STAT 583: Advanced Theory of Statistical Inference **Spring 2018**

Lecture 0: Some useful inequalities

Lecturer: Fang Han

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Lecturer.*

"The good men of every age are those who go to the roots of the old thoughts and bear fruit with them."

— Friedrich Nietzsche, The Gay Science

0.1 LLN and CLT

LLN, in its simplest format, is one of the easiest proofs we will ever meet in statistics, involving only the Markov's inequality.

Theorem 1 (WLLN)**.** *Suppose* $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ *with* $\mathbb{E}_F|X|^2 < \infty$*, then* $\overline{X}_n \overset{P}{\to} \mathbb{E}_F X$*. In other words, for arbitrary* $\epsilon > 0$*,*

$$\lim_{n\to\infty} P_F(|\overline{X}_n - \mathbb{E}_F X| > \epsilon) = 0.$$

Proof. In-class exercise. □

SLLN is much more interesting. In the following we give its comprehensive version.

Theorem 2 (SLLN)**.** *(i) If* $\mathbb{E}_F|X| < \infty$*, then* $\overline{X}_n \overset{a.s.}{\to} \mathbb{E}_F X$*. In other words, for arbitrary* $\epsilon > 0$*,*

$$P_F(\lim_{n\to\infty} X_n = \mathbb{E}_F X) = 1.$$

(ii) (Zygmund-Marcinkiewicz SLLN.) If for some $0 < \delta < 1$*,* $\mathbb{E}_F|X|^\delta < \infty$*, then we have*

$$n^{-1/\delta} \sum X_i \overset{a.s.}{\to} 0.$$

CLT is a fundamental theory. However, its proof, in its simplest form (moment-based one or characteristic-function-based one), has been out of fashion, since it tells us so little about $\overline{X}_n$. But anyway, let's write down the theorem at first.

Theorem 3 (Lindeberg-Levy CLT)**.** *Given* $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} F$ *with* $\mathbb{E}_F X = \mu$ *and* $\mathbb{E}_F(X-\mu)^2 = \sigma^2 < \infty$*. Then we have*

$$\sqrt{n}(\overline{X}_n - \mu) \overset{d}{\to} N(0, \sigma^2),$$

or in other words, for arbitrary $x \in \mathbb{R}$*,*

$$\lim_{n\to\infty} P(\sqrt{n}(\overline{X}_n - \mu) \le x) = \Phi(x/\sigma). \quad \text{(pointwise convergence)}$$

Proof. In-class exercise. □

0.2 Berry-Esseen theorem and Edgeworth expansion

CLT tells us the limiting behavior of $\overline{X}_n$ as $n \to \infty$. However, it never tells us how fast $\overline{X}_n - \mu$ converges to $N(0, \sigma^2)$. Actually, a result like

$$|P(\sqrt{n}(\overline{X}_n - \mu) \le x) - \Phi(x/\sigma)| = O(1/\log n)$$

would be useless. Gladly, Berry-Esseen Theorem tells us the convergence rate is usually not that disappointing.

Theorem 4 (Berry-Esseen Theorem (Esseen 1956))**.** *Suppose $\mathbb{E}_F|X - \mu|^3 < \infty$. We then have*

$$\sup_x |P(\sqrt{n}(\overline{X}_n - \mu)/\sigma \le x) - \Phi(x)| \le \frac{0.4785 \cdot \mathbb{E}|X - \mu|^3}{\sigma^3\sqrt{n}}.$$

(An interesting story on how the constant is sharpened from 7.59 (the original proof of Esseen) to the current one by Tyurin (2010) could be found at the wikipedia.)

Berry-Esseen theorem gives us the first-order (root-n) approximation for CLT. Its proof is based on the Fourier expansion and is very revealing. I won't cover it in this lecture. However, persons could easily find it online, or drop me an email (I have a proof by myself).

Let's move on to characterizing the higher-order approximation for CLT. This is known as the Edgeworth expansion, and is celebrated for its application to proving the second-order accuracy of the bootstrap.

Theorem 5 (Edgeworth expansion)**.** *Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$. Write*

$$\gamma := \mathbb{E}_F(X - \mu)^3/\sigma^3 \text{ (skewness) } \text{ and } \kappa := \mathbb{E}_F(X - \mu)^4/\sigma^4 \text{ (kurtosis).}$$

We then have

$$\begin{aligned} G_n(x) &:= P_F(\sqrt{n}(\overline{X}_n - \mu)/\sigma \le x) \\ &= \Phi(x) - \phi(x)\left(\frac{\gamma(x^2 - 1)}{6\sqrt{n}} + \frac{(\kappa - 3)(x^3 - 3x)}{24n} + \frac{\gamma^2(x^5 - 10x^3 + 15x)}{72n}\right) + o(1/n). \end{aligned}$$

Proof. See the blackboard, LOL. □

Some implications are immediate.

Remark 6. *When F is symmetric, we have $\gamma = 0$, so that $\Phi(x)$ approximates $G_n(x)$ in the rate $O(1/n)$ (This justifies the intuition that "30 is good enough for CLT to work").*

Remark 7. *When F is asymmetric, generally $\gamma \ne 0$ and the CLT can only attain $O(1/\sqrt{n})$ rate of convergence. However, say, if we are interested in calculating the confidence interval of $\sqrt{n}(\overline{X}_n - \mu)/\sigma$, a balanced interval gives us*

$$G_n(x) - G_n(-x) = \Phi(x) - \Phi(-x) + O(1/n),$$

with the first term cancelled out. This is the intuition why balanced confidence interval is more preferred.

Remark 8. *The first two-order approximation is involved with κ only through $\kappa - 3$. This is the intuition why the excess kurtosis is defined as $\kappa - 3$.*

0.3 A case study: the bootstrap theory built on Berry-Esseen theorem and Edgeworth expansion

Suppose $T(X_1, \dots, X_n; F)$ is a functional (e.g., $T(X_1, \dots, X_n; F) = \sqrt{n}(\overline{X}_n - \mu)$). Each time, the bootstrapped sample $X_1^*, \dots, X_n^*$ is sampled from $X_1, \dots, X_n$ with replacement. In other words, the bootstrap sample is drawn from the ECDF F_n with point mass on $X_1, \dots, X_n$. The corresponding statistic is $T(X_1^*, \dots, X_n^*; F_n)$. It is set up to approximate the true distribution of $T(X_1, \dots, X_n; F)$.

Let's consider the simplest case, where $T(X_1, \dots, X_n; F) = \sqrt{n}(\overline{X}_n - \mu)$. It is truely surprising that we can prove the bootstrap consistency based on such few results we have known. The following proof is due to Professor Anirban DasGupta.

Theorem 9. *Provided $\mathbb{E}_F X^2 < \infty$ and $T(X_1, \dots, X_n; F) := \sqrt{n}(\overline{X}_n - \mu)$, we have*

$$\sup_x |P_F(T_n \le x) - P_*(T_n^* \le x)| \overset{a.s.}{\to} 0,$$

where P_* corresponds to the uniform distribution over all the n^n possible replacement resamples from $(X_1, \dots, X_n)$, and $T_n^* := \sqrt{n}(\sum X_i^*/n - \overline{X}_n)$.

Proof. By triangle inequality, we have

$$\begin{aligned}\sup_x |P_F(T_n \le x) - P_*(T_n^* \le x)| &\le \sup_x |P_F(T_n/\sigma \le x/\sigma) - \Phi(x/\sigma)| + \sup_x |\Phi(x/\sigma) - \Phi(x/s)| \\ &\quad + \sup_x |\Phi(x/s) - P_*(T_n^*/s \le x/s)| \\ &= A_n + B_n + C_n,\end{aligned}$$

where s is the sample standard deviation, and is the standard deviation of $(X_1, \dots, X_n)$ under P_*. Here $A_n \to 0$ by CLT. $B_n \to 0$ by the fact $s \overset{a.s.}{\to} \sigma$ and the continuous mapping theorem. Finally, applying the Berry-Esseen theorem to P_*, we have

$$C_n \le \frac{C}{\sqrt{n}} \cdot \frac{\mathbb{E}_{F_n}(X_1^* - \overline{X}_n)^3}{[\mathrm{Var}_{F_n}(X_1^*)]^{3/2}} = \frac{C}{\sqrt{n}} \cdot \frac{\sum |X_i - \overline{X}_n|^3}{ns^3} \le \frac{8C}{n^{3/2}s^3} \cdot \left(\sum |X_i - \mu|^3 + n|\overline{X}_n - \mu|^3\right),$$

where in the last inequality we use the fact $(a+b)^3 \le 8(a^3 + b^3)$ for any $a, b > 0$. We then continue to have

$$\frac{8C}{n^{3/2}s^3} \cdot \left(\sum |X_i - \mu|^3 + n|\overline{X}_n - \mu|^3\right) \le \frac{C'}{s^3}\left(\frac{1}{n^{3/2}}\sum |X_i - \mu|^3 + \frac{|\overline{X}_n - \mu|^3}{\sqrt{n}}\right).$$

Clearly, these two terms will vanish by Zygmund-Marcinkiewicz SLLN. □

We then move to study the so-called second-order accuracy of the bootstrap. In particular, we aim to rigorously answer the following question: why the bootstrap is more preferred even when CLT-type results are available. For example, even if we know $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ converges to $N(0,1)$, why should we still use the bootstrapped sample to approximate its distribution.

In short, under some assumptions, the bootstrap convergence rate is $O(1/n)$ compared to $O(1/\sqrt{n})$ for CLT. The following argument is due to Eric Lehmann.

Consider $T = \sqrt{n}(\overline{X}_n - \mu)/\sigma$. By Edgeworth expansion, we have

$$\begin{aligned}P_F(T \le x) &= \Phi(x) + \phi(x)(p_1(x|F)/\sqrt{n} + p_2(x|F)/n) + o(1/n) \\ P_{F^*}(T^* \le x) &= \Phi(x) + \phi(x)(p_1(x|F_n)/\sqrt{n} + p_2(x|F_n)/n) + o(1/n) \\ P_F(T \le x) - P_{F^*}(T^* \le x) &= \phi(x)\left(\frac{p_1(x|F) - p_1(x|F_n)}{\sqrt{n}} + \frac{p_2(x|F) - p_2(x|F_n)}{n}\right) + o(1/n),\end{aligned}$$

with

$$p_1(x|F) = \frac{\gamma}{6}(1-x^2), \quad p_2(x|F) = \frac{\kappa-3}{24}(3x-x^3) - \frac{\gamma^2}{72}(x^5-10x^3+15x).$$

Hence, since $\gamma_{F_n} - \gamma_F = O_P(1/\sqrt{n})$, we obtain $O(1/n)$ rate of convergence given the finiteness of the moments, which is called the second-order accuracy, in comparison to the first-order accuracy ($O(1/\sqrt{n})$) in CLT.

However, when we do not standardize the data, the second-order accuracy is lost, since additional effort is required to bound $\Phi(x/\sigma) - \Phi(x/s)$. Therefore, a rule of thumb is as follows:

Proposition 10 (DasGupta)**.** *If $T(X_1, \ldots, X_n; F) \xrightarrow{d} N(0, \tau^2)$ with τ independent of F and an Edgeworth expansion is available to T, then the second order accuracy is likely.*

[Talk about the bootstrap inconsistency.]

0.4 LIL and Cramer's moderate deviation theory

I do not expect that we will have the time to cover these. But they are interesting, and extremely useful in high dimensional statistics (I have quite a few papers built on these two types of theories). So here they are.

0.4.1 LIL

SLLN tells us $\overline{X}_n$ will converge to μ, but it does not tell us how fast the convergence rate is. In this sense, CLT tells us more things, since it intuitively shows us the rate would be very close to root-n. The sharpest result in this sense is LIL, which gives us the exact convergence rate.

Theorem 11 (Law of the iterated lograthim)**.** *Assume mean zero and variance 1. We then have*

$$\limsup_{n\to\infty} \frac{\sum X_i}{\sqrt{2n\log\log n}} = 1 \quad \text{a.s.},$$

and

$$\liminf_{n\to\infty} \frac{\sum X_i}{\sqrt{2n\log\log n}} = -1 \quad \text{a.s.}.$$

The proof is surprisingly simple, and can be found everywhere. This result is relatively less touched recently. However, a very useful result does stem from it, called the Erdos' inequality. It gives us the convergence rate for partial mean:

$$\max_{k\le n} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^{k} X_i \right| = O_P(\sqrt{\log\log n}),$$

which is in contrast to the typical $\sqrt{\log n}$ rate for extreme value. This result is repeatedly used in change point detection.

0.4.2 Cramer's moderate deviation theory

Berry-Esseen and Edgeworth indeed give us a lot of understanding on the performance of the sample mean. However, the usefulness of the bound is very restricted in the tails when $\Phi(x)$ approximates 0 or 1. The reason is because, at this area, $\Phi(x)$ (1-$\Phi(x)$) itself is too small to be considered as a constant. However, concentration inequalities (introduced in the next section) do tell us the convergence to $\Phi(x)$ could even hold in the far tails. This intuition is fixed by the following theorem.

Theorem 12. *(i) Let $X_1, \ldots, X_n$ be i.i.d. with mean zero and $\mathbb{E}\exp(t_0|X|) < \infty$ for some $t_0 > 0$. Then for $x \geq 0$ and $x = o(n^{1/2})$, we have*

$$\frac{P(\sqrt{n}\overline{X}_n/\sigma \geq x)}{1-\Phi(x)} = \exp\left(x^2\lambda\left(\frac{x}{\sqrt{n}}\right)\right)\left(1+O\left(\frac{1+x}{\sqrt{n}}\right)\right),$$

where $\lambda(t)$ is the Cramer series.
(ii) If we further have $\mathbb{E}\exp(t_0\sqrt{|X|}) < \infty$, then

$$\sup_{x\in[0,o(n^{1/6}))} \frac{P(\sqrt{n}\overline{X}_n/\sigma \geq x)}{1-\Phi(x)} \to 1.$$

0.5 Markov and Chernoff

Theorem 13 (Markov)**.** *For any random variable $X \geq 0$ and $t > 0$, we have*

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

Theorem 14 (Chernoff)**.** *For any random variable X, we have*

$$P(X \geq t) \leq \inf_{s>0} \frac{\mathbb{E}\exp(sX)}{\exp(st)}.$$

Example 15. *For $X \sim N(0,1)$, we have*

$$P(X \geq t) \leq \inf_{s>0} \exp(s^2/2 - st) = \exp(-t^2/2) \to P(|X| \geq t) \leq 2\exp(-t^2/2).$$

Another approach gives us:

$$P(|X| > t) \leq 2\exp(-t^2/2)/t$$

by

$$P(X > t) = \int_t^\infty \phi(x)dx \leq \frac{1}{t}\int_t^\infty x\phi(x)dx = -\frac{1}{t}\int_t^\infty \phi'(x)dx = \frac{\phi(t)}{t}.$$

Example 16. *In the last example, consider $\overline{X}_n$ instead of X. Show the advantage of using the second inequality.*

Chernoff gives you a bound, but this bound is almost always a loose one when we consider parametric models.

Now we have been technically ready to prove the SLLN under the simplest case.

Example 17. *Prove the SLLN for Gaussian distributions. For any $t > 0$, we have*

$$\sum_{n=1}^\infty P(|\overline{X}_n| > t) \leq 2\sum_{n=1}^\infty \exp(-nt^2/2) \leq 2(1-e^{-1})^{-1}\exp(-t^2/2) < \infty. \tag{0.1}$$

This implies $\overline{X}_n \overset{a.s.}{\to} 0$ by the Borel-Cantelli Lemma.

Definition 18. *A r.v. X is said to be subgaussian of subgaussian (David Pollard calls it "scale") constant σ if for any $t \in \mathbb{R}$,*

$$\mathbb{E}\exp(t(X - \mathbb{E}X)) \leq \exp(\sigma^2 t^2/2).$$

Gaussian distribution $X \sim N(\mu, \sigma^2)$ has subgaussian constant σ.

It is obvious, if X is subgaussian of subgaussian constant σ, by Chernoff,

$$P(|X| > t) \leq 2\exp(-t^2/2\sigma^2).$$

0.6 Hoeffding and McDiarmid

We then consider bounded r.v. $X \in [a, b]$ for some constants $a, b \in \mathbb{R}$. Hoeffding managed to prove that X is subgaussian of subgaussian constant $(b - a)/2$.

Lemma 19 (Hoeffding's Lemma)**.** *Suppose $\mathbb{E}X = 0$ (why WLOG?) and $\mathbb{P}(X \in [a, b]) = 1$. Then*

$$\mathbb{E}\exp(tX) \leq \exp(t^2(b - a)^2/8).$$

Proof. By the convexity of the exponential function, we have

$$\mathbb{E}\exp(tX) \leq -\frac{a}{b-a}\exp(tb) + \frac{b}{b-a}\exp(ta) = \exp(g(u))$$

where $u = t(b - a)$, $g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u)$, and $\gamma = -a/(b - a)$. By Taylor expansion, there exists $\xi \in (0, u)$ such that

$$g(u) = g(0) + ug'(0) + \frac{u^2}{2}g''(\xi).$$

By calculation, we have $g(0) = g'(0) = 0$ and $g''(u) \leq 1/4$ for all $u > 0$. Therefore, we continue the above equation to have $g(u) \leq u^2/8 = t^2(b - a)^2/8$. This completes the proof. □

Theorem 20 (Hoeffding's Inequality)**.** *For $X_1, \ldots, X_n$ as independent (not necessarily identically distributed) r.v.'s satisfying $X_i \in [a, b]$ for $i = 1, \ldots, n$. for any $t > 0$, we have*

$$P\left(\left|\overline{X}_n - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i\right| > t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

Proof. In-class exercise. □

We then proceed to a generalization of the Hoeffding, called the McDiarmid. We will repeatedly use it in the future.

Theorem 21 (McDiarmid's inequality)**.** *Let $X_1, \ldots, X_n$ be n independent r.v.'s (again, not necessarily identically distributed) taking values from a general set A. Let $f : A^n \to \mathbb{R}$ satisfy the following bounded gap condition:*

$$\sup_{x_1,\ldots,x_n,x_i'} |f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq M_i \quad \text{for } i = 1, \ldots, n.$$

Then for any $t > 0$, we have

$$P(f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n) \geq t) \leq \exp\left(-\frac{2t^2}{\sum M_i^2}\right).$$

Example 22. *Show that McDiarmid can imply Hoeffding.*

Proof. We write $X = \{X_1, \ldots, X_n\}$, $Z_0 = \mathbb{E}f(X)$, $Z_i = \mathbb{E}(f(X)|X_1, \ldots, X_i)$, and $Z_n = f(X)$ (Is $Z_1, \ldots, Z_n$ a martingale?). Then by the bounded gap condition, it is easy to derive, for any $s > 0$,

$$\mathbb{E}(\exp(s(Z_k - Z_{k-1}))|X_1, \ldots, X_{k-1}) \leq \exp(s^2 M_k^2/8).$$

(Prove it by yourself, LOL)

We then have

$$\begin{aligned}
P(f(X) - \mathbb{E}f(X) \geq t) &\leq \exp(-st)\mathbb{E}(\exp(s(f(X) - \mathbb{E}f(X)))) \\
&= \exp(-st)\mathbb{E}\exp(s(Z_n - Z_{n-1} + Z_{n-1} - Z_0)) \\
&= \exp(-st)\mathbb{E}\exp(s\sum(Z_i - Z_{i-1})) \\
&= \exp(-st)\mathbb{E}\left\{\exp s\sum_{i=1}^{n-1}(Z_i - Z_{i-1})\mathbb{E}(\exp(s(Z_n - Z_{n-1}))|X_1, \ldots, X_{n-1})\right\} \\
&\leq \exp(-st)\exp(s^2 M_n^2/8)\mathbb{E}\exp(s\sum_{i=1}^{n-1}(Z_i - Z_{i-1})) \\
&\leq \exp(-st)\prod\exp(s^2 M_i^2/8).
\end{aligned}$$

The rest is straightforward. □

Example 23. *For i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n) \sim (X, Y)$, the correlation between X, Y is calculated by Kendall's tau:*

$$\widehat{\tau} := \frac{2}{n(n-1)}\sum \operatorname{sign}(X_i - X_{i'})\operatorname{sign}(Y_i - Y_{i'}).$$

Prove the concentration for $\widehat{\tau} - \mathbb{E}\widehat{\tau}$ by using the McDiarmid's inequality. [In-class exercise]

0.7 Subgaussian and subexponential

In the previous sections, we have introduced the subgaussian distribution. In this section, we will present some more fundamental relations between subgaussian and the r.v.'s L_p norms. To this end, we first define the subgaussian norm:

Definition 24 (subgaussian norm)**.** *The subgaussian norm of a r.v. $X \in \mathbb{R}$ is defined as follows:*

$$\|X\|_{\psi_2} := \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}.$$

Lemma 25. *The following assertions are equivalent:*

1. $P(|X - \mathbb{E}X| > t) \leq -\exp(-t^2/C_1^2)$ *for all* $t \geq 0$*;*

2. $(\mathbb{E}|X - \mathbb{E}X|^p)^{1/p} \leq C_2\sqrt{p}$ *for all* $p \geq 1$*;*

3. $\mathbb{E}\exp((X - \mathbb{E}X)^2/C_3^2) \leq e$*;*

4. $\mathbb{E}\exp(t(X - \mathbb{E}X)) \leq \exp(C_4^2 t^2)$.

Lemma 25 shows $\|X\|_{\psi_2} < \infty$ is equivalent to saying X is subgaussian. The proof is the proof of Lemma 5.5 in Roman Vershynin's note "Introduction to the non-asymptotic analysis of random matrices".

In statistics, we are not only interested in deriving Gaussian sample mean estimation, but also Gaussian sample variance estimation. Accordingly, we also need to study the chi-square-type distributions. It is well known that the chi-square distribution is much more tricky than the Gaussian distribution. Fortunately, we have a unified framework to tackle it.

Let's first define the subexponential norm.

Definition 26 (subexponential norm)**.** *The subexponential norm of a r.v. $X \in \mathbb{R}$ is defined as follows:*

$$\|X\|_{\psi_1} := \sup_{p \geq 1} p^{-1} (\mathbb{E}|X|^p)^{1/p}.$$

Similar to the subgaussian, a r.v. X is said to be subexponential iff $\|X\|_{\psi_1} < \infty$.

Example 27. *Show that chi-square distribution is subexponential.*

Lemma 28. *A r.v. X is subgaussian if and only if X^2 is subexponential:*

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2.$$

The most useful result regarding subexponential distributions is that they have similar tail performance as the subgaussian.

Lemma 29. *Assume X is subexponential. Then, for any t s.t. $|t| \leq c/\|X - \mathbb{E}X\|_{\psi_1}$, we have*

$$\mathbb{E}\exp(t(X - \mathbb{E}X)) \leq \exp(Ct^2\|X - \mathbb{E}X\|_{\psi_1}^2).$$

0.8 Dimension-free and Talagrand concentration inequalities

0.8.1 A motivating example

In many cases, we wish to reduce the data dimension while preserving the topology of the original data. A natural way is to reduce the dimension while preserving the original data's pairwise distances (up to the scale). MDS is a way for doing so deterministically. Another track, called the random projection, aims to achieve this goal using the probabilistic method (Paul Erdos!).

Theorem 30 (Johnson-Lindenstrauss Lemma)**.** *Let $t \in (0, 1/2)$. Let $Q \subset \mathbb{R}^p$ be a set of n points and $k = 20\log n/t^2$. Then there exists a mapping $f : \mathbb{R}^p \to \mathbb{R}^k$ s.t. for any $u, v \in Q$, we have*

$$(1-t)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1+t)\|u - v\|^2.$$

Proof. The proof is constructive and a classic use of the *probabilistic method.* We choose f as follows:

$$f(x) = \frac{1}{\sqrt{k}} Ax$$

where $A \in \mathbb{R}^{k\times d}$ with entries as i.i.d. $N(0,1)$. Then we have

$$\begin{aligned}
&P(\text{there exist } u, v, \text{ s.t. } (1-t)\|u-v\|^2 \le \|\frac{1}{\sqrt{k}}A(u-v)\|^2 \le (1+t)\|u-v\|^2 \text{ does not hold})\\
\le &\sum_{u,v\in Q} P((1-t)\|u-v\|^2 \le \|\frac{1}{\sqrt{k}}A(u-v)\|^2 \le (1+t)\|u-v\|^2 \text{ does not hold})\\
\le &2n^2\exp(-(t^2-t^3)k/4)\\
< &1,
\end{aligned}$$

when we choose $k = 20\log n/t^2$. In the second inequality we use the fact $Ax/\|x\| \sim N(0, I_k)$, the Gaussian dimension-free concentration inequality, and the subexponential concentration inequality introduced in Lemma 29. □

0.8.2 Dimension-free and Talagrand inequalities

Recall that, in establishing the random projection properties, we require the χ^2_k distribution of df k, $\sum_{i=1}^k X_i^2$, to be sub-exponential distributed, where $X_i \overset{i.i.d}{\sim} N(0,1)$ and k is arbitrary (in particular, k could increase to infinity).

At the first sight, this is a very counter-intuitive phenomenon: it implies, though the mean of the χ^2_k increases to infinity, the variance remains stable.

By the relation between subgaussian and subexponential distributions, we know it is equivalent to showing $\|\boldsymbol{X}\|_2$ is subgaussian, where $\boldsymbol{X} = (X_1,\ldots,X_k)^T \sim N_k(\mathbf{0},\mathbf{I}_k)$ and $\|\cdot\|_2$ stands for the Euclidean norm. This property is established by the celebrated Gaussian dimension-free concentration inequality, which is repeatedly used in my research.

Theorem 31 (Gaussian dimension-free concentration inequality)**.** *Let $\boldsymbol{X} = (X_1,\ldots,X_p)^T \sim N_p(\mathbf{0},\mathbf{I}_p)$ and let $f(\cdot):\mathbb{R}^p\to\mathbb{R}$ be a 1-Lipschitz function (i.e., $|f(\boldsymbol{x})-f(\boldsymbol{y})| \le \|\boldsymbol{x}-\boldsymbol{y}\|_2$ for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^p$ for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^p$). Then for any $t>0$, we have*

$$P(|f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X})| \ge t) \le C\exp(-ct^2)$$

for some absolute constants $C, c > 0$.

There are one thousand ways to prove this famous result. The following proof, which is the most elementary one, comes from Maurey and Pisier, stated in "Topics in Random Matrix Theory" by Terrence Tao. I personally like the proof using the entropy method and log-Sobolev inequality the most. Please check the note I wrote years ago on this topic.

Proof. WLOG, we assume $\mathbb{E}f(\boldsymbol{X}) = 0$ and focus on proving

$$P(f(\boldsymbol{X}) \ge t) \le C\exp(-ct^2),$$

or equivalently,

$$\mathbb{E}\exp(tf(\boldsymbol{X})) \le \exp(Ct^2).$$

By a routine limiting argument, we may assume $f(\cdot)$ to be smooth. The Lipschiz condition then is equivalent to the following one:

$$\|\nabla f(\boldsymbol{x})\|_2 \le 1 \text{ for all } \boldsymbol{x}\in\mathbb{R}^p.$$

We introduce a second copy of $\boldsymbol{X}$: $\boldsymbol{Y}$. Since $\mathbb{E}f(\boldsymbol{Y}) = 0$, Jensen's inequality yields

$$\mathbb{E}\exp(-tf(\boldsymbol{Y})) \ge 1$$

and thus

$$\mathbb{E}\exp(tf(\boldsymbol{X})) \leq \mathbb{E}\exp(t(f(\boldsymbol{X}) - f(\boldsymbol{Y}))).$$

Using the routine Ornstein-Uhlenbeck process argument, we write

$$f(\boldsymbol{X}) - f(\boldsymbol{Y}) = \int_0^{\pi/2} \frac{d}{d\theta} f(\boldsymbol{Y}\cos\theta - \boldsymbol{X}\sin\theta) d\theta,$$

where we have

- $\boldsymbol{X}_\theta := \boldsymbol{Y}\cos\theta - \boldsymbol{X}\sin\theta$ is another standard Gaussian;
- $\boldsymbol{X}_\theta' := -\boldsymbol{Y}\sin\theta - \boldsymbol{X}\cos\theta$ is another standard Gaussian and independent of $\boldsymbol{X}_\theta$.

By Jensen's inequality, we have

$$\exp(t(f(\boldsymbol{X}) - f(\boldsymbol{Y}))) \leq \frac{2}{\pi}\int_0^{\pi/2} \exp\left(\frac{\pi t}{2}\frac{d}{d\theta}f(\boldsymbol{X}_\theta)\right) d\theta,$$

(This is via the argument

$$\exp\left(t \cdot \int_a^b f(x)dx\right) = \exp\left(t(b-a)\cdot \int_a^b \frac{f(x)}{b-a}dx\right) \leq \frac{\int_a^b \exp(t(b-a)f(x))dx}{b-a}.$$

) which further implies

$$\mathbb{E}\exp(t(f(\boldsymbol{X}) - f(\boldsymbol{Y}))) \leq \frac{2}{\pi}\int_0^{\pi/2} \mathbb{E}\exp\left(\frac{\pi t}{2}\cdot\langle \nabla f(\boldsymbol{X}_\theta), \boldsymbol{X}_\theta'\rangle\right) d\theta.$$

Conditioning on $\boldsymbol{X}_\theta$ (reminding $\boldsymbol{X}_\theta$ is independent of $\boldsymbol{X}_\theta'$), $\langle \frac{\pi t}{2}\nabla f(\boldsymbol{X}_\theta), \boldsymbol{X}_\theta'\rangle$ is normally distributed with standard deviation at most $\pi t/2$. This further implies

$$\mathbb{E}\exp\left(\frac{\pi t}{2}\cdot\langle \nabla f(\boldsymbol{X}_\theta), \boldsymbol{X}_\theta'\rangle\right) \leq \exp(Ct^2),$$

and thus completes the proof. □

Remark 32. *Show that the Gaussian dimension-free concentration inequality is equivalent to the inequality*

$$P(\boldsymbol{X}\in A)P(\boldsymbol{X}\notin A_t) \leq C\exp(-ct^2), \quad \text{where } \boldsymbol{X}\sim N_p(\boldsymbol{0}, \mathbf{I}_p),$$

holding for all $t > 0$, all measurable set A, and A_t is the t-neighborhood of A.

The Gaussian dimension-free concentration inequality reveals another one of the magics surrounding the Gaussian. However, this inequality, in a weaker version, can apply to a much more general family of distributions. This is via the celebrated Talagrand concentration inequality.

Theorem 33. *(Talagrand concentration inequality) Let $K > 0$ and $X_1, \ldots, X_p$ be independent random variables with $|X_j| \leq K$ for $1 \leq j \leq p$. Let $f(\cdot): \mathbb{R}^p \to \mathbb{R}$ be a 1-Lipschitz convex function. Then for any $t > 0$ one has*

$$P(|f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X})| \geq tK) \leq C\exp(-ct^2).$$

Here $X_1, \ldots, X_p$ could also be complex-valued, and $\mathbb{E}f(\boldsymbol{X})$ could be replaced by the median of $f(\boldsymbol{X})$.

The proof is very simple, yet revealing via the entropy method. Due to the scope limit, I won't cover the details here. Students of interest should refer to Terrence Tao's Random Matrix Theory book (Pages 86 to 91).