

# TIGHT KERNEL QUERY COMPLEXITY OF KERNEL RIDGE REGRESSION AND KERNEL $k$ -MEANS CLUSTERING

Taisuke Yasuda David P. Woodruff

Carnegie Mellon University

## INTRODUCTION

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

$$\boldsymbol{\alpha}_{\text{opt}} = \underset{\boldsymbol{\alpha} \in \mathbb{R}^n}{\operatorname{argmin}} \|\mathbf{K}\boldsymbol{\alpha} - \mathbf{z}\|_2^2 + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}. \quad (1)$$

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

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_{\text{opt}}\|_2 \leq \varepsilon \|\boldsymbol{\alpha}_{\text{opt}}\|_2 = \varepsilon \|(\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{z}\|_2. \quad (2)$$

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

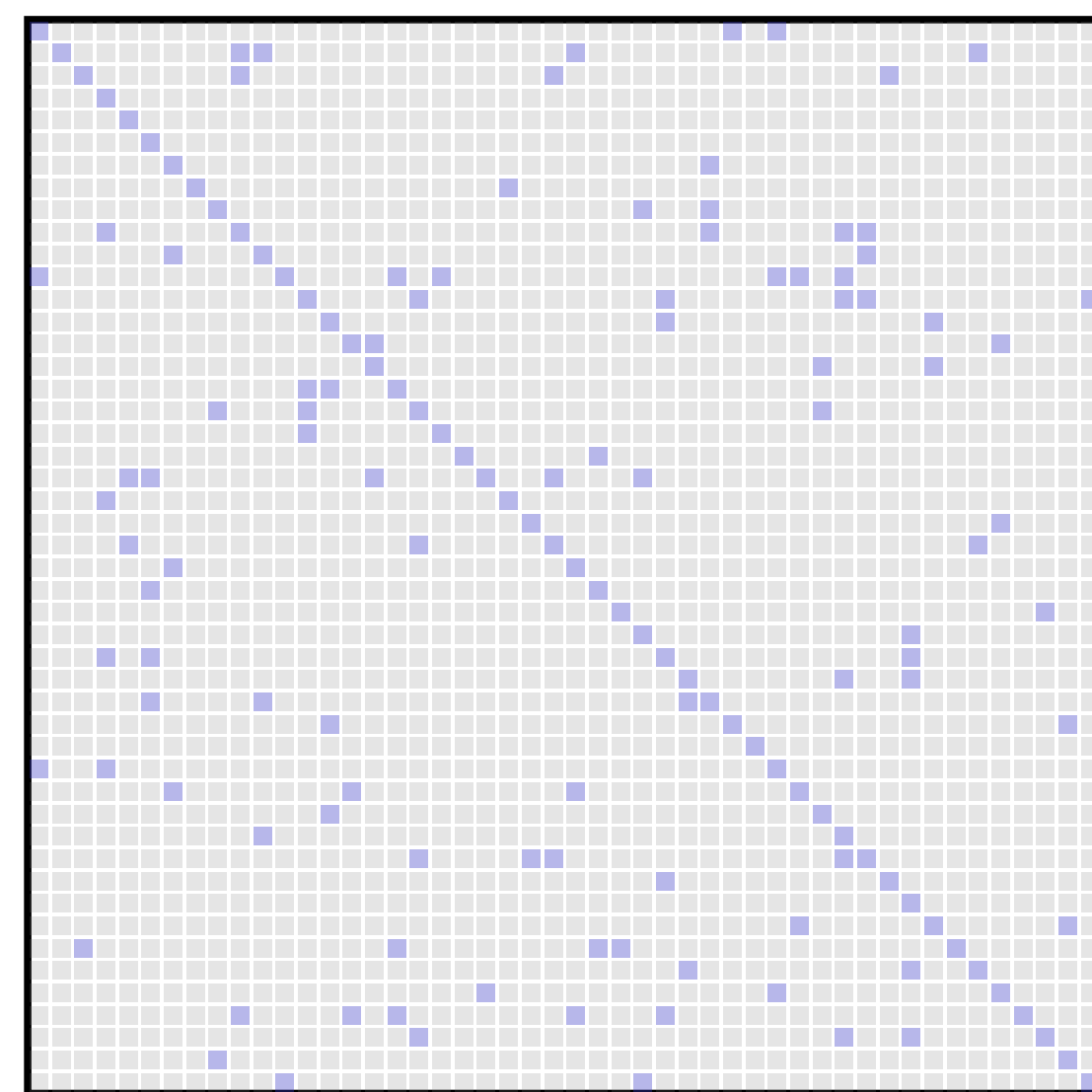


Figure 1: An algorithm might read the diagonal and sample some uniformly random entries.

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

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

For example, the following result is proved in [MM17] using an adaptive sampling technique 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  $O(\frac{nd_{\text{eff}}^\lambda}{\varepsilon} \log \frac{d_{\text{eff}}^\lambda}{\varepsilon})$  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 [FWY19]

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

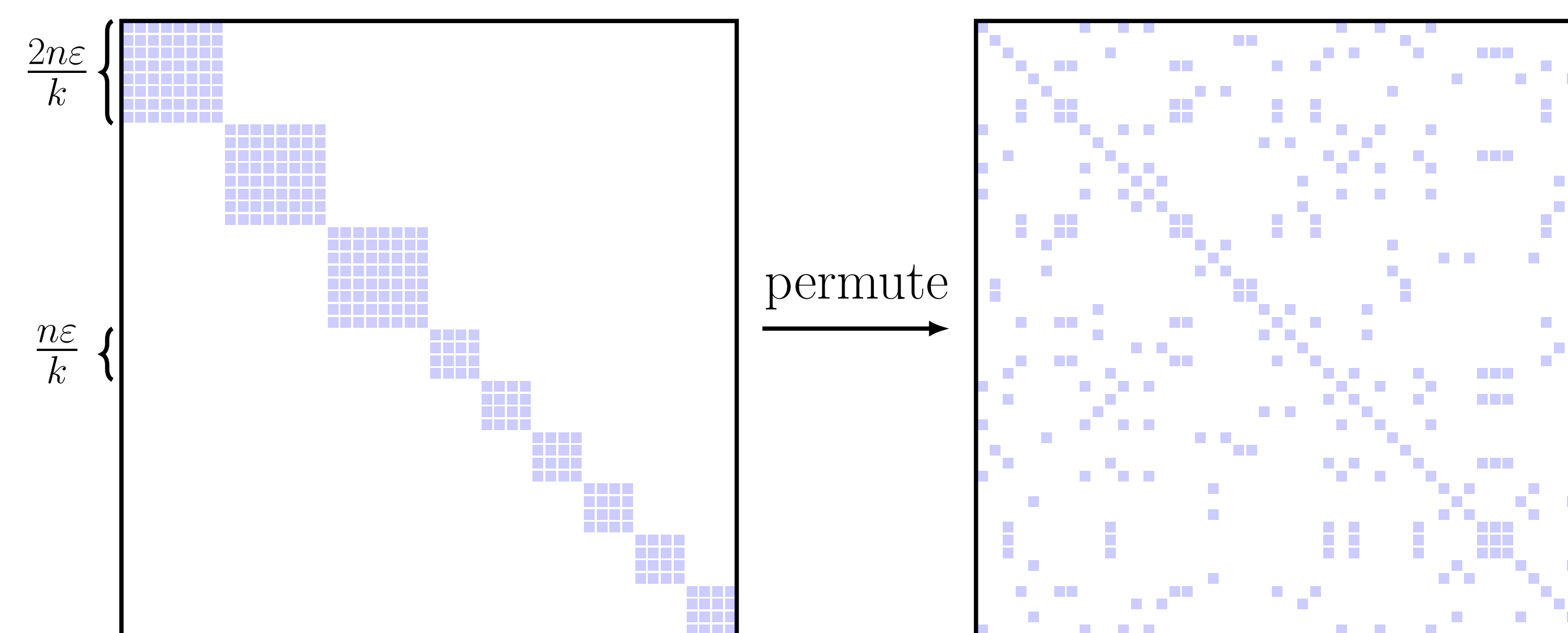


Figure 2: Hard input distribution for KRR: does the  $i$ th row have  $\frac{2n\varepsilon}{k}$  or  $\frac{n\varepsilon}{k}$  ones?

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

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

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

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

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

## MAIN RESULT: KKMC

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

## LOWER BOUND FOR KKMC [FWY19]

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. 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.
- [FWY19] Manuel Fernández, David P Woodruff, and Taisuke Yasuda. Tight kernel query complexity of kernel ridge regression and kernel  $k$ -means clustering. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.
- [MM17] Cameron Musco and Christopher Musco. Recursive sampling for the Nyström method. In *Advances in Neural Information Processing Systems*, pages 3833–3845, 2017.

## ACKNOWLEDGEMENTS

I would like to thank Manuel Fernández for working on this project with me.