Local optimisation of Nyström samples through stochastic gradient descent

Matt Hutchings
in collaboration with Bertrand Gauthier

01/07/2022

Cardiff University - School of Mathematics

Table of contents

1. Introduction to the Nyström method
2. Efficient matrix approximations and the radial SKD criterion
3. Locally optimising the radial SKD
4. Experiments
5. Conclusion

Introduction to the Nyström method

What is the Nyström method?

Low-rank approximation for a symmetric, positive semi-definite (SPSD) matrix $\mathbf{K}$.

Nyström method

1. Sample n columns from $\mathbf{K}$.
2. Construct the Nyström approximation matrix

$$\hat{\mathbf{K}} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^{T} \ :$$

$\mathbf{C}$ is the sample of columns; $\mathbf{W}$ is the principal submatrix of $\mathbf{K}$ indexed by the column sample; † denotes the Moore-Penrose generalised inverse of a matrix. $\hat{\mathbf{K}}$ is an approximation of rank at most n.

Task: find a column sample that defines an efficient approximation for $\mathbf{K}$.

Remark: The optimal rank-n approximation for $\mathbf{K}$ is formed by truncating the spectrum of $\mathbf{K}$; this becomes intractable when N is large.

Nyström method in machine learning

Kernel methods involve representing data sets with a kernel matrix.

- Data set $\mathcal{D} = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$.
- SPSD kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.
- $N \times N$ SPSD kernel matrix $\mathbf{K}$ defined by

$$[\mathbf{K}]_{i,j} = K(x_i, x_j) \text{ for } i, j \in \{1, \dots, N\}.$$

Low-rank approximations can improve efficiency of algorithms.

Sampling n columns from $\mathbf{K}$ is equivalent to sampling n points from $\mathcal{D}$. We call a sample $\mathcal{S} = \{s_1, \dots, s_n\} \subset \mathcal{D}$ a *Nyström sample*, and we refer to the s_i as *landmark points*.

Nyström method in machine learning

In the kernel matrix setting, it is not necessary to restrict $\mathcal{S}$ to a sample from $\mathcal{D}$.

We could instead sample from the ambient space $\mathbb{R}^d$.

Relaxed Nyström method for kernel matrices

1. Sample n points $\mathcal{S} = \{s_1, \dots, s_n\}$ from $\mathbb{R}^d$.
2. Construct the Nyström approximation matrix $\hat{\mathbf{K}}(\mathcal{S})$ as follows:

$$[\hat{\mathbf{K}}(\mathcal{S})]_{i,j} = \mathbf{k}^T(x_i)\mathbf{K}_{\mathcal{S}}^{\dagger}\mathbf{k}(x_j) \text{ for } i, j \in \{1, \dots, N\} :$$

$\mathbf{k}(x) = (K(x, s_1), \dots, K(x, s_n))^T$ for $x \in \mathbb{R}^d$; $\mathbf{K}_{\mathcal{S}}$ is the $n \times n$ kernel matrix defined by $\mathcal{S}$.

How do we choose landmark points that define efficient low-rank approximations?

Remark: In the RKHS framework, the matrix $\hat{\mathbf{K}}(\mathcal{S})$ is defined by the data set $\mathcal{D}$ and the kernel function $K_{\mathcal{S}}$, which is the reproducing kernel of the subspace $\mathcal{H}_{\mathcal{S}} \subset \mathcal{H}$ spanned by the functions $K(s_1, \cdot), \dots, K(s_n, \cdot)$; $\mathcal{H}$ is the RKHS associated with K.

Efficient matrix approximations and the radial SKD criterion

Matrix reconstruction errors

Three classical criteria:

- Trace norm error: $\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_*$
- Frobenius norm error: $\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_\mathrm{F}$
- Spectral norm error: $\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_2$

Costly to evaluate; cheapest is trace norm error with complexity $\mathcal{O}(n^3 + Nn^2)$. Spectral is completely intractable for large N.

Can we find a cheaper alternative?

Radial squared-kernel discrepancy

For $\mathcal{S} = \{s_1, \dots, s_n\} \subset \mathbb{R}^d$, the radial SKD of $\mathcal{S}$ is defined as

$$R(\mathcal{S}) = \sum_{x \in D} \sum_{y \in D} K^2(x, y) - \frac{1}{\sum_{s \in \mathcal{S}} \sum_{t \in \mathcal{S}} K^2(s, t)} \left(\sum_{x \in D} \sum_{s \in \mathcal{S}} K^2(x, s) \right)^2.$$

We have that for all $\mathcal{S} \subset \mathbb{R}^d$,

$$\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_2^2 \leq \|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_{\mathrm{F}}^2 \leq R(\mathcal{S}) \leq \|\mathbf{K}\|_{\mathrm{F}}^2 \text{ and } \frac{1}{N} \|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S})\|_*^2 \leq R(\mathcal{S}).$$

Radial SKD in reproducing kernel Hilbert spaces

The radial SKD criterion can be defined more generally in the context of Hilbert-Schmidt operators on RKHSs, where it enjoys nice properties and has deep connections with integral operator approximation.

Locally optimising the radial SKD

Gradient of the radial SKD

Idea: From an initial Nyström sample $\mathcal{S}^{(0)} \subset \mathbb{R}^d$, find the nearest local minimum of the radial SKD using gradient descent methods.

For $\mathcal{S} = \{s_1, \dots, s_n\} \subset \mathbb{R}^d$, the partial derivative of R at $\mathcal{S}$ with respect to the l-th coordinate of the k-th landmark point s_k is given by

$$\partial_{[s_k]_l} R(\mathcal{S}) = \frac{\left(\sum_{x\in\mathcal{D}}\sum_{s\in\mathcal{S}} K^2(x,s)\right)^2}{\left(\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}} K^2(s,t)\right)^2}\left(\partial^{[\mathrm{d}]}_{[s_k]_l} K^2(s_k,s_k) + 2\sum_{t\in\mathcal{S}\setminus\{s_k\}} \partial^{[1]}_{[s_k]_l} K^2(s_k,t)\right) \\ - 2\,\frac{\sum_{x\in\mathcal{D}}\sum_{s\in\mathcal{S}} K^2(x,s)}{\sum_{s\in\mathcal{S}}\sum_{t\in\mathcal{S}} K^2(s,t)}\left(\sum_{x\in\mathcal{D}} \partial^{[1]}_{[s_k]_l} K^2(s_k,x)\right).$$

Gradient descent iterates converge under reasonable assumptions on K^2 (gradient is Lipschitz continuous).

Evaluation of partial derivatives is $\mathcal{O}(n^2 + nN)$, cheaper than evaluating norm errors.

Stochastic approximations of the gradient

There are still large sums of size N in the partial derivatives.

We can approximate the partial derivatives stochastically by sampling at random from $\mathcal{D}$.

One-sample approximation:
Random sample $X_1, \ldots, X_b$ i.i.d. from $\mathcal{D}$ for some batch size $b \in \mathbb{N}$.

$$\sum_{x \in \mathcal{D}} \sum_{s \in \mathcal{S}} K^2(s, x) = \mathbb{E}\left[\frac{N}{b} \sum_{i=1}^{b} \sum_{s \in \mathcal{S}} K^2(s, X_i) \right];$$

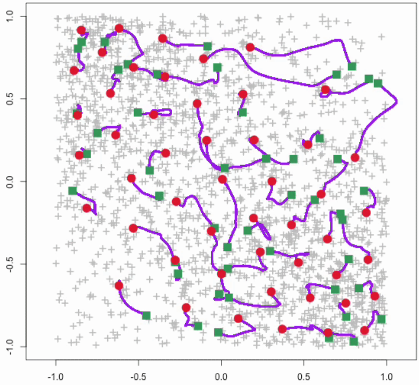
$$\sum_{x \in \mathcal{D}} \partial^{[l]}_{[s_k]_l} K^2(s_k, x) = \mathbb{E}\left[\frac{N}{b} \sum_{i=1}^{b} \partial^{[l]}_{[s_k]_l} K^2(s_k, X_i) \right].$$

Two-sample approximation:
Two independent random samples produce unbiased estimators of partial derivatives. In practice, no significant benefit observed over one-sample approximation.

Experiments

Gradient descent example

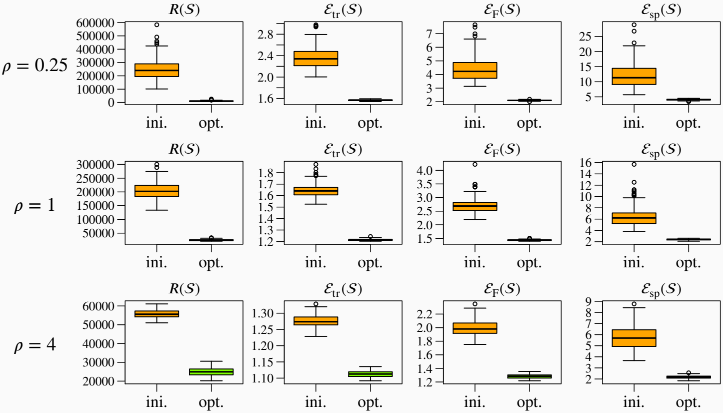
Data: Two Gaussian clusters in dimension 2, $N = 2000$.



Initial Nyström sample of size $n = 50$.

Real data set 1: Abalone

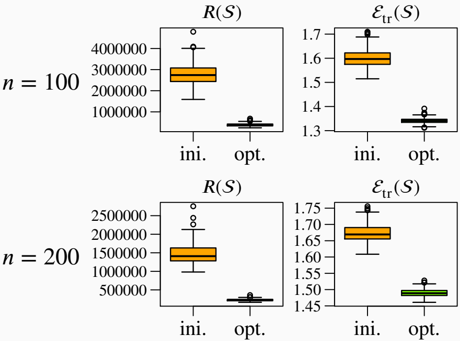
Data: Physical measurements of molluscs. $N = 4{,}175$ observations, $d = 8$ features.

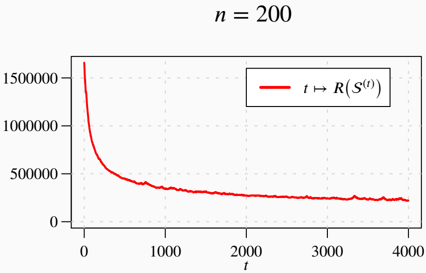


Boxplots show radial SKD and measures of efficiency for Nyström samples pre- and post-optimisation through SGD. Random initialisations of size $n = 50$.

Real data set 2: MAGIC

Data: Monte-Carlo simulated image data for gamma particles in a telescope.
$N = 18{,}905$, $d = 10$.

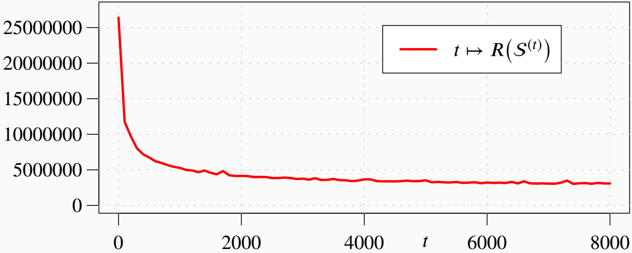




Gaussian kernel with parameter $\rho = 0.2$.

Real data set 3: MiniBooNE

Data: Particle identification data for neutrinos. $N = 129{,}592$, $d = 50$.



$$\left\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S}^{(0)})\right\|_* \approx 63{,}272.7$$

$$\left\|\mathbf{K} - \hat{\mathbf{K}}(\mathcal{S}^{(T)})\right\|_* \approx 53{,}657.2$$

$n = 1{,}000$, $\rho = 0.04$. SGD with $T = 8{,}000$ iterations, taking $1{,}350$ seconds if cost is not recorded.

One trace norm error: $6{,}600$ seconds (nearly 2 hours!)

Conclusion

Conclusion

Radial SKD can

a) be used as an affordable surrogate for classical efficiency criteria.
b) be locally optimised through SGD, resulting in consistently more efficient Nyström approximations.

SGD on radial SKD shown to be tractable for relatively large data sets.

Moving forward

- Algorithm could be made more efficient (parallelisation, adaptive step sizes).
- Different initialisation strategies could be explored (sequential/herding, more sophisticated column sampling).

References

Preprint this talk was based on:

Matthew Hutchings and Bertrand Gauthier. “Local optimisation of Nyström samples through stochastic gradient descent”. In: *arXiv preprint arXiv:2203.13284* (2022)

Some references on Nyström column sampling:

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. “Sampling methods for the Nyström method”. In: *Journal of Machine Learning Research* 13 (2012), pp. 981–1006

Alex Gittens and Michael W. Mahoney. “Revisiting the Nyström method for improved large-scale machine learning”. In: *Journal of Machine Learning Research* 17 (2016), pp. 1–65

Thank you!

Any questions?

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As the
unprocessed data that should have been added to the final page this e
has been added to receive it.

If you rerun the document (without altering it) this surplus page will g
because LaTeX now knows how many pages to expect for this documen