Supplementary materials: Recursive nearest agglomeration (ReNA) – fast clustering for approximation of structured signals

1 Nyström feature mapping

Here, we present the standard implementation of the Nyström approximation for linear kernels. Algorithm 2 allows to build a data-driven feature mapping that is used to reduce the dimensionality of the data matrix. The algorithm is summarized as follows: first, we select images uniformly at random[1], then we calculate the kernel of these samples and use it to normalize the selected images.

Algorithm 2 Nyström: Learning the feature mapping

Require: The training data matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$, number k of components, where $k < n$.
Ensure: The feature mapping $\mathbf{\Phi}_{\mathsf{Nys}} \in \mathbb{R}^{k \times p}$
1: $\mathbf{r} \leftarrow$ Generate uniform sampling of k components
2: $\mathbf{X}_{*,\mathbf{r}} \in \mathbb{R}^{p \times k}$ {Subsample of k columns}
3: $\tilde{\mathbf{K}} = \mathbf{X}_{*,\mathbf{r}}^{\mathsf{T}} \mathbf{X}_{*,\mathbf{r}}$ {Kernel matrix of the subsampled data}
4: $\mathbf{\Phi}_{\mathsf{Nys}} = \tilde{\mathbf{K}}^{-1/2} \mathbf{X}_{*,\mathbf{r}}$ {Normalization: via SVD}
5: **return** $\mathbf{\Phi}_{\mathsf{Nys}}$

2 Proof of Lemma 2.2

In this appendix, we present the arithmetic manipulation necessary to prove the first part of the Lemma 2.2. As $\mathbf{\Phi}_{\mathsf{FG}}^{\mathsf{T}} \mathbf{\Phi}_{\mathsf{FG}}$ is an orthogonal operator, Eq. 6 corresponds to an orthogonal decomposition. But, for the sake of clarity we include it in this appendix.

Corollary 2.1. Let $\mathbf{x} \in \mathbb{R}^p$ be a signal, and $\mathbf{\Phi}_{\mathsf{FG}}$ be a feature-grouping matrix, the following holds

$$\|\mathbf{x}\|_2^2 - \sum_{q=1}^{k} \left\| \mathbf{x}_{\mathcal{C}_q} - \frac{(\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x})_q}{\sqrt{|\mathcal{C}_q|}} \right\|_2^2 = \|\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x}\|_2^2 \leq \|\mathbf{x}\|_2^2. \tag{14}$$

1. We can use other sampling probabilities (e.g. the leverage score [1]), but by sampling uniformly we are assuming a regular structure

Proof 2.1. We start by writing down the ℓ_2 norm of the data vector $\mathbf{x}$ for every point inside all the clusters $\{\mathcal{C}_q\}_{q=1}^k$. Then, we perform simple manipulations, as follows

$$\begin{aligned} \|\mathbf{x}\|_2^2 &= \sum_{q=1}^{k} \sum_{i \in \mathcal{C}_q} \mathbf{x}_i^2 \\ &= \sum_{q=1}^{k} \sum_{i \in \mathcal{C}_q} \left(\mathbf{x}_i^2 + \frac{(\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x})_q^2}{|\mathcal{C}_q|} - \frac{(\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x})_q^2}{|\mathcal{C}_q|} \right) \\ &= \sum_{q=1}^{k} (\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x})_q^2 + \sum_{q=1}^{k} \sum_{i \in \mathcal{C}_q} \left(\mathbf{x}_i^2 - \frac{(\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x})_q^2}{|\mathcal{C}_q|} \right) \\ &= \|\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x}\|_2^2 + \sum_{q=1}^{k} \left\| \mathbf{x}_{\mathcal{C}_q} - \frac{(\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x})_q}{\sqrt{|\mathcal{C}_q|}} \right\|_2^2. \end{aligned} \tag{15}$$

Finally, the right hand inequality of Eq.14 comes naturally after scaling each cluster (the non-zero singular values are set to 1). □

Proof 2.2. The corollary 2.1 shows that the lower bound of the representation is only affected by the inertia (see left hand side of Eq.14). So, the worst case corresponds to the upper bound of the inertia $M(\mathbf{x}) = \sum_{q=1}^{k} m_q(\mathbf{x})$. Then, the inertia of each cluster can be bounded as follows

$$\begin{aligned} m_q(\mathbf{x}) &= \left\| \mathbf{x}_{\mathcal{C}_q} - \frac{(\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x})_q}{\sqrt{|\mathcal{C}_q|}} \right\|_2^2 \\ &= \sum_{j \in \mathcal{C}_q} \left(\mathbf{x}_j - \frac{1}{|\mathcal{C}_q|} \sum_{i \in \mathcal{C}_q} \mathbf{x}_i \right)^2 \\ &\leq |\mathcal{C}_q| \max_{j \in \mathcal{C}_q} \left| \mathbf{x}_j - \frac{1}{|\mathcal{C}_q|} \sum_{i \in \mathcal{C}_q} \mathbf{x}_i \right|^2 \end{aligned} \tag{16}$$

The last inequality corresponds to the worst case for the sum inside the cluster.

Let constrain our analysis to L-smooth signals $\mathbf{x} \in \mathbb{R}^p$ structured by graph $\mathcal{G}$ (see Definition 2.1). Under this assumption, we can bound the inertia of each cluster as

follows:

$$\begin{aligned} m_q(\mathbf{x}) &\leq |\mathcal{C}_q| \max_{j\in\mathcal{C}_q} \left| \mathbf{x}_j - \frac{1}{|\mathcal{C}_q|}\sum_{i\in\mathcal{C}_q}\mathbf{x}_i \right|^2 \\ &\leq |\mathcal{C}_q|\, L^2 \sup_{i,j\in\mathcal{C}_q} \mathrm{dist}_{\mathcal{G}}(v_i,v_j)^2 \\ &= L^2\, |\mathcal{C}_q|\, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2, \end{aligned} \tag{17}$$

where the second inequality follows from the pairwise L-Lipschitz condition. Finally, plugging Eq.17 into Eq.15 we have:

$$\|\mathbf{x}\|^2 - L^2 \sum_{q=1}^{k} |\mathcal{C}_q|\, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2 \leq \|\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x}\|^2 \leq \|\mathbf{x}\|_2^2. \tag{18}$$

□

Corollary 2.2. Let L_q be the smoothness index inside cluster C_q, for all $q \in [k]$. This is the minimum L_q such that:

$$|\mathbf{x}_i - \mathbf{x}_j| \leq L_q\, \mathrm{dist}_{\mathcal{G}}(v_i,\, v_j), \quad \forall (i,j) \in C_q^2.$$

Then the following two inequalities hold:

$$\begin{aligned} \|\mathbf{x}\|_2^2 - \sum_{q=1}^{k} |\mathcal{C}_q| \sup_{\mathbf{x}_i,\mathbf{x}_j\in\mathbf{x}_{\mathcal{C}_q}} |\mathbf{x}_i - \mathbf{x}_j|_2^2 \leq \\ \|\mathbf{x}\|_2^2 - \sum_{q=1}^{k} L_q^2\, |\mathcal{C}_q|\, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2 \leq \|\mathbf{\Phi}_{\mathsf{FG}}\, \mathbf{x}\|_2^2. \end{aligned} \tag{19}$$

Proof 2.3. As in Eq.17, the second inequality is consequence of adding the local L-smoothness condition. □

3 Denoising properties

In this section, we analyze the regularity condition of the signal of interest $\mathbf{s}$, its relation with the noise and maximum cluster size. Let $\mathbf{x}$ be the acquired signal, which is a fixed signal of interest $\mathbf{s}$ contaminated with an i.i.d. zero-mean Gaussian noise $\mathbf{n}$ with variance σ^2, $\mathbf{x} = \mathbf{s} + \mathbf{n}$.

With the purpose of ensuring clarity, we define $\mathbf{A} = \mathbf{\Phi}_{\mathsf{FG}}^{\mathsf{T}}\mathbf{\Phi}_{\mathsf{FG}}$. Let $\mathrm{MSE}_{\mathrm{appox}} = \mathbb{E}_{\mathbf{n}}\left[\|\mathbf{s} - \mathbf{A}\,\mathbf{x}\|_2^2\right]$ and $\mathrm{MSE}_{\mathrm{orig}} = \mathbb{E}_{\mathbf{n}}\left[\|\mathbf{n}\|_2^2\right]$ be the mean squared error with and without approximation, receptively. As we are dealing with Gaussian noise, the risk of the raw data is $\mathrm{MSE}_{\mathrm{orig}} = p\,\sigma^2$.

Given that $\|\mathbf{s}\|_2^2$ is fixed, it is enough to show $\mathrm{MSE}_{\mathrm{approx}} \leq \mathrm{MSE}_{\mathrm{orig}}$ to ensure an increase in the SNR.

Proposition 3.1. Let $\mathbf{x} = \mathbf{s} + \mathbf{n} \in \mathbb{R}^p$ be an acquired signal, where $\mathbf{s}$ is a fixed smooth L-Lipschitz signal and $\mathbf{n}$ an i.i.d. zero-mean Gaussian noise with variance σ^2. Then, for a given grouping matrix $\mathbf{\Phi}_{\mathsf{FG}} \in \mathbb{R}^{k\times p}$ the mean squared error of the approximation ($\mathrm{MSE}_{\mathrm{approx}}$) is upper-bounded by

$$\mathrm{MSE}_{\mathrm{approx}} \leq L^2 \sum_{q=1}^{k} |\mathcal{C}_q|\, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2 + \frac{k}{p}\,\mathrm{MSE}_{\mathrm{orig}}. \tag{20}$$

Proof 3.1. We start by writing down the $\mathrm{MSE}_{\mathrm{approx}}$, then we separate the components thanks to the i.i.d assumption and plug the upper-bound of the inertia, as follows

$$\begin{aligned} \mathrm{MSE}_{\mathrm{approx}} &= \mathbb{E}_{\mathbf{n}}\left[\|\mathbf{s} - \mathbf{A}\,\mathbf{x}\|_2^2\right] \\ &= \|(\mathbf{I} - \mathbf{A})\,\mathbf{s}\|_2^2 + \mathbb{E}_{\mathbf{n}}\left[\|\mathbf{A}\,\mathbf{n}\|_2^2\right] \\ &= \|(\mathbf{I} - \mathbf{A})\,\mathbf{s}\|_2^2 + k\,\sigma^2 \\ &\leq L^2 \sum_{q=1}^{k} |\mathcal{C}_q|\, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2 + k\,\sigma^2. \end{aligned}$$

□

Corollary 3.1. Let $\mathbf{x} = \mathbf{s} + \mathbf{n} \in \mathbb{R}^p$ be an acquired signal, where $\mathbf{s}$ is a fixed a pairwise smooth L-Lipschitz signal and $\mathbf{n}$ is an i.i.d. zero-mean Gaussian noise with variance σ^2. For a given grouping matrix $\mathbf{\Phi}_{\mathsf{FG}} \in \mathbb{R}^{k\times p}$, the noise after approximation is reduced, $\mathrm{MSE}_{\mathrm{approx}} \leq \mathrm{MSE}_{\mathrm{orig}}$, only if the L^2 smoothness parameter satisfy

$$L^2 \leq \frac{(p-k)}{\sum_{q=1}^{k} |\mathcal{C}_q|\, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2}\, \sigma^2. \tag{21}$$

Proof 3.2. This is a direct result of the proposition3.1 after some arithmetic manipulations,

$$\begin{aligned} \mathrm{MSE}_{\mathrm{approx}} &\leq L^2 \sum_{q=1}^{k} |\mathcal{C}_q|\, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2 + p\,\sigma^2 - (p-k)\,\sigma^2 \\ &\leq L^2 \sum_{q=1}^{k} |\mathcal{C}_q|\, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2 + \mathrm{MSE}_{\mathrm{orig}} - (p-k)\,\sigma^2. \end{aligned}$$

Then, to satisfy $\mathrm{MSE}_{\mathrm{approx}} \leq \mathrm{MSE}_{\mathrm{orig}}$, we must have:

$$L^2 \sum_{q=1}^{k} |\mathcal{C}_q|\, \mathrm{diam}_{\mathcal{G}}(\mathcal{C}_q)^2 \leq (p-k)\,\sigma^2,$$

which lead us to the upper-bound of the Lipschitz constant. □

Cluster of the same size: This is a particular case, where we assume that the clusters $\mathcal{P} = \{\mathcal{C}_q\}_{q=1}^k$ have the same size, $\frac{p}{k}$. Under this assumption, the following holds:

$$\mathrm{MSE}_{\mathrm{approx}} \leq p\left(\frac{L}{k}\right)^2 + \frac{k}{p}\,\mathrm{MSE}_{\mathrm{orig}} = O\left(\max\left\{\frac{p}{k^2}, \frac{k}{p}\right\}\right). \tag{22}$$

We need to balance both therms in the right-hand side of 22 in order to maximize the rate of decay. This implies that $\frac{p}{k^2} = \frac{k}{p}$ therefore $k = p^{2/3}$ and $\mathrm{MSE}_{\mathrm{approx}} = O(k^{-1/2})$.

4 Experiment description: distortion

To assess the quality of this approximation (see Eq.2 and Eq.8 , we randomly split half of the data to form a train and test clean signals ($\mathbf{S}^{\mathrm{train}}, \mathbf{S}^{\mathrm{test}}$) and a train corrupted data matrix $\mathbf{X}^{\mathrm{train}}$. We learn $\mathbf{\Phi}$ on the train corrupted data $\mathbf{X}^{\mathrm{train}}$. On the test data, we fit a proportionality constant η that relates the distances in reductions of the corrupted data with the corresponding distances in the clean signals.

We denote $\boldsymbol{\delta}_{(i,j)}^{\mathrm{orig}}$ the norm of the difference of the i and j uncorrupted signals, $\|\mathbf{S}_{*,i}^{\mathrm{test}} - \mathbf{S}_{*,j}^{\mathrm{test}}\|_2$, and $\boldsymbol{\delta}_{(i,j)}^{\mathrm{noisy}}$ the norm of the difference of the i and j scaled noisy signals, $\eta\|\mathbf{\Phi}\,\mathbf{X}_{*,i}^{\mathrm{test}} -$

$\mathbf{\Phi}\,\mathbf{X}^{\text{test}}_{*,j}\|_2$, for all $(i,j) \in \left[\left\lfloor \frac{n}{2} \right\rfloor\right]^2$ (note that $\boldsymbol{\delta} \in \mathbb{R}^{n(n-1)/8}$). We then use the relative distortion (RD) between $\boldsymbol{\delta}^{\text{orig}}$ and $\boldsymbol{\delta}^{\text{noisy}}$ to quantity the denoising effect of each method:

$$\text{RD}(\boldsymbol{\delta}^{\text{orig}}, \boldsymbol{\delta}^{\text{noisy}})(\text{dB}) = -10 \log_{10} \frac{\|\boldsymbol{\delta}^{\text{noisy}} - \boldsymbol{\delta}^{\text{orig}}\|_2^2}{\|\boldsymbol{\delta}^{\text{orig}}\|_2^2}. \quad (23)$$

This measure gives us an insight on the distortion and possibly denoising effect. In particular, it shows us for which fraction of the signal the condition of Eq.7 is satisfied.

REFERENCES

[1] M. W. Mahoney, "Randomized algorithms for matrices and data," *Found. Trends Mach. Learn.*, vol. 3, pp. 123–224, 2011.