Tight Kernel Query Complexity of Kernel Ridge Regression and Kernel k-means Clustering

Manuel Fernández V David P. Woodruff Taisuke Yasuda

Carnegie Mellon University

Introduction

Given a kernel matrix $\mathbf{K}$, a regularization parameter λ, and a target vector $\mathbf{z}$, the kernel ridge regression (KRR) problem asks for the minimizer of the following objective function:

$$\boldsymbol{\alpha}_{\mathrm{opt}} = \operatorname*{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^n} \|\mathbf{K}\boldsymbol{\alpha} - \mathbf{z}\|_2^2 + \lambda\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha}. \tag{1}$$

We consider the problem of outputting *approximate solutions*, i.e.

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_{\mathrm{opt}}\|_2 \le \varepsilon\|\boldsymbol{\alpha}_{\mathrm{opt}}\|_2 = \varepsilon\left\|(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{z}\right\|_2. \tag{2}$$

How many kernel entries do we need to read in order to approximate KRR?

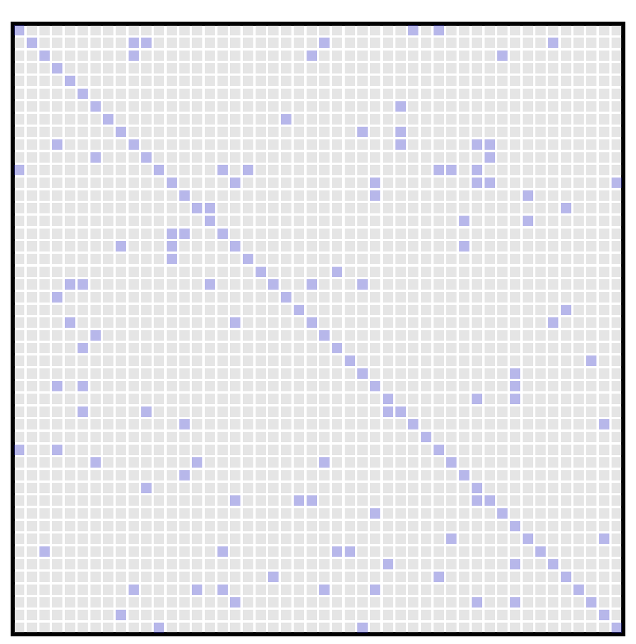

Figure 1: An algorithm might read the diagonal and sample some uniformly random entries.

The best known algorithms have kernel query complexity depending on the *effective statistical dimension*

$$d_{\mathrm{eff}}^{\lambda}(\mathbf{K}) := \operatorname{tr}\left(\mathbf{K}(\mathbf{K}+\lambda\mathbf{I}_n)^{-1}\right) = \sum_{i=1}^{\operatorname{rank}(\mathbf{K})} \frac{\sigma_i^2}{\sigma_i^2+\lambda}. \tag{3}$$

For example, the following result is proved in [MM17] using an adaptive sampling techinque known as *ridge leverage score sampling*:

Algorithm for KRR [MM17]

There is an algorithm computing a $(1+\varepsilon)$ relative error KRR solution with probability at least $2/3$ making $O(\frac{nd_{\mathrm{eff}}^{\lambda}}{\varepsilon}\log\frac{d_{\mathrm{eff}}^{\lambda}}{\varepsilon})$ kernel queries.

Main Result: KRR

The following open question was posed by El Alaoui and Mahoney [EAM15]:

Is the effective statistical dimension a lower bound on the kernel query complexity of KRR?

We answer this question affirmatively in our main result for KRR:

Lower Bound for KRR

Any algorithm computing a $(1+\varepsilon)$ relative error KRR solution with probability at least $2/3$ makes at least $\Omega(\frac{nd_{\mathrm{eff}}^{\lambda}}{\varepsilon})$ kernel queries.

Proof sketch. To show the result, we consider a reduction to the problem of labeling the block size of each row of the following kernel matrix:

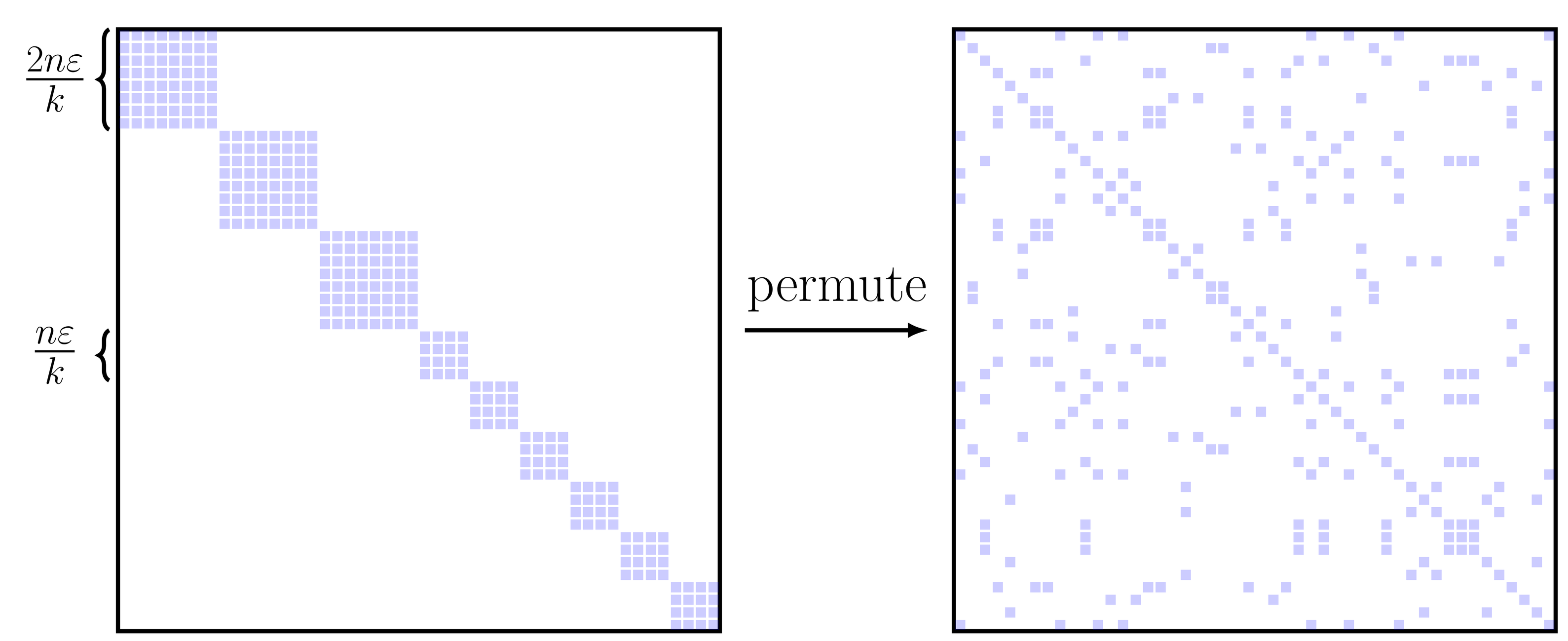

Figure 2: Hard input distribution for KRR: does the ith row have $\frac{2n\varepsilon}{k}$ or $\frac{n\varepsilon}{k}$ ones?

By standard arguments, this problem requires $\Omega(nk/\varepsilon)$ kernel queries. It is well-known that the exact solution to the KRR problem is

$$\boldsymbol{\alpha}_{\mathrm{opt}} = (\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{z}. \tag{4}$$

If we set $\mathbf{z} = \mathbf{1_n}$ and $\lambda = n/k$, then the ith coordinate of $\boldsymbol{\alpha}_{\mathrm{opt}}$ is

$$\mathbf{e}_i^\top\boldsymbol{\alpha}_{\mathrm{opt}} = \begin{cases} (2n\varepsilon/k + n/k)^{-1} = \frac{k/n}{1+2\varepsilon} & \text{if row } i \text{ has block size } 2n\varepsilon/k \\ (n\varepsilon/k + n/k)^{-1} = \frac{k/n}{1+\varepsilon} & \text{if row } i \text{ has block size } n\varepsilon/k \end{cases} \tag{5}$$

so a $(1+\varepsilon)$ relative error solution can distinguish these two cases. We also have $d_{\mathrm{eff}}^{\lambda} = \Theta(k)$ so we conclude. □

Main Result: KKMC

Next, we present our main result for kernel k-means clustering:

Lower Bound for KKMC

Any algorithm computing a $(1+\varepsilon)$ relative error KKMC solution with probability at least $2/3$ makes at least $\Omega(\frac{nk}{\varepsilon})$ kernel queries.

This result uses similar reductions as our KRR result, but the cost computations are much more involved. The hard distribution samples from a distribution supported on pairwise sums of standard basis vectors $\mathbf{e}_i + \mathbf{e}_j$ in $\mathbb{R}^{k/\varepsilon}$:

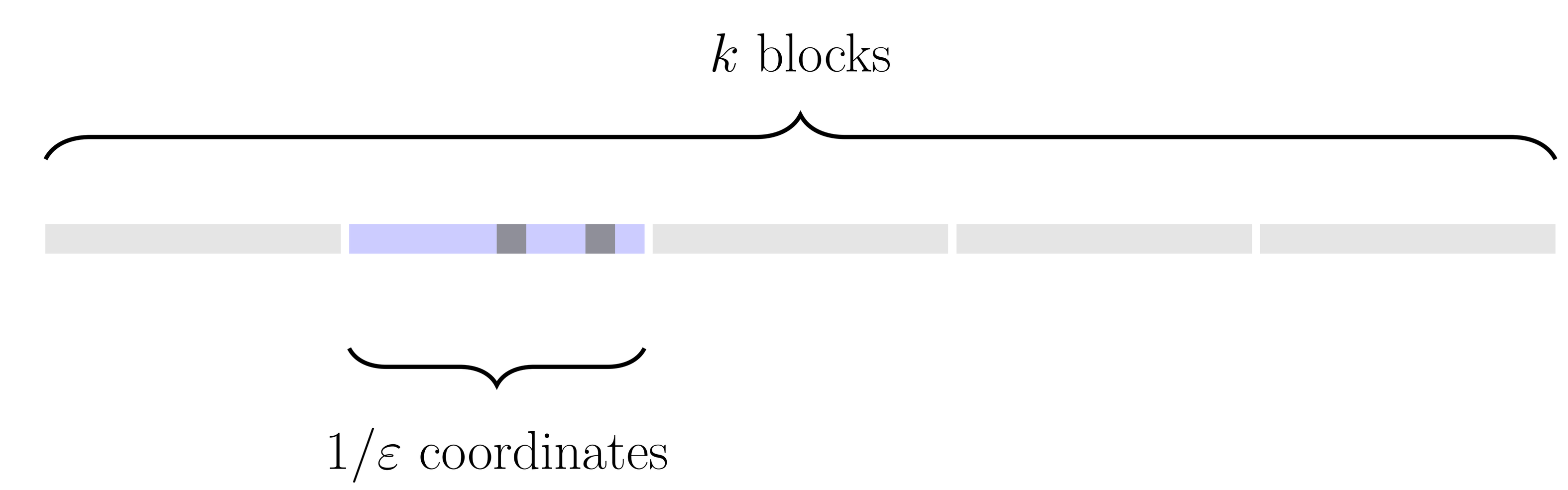

Figure 3: Hard input distribution for KKMC

Our result matches the following algorithmic result of [MM17] (also based on ridge leverage score sampling) up to log factors:

Algorithm for KKMC [MM17]

There is an algorithm computing a $(1+\varepsilon)$ relative error KKMC solution with probability at least $2/3$ making $O(\frac{nk}{\varepsilon}\log\frac{k}{\varepsilon})$ kernel queries.

References

[EAM15] Ahmed El Alaoui and Michael W Mahoney.
Fast randomized kernel methods with statistical guarantees.
In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

[MM17] Cameron Musco and Christopher Musco.
Recursive sampling for the Nyström method.
In *Advances in Neural Information Processing Systems*, pages 3833–3845, 2017.