Random Moments for Sketched Statistical Learning

G. Blanchard

Universität Potsdam, Institut für Mathematik

Joint work with: R. Gribonval, N. Keriven, Y. Traonmilin (INRIA Rennes)

PLAN

1. The sketched learning approach
2. A framework for sketched learning
3. Two examples
   - Sketched PCA
   - Sketched clustering
4. How to construct a sketching operator

OUTLINE

1. The sketched learning approach
2. A framework for sketched learning
3. Two examples
   - Sketched PCA
   - Sketched clustering
4. How to construct a sketching operator

Classical model for learning

- Each training data point stored as a d-vector
- Training collection $\mathbf{X} = (x_1, \ldots, x_n)$ seen as a (d, n) matrix

CLASSICAL MODEL FOR LEARNING

- Each training data point stored as a d-vector
- Training collection $\mathbf{X} = (x_1, \ldots, x_n)$ seen as a (d, n) matrix
- Usual abstract approach (decision theory):
  - Want to find a predictor ("hypothesis") $h \in \mathcal{H}$ suited to data
  - Performance on data point x measured by loss function $\ell(x, h)$
  - Goal is to minimize averaged loss and approximate the minimizer

$$h^* = \operatorname*{Arg\,Min}_{h \in \mathcal{H}} \mathcal{R}(h) = \operatorname*{Arg\,Min}_{h \in \mathcal{H}} \mathbb{E}[\ell(X, h)]$$

Classical model for learning

- Each training data point stored as a d-vector
- Training collection $\mathbf{X} = (x_1, \ldots, x_n)$ seen as a (d, n) matrix
- Usual abstract approach (decision theory):
  - Want to find a predictor ("hypothesis") $h \in \mathcal{H}$ suited to data
  - Performance on data point x measured by loss function $\ell(x, h)$
  - Goal is to minimize averaged loss and approximate the minimizer

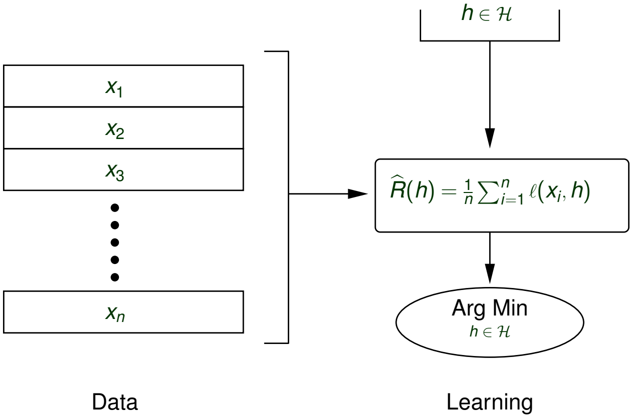
$$h^* = \operatorname*{Arg\,Min}_{h \in \mathcal{H}} \mathcal{R}(h) = \operatorname*{Arg\,Min}_{h \in \mathcal{H}} \mathbb{E}[\ell(X, h)]$$

  - Assuming $(x_1, \ldots, x_n)$ are drawn i.i.d., natural proxy is empirical risk minimizer

$$\widehat{h}_{ERM} = \min_{h \in \mathcal{H}} \widehat{\mathcal{R}}(h) = \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, h)$$

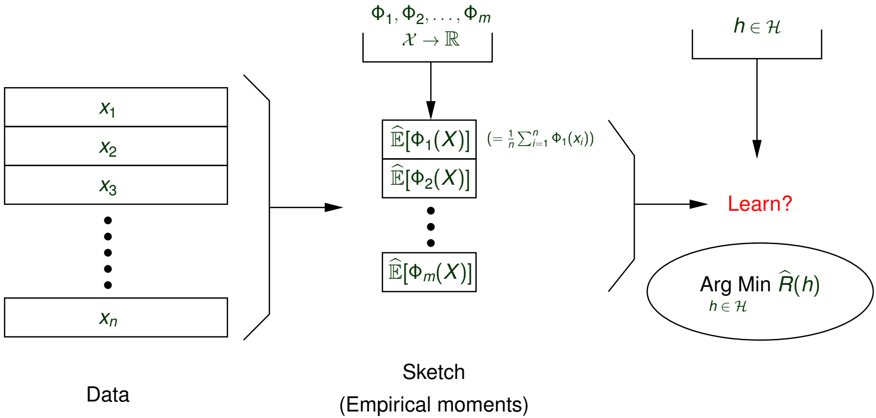
(can possibly be combined with regularization)

CLASSICAL FRAMEWORK



- Storage cost: $O(nd)$
- Computation cost: $O((nd)^\kappa)$
- Stochastic gradient can improve computation bottlenecks but usually requires several data passes

SKETCHED LEARNING APPROACH



- Storage cost after sketching: $O(m)$
- Computation cost: hopefully polynomial in m
- Sketch can be updated very easily
- Which moments Φ_i? How large should m be?

First considerations

- In the classical approach, learning theory guarantees are of the form

$$\sup_{h\in\mathcal{H}}\left|\mathcal{R}(h)-\widehat{\mathcal{R}}(h)\right| \leq \varepsilon(n)\,,$$

with high probability, e.g. $\varepsilon(n) = O\left(\sqrt{\frac{\gamma}{n}}\right)$ for a hypothesis space of metric dimension γ.

FIRST CONSIDERATIONS

- In the classical approach, learning theory guarantees are of the form

$$\sup_{h \in \mathcal{H}} \left| \mathcal{R}(h) - \widehat{\mathcal{R}}(h) \right| \leq \varepsilon(n) \,,$$

  with high probability, e.g. $\varepsilon(n) = O\left(\sqrt{\frac{\gamma}{n}}\right)$ for a hypothesis space of metric dimension γ.
- This implies that the ERM estimator satisfies the risk bound

$$\mathcal{R}(\widehat{h}_{ERM}) \leq \mathcal{R}(h^*) + \varepsilon(n).$$

FIRST CONSIDERATIONS

- In the classical approach, learning theory guarantees are of the form

$$\sup_{h \in \mathcal{H}} \left| \mathcal{R}(h) - \widehat{\mathcal{R}}(h) \right| \leq \varepsilon(n) \,,$$

with high probability, e.g. $\varepsilon(n) = O\left(\sqrt{\frac{\gamma}{n}}\right)$ for a hypothesis space of metric dimension γ.

- This implies that the ERM estimator satisfies the risk bound

$$\mathcal{R}(\widehat{h}_{ERM}) \leq \mathcal{R}(h^*) + \varepsilon(n).$$

- To preserve this property up to constant factor for an estimator $\widetilde{h}_{Sketched}$ it is sufficient to ensure that

$$\left| \mathcal{R}(\widehat{h}_{ERM}) - \mathcal{R}(\widetilde{h}_{Sketched}) \right| \lesssim \sup_{h \in \mathcal{H}} \left| \mathcal{R}(h) - \widehat{\mathcal{R}}(h) \right|.$$

A NAIVE APPROACH

- A first thought is to discretize the hypothesis space into $h_1, \ldots, h_m$ and take $\Phi_i(x) = \ell(x, h_i), i = 1, \ldots, m$.
- Then we simply have

$$\mathbb{E}[\Phi_i(X)] = \frac{1}{n}\sum_{j=1}^{n} \ell(x_j, h_i) = \widehat{R}(h_i), \qquad i = 1, \ldots, m.$$

- With the moment information, we can replace ERM by "discretized ERM" over $h_1, \ldots, h_m$.

A NAIVE APPROACH

- A first thought is to discretize the hypothesis space into $h_1, \ldots, h_m$ and take $\Phi_i(x) = \ell(x, h_i), i = 1, \ldots, m$.
- Then we simply have

$$\mathbb{E}[\Phi_i(X)] = \frac{1}{n} \sum_{j=1}^{n} \ell(x_j, h_i) = \widehat{R}(h_i), \qquad i = 1, \ldots, m.$$

- With the moment information, we can replace ERM by "discretized ERM" over $h_1, \ldots, h_m$.
- To ensure $\left| \mathcal{R}(\widehat{h}_{ERM}) - \mathcal{R}(\widetilde{h}_{disc.ERM}) \right| \leq \varepsilon(n)$, require $(h_1, \ldots, h_m)$ to be an $\varepsilon(n)$-covering of the space $\mathcal{H}$ (say for supremum norm).
- If $\mathcal{H}$ is of metric dimension γ a covering typically requires $m = O(\varepsilon^{-\gamma}) = O(n^{\gamma/2})$, seems hopeless!

SOME HOPE (1)

- Consider "trivial" example $\ell(x,h) = \|x-h\|^2$, goal is to learn mean $h^* = \mathbb{E}[X]$; obviously only need to store only the empirical mean $\widehat{\mathbb{E}}[h(X)] = \frac{1}{n}\sum_{i=1}^n x_i$ i.e. $m = 1$!

SOME HOPE (1)

- Consider "trivial" example $\ell(x, h) = \|x - h\|^2$, goal is to learn mean $h^* = \mathbb{E}[X]$; obviously only need to store only the empirical mean $\widehat{\mathbb{E}}[h(X)] = \frac{1}{n}\sum_{i=1}^n x_i$ i.e. $m = 1$!
- Can this phenomenon be generalized?

SOME HOPE (2)

- **Example 2:** PCA. Since we only need the estimated (covariance) matrix to find PCA directions, we only need to keep moments of order 2 ($m = O(d^2)$).

- We can even hope do to better by using low-rank approximations of the covariance. Using random projections on Gaussian vectors is a well-known mean to this goal.

TOWARDS SKETCHED CLUSTERING

- **Example 3:** We will be interested in learning goals where the target cannot be easily represented in terms of moments, i.e. k-means/k-medians.

OUTLINE

1. The sketched learning approach
2. A framework for sketched learning
3. Two examples
   - Sketched PCA
   - Sketched clustering
4. How to construct a sketching operator

AN ABSTRACT FRAMEWORK

- Let $\mathfrak{M}$ denote the set of probability measures on $\mathcal{X} = \mathbb{R}^d$.
- Define the Risk Operator

$$\mathcal{R}(\pi, h) = \mathbb{E}_{X \sim \pi}[\ell(X, h)].$$

  Note that the empirical risk is

$$\widehat{\mathcal{R}}(h) = \mathcal{R}(\widehat{\pi}_n, h), \text{ with } \widehat{\pi}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \text{ (empirical measure).}$$

- Observe that $\mathcal{R}(\pi, h)$ is linear in π.

AN ABSTRACT FRAMEWORK

- Let $\mathfrak{M}$ denote the set of probability measures on $\mathcal{X} = \mathbb{R}^d$.
- Define the Risk Operator

$$\mathcal{R}(\pi, h) = \mathbb{E}_{X\sim\pi}[\ell(X, h)].$$

Note that the empirical risk is

$$\widehat{\mathcal{R}}(h) = \mathcal{R}(\widehat{\pi}_n, h), \quad \text{with } \widehat{\pi}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \text{ (empirical measure).}$$

- Observe that $\mathcal{R}(\pi, h)$ is linear in π.
- Given $\Phi(x) = (\Phi_1(x), \ldots, \Phi_m(x))$ define the sketching operator

$$\mathcal{A}_\Phi(\pi) = \mathbb{E}_{X\sim\pi}[\Phi(X)].$$

The data sketch is $s = \widehat{\mathbb{E}}[\Phi(X)] = \mathcal{A}_\Phi(\widehat{\pi}_n)$.

- Note that $\mathcal{A}_\Phi$ is a linear operator on probability measures.

Approach (formal version)

- Sketch step:

$$s = \mathcal{A}_\phi(\widehat{\pi}_n) \in \mathbb{R}^m.$$

Approach (formal version)

- Sketch step:

$$s = \mathcal{A}_\phi(\widehat{\pi}_n) \in \mathbb{R}^m.$$

- Reconstruction ("decoding") from sketch step:

$$s \mapsto \Delta[s] =: \widetilde{\pi} \in \mathfrak{M}.$$

This formally reconstructs a probability distribution $\widetilde{\pi}$ by applying the "decoder" Δ to the sketch.

Approach (formal version)

- Sketch step:

$$s = \mathcal{A}_\phi(\widehat{\pi}_n) \in \mathbb{R}^m.$$

- Reconstruction ("decoding") from sketch step:

$$s \mapsto \Delta[s] =: \widetilde{\pi} \in \mathfrak{M}.$$

This formally reconstructs a probability distribution $\widetilde{\pi}$ by applying the "decoder" Δ to the sketch.

- Approximate learning step:

$$\widetilde{h} = \operatorname*{Arg\,Min}_{h \in \mathcal{H}} \mathcal{R}(\widetilde{\pi}, h).$$

GOAL FOR THEORY

- Remember from initial considerations we aim (ideally) at

$$\left|\mathcal{R}(\widehat{h}_{ERM}, \pi) - \mathcal{R}(\widetilde{h}_{Sketched}, \pi)\right| \lesssim \sup_{h \in \mathcal{H}} \left|\mathcal{R}(h, \pi) - \mathcal{R}(h, \widehat{\pi}_n)\right|.$$

Goal for theory

- Remember from initial considerations we aim (ideally) at

$$\left|\mathcal{R}(\widehat{h}_{ERM}, \pi) - \mathcal{R}(\widetilde{h}_{Sketched}, \pi)\right| \lesssim \sup_{h \in \mathcal{H}} |\mathcal{R}(h, \pi) - \mathcal{R}(h, \widehat{\pi}_n)|.$$

- Since $\widehat{h}_{ERM}$ and $\widetilde{h}_{Sketched}$ are two ERMs based on the true empirical $\widehat{\pi}_n$ and its reconstruction $\widetilde{\pi}$, a sufficient condition for the above is

$$\sup_{h \in \mathcal{H}} |\mathcal{R}(h, \pi) - \mathcal{R}(h, \widetilde{\pi})| \lesssim \sup_{h \in \mathcal{H}} |\mathcal{R}(h, \pi) - \mathcal{R}(h, \widehat{\pi}_n)|.$$

GOAL FOR THEORY

- Remember from initial considerations we aim (ideally) at

$$\left|\mathcal{R}(\widehat{h}_{ERM},\pi)-\mathcal{R}(\widetilde{h}_{Sketched},\pi)\right| \lesssim \sup_{h\in\mathcal{H}}|\mathcal{R}(h,\pi)-\mathcal{R}(h,\widehat{\pi}_n)|.$$

- Since $\widehat{h}_{ERM}$ and $\widetilde{h}_{Sketched}$ are two ERMs based on the true empirical $\widehat{\pi}_n$ and its reconstruction $\widetilde{\pi}$, a sufficient condition for the above is

$$\sup_{h\in\mathcal{H}}|\mathcal{R}(h,\pi)-\mathcal{R}(h,\widetilde{\pi})| \lesssim \sup_{h\in\mathcal{H}}|\mathcal{R}(h,\pi)-\mathcal{R}(h,\widehat{\pi}_n)|.$$

  Using notation $\|\rho\|_{\mathcal{L}(\mathcal{H})} := \sup_{h\in\mathcal{H}}|\mathcal{R}(h,\rho)|$, rewrite as

$$\|\pi-\Delta(\mathcal{A}_\Phi(\pi'))\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\pi-\pi'\|_{\mathcal{L}(\mathcal{H})}.$$

- Since the reconstruction is obtained from the sketch information only, it is reasonable to aim at

$$\|\pi-\Delta(\mathcal{A}_\Phi(\pi'))\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\mathcal{A}_\Phi(\pi-\pi')\|_2.$$

Abstract compression/decoding results

- Assume we have a "model" $\mathfrak{S} \subset \mathfrak{M}$ so that the sketching operator satisfies the following lower restricted isometry property:

$$\forall \pi, \pi' \in \mathfrak{S} \qquad \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \leq C_{\mathcal{A}} \|\mathcal{A}(\pi - \pi')\|_2 . \qquad \text{(LRIP)}$$

ABSTRACT COMPRESSION/DECODING RESULTS

- Assume we have a "model" $\mathfrak{S} \subset \mathfrak{M}$ so that the sketching operator satisfies the following lower restricted isometry property:

$$\forall \pi, \pi' \in \mathfrak{S} \qquad \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \leq C_{\mathcal{A}} \|\mathcal{A}(\pi - \pi')\|_2 . \qquad \text{(LRIP)}$$

- Then the "ideal decoder"

$$\Delta(s) = \operatorname*{Arg\,Min}_{\pi \in \mathfrak{S}} \|s - \mathcal{A}(\pi)\|_2$$

satisfies the following instance optimality property for any π, π':

$$\|\pi - \Delta(\mathcal{A}(\pi'))\|_{\mathcal{L}(\mathcal{H})} \lesssim d(\pi, \mathfrak{S}) + \|\mathcal{A}(\pi - \pi')\|_2 ,$$

with

$$d(\pi, \mathfrak{S}) = \inf_{\sigma \in \mathfrak{S}} \left(\|\pi - \sigma\|_{\mathcal{L}(\mathcal{H})} + 2C_{\mathcal{A}} \|\mathcal{A}(\pi - \sigma)\|_2 \right) .$$

- (Conversely, the above property implies a LRIP inequality).

(Bourrier et al, 2014)

Blueprint for sketched learning method

- Define suitable restricted model for distributions $\mathfrak{S}$. Generally it should include distributions whose risk vanishes.

BLUEPRINT FOR SKETCHED LEARNING METHOD

- Define suitable restricted model for distributions $\mathfrak{S}$. Generally it should include distributions whose risk vanishes.
- Find suitable sketching dimension m and features Φ so that the corresponding sketching operator $\mathcal{A}_\Phi$ satisfies a LRIP inequality, restricted to model $\mathfrak{S}$.

Blueprint for sketched learning method

- Define suitable restricted model for distributions $\mathfrak{S}$. Generally it should include distributions whose risk vanishes.
- Find suitable sketching dimension m and features Φ so that the corresponding sketching operator $\mathcal{A}_\Phi$ satisfies a LRIP inequality, restricted to model $\mathfrak{S}$.
- Define the ideal decoder from sketch s

$$\Delta(s) = \operatorname*{Arg\,Min}_{\pi \in \mathfrak{S}} \|s - \mathcal{A}_\Phi(\pi)\|_2 \, .$$

- **For theory:** interpret the resulting instance optimality bound in terms of the learning risk.
- **For practice:** find suitable approximation of the ideal decoder if it is computationally too demanding.

Outline

Warm up: sketched PCA

- The risk is the PCA reconstruction error

$$\mathcal{R}_{PCA}(\pi, h) = \mathbb{E}_{X\sim\pi}\left[\|X - P_h X\|_2^2\right],$$

where hypothesis space $\mathcal{H}$ = linear subspaces of dimension k and P_h = orthogonal projector onto h.

WARM UP: SKETCHED PCA

- The risk is the PCA reconstruction error

$$\mathcal{R}_{PCA}(\pi, h) = \mathbb{E}_{X\sim\pi}\left[\|X - P_h X\|_2^2\right],$$

where hypothesis space $\mathcal{H}$ = linear subspaces of dimension k and P_h = orthogonal projector onto h.

- To construct $\mathcal{A}_\Phi$, use a linear operator $\mathcal{M}$ to $\mathbb{R}^m$ satisfying the RIP

$$1 - \delta \leq \frac{\|\mathcal{M}(M)\|_2^2}{\|M\|_{Frob}^2} \leq 1 + \delta$$

for all matrices M of rank less than k.
($m = O(kd)$ using random linear operator, Candès and Plan 2011)

WARM UP: SKETCHED PCA

- The risk is the PCA reconstruction error

$$\mathcal{R}_{PCA}(\pi, h) = \mathbb{E}_{X\sim\pi}\left[\|X - P_h X\|_2^2\right],$$

  where hypothesis space $\mathcal{H}$ = linear subspaces of dimension k and P_h = orthogonal projector onto h.
- To construct $\mathcal{A}_\Phi$, use a linear operator $\mathcal{M}$ to $\mathbb{R}^m$ satisfying the RIP

$$1 - \delta \leq \frac{\|\mathcal{M}(M)\|_2^2}{\|M\|_{Frob}^2} \leq 1 + \delta$$

  for all matrices M of rank less than k.
  ($m = O(kd)$ using random linear operator, Candès and Plan 2011)
- **Sketch:** $\mathcal{A}_\Phi(\widehat{\pi}_n) = \mathcal{M}(\widehat{\Sigma}_n)$ (apply $\mathcal{M}$ to empirical covar. matrix $\widehat{\Sigma}$.)
- **Reconstruct from a sketch s:** find

$$\widetilde{\Sigma} = \operatorname*{Arg\,Min}_{\operatorname{rank}(M)\leq k} \|s - \mathcal{M}(M)\|_2 .$$

- **Output:** $\widetilde{h}$ = space spanned by k first eigenvectors of $\widetilde{\Sigma}$.

THEORETICAL GUARANTEE

For any distribution π on $B(0, R)$, we have the bound (w.h.p. over data sampling)

$$\mathcal{R}_{PCA}(\pi, \widetilde{h}) - \mathcal{R}_{PCA}(\pi, h^*) \leq C\left(\sqrt{k}\mathcal{R}_{PCA}(\pi, h^*) + R^2\sqrt{\frac{k}{n}}\right).$$

- independent of total data dimension
- the first factor $\sqrt{k}$ may be spared using more precise results from low rank matrix sensing (also convex relaxation of reconstruction program for better computational efficiency)

Sketched clustering: setting

- Consider k-means or k-medians. Assume data is bounded by R.
- **Hypothesis space:** $\mathcal{H} = \mathcal{H}_{k,2\varepsilon,R}$, set of cluster centroids $h = (c_1, \ldots, c_k)$ that are R-bounded and pairwise 2ε-separated.
- Loss function

$$\ell(x,h) = \min_{1 \le i \le k} \|x - c_i\|_2^p ,$$

with $p = 1$ for k-medians, $p = 2$ for k-means.

SKETCHED CLUSTERING: SETTING

- Consider k-means or k-medians. Assume data is bounded by R.
- **Hypothesis space:** $\mathcal{H} = \mathcal{H}_{k,2\varepsilon,R}$, set of cluster centroids $h = (c_1, \ldots, c_k)$ that are R-bounded and pairwise 2ε-separated.
- Loss function
$$\ell(x, h) = \min_{1 \le i \le k} \|x - c_i\|_2^p \,,$$
with $p = 1$ for k-medians, $p = 2$ for k-means.
- **Restricted model:** $\mathfrak{S} = \mathfrak{S}_{k,2\varepsilon,R}$ set of k-point distributions whose support is in $\mathcal{H}_{k,2\varepsilon,R}$.

SKETCHED CLUSTERING: SKETCHING

- **Fourier features**: consider scaled Fourier features

$$\Phi_\omega(x) = \frac{C_\omega}{\sqrt{m}} e^{i\omega^t x} ,$$

where $C_\omega \simeq d/((1 + \varepsilon\|\omega\|) \log k)$.

SKETCHED CLUSTERING: SKETCHING

- **Fourier features**: consider scaled Fourier features

$$\Phi_\omega(x) = \frac{C_\omega}{\sqrt{m}} e^{i\omega^t x} ,$$

where $C_\omega \simeq d/((1 + \varepsilon\|\omega\|) \log k)$.

- **Random frequency vectors**: draw $\omega_1, \ldots, \omega_m$ i.i.d. in $\mathbb{R}^d$ from the distribution with density

$$\Lambda(\omega) \propto (1 + \varepsilon^2\|\omega\|^2) \exp(-\varepsilon^2\|\omega\|^2/(2 \log k)) .$$

- The sketching operator $\mathcal{A}_\Phi$ corresponds to the random Fourier features (Φ_{ω_i}), $i = 1, \ldots, m$.

Sketched clustering: Reconstruction

- **Reconstruct from a sketch s:** find

$$\widetilde{\pi} = \operatorname*{Arg\,Min}_{\pi \in \mathfrak{S}_{k,2\varepsilon,R}} \| s - \mathcal{A}_\Phi(\pi) \|_2 \,.$$

- **Output:** centroids given by support of $\widetilde{\pi}$.

SKETCHED CLUSTERING: RECONSTRUCTION

- **Reconstruct from a sketch s:** find

$$\widetilde{\pi} = \operatorname*{Arg\,Min}_{\pi \in \mathfrak{S}_{k,2\varepsilon,R}} \| s - \mathcal{A}_{\Phi}(\pi) \|_2 \,.$$

- **Output:** centroids given by support of $\widetilde{\pi}$.

- **Theoretical guarantee on reconstruction:** if

$$m \geq k^2 d^3 \texttt{polylog}(k,d) \log\left(\frac{R}{\varepsilon}\right),$$

then for any distribution π on $\mathcal{B}(0,R)$, with high probability on the draw of frequencies and of the data, it holds

$$\mathcal{R}(\pi, \widetilde{h}) - \mathcal{R}(\pi, h^*) \lesssim \frac{R^p \sqrt{k \log k}}{\varepsilon} \mathcal{R}(\pi, h^*)^{\frac{1}{p}} + \frac{R^p d \sqrt{k} \log k}{\sqrt{n}}.$$

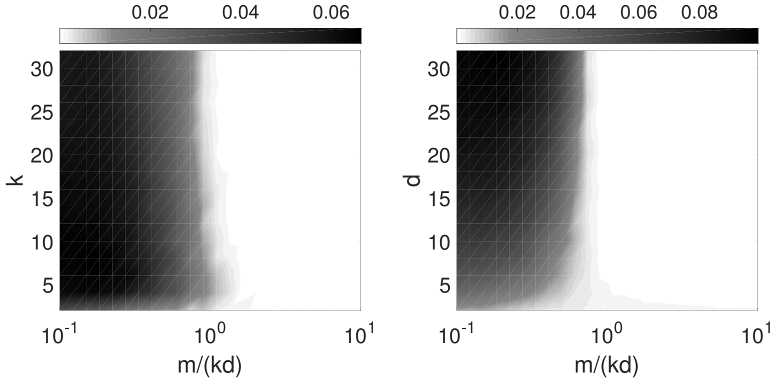
SKETCHED CLUSTERING: EXPERIMENTS

Simplifications (or cut corners. . .) for experiments:

- Use regular Gaussian density for frequency drawing (no weighting)
- Use heuristic greedy search for the reconstruction operator
- Ignore the 2ε-separation constraint for reconstruction

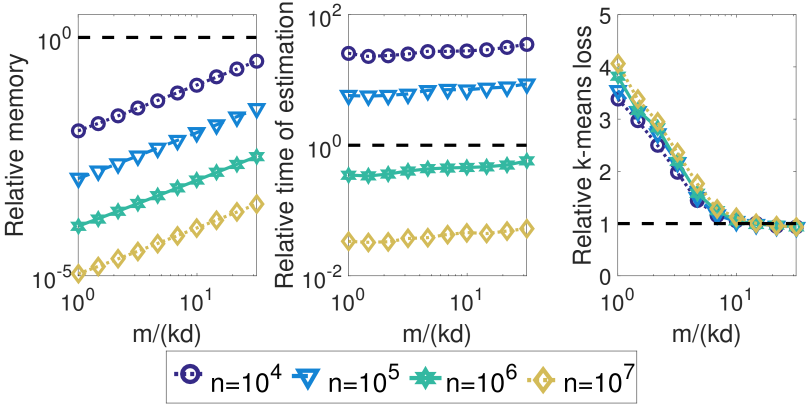
SKETCHED CLUSTERING: EXPERIMENTS

Data: mixture of 10 Gaussians with uniform weights and centers drawn from a Gaussian



Normalized k-means risk, on $n = 10^4 k$ points uniformly drawn in $[0, 1]^d$, $d = 10$ (left), $k = 10$ (right).

Sketched clustering: experiments



Relative time, memory and k-means risk of CKM with respect to k-means (10^0 represents the k-means result). ($d = 10$)

Outline

CONSTRUCTING A SUITABLE SKETCHING OPERATOR

- **Core of approach:** finding a sketching operator $\mathcal{A}_\Phi$ satisfying LRIP.

Constructing a suitable sketching operator

- **Core of approach:** finding a sketching operator $\mathcal{A}_\Phi$ satisfying LRIP.
- Use as intermediary a kernel Hilbert norm $\|.\|_\kappa$ satisfying LRIP:

$$\forall \pi, \pi' \in \mathfrak{S} \qquad \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\pi - \pi'\|_\kappa,$$

where κ is a reproducing kernel and $\|\pi\|_\kappa^2 = \mathbb{E}_{X,X' \sim \pi^{\otimes 2}}[\kappa(X, X')]$.

Constructing a suitable sketching operator

- **Core of approach:** finding a sketching operator $\mathcal{A}_\Phi$ satisfying LRIP.
- Use as intermediary a kernel Hilbert norm $\|.\|_\kappa$ satisfying LRIP:

$$\forall \pi, \pi' \in \mathfrak{S} \qquad \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\pi - \pi'\|_\kappa,$$

where κ is a reproducing kernel and $\|\pi\|_\kappa^2 = \mathbb{E}_{X,X'\sim\pi^{\otimes 2}}[\kappa(X, X')]$.

- Assume on the other hand the following representation holds:

$$\kappa(x, x') = \mathbb{E}_{\omega\sim\Lambda}\left[\phi_\omega(x)\overline{\phi_\omega(x')}\right],$$

where (ϕ_ω) is a family of complex-valued feature functions.

CONSTRUCTING A SUITABLE SKETCHING OPERATOR

- **Core of approach:** finding a sketching operator $\mathcal{A}_\Phi$ satisfying LRIP.
- Use as intermediary a kernel Hilbert norm $\|.\|_\kappa$ satisfying LRIP:

$$\forall \pi, \pi' \in \mathfrak{S} \qquad \|\pi - \pi'\|_{\mathcal{L}(\mathcal{H})} \lesssim \|\pi - \pi'\|_\kappa ,$$

  where κ is a reproducing kernel and $\|\pi\|_\kappa^2 = \mathbb{E}_{X,X' \sim \pi^{\otimes 2}}[\kappa(X, X')]$.

- Assume on the other hand the following representation holds:

$$\kappa(x, x') = \mathbb{E}_{\omega \sim \Lambda}\left[\phi_\omega(x)\overline{\phi_\omega(x')}\right],$$

  where (ϕ_ω) is a family of complex-valued feature functions.

- **Strategy:** sample random features $\omega_i \sim \Lambda$, ensuring (w.h.p.) the corresponding sketching operator delivers good enough approximation to $\|.\|_\kappa$ i.e.

$$\forall \pi, \pi' \in \mathfrak{S} \qquad \|\pi - \pi'\|_\kappa \lesssim \|\mathcal{A}_\Phi(\pi - \pi')\|_2 .$$

Dimension of sketch required

- Uniform approximation of the kernel norm by the sketching norm obtained via Bernstein's inequality + covering argument on the normalized **secant set**

$$\mathcal{S}_{\|.\|_\kappa}(\mathfrak{S}) = \left\{ \frac{\pi - \pi'}{\|\pi - \pi'\|_\kappa} \middle| \pi, \pi' \in \mathfrak{S} \right\}.$$

DIMENSION OF SKETCH REQUIRED

- Uniform approximation of the kernel norm by the sketching norm obtained via Bernstein's inequality + covering argument on the normalized **secant set**

$$\mathcal{S}_{\|.\|_\kappa}(\mathfrak{S}) = \left\{ \frac{\pi - \pi'}{\|\pi - \pi'\|_\kappa} \middle| \pi, \pi' \in \mathfrak{S} \right\}.$$

- More precisely we find the sufficient condition

$$m \gtrsim \log \mathcal{N}(\mathcal{S}_{\|.\|_\kappa}(\mathfrak{S}), d_{\mathcal{F}}, 1/2)\,,$$

where $d_{\mathcal{F}}(\pi, \pi') = \sup_\omega \left| \left|\mathbb{E}_{X\sim\pi}[\Phi_\omega(X)]\right|^2 - \left|\mathbb{E}_{X\sim\pi'}[\Phi_\omega(X)]\right|^2 \right|.$

DIMENSION OF SKETCH REQUIRED

- Uniform approximation of the kernel norm by the sketching norm obtained via Bernstein's inequality + covering argument on the normalized **secant set**

$$\mathcal{S}_{\|.\|_\kappa}(\mathfrak{S}) = \left\{ \frac{\pi - \pi'}{\|\pi - \pi'\|_\kappa} \middle| \pi, \pi' \in \mathfrak{S} \right\}.$$

- More precisely we find the sufficient condition

$$m \gtrsim \log \mathcal{N}(\mathcal{S}_{\|.\|_\kappa}(\mathfrak{S}), d_{\mathcal{F}}, 1/2) \,,$$

where $d_{\mathcal{F}}(\pi, \pi') = \sup_\omega \left| |\mathbb{E}_{X\sim\pi}[\Phi_\omega(X)]|^2 - |\mathbb{E}_{X\sim\pi'}[\Phi_\omega(X)]|^2 \right|.$

- Finally, the vectorial form of Bernstein's inequality can be used again (this time on the data) to control the estimation noise $\|\mathcal{A}_\Phi(\pi - \widehat{\pi}_n)\|_2$.

Application to mixtures and clustering

Overview of remaining steps to obtain bound on risk and sketch dimension:

- **Establish** the LRIP between the risk norm $\|.\|_{\mathcal{L}(\mathcal{H})}$ and the kernel norm $\|.\|_{\kappa}$ on the model $\mathfrak{S}$.
  - Results obtained for general family of RBF-type kernels and models given by k-mixtures of distributions

APPLICATION TO MIXTURES AND CLUSTERING

Overview of remaining steps to obtain bound on risk and sketch dimension:

- **Establish** the LRIP between the risk norm $\|.\|_{\mathcal{L}(\mathcal{H})}$ and the kernel norm $\|.\|_{\kappa}$ on the model $\mathfrak{S}$.
  - Results obtained for general family of RBF-type kernels and models given by k-mixtures of distributions
- **Bound the (log) covering numbers**: requires some classical inequalities between covering numbers

Application to mixtures and clustering

Overview of remaining steps to obtain bound on risk and sketch dimension:

- **Establish** the LRIP between the risk norm $\|.\|_{\mathcal{L}(\mathcal{H})}$ and the kernel norm $\|.\|_{\kappa}$ on the model $\mathfrak{S}$.
  - Results obtained for general family of RBF-type kernels and models given by k-mixtures of distributions
- **Bound the (log) covering numbers**: requires some classical inequalities between covering numbers
- Once the instance optimality inequality is obtained, relate back the terms of the bound to the learning task (learning risk).

CONCLUSION

- The sketched learning framework holds promise to reduce computation and memory burden
- General theoretical framework based on:
  - LRIP/compressed sensing recovery principles
  - Kernel embeddings and random features
- Theoretical recovery guarantees and bounds on the sketch dimension needed
- Applications:
  - sketched PCA
  - sketched clustering
  - skteched mixture of Gaussians estimation
  - . . . more to come?

SketchML matlab toolbox available:
(large-scale mixture learning using sketches)

`http://sketchml.gforge.inria.fr/`

ArXiv Preprint:

Compressive Statistical Learning with Random Feature Moments
R. Gribonval, G. Blanchard, N. Keriven, Y. Traonmilin
`https://arxiv.org/abs/1706.07180`