[(f - \inf f)(e^f - e^{\inf f})]$$

and since e^x is convex, the first-order condition gives

$$e^{\inf f} \geq e^f + e^f(\inf f - f) \implies e^f - e^{\inf f} \leq e^f(f - \inf f)$$

and substituting above gives the result.

Now, by defining the one-sided differences

$$D_k^- f(x) = f(x_1, \dots, x_n) - \inf_z f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n)$$
$$D_k^+ f(x) = \sup_z f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n) - f(x_1, \dots, x_n)$$

we can use the discrete modified log-Sobolev inequality on each of them and then tensorize to get the following.

Theorem 4.5 (Bounded Difference Inequality)

For all $t \geq 0$,

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

whenever $X_1, \dots, X_n$ are independent. In particular, $f(X_1, \dots, X_n)$ is subgaussian with variance proxy $2||\sum_{k=1}^n |D_k f|^2||_\infty$, where $D_k f = \sup_z f - \inf_z f$.

4.4 Modified log-Sobolev Inequalities

Theorem 4.6 (Modified log-Sobolov Inequality)

Let P_t be a Markov semigroup with stationary measure μ. The following are equivalent:
1. $\text{Ent}_\mu[f] \leq c\mathcal{E}(\log f, f)$ for all f (modified log-Sobolev inequality).
2. $\text{Ent}_\mu[P_t f] \leq e^{-t/c}\,\text{Ent}_\mu[f]$ for all f, t (entropic exponential ergodicity).

Moreover, if $\text{Ent}_\mu[P_t f] \to 0$ as $t \to +\infty$, then

$$\mathcal{E}(\log P_t f, P_t f) \leq e^{-t/c}\mathcal{E}(\log f, f) \text{ for all } f, t$$

implies 1 and 2 above.

5 Lipschitz Concentration and Transportation Inequalities

5.1 Concentration in Metric Spaces

Recall what a Lipschitz function is.

Definition 5.1 (Lipschitz Function)

Let (X, d) be a matrix space. A function $f : X \to \mathbb{R}$ is called L-**Lipschitz** if $|f(x) - f(y)| \leq L\, d(x, y)$ for all $x, y \in X$. The family of all 1-Lipschitz functions is denoted $\text{Lip}(X)$.

Remember that given iid $X_1, \dots, X_n \sim N(0, 1)$, Gaussian concentration states that the random variable is $||||\nabla f||^2||_\infty$-subgaussian. But we can write it in an equivalent way in terms of a Lipschitz property.

Lemma 5.1 ()

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1 function. Then, $||||\nabla f||^2||_\infty \leq L^2$ if and only if f is L-lipschitz.

Therefore, if given random vector $X \sim N(0, I)$, then $f(X)$ is 1-subgaussian for every $f \in \text{Lip}(\mathbb{R}^n, ||\cdot||)$.