Randomized spectral density estimation applied to neural network optimization

Fabio Matti† Hei Yin Lam† Haoze He† Daniel Kressner†

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

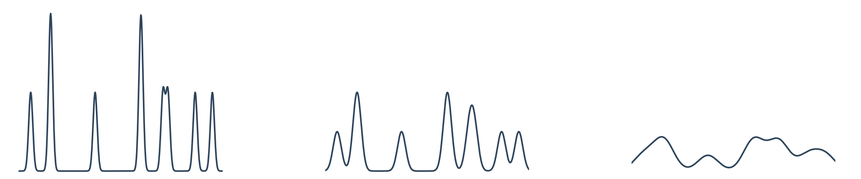
We approximate the **spectral density** of matrices which are only accessed through matrix-vector products. This boils down to estimating the trace of a certain **parameter-dependent** matrix $\boldsymbol{B}(t)$ for a parameter $t \in [a, b]$. Therefore, we consider three well established **randomized trace estimators** for constant matrices $\boldsymbol{B}$ when they are straightforwardly applied to parameter-dependent matrices $\boldsymbol{B}(t)$.

Smoothed spectral density

The **smoothed spectral density** of a symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ is given by the density function

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

for a Gaussian g_σ of width $\sigma > 0$. This is a **parameter-dependent trace estimation problem** for the positive semi-definite $\boldsymbol{B}(t) = g_\sigma(t\boldsymbol{I}_n - \boldsymbol{H})$.



(a) ϕ_σ with small σ (b) ϕ_σ with medium σ (c) ϕ_σ with large σ

Fast Hessian-vector products

The **Hessian matrix** $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ of a neural network parametrized by $\boldsymbol{w} = (w_1, \dots, w_n)^\top \in \mathbb{R}^n$ and with respect to a loss function $\mathcal{L}$ has entries

$$H_{ij} = \frac{\partial^2 \mathcal{L}}{\partial w_i \partial w_j}.$$

The expansion of the gradient reveals a useful formula for computing **Hessian-vector products** $\boldsymbol{H}\boldsymbol{v}$ for any $\boldsymbol{v} \in \mathbb{R}^n$:

$$\nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}+h\boldsymbol{v}) = \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}) + h\boldsymbol{H}\boldsymbol{v} + \mathcal{O}(h^2) \overset{h\to 0}{\implies} \boldsymbol{H}\boldsymbol{v} = \frac{\partial}{\partial h}\Big|_{h=0} \nabla_{\boldsymbol{w}}\mathcal{L}(\boldsymbol{w}+h\boldsymbol{v}).$$

Because $f \mapsto \frac{\partial}{\partial h}\big|_{h=0} f(\cdot + h\boldsymbol{v})$ is a **differential operator**, it can be computed similarly to $\nabla_{\boldsymbol{w}}\mathcal{L}$ in $\mathcal{O}(n)$ operations with a **backpropagation scheme** [1].

1: Girard-Hutchinson estimator

We define the **Girard-Hutchinson estimator** as

$$\operatorname{Tr}_{\boldsymbol{\Psi}}(\boldsymbol{B}(t)) = \frac{1}{n_{\boldsymbol{\Psi}}}\sum_{j=1}^{n_{\boldsymbol{\Psi}}} \boldsymbol{\psi}_j^\top \boldsymbol{B}(t)\boldsymbol{\psi}_j$$

for standard Gaussian $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_{n_{\boldsymbol{\Psi}}}$.

Theorem: Girard-Hutchinson error [2]

If $\boldsymbol{B}(t) \in \mathbb{R}^{n\times n}$ is symmetric, then $\int_a^b |\operatorname{Tr}(\boldsymbol{B}(t)) - \operatorname{Tr}_{\boldsymbol{\Psi}}(\boldsymbol{B}(t))|\mathrm{d}t \le \varepsilon \int_a^b \|\boldsymbol{B}(t)\|_F \mathrm{d}t$ with probability $\ge 1-\delta$ if $n_{\boldsymbol{\Psi}} = \mathcal{O}(\varepsilon^{-2}\log(\delta^{-1}))$.

Downside: Convergence with $n_{\boldsymbol{\Psi}}$ is slow.

2: Nyström estimator

The **Nyström approximation** is the matrix

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

for a standard Gaussian $\boldsymbol{\Omega} \in \mathbb{R}^{n\times n_{\boldsymbol{\Omega}}}$. The error of this approximation depends on the **singular value decay** of $\boldsymbol{B}(t)$ [2].

We approximate the trace $\operatorname{Tr}(\boldsymbol{B}(t))$ with

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

thanks to the cyclic property of the trace.

Downside: Poor approximation for matrices with slow singular value decay.

1 & 2: Nyström++ estimator

The **Nyström++ estimator** computes the Nyström trace estimator and corrects it with the Girard-Hutchinson estimate of the residual

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

Theorem: Nyström++ error [2]

If $\boldsymbol{B}(t) \in \mathbb{R}^{n\times n}$ is symmetric positive semi-definite, then $\int_a^b |\operatorname{Tr}(\boldsymbol{B}(t)) - \operatorname{Tr}_{\boldsymbol{\Psi},\boldsymbol{\Omega}}(\boldsymbol{B}(t))|\mathrm{d}t \le \varepsilon \int_a^b \operatorname{Tr}(\boldsymbol{B}(t))\mathrm{d}t$ with probability $\ge 1-\delta$ if $n_{\boldsymbol{\Psi}} = n_{\boldsymbol{\Omega}} = \mathcal{O}(\varepsilon^{-1}\log(\delta^{-1}))$.

Chebyshev-Nyström++ algorithm

Expand $g_\sigma(t\boldsymbol{I}_n - \boldsymbol{H})$ in the first m **Chebyshev polynomials** $T_0, \dots, T_m$ [3]

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

At each t, the coefficients $\boldsymbol{\mu} = \{\mu_l(t)\}_{l=0}^m$ and the function evaluations $\boldsymbol{g} = \{g_\sigma(t - \cos(\pi l/m))\}_{l=0}^m$ are related by the **discrete cosine transform (DCT)**

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

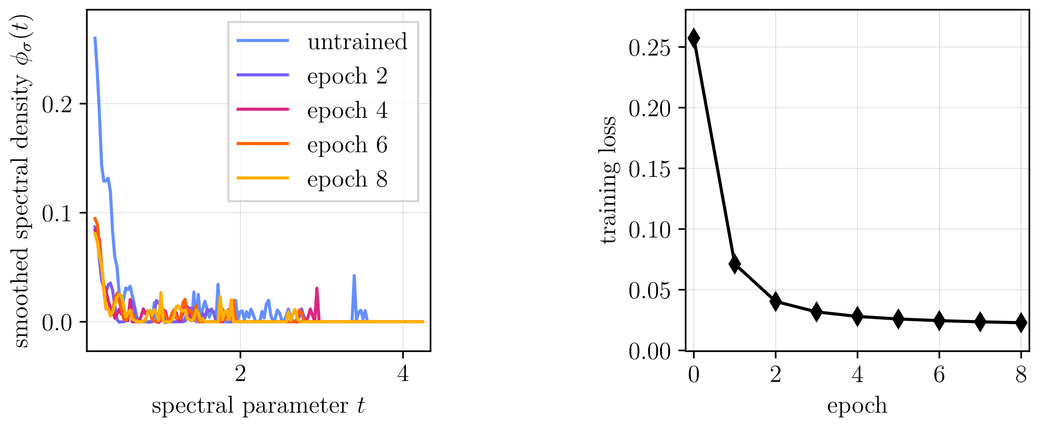
This is used to square Chebyshev expansions in $\mathcal{O}(m\log(m))$ operations:

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

The **Chebyshev-Nyström++ algorithm** applies this observation to the Nyström++ estimator to accurately approximate the spectral density [2].

Application: Sharpness-aware model selection

The smoothed spectral density of the Hessian matrix $\boldsymbol{H}$ of a neural network is used to monitor its **generalizability**. The presence of large eigenvalues indicates a **sharp minimum**, hence, unfavorable generalization properties.



Conclusion

Despite reusing the same randomization for computing the estimate at each value of the parameter t, the proposed estimators **closely match the corresponding results for constant matrices**.

In practice, the estimators, paired with a **rigorous approach for expanding parameter-dependent matrix functions** in terms of Chebyshev polynomials, can be employed to accurately and efficiently approximate smoothed spectral densities. This can effectively be used to **assess the generalizability of minima** found in a neural network optimization process.

References

[1] B. A. Pearlmutter, "Fast exact multiplication by the Hessian," *Neural Comput.*, vol. 6, no. 1, pp. 147–160, 1994.

[2] 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.

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