Generic Chaining and the singular values of random matrices with iid rows

Shahar Mendelson

The basic set-up

Let F be a set of mean-zero functions on (Ω, μ) and let $\sigma = (X_1, ..., X_N)$ be independent, distributed according to μ^N.

If $\phi : \mathbb{R} \to \mathbb{R}$ is a reasonable function, is

$$\sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \phi(f)(X_i) - \mathbb{E}\phi(f) \right|$$

small, and if so, why?

- $\phi(t) = t \implies$ Uniform law of large numbers
- $\phi(t) = t^2 \implies$ Uniform CLT

The goal

- We will focus on $\phi(t) = t^2$.

Question: What is the right notion of complexity $\mathrm{Comp}(F, \mu)$ and the right norm $\| \ \|$ for which

$$\mathbb{E} \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} f^2(X_i) - \mathbb{E} f^2 \right| \lesssim (\sup_{f \in F} \|f\|) \frac{\mathrm{Comp}(F, \mu)}{\sqrt{N}} + \frac{\mathrm{Comp}^2(F, \mu)}{N}?$$

Question: Does the correct notion of complexity have a geometric interpretation?

A class of particular interest is $F = \{\langle t, \cdot \rangle : t \in S^{n-1}\}$.

What can we hope for?

Recall that for a metric space (T, d),

$$\gamma_\alpha(T,d) = \inf \sup_{t\in T} \sum_{s\geq 0} 2^{s/\alpha} d(t, \pi_s(t)),$$

where the infimum is taken with respect to all admissible sequences of T.

- Very easy (contraction):

$$\mathbb{E} \sup_{f\in F} \left| \frac{1}{N} \sum_{i=1}^N f^2(X_i) - \mathbb{E} f^2 \right| \lesssim (\sup_{f\in F} \|f\|_{L_\infty}) \frac{\gamma_2(F,\psi_2)}{\sqrt{N}}.$$

- $\gamma_2(F,\psi_2)$ cannot be improved to any $\gamma_2(F,\psi_\alpha)$, $\alpha < 2$.

- $\sup_{f\in F} \|f\|_{L_\infty}$ cannot be improved beyond $\sup_{f\in F} \|f\|_{L_{4+\varepsilon}}$ (depending on what we mean by "high probability").

What can we hope for?

So the best we can hope for: with high probability and in expectation,

$$\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f^2(X_i) - \mathbb{E}f^2\right| \lesssim \sup_{f\in F}\|f\|_{L_q}\frac{\mathrm{Comp}(F,\mu)}{\sqrt{N}} + \frac{\mathrm{Comp}(F,\mu)}{N},$$

where $q > 4$ and $\mathrm{Comp}(F,\mu)$ is not very far from $\gamma_2(F,\psi_2)$.

Examples I

Special case: If μ is the gaussian measure on $\mathbb{R}^n$, $T \subset \mathbb{R}^n$ and $F = \{\langle t, \cdot \rangle : t \in T\}$ then

$$\mathrm{Comp}(F, \mu) \lesssim \gamma_2(F, \psi_2(\mu)) \sim \gamma_2(F, L_2(\mu)).$$

By the Majorizing Measures Theorem, $\gamma_2(F, L_2(\mu)) \sim \mathbb{E} \sup_{f \in F} G_f$, (if $f = \langle t, \cdot \rangle$, $g_f = \sum_{i=1}^n g_i t_i$). Thus

$$\mathbb{E} \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^N f^2(X_i) - \mathbb{E} f^2 \right| \lesssim (\sup_{f \in F} \|f\|_{L_2(\mu)}) \frac{\mathbb{E} \sup_{f \in F} G_f}{\sqrt{N}} + \frac{(\mathbb{E} \sup_{f \in F} G_f)^2}{N}.$$

So $\mathrm{Comp}(F, \mu)$ is equivalent to $\gamma_2(F, \psi_2)$ and has a geometric meaning - the "mean-width" relative to the gaussian measure.

But is something like this always possible?

Examples II - Special families of measures

Theorem (M-Pajor-Tomczak Jaegermann 07): If F is a class of functions, then in expectation and with high probability,

$$\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f^2(X_i) - \mathbb{E}f^2\right| \lesssim \sup_{f\in F}\|f\|_{\psi_2}\frac{\gamma_2(F,\psi_2)}{\sqrt{N}} + \frac{\gamma_2^2(F,\psi_2)}{N}.$$

If $F \cup \{0\}$ is L-subgaussian relative to μ (i.e. $\|f-h\|_{\psi_2(\mu)} \le L\|f-h\|_{L_2(\mu)}$) then

$$\mathrm{Comp}(F,\mu) \lesssim \gamma_2(F,\psi_2) \lesssim_L \gamma_2(F,L_2) \sim \mathbb{E}\sup_{f\in F} G_f,$$

and the "subgaussian complexity" is dominated by the "gaussian" one.

Examples III - Special families of measures

Theorem (M-2010): If F is a class of functions then in expectation and with high probability,

$$\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f^2(X_i) - \mathbb{E}f^2\right| \lesssim \sup_{f\in F}\|f\|_{\psi_1}\frac{\gamma_2(F,\psi_2)}{\sqrt{N}} + \frac{\gamma_2^2(F,\psi_2)}{N}.$$

Here, even if the ψ_1 and L_2 norms are equivalent on F, then

$$\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f^2(X_i) - \mathbb{E}f^2\right| \lesssim L\sup_{f\in F}\|f\|_{L_2}\frac{\gamma_2(F,\psi_2)}{\sqrt{N}} + \frac{\gamma_2(F,\psi_2)}{N},$$

but here $\gamma_2(F,\psi_2)$ does not have a clear geometric interpretation!

Some improvement...[M-Paouris]

- Note that $\|f\|_{\psi_2} \sim \sup_{p\geq 2} \frac{\|f\|_{L_p}}{\sqrt{p}}$.
- Let

$$\bar{\gamma}_2(F,\mu) = \inf \sup_{f\in F} \sum_{s\geq 0} 2^{s/2} \|f - \pi_s(f)\|_{(2^s)},$$

where $\|f\|_{(2^s)} = \sup_{1\leq p\leq 2^s} \|f\|_{L_p}/\sqrt{p}$.
Note that $\bar{\gamma}_2(F,\mu) \lesssim \gamma_2(F,\psi_2)$.

It is relatively easy to extend the previous result and replace $\gamma_2(F,\psi_2)$ by $\bar{\gamma}_2(F,\mu)$:

$$\sup_{f\in F} \left| \frac{1}{N} \sum_{i=1}^{N} f^2(X_i) - \mathbb{E}f^2 \right| \lesssim \sup_{f\in F} \|f\|_{\psi_1} \frac{\bar{\gamma}_2(F,\mu)}{\sqrt{N}} + \frac{\bar{\gamma}_2^2(F,\mu)}{N},$$

and in some cases this has a geometric meaning!

Geometry again

Claim: If μ is isotropic, log-concave and 1-unconditional on $\mathbb{R}^n$, then

$$\bar{\gamma}_2(T,\mu) \lesssim \mathbb{E}\sup_{t\in T}\langle X,t\rangle \equiv E(T),$$

where X has iid exponentials coordinates.

If $T \subset \mathbb{R}^n$, then in expectation and with high probability,

$$\sup_{t\in T}\left|\frac{1}{N}\sum_{i=1}^{N}\langle X_i,t\rangle^2 - |t|^2\right| \lesssim \sup_{t\in T}|t|\frac{E(T)}{\sqrt{N}} + \frac{E^2(T)}{N}.$$

Question: Is there a stronger result hidden here? When is

$$\bar{\gamma}_2(T,\mu) \sim \mathbb{E}_\mu \sup_{t\in T}\langle t,\cdot\rangle?$$

But this is a topic for another lecture.....

Heavy tails - a general class

Theorem (M-Paouris): For every $q > 4$ there exist constants c_1, c_2, c_3 that depend only on q and for which the following holds. If $F \subset L_q$, and $u \geq c_1$, then with μ^N-probability at least $1 - c_2/(uN)^{c_3}$,

$$\sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} f^2(X_i) - \mathbb{E} f^2 \right| \lesssim_u \left(\sup_{f \in F} \|f\|_{L_q} \frac{\bar{\gamma}_2(F, \mu)}{\sqrt{N}} + \frac{\bar{\gamma}_2^2(F, \mu)}{N} \right).$$

Heavy tails–main ideas

• No hope of balancing the tails of each $N^{-1}\sum_{i=1}^N f^2(X_i) - \mathbb{E}f^2$ and the complexity of the class....

Therefore, the method will be based on

• Symmetrization:

$$\mathbb{E}\sup_{f\in F}\left|\sum_{i=1}^N (f^2(X_i) - \mathbb{E}f^2)\right| \sim \mathbb{E}\sup_{f\in F}\left|\sum_{i=1}^N \varepsilon_i f^2(X_i)\right|.$$

• Analysis of the structure of $P_\sigma F = \left\{(f(X_i))_{i=1}^N : f \in F\right\}$.

Heavy tails II - main ideas

Consider $V = P_\sigma F = \left\{ (f(X_i))_{i=1}^N : f \in F \right\}$

- Note that for every $v \in V$ and every $k \leq N$, with probability at least $1 - \exp(-ct^2)$

$$|\sum_{i=1}^N \varepsilon_i v_i^2| \leq \sum_{i=1}^k (v_i^*)^2 + t(\sum_{i=k+1}^N (v_i^4)^*)^{1/2}.$$

- So we need high probability estimates on

$$\sup_{f \in F_s} \left(\sum_{i=1}^k (f^2(X_i))^* \right)^{1/2} \quad \text{and} \quad \sup_{f \in F_s} \left(\sum_{i=k+1}^N (f^4(X_i))^* \right)^{1/2},$$

- F_s comes from the chaining process and $k = k_s$.

Question: how does one generate "high enough" probability without concentration?

Heavy tails III - main ideas

- Let $|F| \leq \binom{N}{k}$ and assume that F is bounded in L_q for some $q > 4$.
- Fix $f \in F$ and consider $(f(X_1), ..., f(X_N)) \in P_\sigma F$.

Claim: Let $\varepsilon = q/4 - 1$. With probability at least $1 - \exp(-c\varepsilon tk \log(eN/k))$, for every $f \in F$, and every $j \geq tk$,

$$(f(X_i))_j^* \lesssim \|f\|_{L_q} \left(\frac{N}{j}\right)^{(1+\varepsilon)/q} \qquad \left(\Longrightarrow \left((f(X_i))_j^*\right)_{j \geq tk} \subset c_q' \|f\|_{L_q} B_{L_4^N}\right)$$

Indeed, this is true since

$$Pr(Y_j^* \geq u\|Y\|_{L_q}) \leq \binom{N}{j} Pr^j(|Y| \geq u\|Y\|_{L_q}) \leq \left(\frac{eN}{j}\right)^j \cdot u^{-qj}.$$

- It remains to bound $\sup_{f \in F}(\sum_{i=1}^k (f^*(X_i))^2)^{1/2}$ – the "global complexity".

Heavy tails and chaining

It is possible to incorporate this idea in a chaining argument:

With high probability, $P_\sigma F \subset V + U$, where

- $U \subset (\sup_{f \in F} \|f\|_{L_q}) B_{L_4^N}$.
- V comes from the "global complexity", is well bounded in ℓ_2^N and leads to the complexity term $\bar{\gamma}_2(F, \mu)$.
- U comes from all the "small" coordinates of each link in the chain $\pi_s(f) - \pi_{s-1} f$ and is determined by the "diameter" of F in L_q.
- All the structure of F should be captured in V.

Special case: $T = S^{n-1}$

Consider the following situation: fix $N \geq n$ and assume that

- μ is an isotropic measure on $\mathbb{R}^n$.
- $|X| \leq (Nn)^{1/4}$ almost surely.
- $\sup_{t \in S^{n-1}} \|\langle t, \cdot \rangle\|_{\psi_1} \leq L$.

Theorem (ALLPT): Under the assumptions above, with probability at least $1 - 2\exp(-c\sqrt{n})$,

$$\sup_{t \in S^{n-1}} \left| \frac{1}{N} \sum_{i=1}^{N} \langle X_i, t \rangle^2 - 1 \right| \lesssim \sqrt{\frac{n}{N}}.$$

A proof using similar ideas was recently given by Talagrand, with a better probability estimate of $1 - 2\exp(-c(Nn)^{1/4}) - 2\exp(-cn)$.

$T = S^{n-1}$ - heavy tails

Theorem (M-Paouris): if $\sup_{t \in S^{n-1}} \|\langle t, \cdot \rangle\|_{L_q} \leq L$ for some $q > 8$, then w.p. at least $1 - c_1 \left(\frac{1}{N^{c_2}} + \exp(-c_3 n) \right)$,

$$\sup_{t \in S^{n-1}} |\frac{1}{N} \sum_{i=1}^{N} \langle X_i, t \rangle^2 - 1| \leq c_4 \sqrt{\frac{n}{N}}.$$

- The same proof recovers the previous result under the ψ_1 assumption.
- Under a stronger assumption on X (unconditionality + slightly stronger boundedness), a similar result holds for $4 < q \leq 8$, because in this case

$$\bar{\gamma}_2(S^{n-1}, \mu) \lesssim \sqrt{n}.$$

- Gives optimal rate up to a factor of $\log(eN/n)$ for $2 < q \leq 4$.

Main ideas of the proof

- The sphere is simple - full chaining should not be different than the union bound, which is "one step" chaining.....

- The decomposition for set of $\exp(cn)$ points: "small coordinates" when $\log(\binom{N}{k_0}) \sim n$.

- "small coordinates" are controlled since $\sup_{t\in S^{n-1}} \|\langle t, \cdot\rangle\|_{L_q} \leq L$.

- The "global complexity" is

$$(\sup_{t\in S^{n-1}} \sum_{i=1}^{k_0} (\langle X_i, t\rangle^2)^*)^{1/2} = \sup_{a\in S^{n-1},\ |\mathrm{supp(a)}|\leq k_0} |\sum_{i=1}^{n} a_i X_i|.$$

This is exactly A_{k_0} defined in [ALPT] and [ALLPT]!

Main ideas of the proof II

- The key point: one can show that w.h.p. $A_{k_0} \lesssim (Nn)^{1/4}$.

- The rest follows from analysis of the Bernoulli process: with probability at least $1 - \exp(-ct^2)$

$$|\sum_{i=1}^{N} \varepsilon_i v_i^2| \leq \sum_{i=1}^{k_0} (v_i^*)^2 + t(\sum_{i=k_0+1}^{N} (v_i^4)^*)^{1/2}.$$

The first term is controlled by A_{k_0} and the second by the diameter in L_4^N – which is bounded by the L_q diameter of the sphere....