Ledoux Minerva Lecture Notes

1 Introduction and Historical Background

1.1 Dvoretzky's Theorem

The following is a theorem in the field of "Asymptotic Geometric Analysis" (think of $\mathbb{R}^n$ where n is large):

Theorem 1. *(Dvoretzky's Theorem)* $\forall \epsilon > 0, \exists \delta(\epsilon) > 0$ *such that for any convex* $K \subset \mathbb{R}^n$*, there exists a subspace* F *of* $\mathbb{R}^n$ *with* $\dim(F) \geq \delta(\epsilon) \log n$ *and an ellipsoid* ξ *in* F *so that* $(1-\epsilon)\xi \subset F \cap K \subset (1+\epsilon)\xi$

This theorem was first proven by Aryeh Drovertzky in the early 1960's. The surprising and interesting connection to probability is the following. Suppose we choose a subspace F at random in a certain way. Then, because of the way that certain functions "concentrate" their mass in high dimensions, with high probability the randomly chosen F will satsifie Dvoretzky's theorem. The following is the concentration result you can use, which we will get to later on:

Theorem 2. *Let* σ_n *be the uniform measure on the n-sphere* $S^n \subset \mathbb{R}^{n+1}$*. Suppose* F *is a 1-Lipshitz function on* S^n*. Let* m *be either the mean or median of* F*. Then for* $r \geq 0$ *there are constants* c *and* C *so that:*

$$\mathbf{P}\left(|F - m| \geq r\right) \leq C \exp\left(-\frac{nr^2}{c}\right)$$

To prove Dvortezky's theorem with this theorem, let F be the norm induced by the convex body K. This approach was first discovered by Vitali Milman in 1971. Since then, this kind of concentration result has found many applications.

1.2 The Three Laws of Classical Probability

Suppose $X : (\Omega, \mathcal{B}, \mathbf{P}) \to \left(\mathbb{R}^d, \|\cdot\|\right)$ is a random variable taking values in $\mathbb{R}^d$ equipped with a norm $\|\cdot\|$. Let $X_1, X_2, \ldots$ be iid copies of X and let $S_n = X_1 + X_2 + \ldots + X_n$. There are the classical three laws of probability:

Theorem 3. *(Strong Law of Large Numbers) If and only if* $\mathbf{E}(X) < \infty$ *we have the following convergence almost surely:*

$$\frac{S_n}{n} \to \mathbf{E}(X) \; a.s.$$

Theorem 4. *(Central Limit Theorem) Suppose now that* $\mathbf{E}(X) = 0$*. If* $\mathbf{E}\left(\|X\|^2\right) < \infty$ *then we have the following convergence in distribution:*

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} G$$

Where G *is a Gaussian random variable with the same covariance structure as* X*. The converse also holds, in the sense that if you have this convergence, then* X *must have a finite second moment.*

Theorem 5. *(Law of the Iterated Logarithm) Suppose* $\mathbf{E}(X) = 0$ *again. If* $\mathbf{E}\left(\|X\|^2\right) < \infty$ *then we have that:*

$$\limsup_{n \to \infty} \frac{\|S_n\|}{\sqrt{n}\sqrt{2\log(\log n)}} < \infty \; a.s$$

The converse is also true, in the sense that if the above $\limsup$ *is finite, then* X *must have a finite second moment.*

How do these laws behaive in infinite dimensions instead of $\mathbb{R}^d$? Suppose that $X : \Omega \to \mathcal{H}$ where $\mathcal{H}$ is a seperable Hilbert space. It turns out that the Strong Law and the C.L.T. carry over to this setting without much change (This goes back to work by Varadhan in the '60s) but the L.I.L. becomes more exotic. Instead of simply having a second moment, you need:

$$\mathbf{E}\left(\frac{\|X\|^2}{\log(\log(3 + \|X\|))}\right) < \infty \text{ and } \mathbf{E}\left(\langle y, X\rangle^2\right) < \infty \forall y \in \mathcal{H}$$

(This is from work by Pisier, Zinn and others.) The first condition is a usual moment condition. However, the second is a type of "weak" moment condition. We will see that weak moments come up naturally in our studies.

1.3 Exponential Inequalities for Sums of Independent Random Variables

Let S be a set and suppose that $X_1, X_2, \ldots$ are random variables taking values in S. The *empirical measure* induced by $X_1, \ldots, X_n$ is the measure $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Typically, when studying empircal measures, we choose a family of functions $\mathcal{F} = \{f : S \to \mathbb{R}\}$ (the choice of $\mathcal{F}$ will depend on our needs) and we consider the expectation of functions in $\mathcal{F}$ under the empircal measure. One thing we typically look for are sharp inequalites for the following quantity:

$$\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(X_i)$$

Below are three exponential inequalities that are useful.

Proposition 6. *(Hoeffding Inequality) Say $X_1, X_2, \ldots, X_n$ are independent real valued random variables. Let $S = X_1 + \ldots + X_n$. Suppose there are constants b_i so that $|X_i| \le b_i$ for each $i = 1, 2, \ldots, n$. Then for all $t > 0$:*

$$\mathbf{P}\left(S \ge \mathbf{E}(S) + t\right) \le \exp\left(\frac{-t^2}{2\sum_{i=1}^n b_i^2}\right)$$

Remark 7. By doing the swap $X \leftrightarrow -X$ and applying Hoeffding Inequality to the X's and then the $-X$'s, you can get the inequality $\mathbf{P}\left(|S - \mathbf{E}(S)| \ge t\right) \le 2\exp\left(\frac{-t^2}{2\sum_{i=1}^n b_i^2}\right)$. One weakness of the Hoeffding inequality is that it depends on n through the sum $\sum_{i=1}^n b_i$, so it may become a very bad estimate as $n \to \infty$. The next inequality improves in this department.

Proposition 8. *(Bernstein Inequality) Say $X_1, X_2, \ldots, X_n$ are independent and $S = X_1 + \ldots + X_n$. Suppose again there are constants b_i so that $|X_i| \le b_i$ for each $i = 1, 2, \ldots, n$. Define:*

$$\begin{aligned} b &:= \max_{i=1}^n b_i \\ v^2 &:= \sum_{i=1}^n \mathbf{E}\left(X_i^2\right) \end{aligned}$$

Then for all $t > 0$:

$$\mathbf{P}\left(S \ge \mathbf{E}(S) + t\right) \le \exp\left(\frac{-t^2}{2(v^2 + bt)}\right)$$

Proposition 9. *(Bennet Inequality) Take X_i's, S, b, and v as above. Let $h(u) = (1+u)\log(1+u) - u$. Then:*

$$\mathbf{P}\left(S \ge \mathbf{E}(S) + t\right) \le \exp\left(-\frac{v^2}{b^2} h\left(\frac{bt}{v^2}\right)\right)$$

Remark 10. Under the simplification $\mathbf{E}(X_i) = 0$, the above inequalites all rely on some form of the orthogonality between independent random variables, namely that

$$\mathbf{E}\left(\left|\sum_{i=1}^n X_i\right|\right) = \sum_{i=1}^n \mathbf{E}\left(|X_i|^2\right)$$

This type of thing will work in a Hilbert space too, but it has no chance to work in a Banach space or other settings that are more general than sums.

For this reason, in his paper "A new look at independence" from Ann. Prob. by Talagrand proposted that instead of focusing on sums $S = X_1 + \ldots + X_n$ we should look at functions $F(X_1, \ldots, X_n)$ which are regular in some sense. This generalization leads us to the heart of concentration of measure. With this new mindset, here is a possible reformulation of Hoeffding's inequality:

Proposition 11. *(Bounded Difference Inequality) Say $X_i : \Omega \to E_i$ with $i = 1, 2, \ldots, n$ and that the X_i's are independent. Let $E = E_1 \times \ldots \times E_n$ and suppose $F : E \to \mathbb{R}$ is a function that satisfies the following bounded difference inequality for some constants $\alpha_1, \alpha_2 \ldots, \alpha_n$:*

$$|F(x_1, \ldots, x_n) - F(y_1, \ldots, y_n)| \le \sum_{i=1}^n \alpha_i \mathbf{1}_{\{x_i \ne y_i\}}$$

(This is saying that a change in the i-th component can yield at most a change of α_i in the value of F). Then with $X = (X_1, \ldots, X_n) \in E$ we have the following inequality:

$$\mathbf{P}\left(F(X) \ge \mathbf{E}\left(F(X)\right) + t\right) \le \exp\left(-\frac{t^2}{2\sum_{i=1}^n \alpha_i^2}\right)$$

Remark 12. The strength of this proof is disguised in the fact that hypothesis has inequality with the first power ofα_i^1, while the conclusion is something involving the second power α_i^2. This is what makes this result non-trivial enough to be interesting. (Of course the power of t and α have to be the same for the inequality to make sense, so this is just saying that its subgaussian in a sense) As we will see in the proof, moving from α_i to α_i^2 is the sneaky part of the proof where the magic happens. This will be related to the following little (totally heuristic)observation. Say W is a mean 0 random variable, and compare the Taylor series of $\mathbf{E}\left(e^W\right)$ and $\mathbf{E}\left(e^{W^2}\right)$. $\mathbf{E}(e^W)$ is like: $1+\mathbf{E}(W)+\frac{1}{2}\mathbf{E}(W^2)+\ldots$ while $\mathbf{E}(e^{W^2})$ is like: $1+\mathbf{E}(W^2)+\ldots$ Since W is mean 0 however, these actually agree to second order! So in some sense $\mathbf{E}\left(e^W\right)$ "looks like" $\mathbf{E}\left(e^{W^2}\right)$. This is the kind of idea that we will use to replace the α_i with α_i^2 in the theorem.

Proof. Let $\mu_i = \mathcal{L}(X_i)$ be the law of X_i, a measure on the space E_i, and let $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ be the measure of $X = (X_1, \ldots, X_n)$ on the space E. Denote $\mathbf{E}_\mu(F) = \int F \mathrm{d}\mu$ by the letter m_F for notational convenience.

Claim: $\mathbf{E}_\mu\left[\exp\left(\lambda\left(F-m_F\right)\right)\right] \leq \exp\left(\lambda^2\left(\sum_{i=1}^n \alpha_i^2\right)/2\right)$

Pf: (By induction on n)

BASE CASE: When $n=1$, $\mu=\mu_1$ and we have that:

$$\begin{aligned}\mathbf{E}_\mu\left[\exp\left(\lambda\left(F-m_F\right)\right)\right] &= \int \exp\left(\lambda\left(F(x_1)-\int F(\tilde{x_1}\right)\mathrm{d}\mu_1(\tilde{x_1})\right)\mathrm{d}\mu_1(x_1)\\ &\leq \int\int \exp\left(\lambda\left(F(x_1)-F(\tilde{x}_1)\right)\right)\mathrm{d}\mu_1\left(\tilde{x}_1\right)\mathrm{d}\mu_1(x_1)\end{aligned}$$

The last inequality is an application of Jensons Inequality using the fact that e^{-x} is convex. Let's write $F = F(x_1)$ and $\tilde{F} = F(\tilde{x}_1)$ for shorthand and examine the integrand here in some detail. By the Taylor series for $e^{\lambda x}$, we have:

$$e^{\lambda(F-\tilde{F})} = 1+\lambda(F-\tilde{F})+\frac{\lambda^2}{2!}\left(F-\tilde{F}\right)^2+\frac{\lambda^3}{3!}\left(F-\tilde{F}\right)^3+\ldots$$

The first non-constant term, $\lambda(F-\tilde{F})$, integrates out to $\mathbf{E}(F)-\mathbf{E}(\tilde{F})$ =0 and vanishes! We can ignore this term then (We'll write a "0" to remind of this term that integrates out to 0). We will also now employ the hypothesis that $F-\tilde{F}\leq\alpha_1$ on the second term to get $\left(F-\tilde{F}\right)^2<\alpha_1^2$. In the same vein, all the *odd* terms will vanish in the expansion, as they are all $\left(F-\tilde{F}\right)^{2k+1} \leq \alpha_1^{2k}\left(F-\tilde{F}\right)$ which integrates to 0. The even terms we contol by $\left(F-\tilde{F}\right)^{2k}<\alpha_1^{2k}$ We remain with:

$$\begin{aligned}e^{\lambda(F-\tilde{F})} &\leq 1+\text{"0"}+\frac{\alpha_1^2\lambda^2}{2!}+\text{"0"}+\frac{\alpha_1^4\lambda^4}{4!}+\ldots\\ &\leq 1+\left(\frac{\alpha_1^2\lambda^2}{2}\right)+\frac{1}{2!}\left(\frac{\alpha_1^2\lambda^2}{2}\right)^2+\ldots+\text{"0"}\\ &= \exp\left(\frac{\alpha_1^2\lambda^2}{2}\right)+\text{"0"}\end{aligned}$$

So after integration over $\mathrm{d}\mu_1(x)\mathrm{d}\mu_1(\tilde{x})$ we get exactly the base case!

INDUCTION STEP: Let $z=(x_2,\ldots,x_n)$, and split up $x=(x_1,z)$. For notational convenience write $m_F(z) = \int_{E_1} F(x_1,z)\mathrm{d}\mu_1(x_1)$. (This is is the average of F when the last $n-1$ components are fixed to z) Have:

$$\begin{aligned}\mathbf{E}_\mu\left[\exp\left(\lambda\left(F-m_F\right)\right)\right] &= \mathbf{E}_{\mu_{2,\ldots,n}}\left[\mathbf{E}_{\mu_1}\left[\exp\left(\lambda\left(F(x_1,z)-m_F(z)\right)\right)\exp\left(\lambda\left(m_F(z)-m_F\right)\right)\right]\right]\\ &= \mathbf{E}_{\mu_{2,\ldots,n}}\left[\mathbf{E}_{\mu_1}\left[\exp\left(\lambda\left(F(x_1,z)-m_F(z)\right)\right)\right]\exp\left(\lambda\left(m_F(z)-m_F\right)\right)\right]\end{aligned}$$

Now, by the proof of the base case, we know that for any value z, $\mathbf{E}_{\mu_1}\left[\exp\left(\lambda\left(F(x_1,z)-m_F(z)\right)\right)\right] \leq \exp\left(\frac{\alpha_1^2\lambda^2}{2}\right)$. (Indeed, one could define $G(\cdot)=F(\cdot,z)$). So we have now:

$$\mathbf{E}_\mu\left[\exp\left(\lambda\left(F-m_F\right)\right)\right] \leq \exp\left(\frac{\alpha_1^2\lambda^2}{2}\right)\mathbf{E}_{\mu_{2,\ldots,n}}\left[\exp\left(\lambda\left(m_F(z)-m_F\right)\right)\right]$$

The induction step applies to the function $m_F(z)$ now! Indeed $m_F(z)$ is a function of the $n-1$ variables $x_2,\ldots,x_n$ whose average is m_F. Hence $\mathbf{E}_{\mu_{2,\ldots,n}}\left[\exp\left(\lambda\left(m_F(z)-m_F\right)\right)\right] \leq \exp\left(\lambda^2\left(\sum_{i=2}^n\alpha_i^2\right)/2\right)$. Plugging in this inequality gives the result!

So the claim is proven by induction. Finally, we use the classic Chebyshev type inequality, $\mathbf{P}(X \geq t) \leq e^{-\lambda t}\mathbf{E}\left(e^{\lambda X}\right)$ which holds for any λ, we get that:

$$\begin{aligned}\mathbf{P}(F - \mathbf{E}(F) \geq t) &\leq e^{-\lambda t}\mathbf{E}(\exp\lambda(F - \mathbf{E}(F))) \\ &\leq \exp\left(-\lambda t + \lambda^2\left(\sum_{i=1}^n \alpha_i^2\right)/2\right)\end{aligned}$$

Choosing $\lambda = \frac{t}{\sum \alpha_i^2}$ to minimize the quadratic gives us finally that $\mathbf{P}(F(X) \geq \mathbf{E}(F(X)) + t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \alpha_i^2}\right)$ as desired. □

Example 13. (Using Bounded Difference inequality on sums: essential the Hoeffding inequality with norms)

Let $F(x_1, \ldots, x_n) = \|\sum_{i=1}^n x_i\|$. If $X_1, \ldots, X_n$ are independent random variables with $\|X_i\| \leq b_i$ a.s. then we satisfy the conditions of the bounded difference inequality with $\alpha_i = 2b_i$ and we have $\mathbf{P}(\|\sum X_i\| \geq \mathbf{E}(\|\sum X_i\|) + t) \leq \exp\left(-\frac{t^2}{8\sum b_i^2}\right)$, which is essentially Hoeffding's inequality.

Example 14. (Bin-Packing: This is an example where you can see that for certain functions F and certain distributions X, the Bounded Difference Inequality is the best possible result)

Imagine we have n objects, all of unit height and with widths $x_1, \ldots, x_n \in [0, 1]$, and that we would like to pack these objects into m bins which are unit squares. We can pack objects $x_1, x_2, \ldots, x_k$ all into a single bin as long as $x_1 + \ldots + x_k \leq 1$. In general we will have some "wasted" space in each bin that is too small to fit any one of our objects. Let $F(x_1, \ldots, x_n)$ be the minimum number of bins needed to pack all our objects, (This is a well known problem in operations research; it turns out computing F is NP hard)

F has the bounded difference criteria because if you replace the the object x_i by an object $\tilde{x}_i$ you will not increase F by more than 1: indeed, you could just give the new object $\tilde{x}_i$ its very own bin. This is exactly the bounded difference criteria with $\alpha_i = 1$. The Bounded difference inequality then tells us that as long as X_i are independent variables in $[0, 1]$ that:

$$\mathbf{P}(F \geq \mathbf{E}(F) + t) \leq \exp\left(-\frac{t^2}{2n}\right)$$

In the particular case where the X_i's are Bernoulli random variables (so $X_i = 1$ with probability p and $X_i = 0$ with probability $1 - p$) then $F(X_1, \ldots, X_n) = \sum X_i$ will be Binomial distributed. The central limit theorem will show us that asymptotically the bounded difference inequality is an equality in this case.

2 The Gaussian model (with examples)

Gaussian measures are a nice family where concentration inequalities work well. We will look at them and some examples. Here is a starting question we will try to answer:

Problem 15. Suppose $X = (X_1, X_2, \ldots)$ are a joint Gaussian family (in the sense that any finite subset is a joint Gaussian). For simplicity, lets assume that each X_k is $X_k \sim N(0, \sigma_k^2)$ but the covariance structure is unknown. Suppose that $\sup_k |X_k| < \infty$ almost surely. What can we say about the integrability properties of X under these conditions (i.e. what moments of X exists and so on?)

Example 16. If the sequence is *finite* i.e. $X = (X_1, \ldots, X_n)$ then we know that $\mathbf{E}\left(e^{\alpha\|X\|^2} < \infty\right)$ for some $\alpha > 0$ and we know that moments of all order exist. (The behavior for finitely many and for a single X_1 is not very different)

2.1 The Gaussian Concentration Inequality

Proposition 17. *(Gaussian Concentration Inequality) Let $d\gamma(x) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{|x|^2}{2}} dx$ be the measure of standard normal $N(0, Id_n)$ random variable. Let $F : \mathbb{R}^n \to \mathbb{R}$ be a Lipshitz function with Lipshitz constant L. Then:*

$$\mathbf{E}_\gamma\left(e^{\lambda F}\right) \leq e^{\lambda \mathbf{E}_\gamma(F) + \lambda^2 L^2/2}$$

Corollary 18. *By using the usual the Chebyshev type inequality, $\mathbf{P}(X \geq t) \leq e^{-\lambda t}\mathbf{E}\left(e^{\lambda X}\right)$ and optimizing over λ we get the inequality:*

$$\mathbf{P}_\gamma(F \geq \mathbf{E}_\gamma(F) + r) \leq \exp\left(-\frac{r^2}{2L^2}\right)$$

Proof. (There are lots of different proofs; we will present the heat flow proof, which is based around the heat kernal)

Let $P_t(x) = \frac{1}{(2\pi t)^{n/2}} \exp\left(-\frac{|x|^2}{2t}\right)$. This is the heat kernal, and it has lots of nice properties. It solve the heat equation and is the semigroup operator for Brownian motion. To review, this is:

$$\begin{aligned}
\partial_t P_t(x) &= \frac{1}{2}\Delta P_t(x) \\
P_0(x) &= \delta(x) \\
P_t h(x) &:= h * P_t(x) \\
&= \int h(x+\sqrt{t}y)\mathrm{d}\gamma(y) \\
&= \mathbf{E}\left(h(B_t)|B_0 = y\right) \\
P_0 f(x) &= f(x) \\
P_1 f(0) &= \mathbf{E}_\gamma(f)
\end{aligned}$$

Now, to tackle our concentration inequality we will assume that F is smooth (you could always convolve it with something to smooth it out a bit) and we will consider the quantity:

$$\varphi(t) = P_t\left(\exp\left(\lambda P_{1-t}F\right)\right)(0)$$

This object is of interest to us because when $t = 0$ we get one piece of the RHS of the inequality we are trying to prove:

$$\begin{aligned}
\varphi(0) &= P_0\left(\exp\left(\lambda P_1 F\right)\right)(0) \\
&= \exp\left(\lambda P_1 F(0)\right) \\
&= \exp\left(\lambda \mathbf{E}_\gamma(F)\right)
\end{aligned}$$

At $t = 1$ we get the LHS of the inequality we are working on:

$$\begin{aligned}
\varphi(1) &= P_1\left(\exp\left(\lambda P_0 F\right)\right)(0) \\
&= \mathbf{E}_\gamma\left(\exp\left(\lambda F\right)\right)
\end{aligned}$$

So indeed, if we can understand how φ is changing from $t = 0$ to $t = 1$ we can understand the inequality. Lets compute $\varphi'(t)$. By chain rule:

$$\begin{aligned}
\varphi'(t) &= \partial_t P_t\left(\exp\left(\lambda P_{1-t}F\right)\right)(0) + P_t\left(\partial_t \exp\left(\lambda P_{1-t}F\right)\right) \\
&= \frac{1}{2}P_t\left(\Delta \exp\left(\lambda P_{1-t}F\right)\right)(0) + (-1)\frac{1}{2}P_t\left(\Delta P_{1-t}\lambda F \exp\left(\lambda P_{1-t}F\right)\right)
\end{aligned}$$

Let $g = \lambda P_{1-t}F$ now, and simplify the above to get:

$$\begin{aligned}
\varphi'(t) &= \frac{1}{2}P_t\left(\Delta(e^g) - \Delta\left(ge^g\right)\right)(0) \\
&= \frac{1}{2}P_t\left(|\nabla g|^2 e^g\right)(0) \text{ (vector calc.)}
\end{aligned}$$

Now, since $P_t$and ∇commute, we have that$|\nabla g|^2 = \lambda^2 P_{1-t}\left(|\nabla F|^2\right) \leq \lambda^2 L^2$ where L is the Lipshitz constant for F. Have then:

$$\begin{aligned}
\varphi'(t) &\leq \frac{1}{2}\lambda^2 L^2 P_t\left(e^{\lambda P_{1-t}F}\right) \\
&= \frac{1}{2}\lambda^2 L^2 \varphi(t)
\end{aligned}$$

Hence, φ must lie below the solution of the ODE $\phi' = \frac{1}{2}\lambda^2 L^2 \phi$. To see this precisely in this set up, we can manipulate the above:

$$\begin{aligned}
\frac{\varphi'(t)}{\varphi(t)} &\leq \frac{1}{2}\lambda^2 L^2 \\
\left(\log(\varphi(t)\right)' &\leq \frac{1}{2}\lambda^2 L^2 \\
\log(\varphi(1)) - \log(\varphi(0)) &\leq 1 \cdot \frac{1}{2}\lambda^2 L^2
\end{aligned}$$

Taking exponential of both sides and recalling that $\varphi(0) = \exp\left(\lambda \mathbf{E}_\gamma(F)\right)$ and that $\varphi(1) = \mathbf{E}_\gamma\left(\exp\left(\lambda F\right)\right)$ are our quantities of interest, we get the desired result. □

Remark 19. There are many different approaches to the above proof. One other way that is similar is to use the Ornstein-Uhlenbeck semigroup, with generator $L = \Delta - x \cdot \nabla$. This has the advantage that as $t \to \infty$ the steady state for the O-U process is a Gaussian, so that $P_t f \to \mathbf{E}_\gamma(f)$. This method can be adapted to other measures μ with $\mathrm{d}\mu(x) = e^{-V(x)}\mathrm{d}x$ when $V(x)$ is a potential that satisfies certain necessary properties.

2.2 Some Examples and Applications

Example 20. Lets us use the Gaussian concentration inequality to answer our previous question. Suppose $(X_1, X_2, \ldots)$ are a joint Gaussian family with $X_k \sim N(0, \sigma_k^2)$ and $\|X\| := \sup_k |X_k| < \infty$ a.s.

Firstly, we see that the second moment of the X_ks must be uniformly bounded. Take m so that $\mathbf{P}(\|X\| \leq m) \geq \frac{1}{2}$. Then $\mathbf{P}(|X_k| \leq m) \geq \frac{1}{2}$ for all $k \geq 1$. Now, since X_k is *Gaussian,* this give an upper bound on the variance of X_k. In particular, one can check that $\mathbf{E}\left(X_k^2\right) \leq 4m$ from this. But then $\mathbf{E}\left(X_k^2\right)$ is uniformly bounded. Let $\sigma^2 := \sup_k \mathbf{E}\left(X_k^2\right) < \infty$ be this uniform bound.

Now, fix any n and consider the first n variables $X = (X_1, X_2, \ldots, X_n)$. Let us denote by $M^T M$ the covariance structure of this Gaussian vector so that $X = MG$ where G is a standard Gaussian with i.i.d $N(0,1)$ entries with $\mathcal{L}(G) = \gamma$. Define now $F(G) = \max_{k \leq n} |(MG)_k| = \max_{k \leq n} |X_k|$. Since M is a matrix, this is a Lipschitz function of G. Moreover the Lipschitz constant for F is the largest entry of the matrix M. Since $M^T M$ is a covariance matrix, its largest entry occurs on the diagonal and is equal to the largest variance and is no more than σ^2. Hence the Lipschitz constant L of F is no more than σ. By the Gaussian concetration inequality then, we have that:

$$\mathbf{P}\left(\max_{k \leq n} |X_k| \geq \mathbf{E}\left(\max_{k \leq n} |X_k|\right) + t\right) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

This bound does not depend on n! On taking $n \to \infty$ (some minor technical details are omitted here), we get:

$$\mathbf{P}(\|X\| \geq \mathbf{E}(\|X\|) + t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right)$$

From which you can get a lot of information about $\|X\|$ and its moments. For example, you can see that $\mathbf{E}\left(e^{\alpha \|X\|^2}\right) < \infty$ for $\alpha < \frac{1}{2\sigma^2}$.

Remark 21. In this example, σ is like a "weak" moment, (it is somehow looking at each X_k individually) while $\mathbf{E}(\|X\|)$ is like a "strong" moment (it looks at all the $X_k's$ together). You need both to be able to tackle the problem in these infinite settings. For example, if the X_k are all iid $N(0,1)$, then $\max_{k \leq n} X_k \approx \sqrt{\log(n)}$ behaves badly as $n \to \infty$ even though $\sigma = 1$.

Example 22. (Relationship to Isoperimetric Inequality)

Fix a set $A \subset \mathbb{R}^n$ and let $F(x) = \mathrm{d}(x, A) = \inf_{y \in A} \mathrm{d}(x, y)$. By the triangle inequality, F is a Lipschitz function with Lipschitz constant 1. Define $A_r = \{x : \mathrm{d}(x, A) < r\} = \{F < r\}$. If we put the Gaussian measure on $\mathbb{R}^n$ and apply the Gaussian concentration inequality we just proved, we get:

$$\mathbf{P}_\gamma(F \geq \mathbf{E}_\gamma(F) + r) \leq \exp\left(-\frac{r^2}{2}\right)$$

We will now play around with this a bit more until we get to the following inequality:

$$\gamma(A_r) \geq 1 - \exp\left(-\frac{\gamma(A)^2 r^2}{2}\right)$$

To do this, fix an r, and put $F_r = \max(F, r)$. Now F_r is still 1-Lipshitz, and we have gained an upper bound on $\mathbf{E}_\gamma(F_r)$, namely:

$$\begin{aligned}
\mathbf{E}_\gamma(F_r) &= \int F_r \mathrm{d}\gamma \\
&= \int_{A^c} F_r \mathrm{d}\gamma \text{ since } F = 0 \text{ on } A \\
&\leq \int_{A^c} r \mathrm{d}\gamma \\
&= r\gamma(A^c) \\
&= r(1 - \gamma(A))
\end{aligned}$$

Using this inequality, we have that $\{F_r < \mathbf{E}_\gamma(F_r) + r\gamma(A)\} \Rightarrow \{F_r < r\} = \{F < r\}$. Hence, by using the Gaussian concentration inequality, we get:

$$\begin{aligned}\mathbf{P}_\gamma(F \geq r) &\leq \mathbf{P}_\gamma(F_r < \mathbf{E}_\gamma(F_r) + r\gamma(A)) \\ &\leq \exp\left(-\frac{\gamma(A)^2 r^2}{2}\right)\end{aligned}$$

Since $\mathbf{P}_\gamma(F \geq r) = 1 - \mathbf{P}_\gamma(F < r) = 1 - \gamma(A_r)$, we get exactly the desired inequality.

Remark 23. It is possible to prove that the half spaces, $H = \{x_1 < c\}$ are the "isoperimetric sets" for the Gaussian measure. That is to say, that if A is any set with $\gamma(A) = \gamma(H)$, then H has at least as much "perimeter" as A, that is:

$$\gamma(A_r) \geq \gamma(H_r) \; \forall r > 0$$

Our above result is getting close to this fact. When $\gamma(A) = \gamma(H) = \frac{1}{2}$ you have $\gamma(H_r) = 1 - \int_r^\infty \exp\left(-\frac{x^2}{2}\right)\frac{\mathrm{d}x}{\sqrt{2\pi}}$ and you can try to make some explicit calculations using Mill's ratio or other tools.

Remark 24. Notice that our inequality did not depend on the dimension n of the space. This gives us hope that we can extend the result to infinite dimensional spaces. For example if we look at the Weiner measure μ on the $C([0,1])$. If we define now $A_r = A + rK$ where K is the unit ball of the Cameron-Martin Thereom, then we will get a similar inequality for $\mu(A_r) \geq 1 - \exp\left(-\frac{1}{2}\mu(A)^2 r^2\right)$. This inequality is one half of the Large Deviation principle in this setting.

Proposition 25. *(Johnson-Lindenstrauss dimension reduction lemma) Suppose $p_1, \ldots, p_N$ are N points in $\mathbb{R}^n$ and $\epsilon > 0$. Choose any k so that $k > \frac{4}{\epsilon^2}\log(N)$. Then there exists a linear map $\pi : \mathbb{R}^n \to \mathbb{R}^k$ such that for every $1 \leq i, j \leq n$ we have:*

$$(1-\epsilon)\left|p_i - p_j\right|_{\mathbb{R}^n} \leq \left|\pi(p_i) - \pi(p_j)\right|_{\mathbb{R}^k} \leq (1+\epsilon)\left|p_i - p_j\right|_{\mathbb{R}^n}$$

Remark 26. The proof we present here using Gaussian concentration is an example of the probabilistic method. We will prove that by choosing the map π is a random way, that there is some positive probability of such a map existsing. Hence, the set of maps π that achieve the result is non-empty.

Proof. (Sketch) Consider an $k \times n$ matrix X which we think of in the usual way as a map $X : \mathbb{R}^n \to \mathbb{R}^k$. Let us choose the entries of X to be i.i.d. $N(0,1)$ Gaussian entries. Now, fix any $u \in \mathbb{R}^n$ and condier the map $F_u : \mathbb{R}^{nk} \to \mathbb{R}$ by $F_u(X) = |Xu|$ .It is easily verified that F is Lipschitz with Lipschitz constant $L \leq |u|$. By using the two-sided Gaussian concentration inequality, we have that

$$\mathbf{P}\left(|F - \mathbf{E}F| \geq r|u|\right) \leq 2\exp\left(-\frac{r^2}{2}\right)$$

Now, a bit of analysis (which we will skip here) is required to show that $\mathbf{E}F \approx \sqrt{k}|u|$. Once this is established, we get that:

$$\mathbf{P}\left(\left||Xu| - \sqrt{k}u\right| \geq r|u|\right) \leq 2\exp\left(-\frac{r^2}{2}\right)$$

Choose $r = \sqrt{k}\epsilon$ to get:

$$\mathbf{P}\left(\frac{|Xu|}{\sqrt{k}} \notin \left[(1-\epsilon)|u|, (1+\epsilon)|u|\right]\right) \leq 2\exp\left(-\frac{k\epsilon^2}{2}\right)$$

Now, since this argument holds for every vector u, let $\ell = \binom{N}{2}$ and let $u_1, \ldots, u_\ell$ enumerate all the differences $p_i - p_j$ with $1 \leq i, j \leq N$. Now, by applying the result to the list $u_1, \ldots, u_\ell$ and applying a union bound, we have:

$$\mathbf{P}\left(\exists i \text{ s.t. } \frac{|Xu_i|}{\sqrt{k}} \notin \left[(1-\epsilon)|u_i|, (1+\epsilon)|p_i|\right]\right) \leq 2\ell\exp\left(-\frac{k\epsilon^2}{2}\right)$$

Since $k > 4\epsilon^{-2}\log(N) > c4\epsilon\log\left((2\ell)^{\frac{1}{2}}\right) = 2\epsilon\log(2\ell)$ and we have:

$$\begin{aligned}\mathbf{P}\left(\exists i \text{ s.t. } \frac{|Xu_i|}{\sqrt{k}} \notin \ldots\ldots\ldots\right) &< 2\ell\exp\left(-2\epsilon^{-2}\log(2\ell)\frac{\epsilon^2}{2}\right) \\ &= 2\ell(2\ell)^{-1} = 1\end{aligned}$$

So taking complements, we have that:

$$\mathbf{P}\left(\forall i, \frac{|Xu_i|}{\sqrt{k}} \in \left[(1-\epsilon)|u_i|, (1+\epsilon)|u_i|\right]\right) > 0$$

Since this probability is non-negative, we know that there is at least one point ω in this set. Choosing the transform $\pi = \frac{X(\omega)}{\sqrt{k}}$ then gives us the desired property we want our transform to have. □

2.3 Variance Bounds

We will now look at some variance bounds, which is a type of concentration at the level of the L^2 norm. If we have sub-gaussian concentration of the type we have been studying so far, then we get a variance bound for free.

Example 27. Let $X = (X_1, X_2, \ldots, X_n)$ be a centered Gaussian vector $\mathbf{E}(X_i) = 0$ for every i, and let $\|X\| = \max_{1\leq k\leq n} |X_k|$. Let $\sigma^2 = \max_{1\leq k\leq n} \mathbf{E}\left(X_k^2\right)$. We have seen already that by using Gaussian concentration, we have:

$$\mathbf{P}\left(|\|X\| - \mathbf{E}\left(\|X\|\right)| \geq r\right) \leq 2\exp\left(-\frac{r^2}{2\sigma^2}\right)$$

If we integrate this in the natural way, we get a bound on the variance of $\|X\|$:

$$\begin{aligned}
\mathbf{Var}\left(\|X\|\right) &= \mathbf{E}\left(\left(\|X\| - \mathbf{E}(\|X\|\right)^2\right) \\
&= \int_0^\infty \mathbf{P}\left(\left(\|X\| - \mathbf{E}(\|X\|\right)^2 > s\right) \mathrm{d}s \\
&\leq \int_0^\infty 2\exp\left(-\frac{s}{2\sigma^2}\right) \mathrm{d}s \\
&= 4\sigma^2
\end{aligned}$$

By the def'n of σ^2, another way to phrase this is to say:

$$\mathbf{Var}\left(\max_{1\leq k\leq n} |X_k|\right) \leq 4 \max_{1\leq k\leq n} \mathbf{Var}\left(X_k\right)$$

With a bit more finess in the inequalities, one can actually show that the same inequality holds with 4 replaced by a 1.

Here is a variance bound that works for Gaussian probability measures:

Proposition 28. *(Poincare Inequality for Gaussian measures) Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function and γ is the standard Gaussian measure on $\mathbb{R}^n$. Then:*

$$\mathbf{Var}_\gamma\left(f\right) \leq \mathbf{E}_\gamma\left(|\nabla f|^2\right)$$

Proof. As for the proof of Gaussian concentration, there are many proofs of this result. As for the proof of Gaussian concentration, there is a slick proof using the Gaussian semigroup P_t and the writing:

$$\begin{aligned}
\mathbf{Var}_\gamma\left(f\right) &= P_1\left(f^2\right) - \left(P_1 f\right)^2 \\
&= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t}\left(P_t\left(P_{1-t}f\right)^2\right) \mathrm{d}t
\end{aligned}$$

The same kind of manipulations as for the proof of Gaussian concentration get us to the desired result. □

Remark 29. If $X \sim N(0, \Gamma)$ is a Gaussian measure with covariance structure Γ, then the Poincare inequality holds in the following form:

$$\mathbf{Var}\left(f(X)\right) \leq \mathbf{E}_\gamma\left(\langle \Gamma\nabla f(X), \nabla f(X)\rangle\right)$$

The proof comes as a change of variables from the case of the standard Gaussian γ.

Example 30. To prove $\mathbf{Var}\left(\max|X_k|\right) \leq \max\left(\mathbf{Var}\left(X_k\right)\right)$ with the Poincare inequality is straightforward. Let $f(x) = \max_k |x_k|$ and notice that although f is not smooth everywhere, we have that:

$$\partial_k f = \mathtt{1}_{A_k}$$

Where $A_k = \{x : \max_j |x_j| = x_k\}$ is the set where the coordinate $x_k$is maximal. Notice that the sets A_k are disjoint and together cover all of $\mathbb{R}^n$ Hence, the RHS of the Poincare inequality becomes:

$$\begin{aligned}
\mathbf{E}_\gamma\left(\langle \Gamma\nabla f, \nabla f\rangle\right) &= \int \sum_{k,l} \Gamma_{kl} \mathtt{1}_{A_k} \mathtt{1}_{A_j} \mathrm{d}\gamma \\
&= \int \sum_{k} \Gamma_{kk} \mathtt{1}_{A_k} \mathrm{d}\gamma \\
&\leq \max_k \Gamma_{kk} \\
&= \max \mathbf{Var}(X_k)
\end{aligned}$$

Example 31. In many cases the Poincare inequality is far from optimal. For example if $X_1, \ldots, X_n$ are all iid $N(0,1)$ random variables, it is known that:

$$\mathbf{Var}\left(\max_{1\leq i\leq n} X_i\right) = O\left(\frac{1}{\log n}\right) \text{ as } n \to \infty$$

The Poincare inequality only gives us an $O(1)$ upper bound.

Example 32. (From Spin Glasses) Let $\xi \in \{-1,1\}^N$ be the configuration of some object that has N spin-up-or-down sites. Let $(g_{ij})_{1\leq i,j\leq n}$ be iid $N(0,1)$ random variables which govern the interactions between the sites. Define the energy of a particular configuration ξ as:

$$H_N(\xi) = \frac{1}{\sqrt{N}} \sum_{i,j} \xi_i \xi_j g_{ij}$$

We can put a so called Gibb's measure on the space of all configurations by putting the probablity of a configuration as $\mathbf{P}(\xi) \propto \exp(\beta H_N(\xi))$. Here β is a parameter, the "inverse temperature" of the Gibbs measure. The normalizing constant is called the partition function and is:

$$Z_N^\beta = \mathbf{E}(\exp(\beta H_N(\xi))) = \frac{1}{2^N} \sum_{\xi} \exp(\beta H_N(\xi))$$

We define the free energy $F_N^\beta = \log\left(Z_N^\beta\right)$. The free energy and the partition function can be thought of as functions of the Gaussian environment (g_{ij}). It is easily verified that the free energy is a Lipshitz function in (g_{ij}) with $\|F\|_{Lip} \leq \beta\sqrt{\frac{N-1}{2}}$. From the Gaussian concentration inequality, we can get:

$$\mathbf{P}(|F_N - \mathbf{E}(F_N)| \geq r) \leq 2\exp\left(-\frac{r^2}{\beta^2(N-1)}\right)$$

The following results actually hold:
For $0 < \beta < 1$ one can show that:

$$\lim_{N\to\infty} \frac{F_N}{N} = \beta^2 \text{ a.s.}$$

For $\beta \geq 1$ (the low temperature regime: this is the so called Talagrand solution of the Parisi formula), one can show that $\frac{F_N}{N}$ converges to something, but it is not known what.

Remark 33. From the Poincare inequality or by integrating the concentration inequality above, one can see that $\mathbf{Var}(F_N) = O(N)$. There is however an even stronger concentration that is happening, known as "superconcentration". Chatterjee used this idea to prove in fact that $\mathbf{Var}(F_N) = O\left(\frac{N}{\log N}\right)$. The proof of this fact involves looking at the proof of the Poincare inequality with Hermite polynomials. For details, see the notes on "Superconcentration and Related Topics" by Chatterjee. Based on simulation, ideas from physics, and conjectured universality results, it is conjectured that $\mathbf{Var}(F_N)$ is actually bounded, but there is no proof of this yet.

Example 34. (Directed Last Passage Percolation) Consider now the grid $\mathbb{Z}_+^2$ and put i.i.d $N(0,1)$ weights g_v at each vertex v. Let Π_n be the set of all Up-Right paths from $(0,0)$ to (n,n). For a path $\pi \in \Pi_n$ define the weight of the path, $L(\pi) = \sum_{v\in\pi} g_v$. The last passage time is $Z = \max_{\pi\in\Pi_n} L(\pi)$. As before, we are interested in controlling $\mathbf{Var}(Z_n)$. The usual Poincare/concentration argument gives:

$$\mathbf{Var}(Z_n) \leq \max\left(\mathbf{Var}\left(\sum_{v\in\Pi} g_v\right)\right) = 2N+1$$

It is conjectured that $\mathbf{Var}(Z_n) = O(n^{2/3})$. This result is known for weights g_v which are exponential distributed. It is believed that the Gaussian case falls in the same universality class (the KPZ universality class) and should have the same order. The best known result is that $\mathbf{Var}(Z_n) = O\left(\frac{n}{\log n}\right)$ which was first proven by Beniamini, Kalai, and Schram. The superconcentration method also works for this.

Proposition 35. *(Talagrand's Improved Poincare Inequality) Here is an improved version of the Poincare inequality:*

$$\mathbf{Var}_\gamma(f) \leq c\sum_{i=1}^{v} \frac{\|\partial_i f\|_2^2}{1 + \log\left(\frac{\|\partial_i f\|_2}{\|\partial_i f\|_1}\right)}$$

Example 36. If $f(x) = \max_{1\le i\le n} x_i$ and say $X \sim N(0, Id_n)$. Then, as we saw before, we can think of $\partial_i f = 1_{A_i}$ where $A_i = \{x_i \text{is maximal}\}$. By symmetry, each $\gamma(A_i) = \frac{1}{n}$. Have then $\|\partial_i f\|_1 = \gamma(A_i) = \frac{1}{n}$ and $\|\partial_i f\|_2 = \sqrt{\gamma(A_i)} = \frac{1}{\sqrt{n}}$. So the improved Poincare inequality says:

$$\mathbf{Var}_\gamma(f) \le c\sum_{i=1}^n \frac{1/n}{1+\log(\sqrt{n})} = \frac{c}{1+\log(\sqrt{n})}$$

Example 37. (Gaussian Free Field with zero boundary condition) I didn't understand this example....

Example 38. (Random Matrices) Suppose that we take a symmetric matrix $(X_{ij})_{1\le i,j\le n}$ and we take $1 \le i \le j \le n$ to be independent centered Gaussian random variables with $\mathbf{E}\left(X_{ij}^2\right) = 1$ for $i \le j$. Let us look at the largest eigenvalue of this ensemble:

$$\lambda_{\max} = \sup_{|u|=1} \langle Xu, u\rangle = \sup_{|u|=1} \sum_{i,j=1}^n X_{ij}u_iu_j$$

By the usual Poincare inequality, we get $\mathbf{Var}(\lambda_{\max}) \le 2$. It turns out that the correct order is $n^{-1/3}$. We will look at random matrices in more detail later.

3 Concentration Inequalities for Product Measures

In this section we will look at a general sequence $X_1, \ldots, X_n$ of independent random variables, and we will be interested in general functionals $F(X_1, \ldots, X_n)$ for functional F with suitable properties. The easiest type of inequality gained is at the variance level.

3.1 Efron Stein Inequality

Theorem 39. *(Efron-Stein Inequality)*

[Remark on Notation: In my phrasing of this result, I use the "absentee hat"ˆ. When you have a list $\{1, 2, \ldots, \hat{i}, \ldots n\}$, the absentee hat $\hat{i}$ means that the element i is NOT on the list. We will also use the notation for this problem that $\mathbf{E}_{\mu_k}$ or $\mathbf{Var}_{\mu_k}$ "integrates out" the k-th variable. In other words $\mathbf{E}_{\mu_k}(f) : E_1 \times \ldots \times \hat{E}_k \times \ldots \times E_n \to \mathbb{R}$ is still a function of the other $n-1$ variables. $\mathbf{E}_{\mu_1,\mu_2}$ means integrate out the first and second space. In this case, since we will be looking at product measures, this can also be thought of as conditioning on the remaining variables, i.e. $\mathbf{E}_{\mu_1}(f) = \mathbf{E}(f|X_2, \ldots, X_n)$. The variance is defined as: $\mathbf{Var}_\mu(f) = \mathbf{E}_\mu(f^2) - \left(\mathbf{E}_\mu(f)^2\right)$.)

Suppose we have n sets $E_1, \ldots, E_n$ and n measures $\mu_1, \ldots, \mu_n$ on these sets. Let $E = E_1 \times \ldots \times E_n$ and $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ be the product measure on this set. Take any function $f : E \to \mathbb{R}$ and suppose $\mathbf{Var}(f) = \mathbf{Var}_{\mu_1,\ldots,\mu_n}(f) < \infty$. Then:

$$\mathbf{Var}_{\mu_1\ldots\mu_n}(f) \le \sum_{i=1}^n \mathbf{E}_{\mu_1\ldots,\hat{\mu_k}\ldots,\mu_n}\left(\mathbf{Var}_{\mu_i}(f)\right)$$

Proof. The presented proof uses the inequality that for functions $f : E_1 \times E_2$ with measure $\mu \otimes \nu$, that $\mathbf{Var}_\mu\left(\mathbf{E}_\nu(f)\right) \le \mathbf{E}_\nu\left(\mathbf{Var}_\mu(f)\right)$. This looks a bit like a Jensen's inequality for the "convex" function $\mathbf{Var}_\mu$, but actually its a bit more subtle here because $\mathbf{Var}_\mu$ outputs functions, not numbers. In any case, I think this inequality is true, but I have been unable to prove it. I will present the proof assuming this assumed "Jensen-type inequality", and I will also present an alternate proof later on.

This is a proof by induction on the dimension n. It is in the same vein as the proof of the bounded difference concentration inequality.

Base Case: When $n = 2$, we have:

$$\begin{aligned}
\mathbf{Var}(f) &= \mathbf{E}(f^2) - \mathbf{E}(f)^2 \\
&= \mathbf{E}_{\mu_2}\left[\mathbf{E}_{\mu_1}(f^2)\right] - \left(\mathbf{E}_{\mu_2}\left[\mathbf{E}_{\mu_1}(f)\right]\right)^2 \\
&= \mathbf{E}_{\mu_2}\left[\mathbf{Var}_{\mu_1}(f) + \mathbf{E}_{\mu_1}(f)^2\right] - \left(\mathbf{E}_{\mu_2}\left[\mathbf{E}_{\mu_1}(f)\right]\right)^2 \\
&= \mathbf{E}_{\mu_2}\left[\mathbf{Var}_{\mu_1}(f)\right] + \left(\mathbf{E}_{\mu_2}\left[\mathbf{E}_{\mu_1}(f)^2\right] - \left(\mathbf{E}_{\mu_2}\left[\mathbf{E}_{\mu_1}(f)\right]\right)^2\right) \\
&= \mathbf{E}_{\mu_2}\left[\mathbf{Var}_{\mu_1}(f)\right] + \mathbf{Var}_{\mu_2}\left(\mathbf{E}_{\mu_1}(f)\right) \\
&\le \mathbf{E}_{\mu_2}\left[\mathbf{Var}_{\mu_1}(f)\right] + \mathbf{E}_{\mu_1}\left(\mathbf{Var}_{\mu_2}(f)\right) \text{ by the Jensen-type ineq.}
\end{aligned}$$

Induction Step: Assume that the inequality holds for $n-1$. Let $g = \mathbf{E}_{\mu_1,\ldots,\mu_{n-1}}(f)$. If we let $\nu = \mu_1 \otimes \ldots \otimes \mu_{n-1}$ then the Jensen-type inequality tells us that $\mathbf{Var}_{\mu_n}(g) = \mathbf{Var}_{\mu_n}\left[\mathbf{E}_\nu(f)\right] \le \mathbf{E}_\nu\left[\mathbf{Var}_{\mu_n}(f)\right] = \mathbf{E}_{\mu_1,\ldots,\mu_{n-1}}\left[\mathbf{Var}_{\mu_n}(f)\right]$. Now, using this inequality and the induction hypothesis, the argument goes in a very similar way to the base case:

$$\begin{aligned}
\mathbf{Var}(f) &= \mathbf{E}(f^2) - \mathbf{E}(f)^2 \\
&= \mathbf{E}_{\mu_n}\left[\mathbf{E}_{\mu_1,\ldots,\mu_{n-1}}(f^2)\right] - \left(\mathbf{E}_{\mu_n}\left[\mathbf{E}_{\mu_1,\ldots,\mu_{n-1}}(f)\right]\right)^2 \\
&= \mathbf{E}_{\mu_n}\left[\mathbf{Var}_{\mu_1\ldots,\mu_{n-1}}(f) + \mathbf{E}_{\mu_1,\ldots,\mu_{n-1}}(f)^2\right] - \left(\mathbf{E}_{\mu_n}[g]\right)^2 \\
&= \mathbf{E}_{\mu_n}\left[\mathbf{Var}_{\mu_1,\ldots,\mu_{n-1}}(f)\right] + \left(\mathbf{E}_{\mu_n}\left[g^2\right] - \left(\mathbf{E}_{\mu_n}[g]\right)^2\right) \\
&= \mathbf{E}_{\mu_n}\left[\mathbf{Var}_{\mu_1,\ldots,\mu_{n-1}}(f)\right] + \mathbf{Var}_{\mu_n}(g) \\
&\leq \mathbf{E}_{\mu_n}\left[\sum_{i=1}^{n-1}\mathbf{E}_{\mu_1\ldots\hat{\mu_i},\ldots,\mu_{n-1}}(\mathbf{Var}_{\mu_i}(f))\right] + \mathbf{E}_{\mu_1,\ldots,\mu_{n-1}}\left[\mathbf{Var}_{\mu_n}(f)\right] \\
&= \sum_{i=1}^{n-1}\mathbf{E}_{\mu_1\ldots,\hat{\mu_k}\ldots,\mu_n}\left(\mathbf{Var}_{\mu_i}(f)\right) + \mathbf{E}_{\mu_1,\ldots,\mu_{n-1},\hat{\mu_n}}\left[\mathbf{Var}_{\mu_n}(f)\right]
\end{aligned}$$

This completes the proof. □

Remark 40. A probabilistic way to think about variance is using "duplication". If X and X' are i.i.d copies of the same random variable, then one can write:

$$\begin{aligned}
\mathbf{Var}\left(f(X)\right) &= \mathbf{E}\left(f(X)^2\right) - \mathbf{E}\left(f(X)\right)^2 \\
&= \frac{1}{2}\left(\mathbf{E}\left(f(X)^2\right) + \mathbf{E}\left(f(X')^2\right)\right) - \mathbf{E}\left(f(X)f(X')\right) \\
&= \frac{1}{2}\mathbf{E}\left[\left(f(X) - f(X')\right)^2\right]
\end{aligned}$$

The Efron-Stein inequality has a nice phrasing along these lines.

Theorem 41. *(Efron Stein Inequality -Probabilistic Formulation) Suppose $X_1,\ldots,X_n,X_1',\ldots,X_n'$ are independent with $X_i' \stackrel{d}{=} X_i$. Let $X = (X_1,\ldots,X_n)$ and $X^{(i)} = (X_1,\ldots,X_{i-1},X_i',X_{i+1},\ldots,X_n)$. Then:*

$$\mathbf{Var}(f(X)) \leq \frac{1}{2}\sum_{i=1}^{n}\mathbf{E}\left[\left(f(X) - f(X^{(i)})\right)^2\right]$$

Remark 42. This is exactly the same as the first inequality, simply reformulated a bit. By the probabalistic way of thinking about variance, we have $\mathbf{Var}_{\mu_i}(f(X)) = \frac{1}{2}\mathbf{E}_{\mu_i}\left[\left(f(X) - f(X^{(i)})\right)^2\right]$ which makes the identification between the two formulations clearer.

Proof. (I took this proof and the phrasing of the Efron-Stein inequality from the Berkeley website of Sourav Chatterjee). Let $X' = (X_1',\ldots,X_n')$ and $X^{[i]} = (X_1',\ldots,X_i',X_{i+1},\ldots,X_n)$. Notice $X^{[0]} = X$ and $X^{[n]} = X'$. Now consider:

$$\begin{aligned}
\mathbf{Var}(f(X)) &= \mathbf{E}\left[f(X)^2\right] - \left(\mathbf{E}\left[f(X)\right]\right)^2 \\
&= \mathbf{E}\left[f(X)^2\right] - \mathbf{E}\left[f(X)f(X')\right] \\
&= \mathbf{E}\left[f(X)\left(f(X) - f(X')\right)\right] \\
&= \mathbf{E}\left[f(X)\left(f(X^{[0]}) - f(X^{[1]}) + f(X^{[1]}) - \ldots, + f(X^{[n-1]}) - f(X^{[n]})\right)\right] \\
&= \sum_{i=1}^{n}\mathbf{E}\left[f(X)\left(f(X^{[i-1]}) - f(X^{[i]})\right)\right]
\end{aligned}$$

Now, consider the map $\varphi_i : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ by

$$\varphi_i\left(X_1,\ldots,X_n,X_1',\ldots,X_n'\right) = \left(X_1,\ldots,X_{i-1},X_i',X_{i+1},\ldots,X_n,X_1',\ldots,X_{i-1}',X_i,X_{i+1}',\ldots,X_n'\right)$$

This simply switches X_i and X_i'. This is a probability preserving transformation (i.e. $\mathbf{P}\left(\varphi_i^{-1}(A)\right) = \mathbf{P}(A)$) because X_i and X_i' are identically distributed. Under this switching, we have by our definitions that $X \stackrel{\varphi_i}{\to} X^{(i)}$, $X^{[i]} \stackrel{\varphi_i}{\to} X^{[i-1]}$, and $X^{[i-1]} \stackrel{\varphi_i}{\to} X^{[i]}$. Because the map is probability preserving, we have then:

$$f(X)\left(f(X^{[i-1]}) - f(X^{[i]})\right) \stackrel{d}{=} f(X^{(i)})\left(f(X^{[i]}) - f(X^{[i-1]})\right)$$

Hence, these two quantities have the same expectation. Since these have the same expectation, we may write:

$$\begin{aligned}\mathbf{E}\left[f(X)\left(f(X^{[i-1]})-f(X^{[i]})\right)\right] &= \frac{1}{2}\left(\mathbf{E}\left[f(X)\left(f(X^{[i-1]})-f(X^{[i]})\right)\right]+\mathbf{E}\left[f(X^{(i)})\left(f(X^{[i]})-f(X^{[i-1]})\right)\right]\right)\\ &= \frac{1}{2}\mathbf{E}\left[\left(f(X)-f(X^{(i)})\right)\left(f(X^{[i]})-f(X^{[i-1]})\right)\right]\\ &\leq \frac{1}{2}\sqrt{\mathbf{E}\left[\left(f(X)-f(X^{(i)})\right)^2\right]\mathbf{E}\left[\left(f(X^{[i]})-f(X^{[i-1]})\right)^2\right]} \text{ by Cauchy-Schwarz}\end{aligned}$$

Now $\left(X, X^{(i)}\right) \stackrel{d}{=} \left(X^{[i]}, X^{[i-1]}\right)$ because both differ by an independent copy in the i-th component. (You can also see this by setting up another probability preserving transformation which switches the first $i-1$ X_k's to X_k''s) In particular then, both expectations under the square root are equal, so we have:

$$\mathbf{E}\left[f(X)\left(f(X^{[i-1]})-f(X^{[i]})\right)\right] \leq \frac{1}{2}\mathbf{E}\left[\left(f(X)-f(X^{(i)})\right)^2\right]$$

This holds for each i. When we sum over i, we get exactly the desired inequality. □

Corollary 43. *Suppose we have n sets $E_1, \ldots, E_n$ and n measures $\mu_1, \ldots, \mu_n$ on these sets. Let $E = E_1 \times \ldots \times E_n$ and $\mu = \mu_1 \otimes \ldots \otimes \mu_n$ be the product measure on this set. Suppose further that each μ_i is supported on $[0,1]$. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex in each coordinate. Then:*

$$\mathbf{Var}_{\mu_1,\ldots,\mu_n}(f) \leq \mathbf{E}_{\mu_1,\ldots,\mu_n}\left(\sum_{i=1}^n (\partial_i f)^2\right)$$

Proof. By application of the Efron Stein inequality, it suffices to prove the result when $n = 1$. To see this let X and X'be iid copies of the same variable and consider:

$$\begin{aligned}\mathbf{Var}(f(X)) &= \frac{1}{2}\mathbf{E}\left[(f(X)-f(X'))^2\right]\\ &= \mathbf{E}\left[(f(X)-f(X'))^2 \mathbf{1}_{\{f(X)>f(X')\}}\right]\end{aligned}$$

Now, since f is convex, we know that $f(x) - f(y) \leq f'(x)(x-y)$. Hence we have:

$$\begin{aligned}\mathbf{Var}(f(X)) &= \mathbf{E}\left[(f(X)-f(X'))^2 \mathbf{1}_{\{f(X)>f(X')\}}\right]\\ &\leq \mathbf{E}\left[f'(X)^2 (X-X')^2 \mathbf{1}_{\{f(X)>f(X')\}}\right]\\ &\leq \mathbf{E}\left[f'(X)^2\right] \text{ since } X - X' \leq 1\text{a.s}\end{aligned}$$

□

Corollary 44. *In the above setting, if f is Lipshitz and convex with Lipshitz constant L then:*

$$\mathbf{Var}(f) \leq L^2$$

Proof. By smoothing we can reduce to the case f is differentiable, and $\sum_{i=1}^n (\partial_i f)^2 = |\nabla f|^2 \leq L^2$ gives the result. □

3.2 Examples

Example 45. (Random Matrices) Suppose $A = (X_{ij})_{1 \leq i,j \leq n}$ is a symmetric random matrix that we get by choosing the subdiagonal entries X_{ij} for $i \leq j$ independently at random with $|X_{ij}| \leq 1$. Let $\lambda_{\max}$ be the largest eigenvalue of the matrix. By the min-max theorem, we know that:

$$\lambda_{\max}(A) = \sup_{|u|=1} \sum_{i,j=1}^n X_{ij} u_i u_j$$

The sum $\sum X_{ij} u_i u_j$ is Lipshitz and convex when thought of as a function of the entries X_{ij}. Indeed, by the Cauchy-Schwarz inequality, we have for $|u| \leq 1$ that:

$$\begin{aligned}\left|\sum X_{ij}u_iu_j - \sum Y_{ij}u_iu_j\right| &= \left|\sum (X_{ij}-Y_{ij})u_iu_j\right|\\ &\leq \sqrt{\sum (X_{ij}-Y_{ij})^2 \sum (u_iu_j)^2}\\ &\leq \sqrt{\sum (X_{ij}-Y_{ij})^2} = \|X-Y\|\end{aligned}$$

When we take the sup in order to get $\lambda_{\max}$, the convexity and Lipschitz property are preserved, so we have by our theorems that:

$$\mathbf{Var}\left(\lambda_{\max}(A)\right) \leq C$$

We will look at random matrices in more detail later.

Example 46. (Last Passage Percolation)

Suppose we have place weights $(X_{ij})_{1\leq i,j\leq n}$ which are iid with $|X_{ij}| \leq 1$ on the lattice $\{1,2,\ldots,n\}^2 \subset \mathbb{Z}_+^2$. For an up-right path π going from $(1,1)$ to (n,n) in $\mathbb{Z}^2$ we define the weight of the path as $\sum_{(i,j)\in\pi} X_{ij}$. The last passage percolation time is:

$$Z = \max_{\pi} \sum_{(i,j)\in\pi} X_{ij}$$

As in the case of the random matrices, it is not too difficult to see that Z is convex and Lipschitz. The Lipschitz constant here is order $\sqrt{n}$ because:

$$\begin{aligned}
\left|\sum_{(i,j)\in\pi} X_{ij} - \sum_{(i,j)\in\pi} Y_{ij}\right|^2 &\leq \left(\sum_{(i,j)\in\pi} |X_{ij} - Y_{ij}|\right)^2 \\
&\leq 4n \sum_{(i,j)\in\pi} |X_{ij} - Y_{ij}|^2 \text{ by Jensen's inequality} \\
&\leq 4n \|X - Y\|^2
\end{aligned}$$

The inequality follows by using Jensen's inequality for $(\cdot)^2$ and the fact that there are exactly $2n$ vertices to be found on the path. Hence we have that:

$$\mathbf{Var}\left(Z\right) \leq Cn$$

3.3 Exponential Inequalities and the Entropic Method

I will type up more notes later! There is approx 1 lecture left to go.