Randomized trace estimation for parameter-dependent matrices applied to spectral density approximation

Fabio Matti Hei Yin Lam Haoze He Daniel Kressner

ANCHP, École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland

We consider three **stochastic trace estimators** when they are applied to matrices $\boldsymbol{B}(t)$ which **continuously depend on a real parameter** $t \in [a, b]$. In particular, we reuse the same random vectors to construct a single estimator which approximates the trace in *all* values of t simultaneously. This makes the methods scale favorably in terms of the number of paramater values at which the estimate is evaluated.

Girard-Hutchinson

We sample standard Gaussian random matrix $\boldsymbol{\Psi} \in \mathbb{R}^{n \times n_\Psi}$ to form the estimate

$$\operatorname{Tr}_{\boldsymbol{\Psi}}(\boldsymbol{B}(t)) = \frac{1}{n_\Psi} \operatorname{Tr}(\boldsymbol{\Psi}^\top \boldsymbol{B}(t) \boldsymbol{\Psi}). \tag{1}$$

Theorem 1: Girard-Hutchinson error

Let $\boldsymbol{B}(t) \in \mathbb{R}^{n \times n}$ be symmetric and continuously depend on $t \in [a, b]$. Then

$$\int_a^b |\operatorname{Tr}(\boldsymbol{B}(t)) - \operatorname{Tr}_{\boldsymbol{\Psi}}(\boldsymbol{B}(t))| \, \mathrm{d}t \leq \varepsilon \int_a^b \|\boldsymbol{B}(t)\|_F \, \mathrm{d}t$$

with probability $\geq 1 - \delta$ if $n_\Psi = \mathcal{O}(\varepsilon^{-2} \log(\delta^{-1}))$.

Issue: The convergence of this estimator is often quite disappointing: if we aim to increase the accuracy of the estimate by one digit, we would need to increase the number of queries by a factor of one hundred.

Nyström

Matrices $\boldsymbol{B}(t)$ whose singular values decay rapidly to zero in each $t \in [a, b]$ can accurately be approximated by **low-rank approximations**. For symmetric positive-semidefinite matrices – ones which we enconuter in our application – the Nyström approximation is computed by sampling a standard Gaussian random matrix $\boldsymbol{\Omega} \in \mathbb{R}^{n \times n_\Omega}$ and forming the approximate

$$\widehat{\boldsymbol{B}}_{\boldsymbol{\Omega}}(t) = (\boldsymbol{B}(t)\boldsymbol{\Omega})(\boldsymbol{\Omega}^\top \boldsymbol{B}(t) \boldsymbol{\Omega})^\dagger (\boldsymbol{B}(t)\boldsymbol{\Omega})^\top.$$

Then we can approximate $\operatorname{Tr}(\boldsymbol{B}(t))$ with

$$\operatorname{Tr}(\widehat{\boldsymbol{B}}_{\boldsymbol{\Omega}}(t)) = \operatorname{Tr}\left((\boldsymbol{\Omega}^\top \boldsymbol{B}(t) \boldsymbol{\Omega})^\dagger (\boldsymbol{\Omega} \boldsymbol{B}(t)^2 \boldsymbol{\Omega}^\top)\right) \tag{2}$$

thanks to the cyclic property of the trace.

Issue: The approximation will generally not work well for matrices with slow singular value decay.

Nyström++

This estimator corrects for eventual inaccuracies in the Nyström trace-approximation (2) by estimating the trace of the approximation residual with the Girard-Hutchinson estimator (1)

$$\operatorname{Tr}_{\boldsymbol{\Psi},\boldsymbol{\Omega}}(\boldsymbol{B}(t)) = \operatorname{Tr}(\widehat{\boldsymbol{B}}_{\boldsymbol{\Omega}}(t)) + \operatorname{Tr}_{\boldsymbol{\Psi}}(\boldsymbol{B}(t) - \widehat{\boldsymbol{B}}_{\boldsymbol{\Omega}}(t)).$$

Theorem 2: Nyström++ error

Let $\boldsymbol{B}(t) \in \mathbb{R}^{n \times n}$ be symmetric positive semidefinite and continuously depend on $t \in [a, b]$. Then

$$\int_a^b |\operatorname{Tr}(\boldsymbol{B}(t)) - \operatorname{Tr}_{\boldsymbol{\Psi},\boldsymbol{\Omega}}(\boldsymbol{B}(t))| \, \mathrm{d}t \leq \varepsilon \int_a^b \operatorname{Tr}(\boldsymbol{B}(t)) \, \mathrm{d}t$$

with probability $\geq 1 - \delta$ if $n_\Psi = n_\Omega = \mathcal{O}(\varepsilon^{-1} \log(\delta^{-1}))$.

This estimator is the parameter-dependent analogue of the Nyström++ estimator [1, 2].

Spectral density

The **smoothed spectral density** of symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ is defined as

$$\phi_\sigma(t) = \frac{1}{n} \sum_{i=1}^n g_\sigma(t - \lambda_i) = \operatorname{Tr}(g_\sigma(t\boldsymbol{I}_n - \boldsymbol{A})) \tag{3}$$

where g_σ is a smoothing kernel, usually a Gaussian of width $\sigma > 0$. Hence, we end up with a **parameter-dependent trace-estimation problem** to which we can apply any of the above methods.

It is often prohibitively expensive to evaluate $g_\sigma(t\boldsymbol{I}_n - \boldsymbol{A})$ directly. Instead, we expand it in a basis of the first m **Chebyshev polynomials** $T_0, \ldots, T_m$

$$g_\sigma^{(m)}(t\boldsymbol{I}_n - \boldsymbol{A}) = \sum_{l=0}^m \mu_l(t) T_l(\boldsymbol{A}). \tag{4}$$

Implementation detail

For each value of the parameter t, the coefficients $\boldsymbol{\mu} = \{\mu_l\}_{l=0}^m$ of the Chebyshev expansion (4) can efficiently be computed through a **discrete cosine transform (DCT)**. If we let $\boldsymbol{g} = \{g_\sigma(t - \cos(\pi l/m))\}_{l=0}^m$, we can convert between $\boldsymbol{g}$ and $\boldsymbol{\mu}$ in $\mathcal{O}(m \log(m))$ operations using

$$\boldsymbol{\mu} = \mathrm{DCT}^{-1}(\boldsymbol{g}) \iff \boldsymbol{g} = \mathrm{DCT}(\boldsymbol{\mu}).$$

This relation can be used to accurately square Chebyshev expansions in $\mathcal{O}(m \log(m))$ operations

$$\Big(\sum_{l=0}^m \mu_l T_l(\boldsymbol{A})\Big)^2 = \sum_{l=0}^{2m} \nu_l T_l(\boldsymbol{A}) \implies \boldsymbol{\nu} = \mathrm{DCT}^{-1}\left(\mathrm{DCT}([\boldsymbol{\mu}, \boldsymbol{0}])^2\right).$$

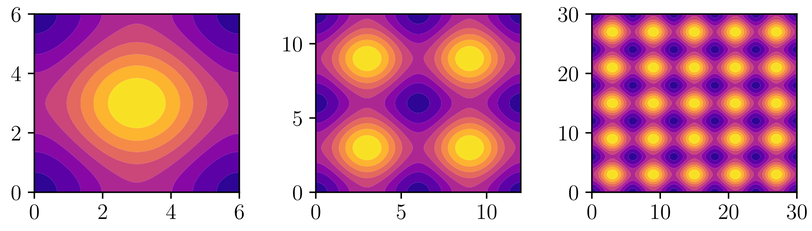
This is crucial for an efficient Nyström trace-approximation (2) of matrix functions approximated by Chebyshev expansions (4).

Model problem

The matrix of interest $\boldsymbol{A}$ comes from the **second order finite difference discretization of the Hamiltonian**

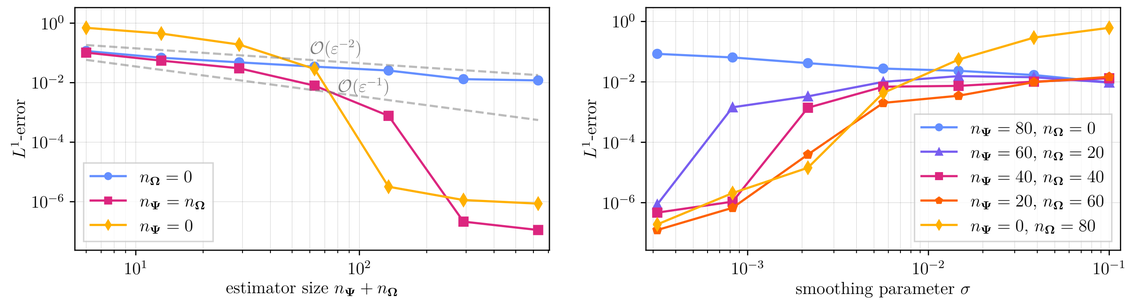
$$\mathcal{H} = -\Delta + V$$

in three dimensions [1]. The potential V interacting with the electrons is generated by Gaussian wells $v(r) = v_0 e^{-\lambda r^2}$ repeated in each spatial dimension. This is an idealized model for the **interaction of nuclei on a regular grid with electrons**. The distribution of the eigenvalues of the Hamiltonian – its spectral density – allows us to interpret the system's energy levels.



Convergence study

For increasing number of random vectors $n_\Psi + n_\Omega$, we plot the L^1 approximation error of the Girard-Hutchinson ($n_\Omega = 0$), the Nyström ($n_\Psi = 0$), and the Nyström++ estimators ($n_\Psi = n_\Omega$) **(left)**. For multiple values of σ and fixed $n_\Psi + n_\Omega$, we plot the L^1 approximation error **(right)**. For small values of the smoothing parameter σ, the matrix function in (3) has low rank and is already well approximated by the Nyström approximation alone, whereas for large σ, the Girard-Hutchinson estimator does most of the work.



Conclusion

We analyze randomized trace estimators when they are applied to parameter-dependent matrices. Despite reusing the same randomization for computing the estimate at each value of the parameter t, we **closely match the corresponding results for constant matrices** [2]. Further, we use a more **rigorous approach for expanding matrix functions** in terms of Chebyshev polynomials than [1], when it comes to constructing approximations to the smoothed spectral density.

References

[1] L. Lin, "Randomized estimation of spectral densities of large matrices made accurate," *Numer. Math.*, vol. 136, no. 1, pp. 183–213, 2017.

[2] D. Persson, A. Cortinovis, and D. Kressner, "Improved variants of the Hutch++ algorithm for trace estimation," *SIAM J. Matrix Anal. Appl.*, vol. 43, no. 3, pp. 1162–1185, 2022.

[3] F. Matti, H. Y. Lam, H. He, and D. Kressner, "Randomized trace estimation for parameter-dependent matrices applied to spectral density approximation," *In preparation*, 2024.