Concentration of measure in probability and high-dimensional statistical learning

Guillaume Aubrun, Aurélien Garivier, Rémi Gribonval

remi.gribonval@inria.fr

http://perso.ens-lyon.fr/remi.gribonval

Dimension reduction - summary

- **Model set** $\Sigma \subset \mathbb{R}^d$
- **Normalized secant set** $\mathcal{S} = \left\{ \frac{z}{\|z\|_2} : z \in \Sigma - \Sigma \right\} \subset \mathbb{S}^{d-1}$
- **Encoder** $E : \mathbb{R}^d \to \mathbb{R}^k$
  - ✓ linear encoder = *kxd* matrix **A**
- **Best RIP constant**

$$\delta^\star = \delta^\star(E, \Sigma) = \sup_{z \in \Sigma - \Sigma} \left| \frac{\|E(z)\|_2^2}{\|z\|_2^2} - 1 \right| = \sup_{u \in \mathcal{S}} \left| \|E(u)\|_2^2 - 1 \right|$$

  - ✓ Characterizes the existence of a stable decoder
  - ✓ Generally hard to compute, easier to bound

Quantitative dimension reduction ?

- Given model set $\Sigma \subset \mathbb{R}^d$, *for which dimension k* does there exist an encoder $E : \mathbb{R}^d \to \mathbb{R}^k$ with

$$\delta^\star(E, \Sigma) < 1$$

- Examples
  - ✓ linear model set — $k \geq \mathtt{dim}(\Sigma)$
  - ✓ finite set — Johnson-Lindenstrauss lemma — $k \geq C \log \sharp\Sigma$
  - ✓ {*s*-sparse vectors} — $k \geq Cs \log(d/s)$
  - ✓ {rank-*r* matrices} $d = p \times p$ — $k \geq Crp$

✦ How to build such encoders ? Where does the « dimensions » come from ?

random projections

covering numbers & « widths »

Random projections & coverings

Gaussian random projections

- **Construction:**
  - ✓ draw *k* i.i.d. vectors
  - ✓ build *kxd* matrix

$$\mathbf{a}_i \sim \mathcal{N}(0, \mathbf{I}_d)$$

$$\mathbf{A} = \frac{1}{\sqrt{k}} \begin{pmatrix} \mathbf{a}_1^\top \\ \ldots \\ \mathbf{a}_k^\top \end{pmatrix}$$

- **Properties**
  - ✓ « isotropy » $\langle \mathbf{a}_i, z \rangle \sim \mathcal{N}(0, \|z\|_2^2), \; \forall z \in \mathbb{R}^d$

$$\mathbb{E}\langle \mathbf{a}_i, z \rangle^2 = \|z\|_2^2$$

  - ✓ energy preservation :
    - ✦ in expectation

$$\|E(z)\|_2^2 = \|\mathbf{A}z\|_2^2 = \frac{1}{k} \sum_{i=1}^{k} \langle \mathbf{a}_i, z \rangle^2 \qquad \mathbb{E}[\|E(z)\|_2^2] = \|z\|_2^2, \; \forall z \in \mathbb{R}^d$$

    - ✦ + pointwise concentration

Gaussian random projections

- **Pointwise concentration**

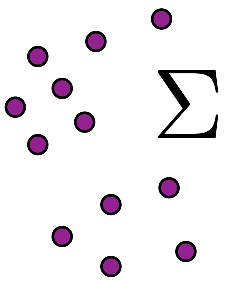
$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{i=1}^{k}\langle \mathbf{a}_i, u\rangle^2 - 1\right| \geq t\right) \leq 2\exp(-kc(t)), \forall u \in \mathcal{S} \subset \mathbb{S}^{d-1}$$

- **Uniform result on normalized secant set ?**

$$\mathbb{P}\left(\sup_{u\in\mathcal{S}}\left|\frac{1}{k}\sum_{i=1}^{k}\langle \mathbf{a}_i, u\rangle^2 - 1\right| \geq t\right) \leq ?$$

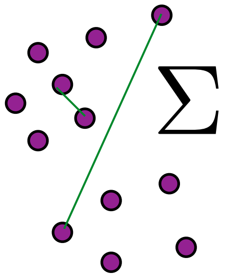
Revisiting Johnson-Lindenstrauss's lemma

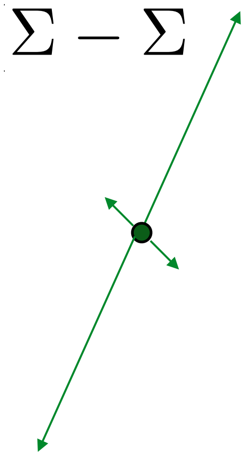
- **Finite model set**



Revisiting Johnson-Lindenstrauss's lemma

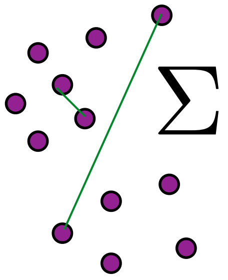
- **Finite model set**
  - ✓ Finite secant $\sharp(\Sigma - \Sigma) \leq (\sharp\Sigma)^2$

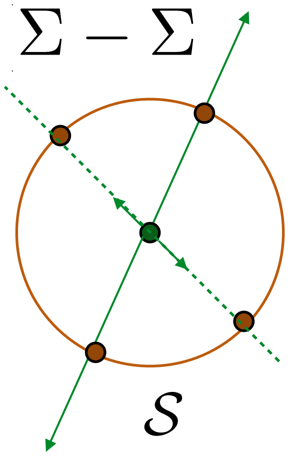




Revisiting Johnson-Lindenstrauss's lemma

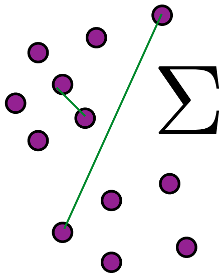
- **Finite model set**
  - ✓ Finite secant $\sharp(\Sigma - \Sigma) \leq (\sharp\Sigma)^2$
  - ✓ Finite normalized secant $\sharp\mathcal{S} \leq (\sharp\Sigma)^2$



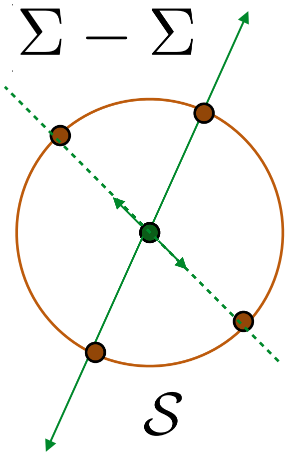


Revisiting Johnson-Lindenstrauss's lemma

- **Finite model set**



  - ✓ Finite secant $\sharp(\Sigma - \Sigma) \leq (\sharp\Sigma)^2$
  - ✓ Finite normalized secant $\sharp\mathcal{S} \leq (\sharp\Sigma)^2$



  - ✓ Johnson-Lindenstrauss lemma:
    - ✦ Pointwise concentration
    - ✦ + union bound

From pointwise concentration to the RIP ?

- **Linear subspace**

Σ

From pointwise concentration to the RIP ?

- **Linear subspace** Σ

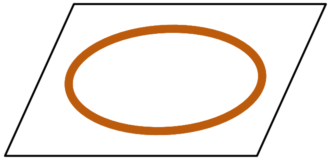
  - ✓ Linear secant $\Sigma - \Sigma = \Sigma$

From pointwise concentration to the RIP ?

- **Linear subspace** Σ

  - ✓ Linear secant $\Sigma - \Sigma$ = Σ

  - ✓ Spherical normalized secant $\mathcal{S}$

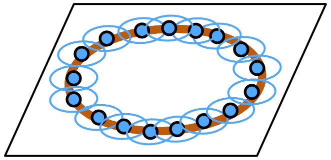


From pointwise concentration to the RIP ?

- **Linear subspace** Σ

  - ✓ Linear secant $\Sigma - \Sigma$ = Σ

  - ✓ Spherical normalized secant $\mathcal{S}$



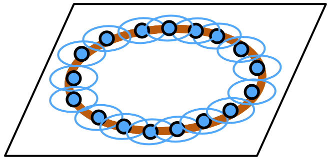
  - ✓ Finite *covering* $\hat{\mathcal{S}} \subset \mathcal{S}$

From pointwise concentration to the RIP ?

- **Linear subspace** Σ

  - ✓ Linear secant $\Sigma - \Sigma = \Sigma$

  - ✓ Spherical normalized secant $\mathcal{S}$



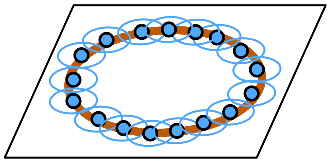
  - ✓ Finite *covering* $\hat{\mathcal{S}} \subset \mathcal{S}$
  - ✓ *Concentration + covering* $\mathbb{P}\left(\max_{u \in \hat{\mathcal{S}}} \left| \|E(u)\|_2^2 - 1 \right| \geq t\right) \leq \sharp 2\hat{\mathcal{S}} \exp(-kc(t))$

From pointwise concentration to the RIP ?

- **Linear subspace** Σ

  - ✓ Linear secant $\Sigma - \Sigma$ = Σ

  - ✓ Spherical normalized secant $\mathcal{S}$



  - ✓ Finite *covering* $\hat{\mathcal{S}} \subset \mathcal{S}$
  - ✓ *Concentration + covering* $\mathbb{P}\left(\max_{u\in\hat{\mathcal{S}}} \left| \|E(u)\|_2^2 - 1\right| \geq t\right) \leq \sharp 2\hat{\mathcal{S}} \exp(-kc(t))$

extension to $\mathcal{S}$? Lipschitz property / chaining (Dudley)

Lipschitz extension

- ϵ-covering of the normalized secant $U = \{u_i, 1 \leq i \leq q\}$
- With high probability, uniformly on *U*

$$\sqrt{1-t} \leq \|\mathbf{A}u_i\| \leq \sqrt{1+t}$$

  - ✓ When this holds, on the normalized secant, there is *i* s.t.

$$\|\mathbf{A}x\| \leq \|\mathbf{A}u_i\| + \|\mathbf{A}(x-u_i)\| \leq \sqrt{1+t} + \|\mathbf{A}\|\epsilon$$

    - ✦ similar lower bound
    - ✦ need to control operator norm = Lipschitz property

- Special case: linear model set

$$\|\mathbf{A}\| = \sup_{x \in \mathcal{S}} \|\mathbf{A}x\| \leq \sqrt{1+t} + \|\mathbf{A}\|\epsilon$$

$$\|\mathbf{A}\| \leq \frac{\sqrt{1+t}}{1-\epsilon}$$

Supremum of empirical processes
-width and complexities-

Supremum of empirical processes

- **Example 1:** the RIP

$$\sup_{u \in \mathcal{S}} \left| \frac{1}{k} \sum_{i=1}^{k} \langle \mathbf{a}_i, u \rangle^2 - 1 \right|$$

- **Example 2:** excess risk of ERM

$$\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \ell(X_i, h) - \mathbb{E}\ell(X, h) \right|$$

- **Goal:** given class $\mathcal{F}$ of real-valued functions and *n* i.i.d. random variables, we want to bound with high probability

$$\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}_{X'} f(X') \right)$$

NB: absolute values handled using $\tilde{\mathcal{F}} = \mathcal{F} \cup (-\mathcal{F})$

Typical approach

- **Step 1**: approximate bound with expected supremum
  - ✓ e.g. with McDiarmid: with probability at least $1-\delta$

$$\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_{X'} f(X')\right) \leq \mathbb{E}_{\mathbf{X}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_{X'} f(X')\right)\right] + B\sqrt{\frac{2\log\delta}{n}}$$

- **Step 2**: symmetrization (cf A. Garrivier's course 5)

$$\mathbb{E}_{\mathbf{X}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_{X'} f(X')\right)\right] \leq 2\mathbb{E}_{\mathbf{X},\boldsymbol{\epsilon}} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)$$

ϵ_i, i.i.d. Rademacher(± 1 with probability $1/2$)

  - ✓ notion of Rademacher complexity

- **Step 3**: bound with Gaussian width

Rademacher complexities

- **Goal**: control $$\mathbb{E}_{\mathbf{X},\epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)$$
- **Definition**: *Empirical Rademacher complexity* of $\mathcal{F}$ with respect to fixed sample $\mathbf{X} = (X_1, \ldots, X_n)$

$$\hat{\mathrm{R}}_{\mathbf{X}}(\mathcal{F}) = \mathbb{E}_{\epsilon} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)$$

  - « correlation » of function class with random (+/-1) noise *on the sample*
  - the higher the more complex the class *on the sample*
- *Rademacher complexity* $\mathbb{E}_{\mathbf{X}}\hat{\mathrm{R}}_{\mathbf{X}}(\mathcal{F})$

Rademacher complexity of a set

- **Geometrically** $T = T_{\mathbf{X}}(\mathcal{F}) = \{(f(X_i))_{i=1}^n : f \in \mathcal{F}\} \subset \mathbb{R}^n$
  - ✓ we can rewrite

$$\hat{\mathbb{R}}_{\mathbf{X}}(\mathcal{F}) = \frac{1}{n} \cdot \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{t \in T} \langle \boldsymbol{\epsilon}, t \rangle$$

- **Definition**: Rademacher complexity of a set $T \subset \mathbb{R}^n$

$$\mathtt{R}(T) = \frac{1}{n} \cdot \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{t \in T} \langle \boldsymbol{\epsilon}, t \rangle, \quad \boldsymbol{\epsilon} \sim \text{Rademacher}$$

  - ✦ see e.g [Shalev-Schwarz & Ben David, 26.1]

Gaussian width

- **Rademacher complexity** $T' = \frac{1}{n} T \subset \mathbb{R}^n$

$$\mathtt{R}(T) = \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{t \in T'} \langle \boldsymbol{\epsilon}, t \rangle, \quad \boldsymbol{\epsilon} \sim \text{Rademacher}$$

- **What if Rademacher replaced with Gaussian ?**
  - ✓ supremum of *Gaussian* process
  - ✓ compatible with Slepian's lemma
  - ✓ invariance wrt rotations of the considered set
- **Definition**: *Gaussian width* of a set $T \subset \mathbb{R}^n$

$$w(T) = \mathbb{E}_{\mathbf{g}} \sup_{t \in T} \langle \mathbf{g}, t \rangle, \quad \mathbf{g} \sim \mathcal{N}(0, \mathbf{Id}_n)$$

« Bounding Rademacher with Gauss »

- **Property:** $\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{t \in T} \langle \boldsymbol{\epsilon}, t \rangle \leq \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{\mathbf{g}} \sup_{t \in T} \langle \mathbf{g}, t \rangle$

- **Proof ingredients**
  - ✓ if $g_i \sim \mathcal{N}(0,1)$ then $\mathbb{E}|g_i| = \sqrt{2/\pi}$
  - ✓ the normal distribution is symmetric
  - ✓ **EXERCISE**

Calculus with widths

Calculus with complexities / width

- **Properties** [Vershinyn, Proposition 7.5.2]
  - $w(T) < \infty$ iff *T* is bounded
  - **Invariance to unitary transformations**: for every unitary matrix **U** and any vector *y*, we have

$$w(\mathbf{U}T + y) = w(T)$$

  - Invariance to convex hulls

$$w(\texttt{conv}(T)) = w(T)$$

  - Minkowski sums and scaling

$$w(T+S) = w(T) + w(S) \qquad w(aT) = |a|w(T)$$

  - Moreover

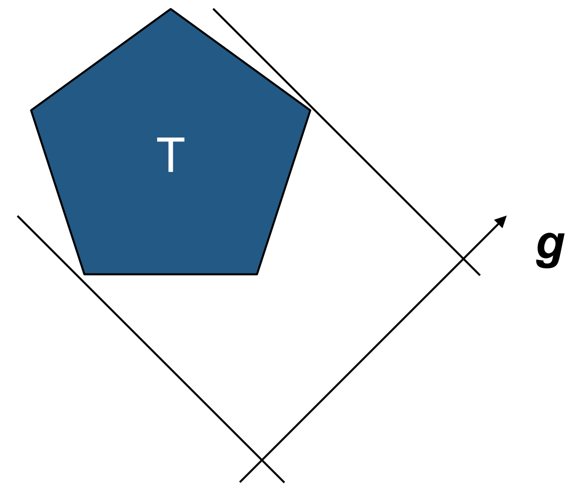
$$w(T) = \frac{1}{2}w(T-T) = \frac{1}{2}\mathbb{E}\sup_{x,y\in T}\langle \mathbf{g}, x-y\rangle \qquad \frac{1}{\sqrt{2\pi}}\texttt{diam}(T) \leq w(T) \leq \frac{\sqrt{n}}{2}\texttt{diam}(T)$$

Proof

- **HOMEWORK**:
  - ✓ prove the first 4 properties
  - ✓ which one(s) also apply to the Rademacher complexity ?
- **EXERCISE**: prove the remaining properties

Geometric interpretation

- **Why « width » ?**
  - ✓ remember the property $$w(T) = \frac{1}{2}w(T-T) = \frac{1}{2}\mathbb{E}\sup_{x,y\in T}\langle \mathbf{g}, x-y\rangle$$



  - ✓ … except that ***g*** is not unit norm

Yet another width

- **Definition :** *spherical width* of a set

$$w_s(T) = \mathbb{E}_{\boldsymbol{\theta}} \sup_{t \in T} \langle \boldsymbol{\theta}, t \rangle \qquad \boldsymbol{\theta} \sim \mathrm{Unif}(\mathbb{S}^{n-1})$$

  - ✓ aka « mean width »
- **Property:**
  - ✓ for each *n* there is a constant such that: $\frac{w(T)}{w_s(T)} = c_n, \ \forall T \subset \mathbb{R}^n$
- **EXERCISE**
  - ✓ prove the property
  - ✓ what can you say about the constant ?

Examples

Examples

- **Exercise**: estimate the Gaussian width of
  - ✓ the Euclidean unit ball
  - ✓ the Euclidean unit sphere
  - ✓ the unit cube $[-1,1]^n$
  - ✓ the unit ball of the L1 norm
  - ✓ a finite set of points

Summary

- **Various notion of complexities and width**
  - ✓ Rademacher complexity $n\mathtt{R}(T) = \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{t \in T} \langle \boldsymbol{\epsilon}, t \rangle, \quad \boldsymbol{\epsilon} \sim \text{Rademacher}$
  - ✓ Gaussian width $w(T) = \mathbb{E}_{\mathbf{g}} \sup_{t \in T} \langle \mathbf{g}, t \rangle, \quad \mathbf{g} \sim \mathcal{N}(0, \mathbf{Id}_n)$
  - ✓ square version $h^2(T) = \mathbb{E}_{\mathbf{g}} \sup_{t \in T} \langle \mathbf{g}, t \rangle^2$
  - ✓ Gaussian complexity $\gamma(T) = \mathbb{E}_{\mathbf{g}} \sup_{t \in T} |\langle \mathbf{g}, t \rangle|$
  - ✓ spherical width $w_s(T) = \mathbb{E}_{\boldsymbol{\theta}} \sup_{t \in T} \langle \boldsymbol{\theta}, t \rangle \quad \boldsymbol{\theta} \sim \text{Unif}(\mathbb{S}^{n-1})$

$$n\mathtt{R}(T) = \frac{n}{2}\mathtt{R}(T-T) \lesssim w(T) = \tfrac{1}{2} w(T-T) = c_n w_s(T) \asymp h(T-T) \asymp \gamma(T-T) \lesssim \gamma(T)$$

That's all folks !