Asymmetric compressive learning guarantees with applications to quantized sketches

Vincent Schellekens

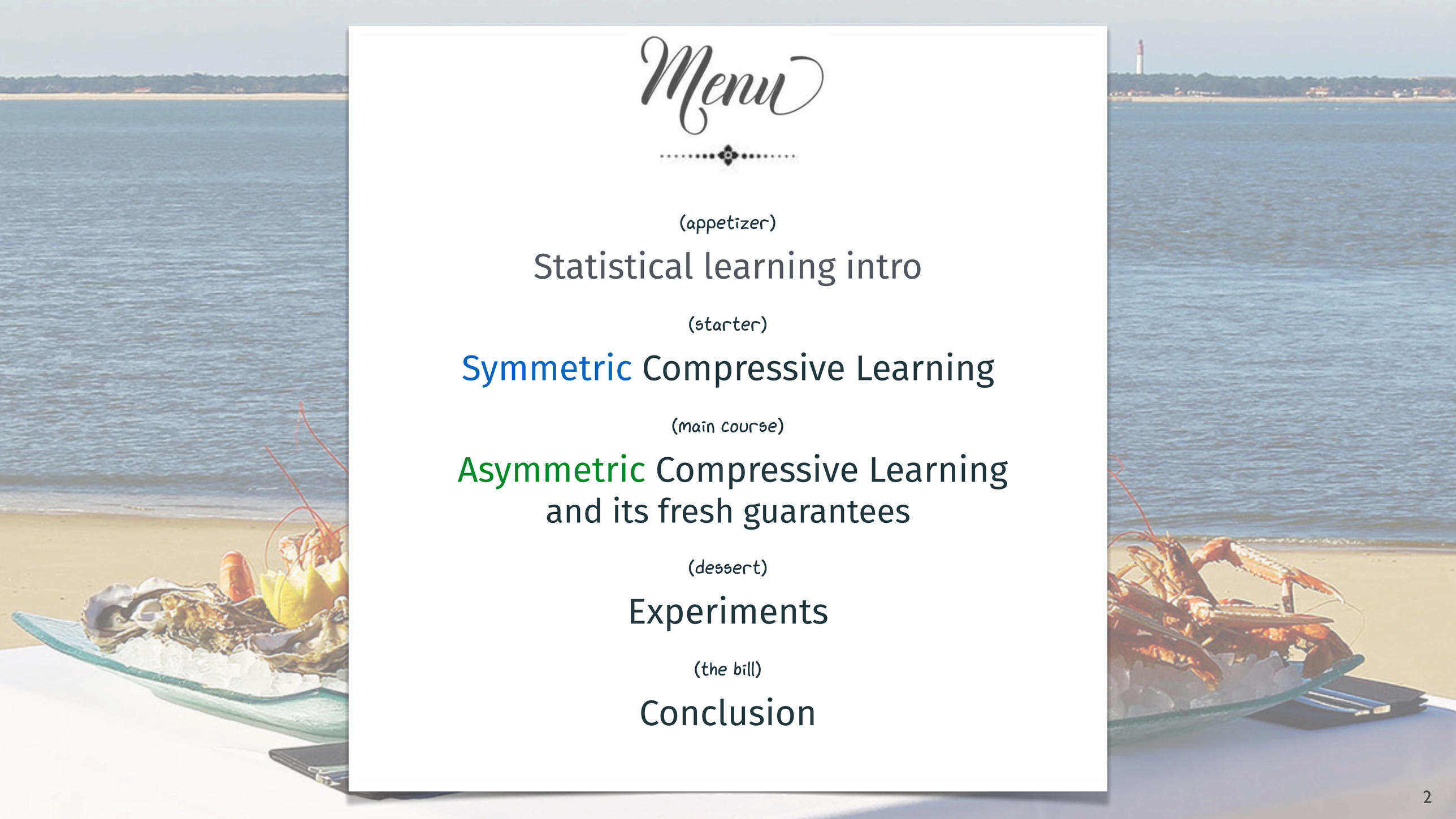
CEA-List, Saclay

Laurent Jacques

INMA, UCLouvain, Belgium

Curves and Surfaces 2022

Arcachon, France, June 20 - June 24



Menu

(appetizer)

Statistical learning intro

(starter)

Symmetric Compressive Learning

(main course)

Asymmetric Compressive Learning
and its fresh guarantees

(dessert)

Experiments

(the bill)

Conclusion

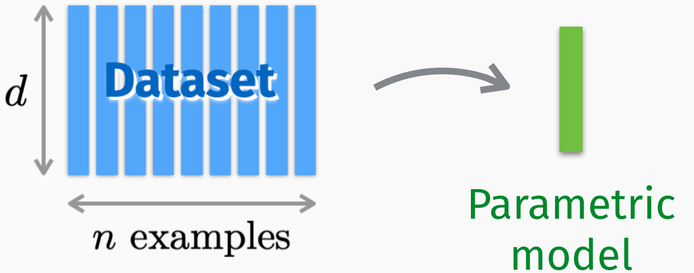
Statistical learning context

Minimize the empirical risk (ERM)

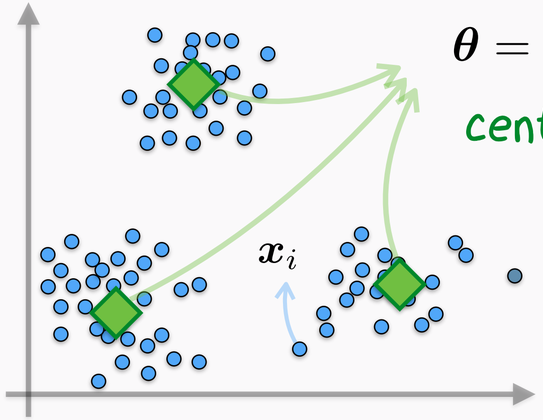
$$\widetilde{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i) \equiv \text{"Pooling of unhappiness"}$$

$\ell \equiv$ "How unhappy $\boldsymbol{x}_i$ is with the model $\boldsymbol{\theta}$"

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n \qquad \boldsymbol{\theta}$$



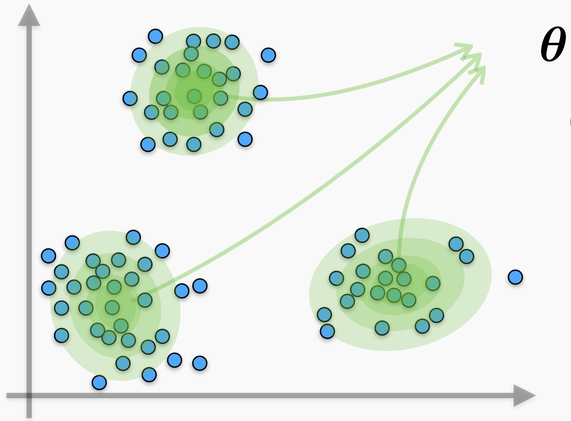
Example: k-means clustering



$$\boldsymbol{\theta} = \{\boldsymbol{c}_k\}_{k=1}^K$$

centroids

Example: GMM fitting



$$\boldsymbol{\theta} = \{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Gamma}_k\}_{k=1}^K$$

Gaussian mixture params.

Sum of Squared Errors (SSE): $\sum_{i=1}^{n} \min_k \|\boldsymbol{x}_i - \boldsymbol{c}_k\|^2$

Maximum likelihood: $\sum_i -\log\left(\sum_k w_k \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Gamma}_k)\right)$

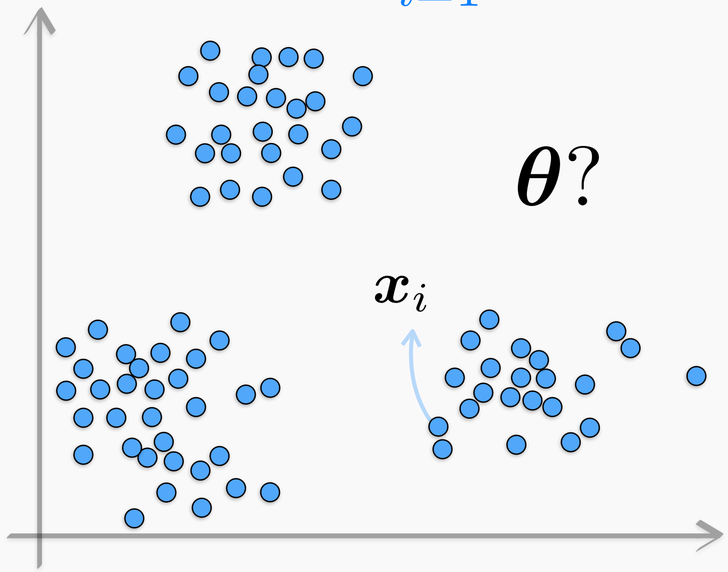
Statistical learning context

Minimize the empirical risk (ERM)

$$\widetilde{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)}_{\mathbb{E}_{\boldsymbol{x}\sim\hat{\mathcal{P}}_{\mathcal{X}}}\ell(\boldsymbol{\theta},\boldsymbol{x})} = \underbrace{\mathcal{R}(\boldsymbol{\theta}; \widehat{\mathcal{P}}_{\mathcal{X}})}_{\text{Risk function}}$$

Empirical distribution

$$\widehat{\mathcal{P}}_{\mathcal{X}} \sim \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}$$



Statistical learning context

Minimize the empirical risk (ERM)

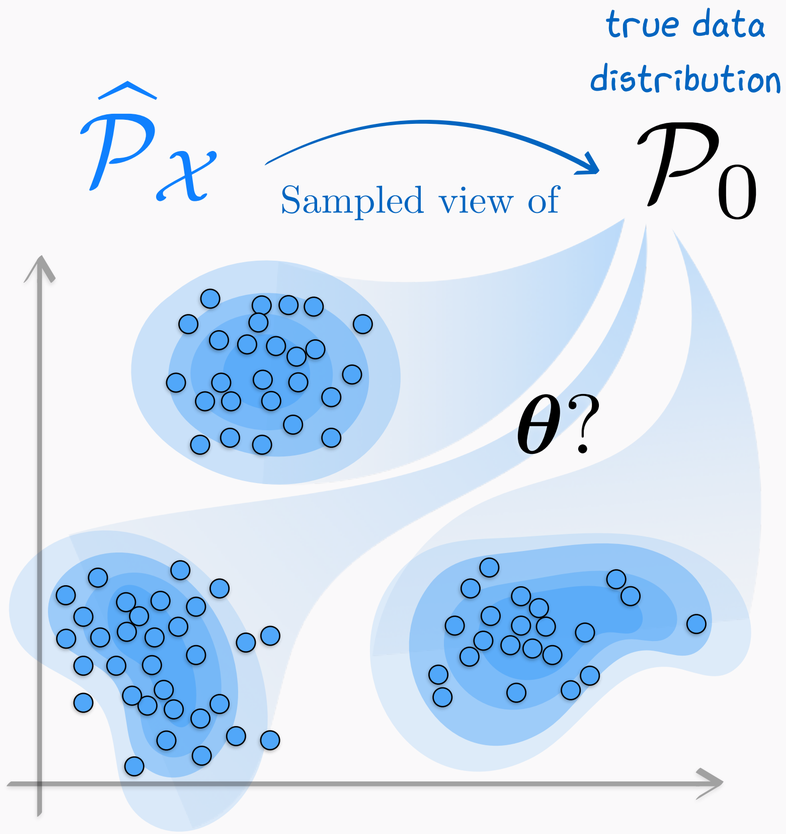
$$\widetilde{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i)}_{\mathbb{E}_{\boldsymbol{x}\sim\widehat{\mathcal{P}}_{\mathcal{X}}}\ell(\boldsymbol{\theta},\boldsymbol{x})} = \underbrace{\mathcal{R}(\boldsymbol{\theta}; \widehat{\mathcal{P}}_{\mathcal{X}})}_{\text{Risk function}}$$

... as a proxy for the true risk:

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}} \mathcal{R}(\boldsymbol{\theta}; \mathcal{P}_0) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_0}\ell(\boldsymbol{\theta}, \boldsymbol{x})$$

$$\text{if } \boldsymbol{x}_i \sim_{\text{iid}} \mathcal{P}_0 \text{ and } n \to +\infty$$

$$\mathcal{R}(\boldsymbol{\theta}, \widehat{\mathcal{P}}_{\mathcal{X}}) \approx \mathcal{R}(\boldsymbol{\theta}, \mathcal{P}_0)$$



→ Target statistical guarantee: Probably Approximately Correct — PAC

$$\mathbb{P}\Big[\underbrace{\mathcal{R}\big(\widetilde{\boldsymbol{\theta}}; \mathcal{P}_0\big) - \mathcal{R}\big(\boldsymbol{\theta}^*; \mathcal{P}_0\big)}_{\text{"excess risk"}} \;\leq\; \eta\Big] \;\geq\; 1-\delta \quad \text{(for small } \eta, \delta > 0\text{)}$$

Statistical learning context

Minimize the empirical risk (ERM)

$$\widetilde{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{\theta}, \boldsymbol{x}_i) = \mathcal{R}(\boldsymbol{\theta}; \widehat{\mathcal{P}}_{\mathcal{X}})$$

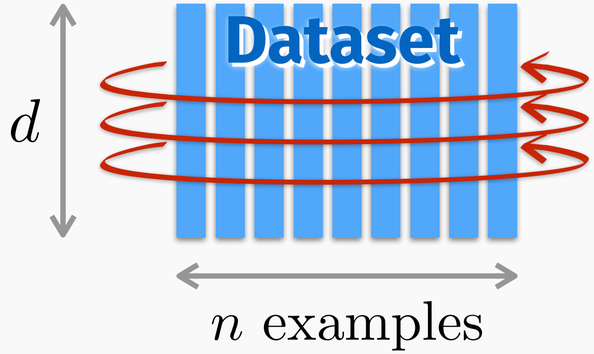
Problem:

$\rightarrow$ Require several passes over $\mathcal{X}$ (*e.g.*, stochastic gradient descent);

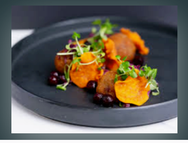
$\rightarrow$ n typically very large;

Computationally expensive
on modern large-scale data

$$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$$

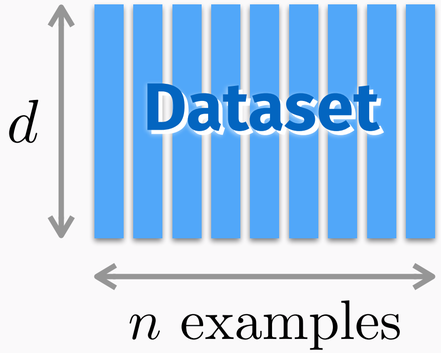


Compressive learning (an introduction)



Break the training down into two "cheaper" steps

$\mathcal{X} = \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$



Sketching phase

Φ

Compress the dataset

sketch $\boldsymbol{z}$

m coefficients

$d < m \ll n$

"Summary"

Learning phase

Φ

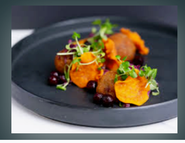
Nonconvex optimization

$\boldsymbol{\theta}$

Parametric model

Gribonval, R., Blanchard, G., Keriven, N., & Traonmilin, Y. (2021).

Gribonval, R., Chatalic, A., Keriven, N., Schellekens, V., Jacques, L., & Schniter, P. (2021)



Sketching phase

Random features moments ("empirical average")

$$\boldsymbol{z} = \mathcal{A}_\Phi(\widehat{P}_\mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n} \Phi(\boldsymbol{x}_i)$$

with random nonlinearity $\Phi : \mathbb{R}^d \to \mathbb{C}^m$

→ *Example*: random Fourier Features (RFF)

$$\Phi_{\mathrm{RFF}}(\boldsymbol{x}) := \frac{1}{\sqrt{m}} \exp(\mathrm{i}\boldsymbol{\Omega}^\top \boldsymbol{x}) \in \mathbb{C}^m$$

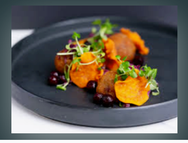
with $\boldsymbol{\Omega} = (\boldsymbol{\omega}_j)_{j=1}^m$, $\boldsymbol{\omega}_j \sim_{\text{i.i.d.}} \Lambda(\boldsymbol{\omega}) = [\mathcal{F}\kappa^\Delta](\boldsymbol{\omega})$.

e.g., Gaussian kernel

Embeds a "nonlinear geometry", *e.g.*,

$$\langle \Phi_{\mathrm{RFF}}(\boldsymbol{x}), \Phi_{\mathrm{RFF}}(\boldsymbol{x}') \rangle \simeq C \underbrace{\exp}_{\kappa_\Delta}\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2}\right)$$

Rahimi, Recht, 07



Sketching phase

Random features moments ("empirical average")

$$\boldsymbol{z} = \mathcal{A}_\Phi(\widehat{P}_\mathcal{X}) = \frac{1}{n}\sum_{i=1}^{n} \Phi(\boldsymbol{x}_i)$$

with random nonlinearity $\Phi : \mathbb{R}^d \to \mathbb{C}^m$

→ *Example*: random Fourier Features (RFF)

$$\Phi_{\mathrm{RFF}}(\boldsymbol{x}) := \frac{1}{\sqrt{m}} \exp(\mathrm{i}\boldsymbol{\Omega}^\top \boldsymbol{x}) \in \mathbb{C}^m$$

with $\boldsymbol{\Omega} = (\boldsymbol{\omega}_j)_{j=1}^m$, $\boldsymbol{\omega}_j \sim_{\text{i.i.d.}} \Lambda(\boldsymbol{\omega}) = [\mathcal{F}\kappa^\Delta](\boldsymbol{\omega})$.

e.g., Gaussian kernel

Embeds a "nonlinear geometry", *e.g.*,

$$\langle \Phi_{\mathrm{RFF}}(\boldsymbol{x}), \Phi_{\mathrm{RFF}}(\boldsymbol{x}') \rangle \simeq C \underbrace{\exp}_{\kappa_\Delta}(-\tfrac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma^2})$$

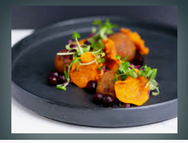
Rahimi, Recht, 07

Sketching with RFF:

$$\boldsymbol{z}_{\Phi_{\mathrm{RFF}}} = \frac{1}{n}\sum_{i=1}^{n} \Phi_{\mathrm{RFF}}(\boldsymbol{x}_i) \simeq \overbrace{[\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_0} e^{\mathrm{i}\boldsymbol{\omega}_j^\top \boldsymbol{x}}]_{j=1}^m}^{\{\mathcal{F}[\mathcal{P}_0](\boldsymbol{\omega}_j)\}_{j=1}^m} =: \mathcal{A}_{\Phi_{\mathrm{RFF}}}(\mathcal{P}_0)$$

"Compressive sensing of prob. distributions"

Random sampling of the characteristic function (aka Fourier transform $\mathcal{F}$) of the data distribution $\mathcal{P}_0$



Learning phase

Minimize a (non-convex) cost function

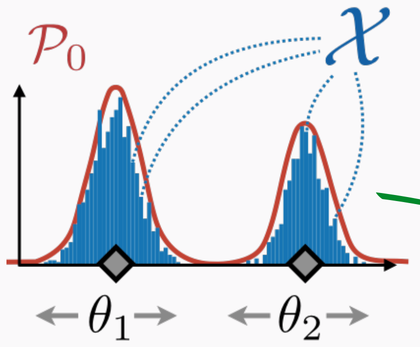
Parametric distribution (model)

Example: k-means

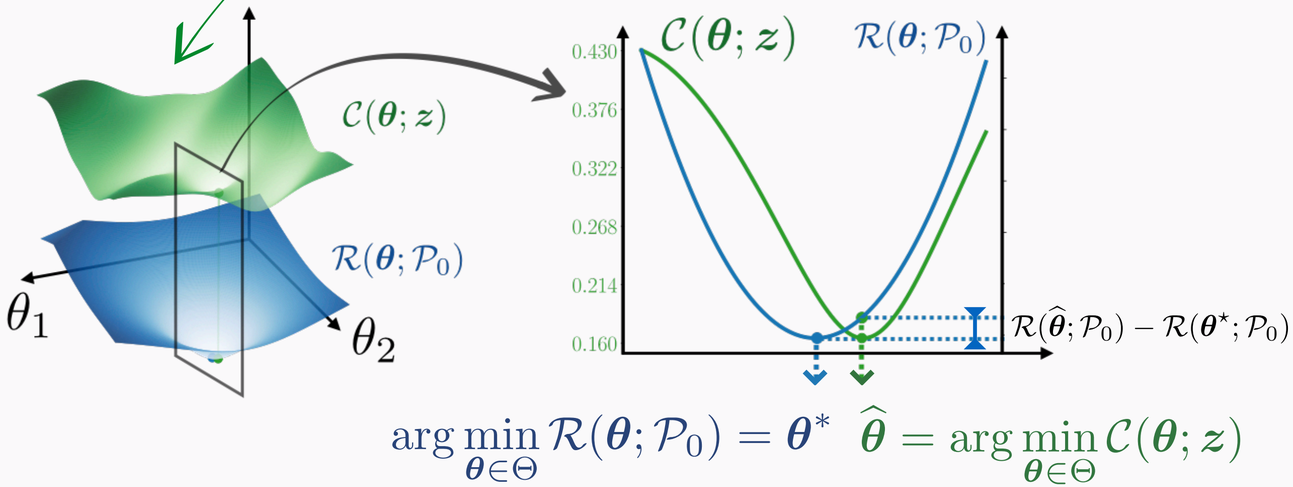
$\mathcal{P}_{\boldsymbol{\theta}} = \sum_{k=1}^{K} \alpha_k \delta_{\boldsymbol{c}_k}$

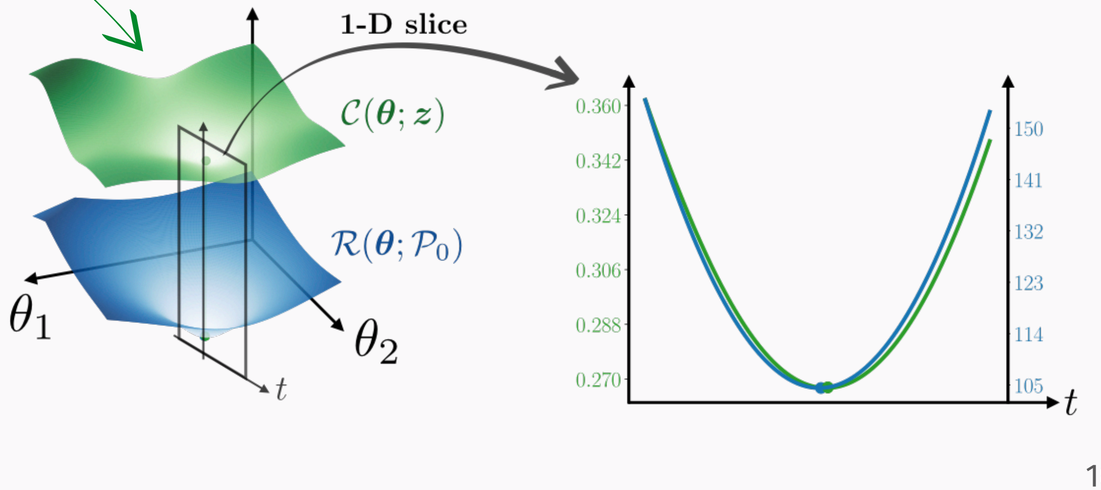
$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{C}(\boldsymbol{\theta}; \boldsymbol{z}) := \|\boldsymbol{z} - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\| \equiv \text{minimize sketch distance, or "matching frequencies"}$$

sketch $\boldsymbol{z}$ (Small)



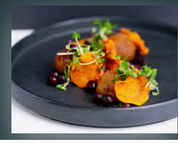
sketch $\boldsymbol{z}$ (Larger)





$$\arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}(\boldsymbol{\theta}; \mathcal{P}_0) = \boldsymbol{\theta}^* \quad \widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{C}(\boldsymbol{\theta}; \boldsymbol{z})$$

Compressive learning (an introduction)



Theoretical guarantees:

(given $\|\mathcal{P} - \mathcal{P}'\|_{\mathcal{R}} := \sup_{\boldsymbol{\alpha} \in \Theta} |\mathcal{R}(\boldsymbol{\alpha}, \mathcal{P}) - \mathcal{R}(\boldsymbol{\alpha}, \mathcal{P}')|$)

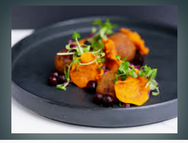
= distance between $\mathcal{P}$ & $\mathcal{P}'$ wrt task-specific risk $\mathcal{R}$

Assumption: The sketch encodes the risk, *i.e.*,

$\mathcal{A}_\Phi$ respects the Lower Restricted Isometry Property (LRIP):

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta, \ \|\mathcal{P}_{\boldsymbol{\theta}} - \mathcal{P}_{\boldsymbol{\theta}'}\|_{\mathcal{R}} \leq \gamma \, \|\mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}}) - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}'})\|_2$$

Compressive learning (an introduction)



Theoretical guarantees:

(given $\|\mathcal{P} - \mathcal{P}'\|_{\mathcal{R}} := \sup_{\boldsymbol{\alpha}\in\Theta} |\mathcal{R}(\boldsymbol{\alpha}, \mathcal{P}) - \mathcal{R}(\boldsymbol{\alpha}, \mathcal{P}')|$)

= distance between $\mathscr{P}$ & $\mathscr{P}'$ wrt task-specific risk $\mathscr{R}$

Assumption: The sketch encodes the risk, *i.e.*,

$\mathcal{A}_\Phi$ respects the Lower Restricted Isometry Property (LRIP):

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ \|\mathcal{P}_{\boldsymbol{\theta}} - \mathcal{P}_{\boldsymbol{\theta}'}\|_{\mathcal{R}} \leq \gamma \|\mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}}) - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}'})\|_2$$

Then, given the estimate $\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \|\overbrace{\mathcal{A}_\Phi(\widehat{\mathcal{P}}_{\mathcal{X}})}^{z} - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\|$, (assuming we can exactly solve it)

(for a certain distance d)

$$\underbrace{\mathcal{R}(\widehat{\boldsymbol{\theta}}; \mathcal{P}_0) - \mathcal{R}(\boldsymbol{\theta}^\star; \mathcal{P}_0)}_{\text{Excess risk}} \leq \underbrace{\inf_{\theta\in\Theta} d(\mathcal{P}_0, \mathcal{P}_\theta)}_{\text{Modeling error}} + \underbrace{4\gamma\|\mathcal{A}_\Phi(\widehat{\mathcal{P}}_{\mathcal{X}}) - \mathcal{A}_\Phi(\mathcal{P}_0)\|_2}_{\text{Sampling error}}$$

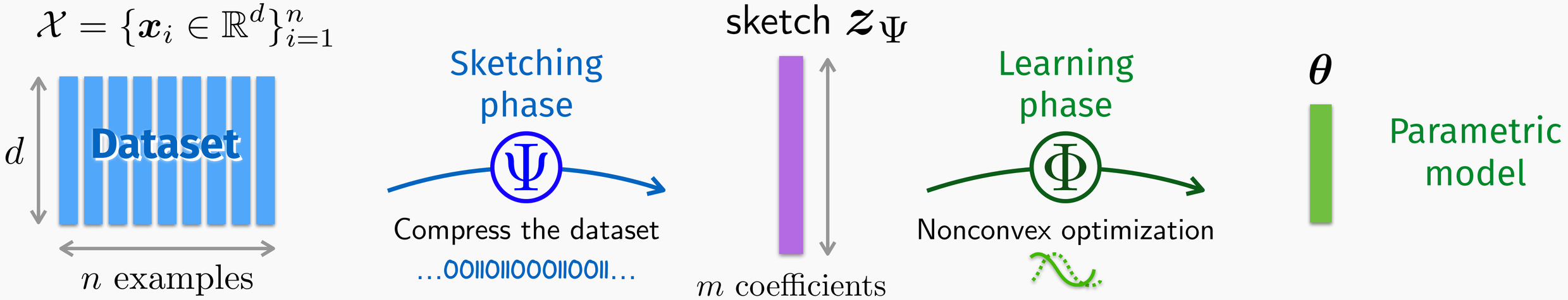
Satisfying the LRIP? (w.h.p)

Task (in $\mathbb{R}^d$)	Sample complexity (up to log.)
k-means	$m \gtrsim k^2 d$
GMM	$m \gtrsim k^2 d$ (or $m \gtrsim k e^d$)

Gribonval, R., Blanchard, G., Keriven, N., & Traonmilin, Y. (2021).

Asymmetric compressive learning

(The question) Decoupling the sketching and learning phases? With $\Psi \neq \Phi$?



Possible advantages:

- avoid complex exponential at sketching;
  - opt for hardware friendly, quantized sketching procedure;
  - reduced transmission cost;
- still differentiable learning cost & relaxed conditions.

Asymmetric CL: sketching phase?

$\Psi \equiv$ random periodic Features (RPF)

$$\boldsymbol{z}_\Psi = \frac{1}{n} \sum_{\boldsymbol{x}_i \in \mathcal{X}} f\left(\boldsymbol{\Omega}^\top \cdot \boldsymbol{x}_i + \boldsymbol{\xi} \right)$$

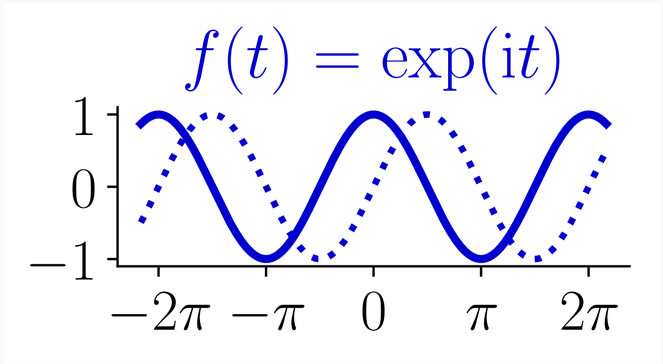
Pooling (average)

$\boldsymbol{\Omega}^\top$ random projections

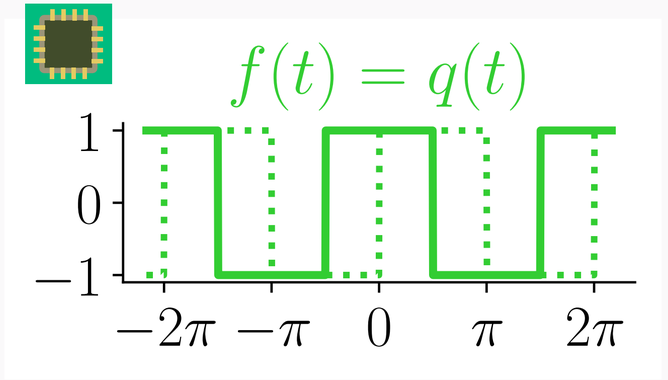
$\boldsymbol{\xi}$ dithering $\sim \mathcal{U}([0,2\pi])$

Dithering "smoothes" quantization/discontinuities (see later)

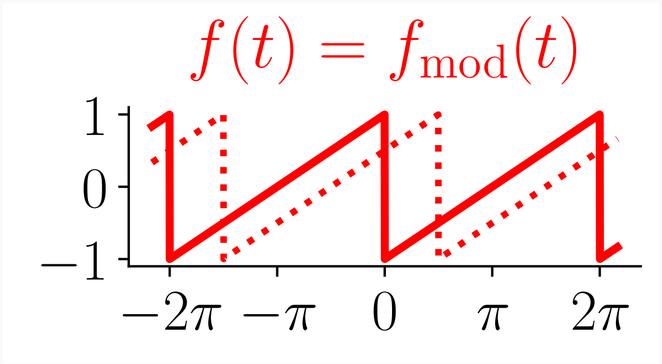
Nonlinear periodic "activation" f:



RFF



Random binary feature



Random modulo feature

Asymmetric CL: learning phase?

Plug the "Ψ" sketch into the "Φ" cost function!

Solve "as before", without changing anything!

(Similar philosophy in non-linear compressive sensing)

Plan, Y., & Vershynin, R. (2016)

$$\widehat{\boldsymbol{\theta}}' \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{C}_{\Phi}(\boldsymbol{\theta}; \boldsymbol{z}_{\Psi}) := \|\boldsymbol{z}_{\Psi} - \mathcal{A}_{\Phi}(\mathcal{P}_{\boldsymbol{\theta}})\|$$

e.g., binary features

e.g., full-precision RFF

Asymmetric CL: learning phase?

Plug the "Ψ" sketch into the "Φ" cost function!
Solve "as before", without changing anything!

(Similar philosophy in non-linear compressive sensing)

Plan, Y., & Vershynin, R. (2016)

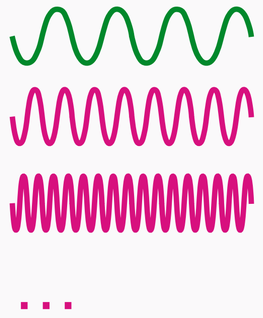
$$\widehat{\boldsymbol{\theta}}' \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{C}_\Phi(\boldsymbol{\theta}; \boldsymbol{z}_\Psi) := \|\boldsymbol{z}_\Psi - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\|$$

e.g., binary features

e.g., full-precision RFF

Intuitive explanation:
(the dither magic trick)

Fourier Series decomposition



Fundamental frequency

Higher frequencies

$$\mathcal{C}_\Phi(\boldsymbol{\theta}; \boldsymbol{z}_\Psi) \approx \overbrace{\|\boldsymbol{z}_\Phi - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\|}^{\mathcal{C}_\Phi(\boldsymbol{\theta}; \boldsymbol{z}_\Phi)} + \sum_{j=1}^{m} \sum_{k\neq 1} C_{k,j} \exp(\mathrm{i}(k-1)\xi_j)$$

(boring expression)

Good ol' CL cost
(what we want)

Artefacts of quantization → 0 in expectation on ξ since $\mathbb{E}_\xi \exp(\mathrm{i}(k-1)\xi) = 0 \quad \forall k \neq 1$

Asymmetric CL: theoretical guarantees?

We do not generalize de LRIP (too hard), we rather complete it! (easier)

Assumption 1: The (*learning*) sketch Φ encodes the risk, *i.e.*,

$\mathcal{A}_\Phi$ respects the Lower Restricted Isometry Property (LRIP): $\exists \gamma > 0$ s.t.

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta, \ \|\mathcal{P}_{\boldsymbol{\theta}} - \mathcal{P}_{\boldsymbol{\theta}'}\|_{\mathcal{R}} \leq \gamma \, \|\mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}}) - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}'})\|_2$$

(as for SCL)

Assumption 2: The (*encoding*) sketch not too much "distorted" along specific directions, *i.e.*,

$\mathcal{A}_\Psi$ and $\mathcal{A}_\Phi$ respect the Limited Projected Distortion (LPD): $\exists \epsilon > 0$ s.t.

$$\forall \mathcal{X}, \forall \boldsymbol{\theta} \in \Theta, \ |\langle \mathcal{A}_\Psi(\widehat{\mathcal{P}}_{\mathcal{X}}) - \mathcal{A}_\Phi(\widehat{\mathcal{P}}_{\mathcal{X}}), \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}}) \rangle| \leq \epsilon$$

(new condition)

Asymmetric CL: theoretical guarantees?

We do not generalize de LRIP (too hard), we rather complete it! (easier)

If both LRIP (with $\gamma > 0$) & LPD (with $\epsilon > 0$) hold then,

given the estimate $\widehat{\boldsymbol{\theta}}' \in \arg\min_{\boldsymbol{\theta}\in\Theta} \|\mathcal{A}_\Psi(\underbrace{\widehat{\mathcal{P}}_\mathcal{X}}_{z_\Psi}) - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\|$, (assuming we can exactly solve it)

$$\underbrace{\mathcal{R}(\widehat{\boldsymbol{\theta}}';\mathcal{P}_0) - \mathcal{R}(\boldsymbol{\theta}^\star;\mathcal{P}_0)}_{\text{Excess risk}} \leq \overbrace{\underbrace{\inf_{\theta\in\Theta} d(\mathcal{P}_0,\mathcal{P}_\theta)}_{\text{Modeling error}} + \underbrace{4\gamma\|\mathcal{A}_\Phi(\widehat{\mathcal{P}}_\mathcal{X}) - \mathcal{A}_\Phi(\mathcal{P}_0)\|_2}_{\text{Sampling error}}}^{\text{(as before)}} + \overbrace{\underbrace{4\gamma\sqrt{\epsilon}}_{\text{Distortion error}}}^{\text{(new term)}}$$

Modeling error (for a certain distance d)

Distortion error (Sketch distortion)

Proving the LPD

What we want : $\mathcal{A}_\Psi$ and $\mathcal{A}_\Phi$ respect the Limited Projected Distortion (LPD): $\exists \epsilon > 0$ s.t.

$$\forall \mathcal{X}, \forall \boldsymbol{\theta} \in \Theta, \ |\langle \mathcal{A}_\Psi(\widehat{\mathcal{P}}_\mathcal{X}) - \mathcal{A}_\Phi(\widehat{\mathcal{P}}_\mathcal{X}), \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}}) \rangle| \leq \epsilon$$

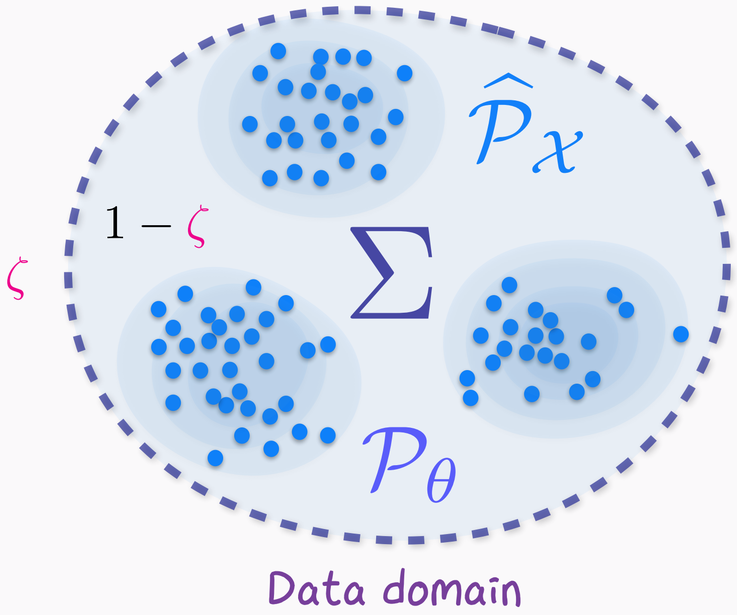
Proving the LPD

What we want : $\mathcal{A}_\Psi$ and $\mathcal{A}_\Phi$ respect the Limited Projected Distortion (LPD): $\exists \epsilon > 0$ s.t.

$$\forall \mathcal{X}, \forall \boldsymbol{\theta} \in \Theta,\ \left|\langle \mathcal{A}_\Psi(\widehat{\mathcal{P}}_\mathcal{X}) - \mathcal{A}_\Phi(\widehat{\mathcal{P}}_\mathcal{X}), \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\rangle\right| \leq \epsilon$$

Assuming $\mathbb{P}[\boldsymbol{x} \in \Sigma] \geq 1 - \zeta$ if $\boldsymbol{x} \sim \widehat{\mathcal{P}}_\mathcal{X}$ or $\mathcal{P}_{\boldsymbol{\theta}}$, we first note that

$$\left|\langle \mathcal{A}_\Psi(\widehat{\mathcal{P}}_\mathcal{X}) - \mathcal{A}_\Phi(\widehat{\mathcal{P}}_\mathcal{X}), \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\rangle\right| \leq \iint \left|\langle \Psi(\boldsymbol{x}) - \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}')\rangle\right| \mathrm{d}\widehat{\mathcal{P}}_\mathcal{X}(\boldsymbol{x}) \mathrm{d}\mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{x}') \leq C\epsilon' + D\zeta,$$



If $\left|\langle \Psi(\boldsymbol{x}) - \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}')\rangle\right| \leq \epsilon',\ \forall \boldsymbol{x}, \boldsymbol{x}' \in \Sigma.$

signal-LPD (sLPD)

so, sLPD ⇒ LPD

Ok, how to prove the sLPD now?

Proving the sLPD (for RPF & RFF)

Considering Random Periodic Features (RPF), with:

$$\Psi_f(\boldsymbol{x}) := \frac{1}{F_1\sqrt{m}} f(\boldsymbol{\Omega}^\top \boldsymbol{x} + \boldsymbol{\xi}) \in \mathbb{C}^m, \;\; \Phi_{\mathrm{RFF}} = \Psi_{\exp(\mathrm{i}\cdot)}, \;\; f(t) = \sum_k F_k e^{\mathrm{i}kt}$$

Key property of f: *mean Lipschitz smoothness* (more permissive than Lipschitz smoothness)

$$\mathbb{E}_{t\sim\mathcal{U}[0,2\pi]} \sup_{r\in[-\delta,+\delta]} |f(t+r) - f(t)| \leq L_f^\mu \cdot \delta$$

"Smooth on average"

(e.g., for q = square wave, $L_q^\mu = \frac{8}{\pi}$)

Sampling condition

$$m \gtrsim \epsilon^{-2}\mathcal{H}_{c\epsilon}(\Sigma)$$

"Dimension" of Σ, e.g., $k\log(k/\epsilon)$ if $\dim(\Sigma) = k$

$\Rightarrow$

signal-LPD (sLPD)

$$\left|\langle \Psi(\boldsymbol{x}) - \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}')\rangle\right| \leq \epsilon, \;\; \forall \boldsymbol{x}, \boldsymbol{x}' \in \Sigma, \text{ w.h.p.}$$

Schellekens, V., & Jacques, L. (2022a).

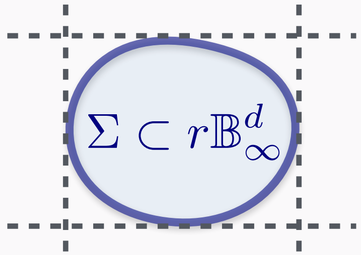
Asymmetric CL: theoretical guarantees with RPF

Schellekens, V., & Jacques, L. (2022b).

Let us consider k-means or GMM fitting, and the estimate

$$\widehat{\boldsymbol{\theta}}' \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{C}_\Phi(\boldsymbol{\theta}; \boldsymbol{z}_\Psi) := \|\boldsymbol{z}_\Psi - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\|$$

Assumptions:
+ bounded data domain (in ℓ_∞)
+ mode variance $\leq S$ (for GMM)
+ LRIP satisfied for the task (with $\gamma > 0$)



If, for k-means, $m \gtrsim \epsilon^{-2} d \log(\frac{c\sqrt{d}r}{\epsilon})$, or, for GMM fitting, $m \gtrsim \epsilon^{-2} d \log(\frac{c\sqrt{d}(2\rho S + r)}{\epsilon})$ (for some $\rho > d$)

Then, with proba $\geq 1 - C\exp(-cm\epsilon^2)$,

$$\underbrace{\mathcal{R}(\widehat{\boldsymbol{\theta}}'; \mathcal{P}_0) - \mathcal{R}(\boldsymbol{\theta}^\star; \mathcal{P}_0)}_{\text{Excess risk}} \leq \underbrace{\inf_{\theta\in\Theta} d(\mathcal{P}_0, \mathcal{P}_\theta)}_{\text{Modeling error}} + \underbrace{4\gamma\|\mathcal{A}_\Phi(\widehat{\mathcal{P}}_{\mathcal{X}}) - \mathcal{A}_\Phi(\mathcal{P}_0)\|_2}_{\text{Sampling error}} + \underbrace{c_f\gamma \begin{cases} \sqrt{\epsilon} & (k\text{-means}) \\ \sqrt{\epsilon} + e^{-\rho^2/4} & (\text{GMM}) \end{cases}}_{\text{Distortion error (Sketch distortion)}}$$

SCL part (as before)

Modeling error (for a certain distance d)

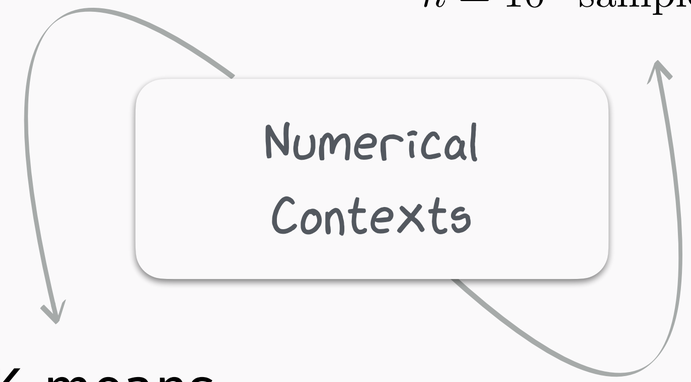
Smoother f (in mean Lipschitz & Fourier)
⇒ better performances

Some numerical results* (synthetic data)

How large should the sketch be to reach desired learning performance?

GMM fitting

$d = 5$ dims,
$K = 10$ modes,
$n = 10^5$ samples.



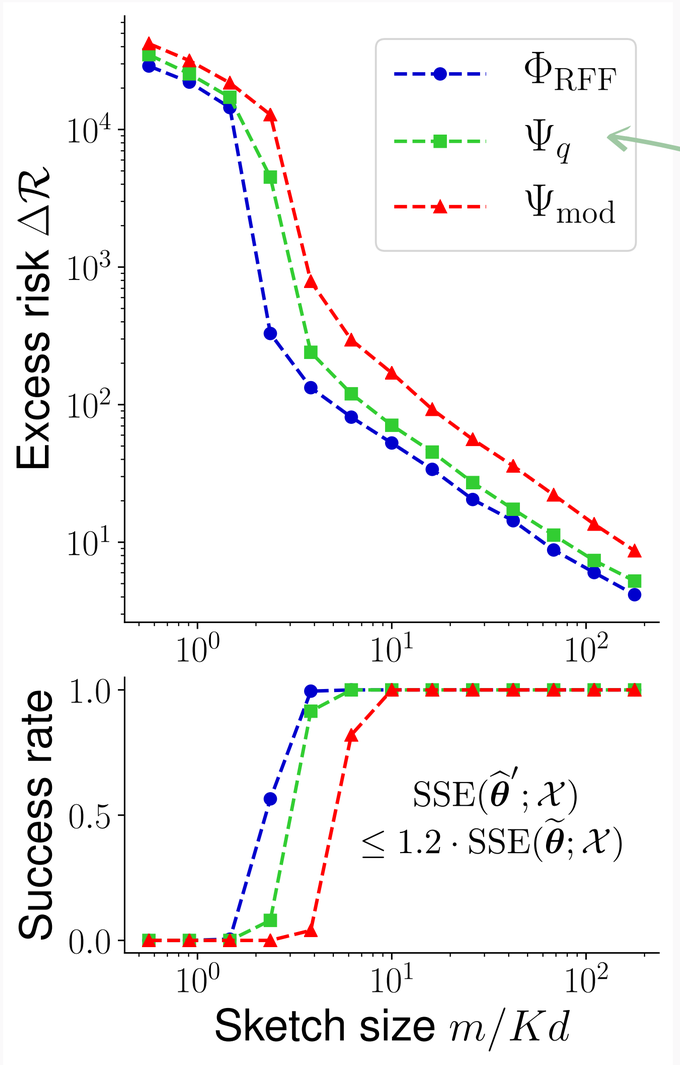
K-means

$d = 5$ dims,
$K = 10$ clusters,
$n = 10^5$ samples.

*: Real data, audio classification task in extra slides

Some numerical results* (synthetic data)

How large should the sketch be to reach desired learning performance?



GMM fitting

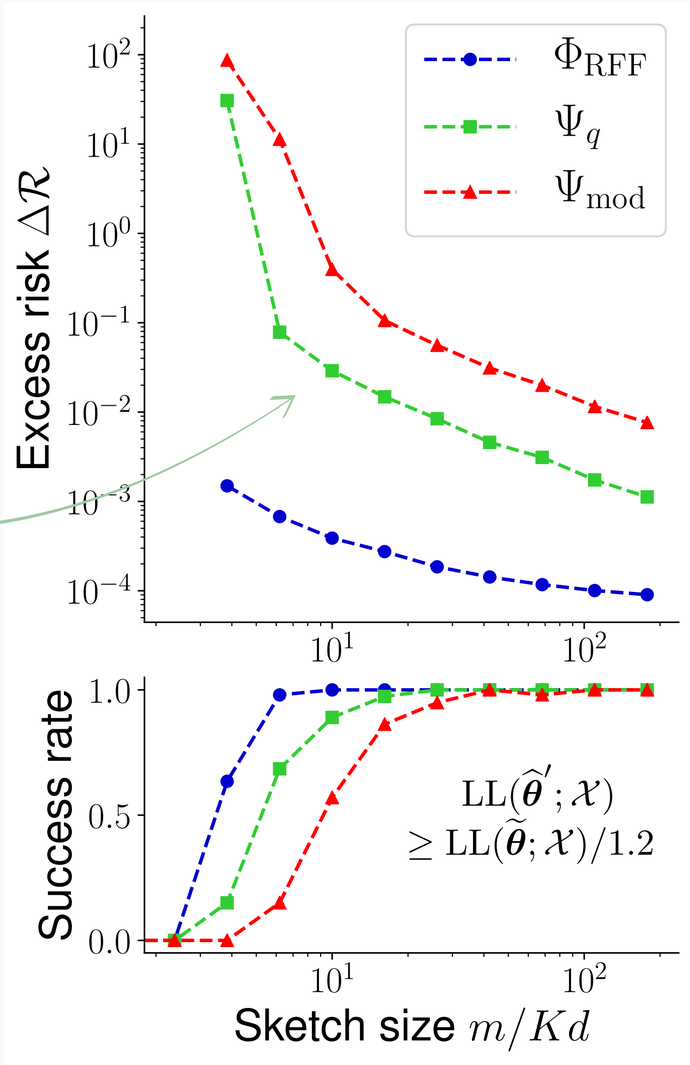
$d = 5$ dims,
$K = 10$ modes,
$n = 10^5$ samples.

Smoother f
better performances

(prediction from theory)

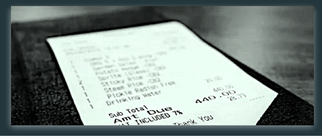
K-means

$d = 5$ dims,
$K = 10$ clusters,
$n = 10^5$ samples.



*: Real data, audio classification task in extra slides

Conclusion



Take away messages:

- We can make CL asymmetric & keep theoretical guarantees (with LPD)
- Relaxed conditions on sketching vs learning
- Hardware friendly sketching (*e.g.*, quantized)
- New sketching are possible (such as modulo)

Limitation & open questions:

- Hidden (large) constants for LRIP and LPD (≡ the "compressive sensing crime")
- Sample complexity for LPD in function of model set dimension (and not Σ)?
- Gap between the theoretical guarantees and empirical performances (SCL & ACL)

Thank you for your attention

Related references:

- Gribonval, R., Chatalic, A., Keriven, N., Schellekens, V., Jacques, L., & Schniter, P. (2021). "Sketching data sets for large-scale learning: Keeping only what you need." *IEEE Signal Processing Magazine*, *38*(5), 12-36.
- Rahimi, A., & Recht, B. (2007). "Random features for large-scale kernel machines." Advances in neural information processing systems, 20.
- Gribonval, R., Blanchard, G., Keriven, N., & Traonmilin, Y. (2021). "Statistical learning guarantees for compressive clustering and compressive mixture modeling." Mathematical Statistics and Learning, 3(2), 165-257.
- Plan, Y., & Vershynin, R. (2016). "The generalized lasso with non-linear observations." IEEE Transactions on information theory, 62(3), 1528-1537.
- Karol J Piczak. "ESC: Dataset for environmental sound classification." In Proceedings of the 23rd ACM international conference on Multimedia, pages 1015–1018, 2015.
- Anurag Kumar and Bhiksha Raj. "Features and kernels for audio event recognition." arXiv preprint arXiv:1607.05765, 2016.
- Schellekens, V., & Jacques, L. (2022b). "Asymmetric compressive learning guarantees with applications to quantized sketches." IEEE Transactions on Signal Processing, 70, 1348-1360.
- Schellekens, V., & Jacques, L. (2022a). "Breaking the waves: asymmetric random periodic features for low-bitrate kernel machines." Information and Inference: A Journal of the IMA, 11(1), 385-421.

Extra slides

Real data, audio classification task

ESC-50 dataset: Karol J Piczak, 2015.

- $J = 2500$ audio clips lasting 5s; C $= 50$ classes (e.g., animals, water sounds, urban noises)
- $\mathcal{X} = \{\boldsymbol{x}_j\}_{j=1}^{n} \subset \mathbb{R}^{d=10} \equiv$ Mel Frequency Cepstral Coefficients (MFCC) of 15ms ($n = 334J = 835\,000$)

Scenario:

- We compute $\boldsymbol{z}_f = \frac{1}{n}\sum_{j=1}^{n} \Psi_f(\boldsymbol{x}_j)$

  for SCL (with $f(\cdot) = \exp(\mathrm{i}\cdot)$) and ACL (with $f = q$)
- Minimizing transmission cost
- Compare (64bits) SCL, and (1bit) ACL

  $\rightarrow$ Objective $\neq$ competitive classification

Method:

- GMM-fit with $K = 32$ modes

  extracted from SCL & ACL.
- Classification using "alpha features" (on a test set)

A. Kumar, B. Raj, 2016.

Schellekens, V., & Jacques, L. (2022b).

Real data, audio classification task

ESC-50 dataset: Karol J Piczak, 2015.

- $J = 2500$ audio clips lasting 5s; C = 50 classes (e.g., animals, water sounds, urban noises)
- $\mathcal{X} = \{\boldsymbol{x}_j\}_{j=1}^n \subset \mathbb{R}^{d=10} \equiv$ Mel Frequency Cepstral Coefficients (MFCC) of 15ms ($n = 334J = 835\,000$)

Scenario:

- We compute $\boldsymbol{z}_f = \frac{1}{n}\sum_{j=1}^n \Psi_f(\boldsymbol{x}_j)$

  for SCL (with $f(\cdot) = \exp(\mathrm{i}\cdot)$) and ACL (with $f = q$)
- Minimizing transmission cost
- Compare (64bits) SCL, and (1bit) ACL

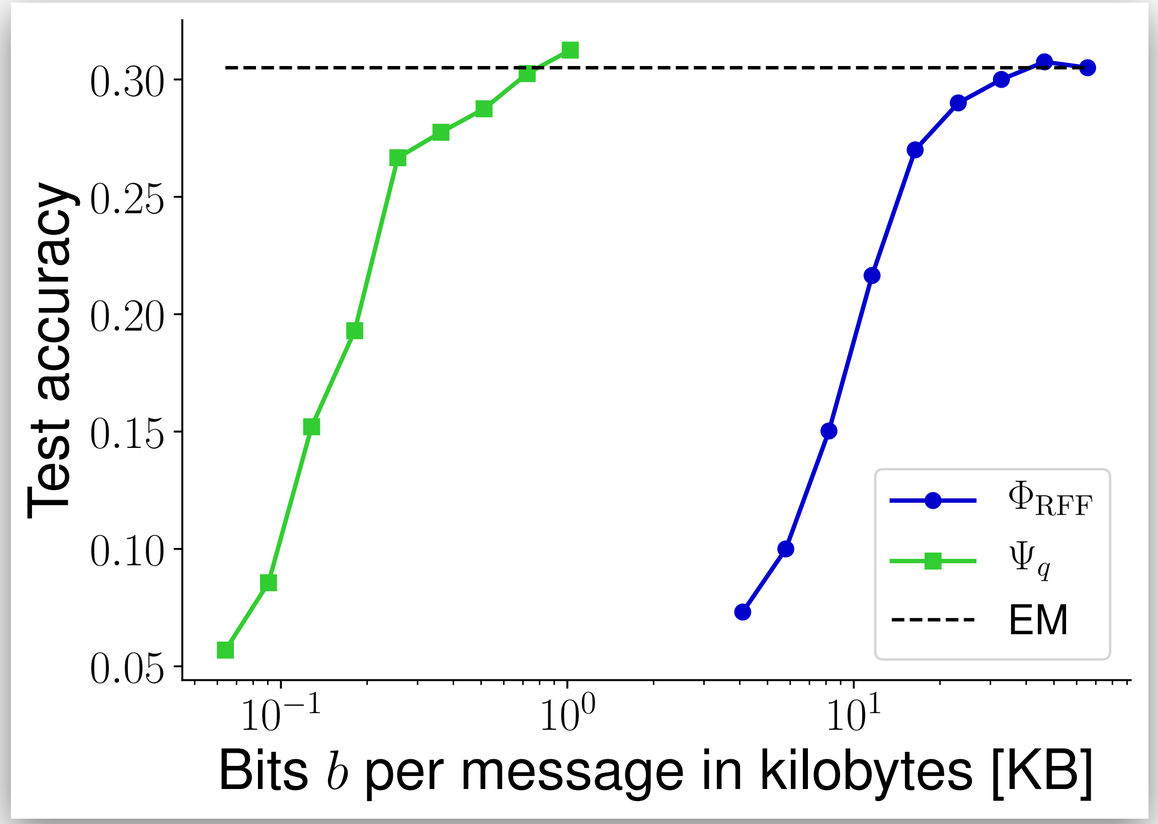
  → Objective ≠ competitive classification

Method:

- GMM-fit with $K = 32$ modes

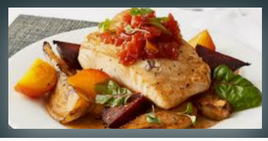
  extracted from SCL & ACL.
- Classification using "alpha features" (on a test set)

  A. Kumar, B. Raj, 2016.



Schellekens, V., & Jacques, L. (2022b).

Basic proof idea



First, show that the LPD and a bit of algebra imply

Suboptimality term
(cost of distortion)

$$\|\mathcal{A}_\Phi(\widehat{\mathcal{P}}_\mathcal{X}) - \mathcal{A}_\Phi(\mathcal{P}_{\widehat{\boldsymbol{\theta}}'})\| \leq \|\mathcal{A}_\Phi(\widehat{\mathcal{P}}_\mathcal{X}) - \mathcal{A}_\Phi(\mathcal{P}_{\widehat{\boldsymbol{\theta}}})\| + 2\sqrt{\epsilon}$$

Asymmetric CL solution

$$\widehat{\boldsymbol{\theta}}' \in \arg\min_{\boldsymbol{\theta}\in\Theta} \|\mathcal{A}_\Psi(\widehat{\mathcal{P}}_\mathcal{X}) - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\|$$

Symmetric CL solution

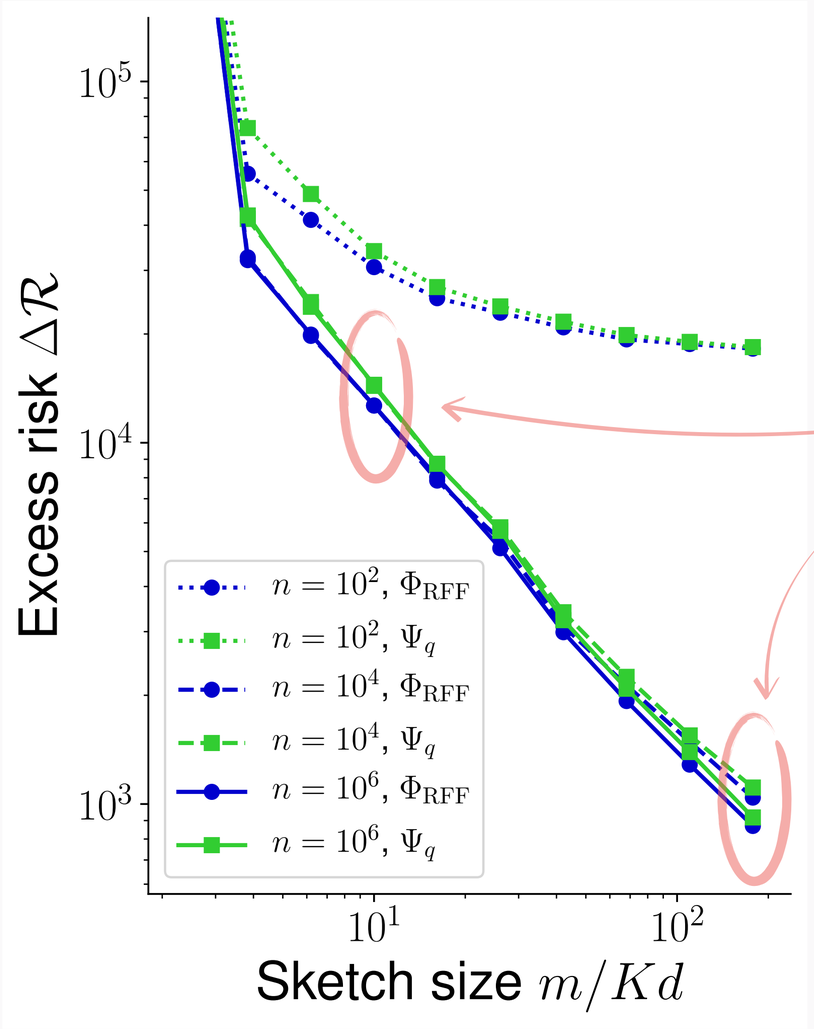
$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \|\mathcal{A}_\Phi(\widehat{\mathcal{P}}_\mathcal{X}) - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\|$$

Then, plug this into the usual LRIP ⇒ performance guarantee proof

30

Some numerical results. (synthetic data)

Influence of the dataset size n ?



GMM fitting

$d = 5$ dims,
$K = 10$ modes,
$n = 10^5$ samples.

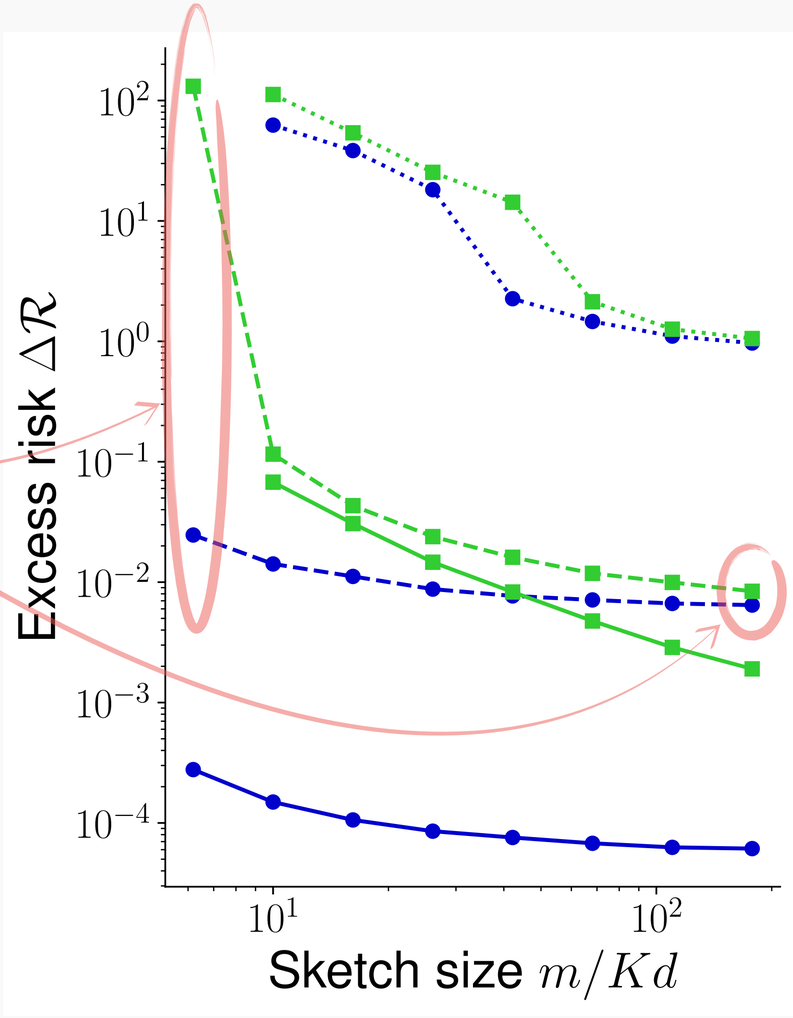
If m large enough,
Quantization effects vanishes

K-means

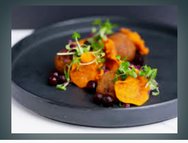
$d = 5$ dims,
$K = 10$ clusters,
$n = 10^5$ samples.

Frequencies:

$\boldsymbol{\omega}_j \sim$ "folded" Gaussian
$\boldsymbol{\omega}_j \boldsymbol{\omega}_j^\top = O(d^{-1})\, \boldsymbol{I}$



Compressive learning (an introduction)



Theoretical guarantees:

(given $\|\mathcal{P} - \mathcal{P}'\|_{\mathcal{R}} := \sup_{\boldsymbol{\alpha}\in\Theta} |\mathcal{R}(\boldsymbol{\alpha}, \mathcal{P}) - \mathcal{R}(\boldsymbol{\alpha}, \mathcal{P}')|$)

= distance between $\mathscr{P}$ & $\mathscr{P}'$ wrt task-specific risk $\mathscr{R}$

Assumption: The sketch encodes the risk, *i.e.*,

$\mathcal{A}_\Phi$ respects the Lower Restricted Isometry Property (LRIP):

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta,\ \|\mathcal{P}_{\boldsymbol{\theta}} - \mathcal{P}_{\boldsymbol{\theta}'}\|_{\mathcal{R}} \leq \gamma \|\mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}}) - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}'})\|_2$$

Then, given the estimate $\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \|\overbrace{\mathcal{A}_\Phi(\widehat{\mathcal{P}}_{\mathcal{X}})}^{z} - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\|$, (assuming we can exactly solve it)

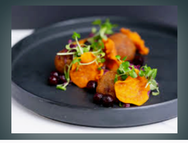
$$\underbrace{\mathcal{R}(\widehat{\boldsymbol{\theta}}; \mathcal{P}_0) - \mathcal{R}(\boldsymbol{\theta}^\star; \mathcal{P}_0)}_{\text{Excess risk}} \leq \underbrace{\inf_{\theta\in\Theta} d(\mathcal{P}_0, \mathcal{P}_\theta)}_{\text{Modeling error}} + \underbrace{4\gamma\|\mathcal{A}_\Phi(\widehat{\mathcal{P}}_{\mathcal{X}}) - \mathcal{A}_\Phi(\mathcal{P}_0)\|_2}_{\text{Sampling error}}$$

(for a certain distance d)

Satisfying the LRIP?

(w.h.p)

Task (in $\mathbb{R}^d$)	Sample complexity	Condition
k-means	$m \gtrsim k^2 d \log \frac{kdR}{\varepsilon}$	$\min_{j\neq k} \|\boldsymbol{c}_j - \boldsymbol{c}_k\| \geq \varepsilon$
GMM	$m \gtrsim k^2 d$ (or $m \gtrsim ke^d$)	sufficiently far modes



Learning phase

Minimize a cost function

Parametric distribution (model)

Example: k-means

$\mathcal{P}_{\boldsymbol{\theta}} = \sum_{k=1}^{K} \alpha_k \delta_{\mathbf{c}_k}$

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{C}(\boldsymbol{\theta}; \boldsymbol{z}) := \|\boldsymbol{z} - \mathcal{A}_\Phi(\mathcal{P}_{\boldsymbol{\theta}})\| \quad \equiv \text{minimize sketch distance (non-convex problem)}$$