When Less is More: Approximating the Quantum Geometric Tensor with Block Structures

Ahmedeo Shokry[1,2,3,4] **Alessandro Santini**[1,2,3,4] **Filippo Vicentini**[1,2,3,4]

[1]CPHT, CNRS, École Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France
[2]Collège de France, Université PSL, 11 place Marcelin Berthelot, 75005 Paris, France
[3]Inria Paris–Saclay, Bâtiment Alan Turing, 1 rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
[4]LIX, CNRS, École Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France

{ahmedeo.shokry, alessandro.santini, filippo.vicentini}@polytechnique.edu

Abstract

The natural gradient is central in neural quantum states optimizations but it is limited by the cost of computing and inverting the quantum geometric tensor, the quantum analogue of the Fisher information matrix. We study a block-diagonal quantum geometric tensor that partitions the metric by network layers, analogous to block-structured Fisher methods such as K-FAC. This layer-wise approximation preserves essential curvature while removing noisy cross-layer correlations, improving conditioning and scalability. Experiments on Heisenberg and frustrated J_1–J_2 models show faster convergence, lower energy, and improved stability.

1 Introduction

A fundamental problem in quantum many-body physics is to determine the ground-state wavefunction that minimizes the system's energy. This is equivalent to solving an eigenvalue problem for an exponentially large Hamiltonian matrix, a task far beyond the reach of exact diagonalization. In practice, variational Monte Carlo (VMC) [1] reformulates this as a stochastic optimization problem: the loss function (the energy) and its gradients are estimated through Monte Carlo sampling. Samples are drawn from the square of the current parametrized wavefunction, either through ancestral sampling or Markov chain Monte Carlo (MCMC).

Neural-network quantum states (NQS) [2, 3] address the core challenge of VMC: representing and optimizing a high-dimensional, complex-valued wavefunction. NQS avoid the exponential memory complexity of quantum mechanics by expressing the wavefunction as a neural network that takes the system configurations as input and outputs its complex amplitude. Within this framework, state-of-the-art accuracy has been demonstrated on challenging benchmarks including the Fermi–Hubbard model [4–7] and the frustrated J_1–J_2 magnet [8–10]. Extensions to variational dynamics [11–13], excited states [14] and subspaces exist [15, 16].

However, standard first-order optimizers such as Adam [17] often plateau before reaching the accuracies required in physics. In the usual formulation, natural-gradient updates are employed to exploit the underlying geometry of the parameter space, an approach known in physics as *stochastic reconfiguration* (SR) [18], and which has rapid convergence guarantees [19]. In the language of machine learning (ML), this corresponds to preconditioning the energy gradient with the inverse of the Fisher information matrix (FIM), known in the literature of complex wave-functions as the *quantum geometric tensor* (QGT). As the FIM is a $n_p \times n_p$ matrix, where n_p is the number of neural network parameters, inverting it quickly becomes untractable. As it's usually ill-conditioned, LU decompositions are unreliable and it is often required to perform a full diagonalization with a complexity scaling as $\mathcal{O}(n_p^3)$ and which cannot be easily parallelized across multiple devices. Recent sample-space reformulations [9, 20] mitigate this by inverting the neural tangent kernel (NTK)-like matrix whose size depends on the number of Monte Carlo samples rather than the parameters [21].

In this paper, inspired by FIM block diagonal approaches such as KFAC and mini block Fisher (MBF) [22–25] we benchmark the block-layer QGT approximation [26] against the full QGT and NTK. In particular, we apply the technique to (i) the Heisenberg chain and (ii) the highly frustrated 2D J_1–J_2 lattice. We find that employing the block approximation we increase the effective rank of the preconditioner, mitigate ill-conditioning [27], achieve better convergence, and yield faster training.

2 Natural Gradient Descent for NQS and Block-Layer QGT

For lattice systems with spin degrees of freedom, the neural network representing the quantum system takes as input a configuration $\mathbf{x} \in \{-1, +1\}^N$ and outputs a complex amplitude $\psi_\theta(\mathbf{x})$, defining a variational wavefunction parameterized by the weights θ. The parameters are optimized to minimize the variational energy, i.e., the expectation value of the Hamiltonian

$$E = \frac{\langle \psi_\theta | H | \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle} = \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x})} \left[\frac{\langle \mathbf{x} | H | \psi_\theta \rangle}{\langle \mathbf{x} | \psi_\theta \rangle} \right], \tag{1}$$

where expectation values are estimated via Monte Carlo sampling from the Born distribution $p_\theta(\mathbf{x}) = |\psi_\theta(\mathbf{x})|^2$. As the Hilbert space grows exponentially with system size, exact evaluation becomes infeasible, making Monte Carlo sampling essential for NQS training.

Training, however, remains challenging: Monte Carlo estimates introduce noise and the energy landscape is highly nonconvex. Methods like stochastic gradient descent and Adam often fail to converge, motivating geometry-aware approaches like natural gradient descent (NGD) [28]. NGD follows updates determined by the solution of the linear system $S \cdot \delta\theta = -F$, where $F_i = \partial_{\theta_i} E$ (the stochastic energy gradient [29]) and S (the QGT) are estimated from the same Monte Carlo samples. For a variational wavefunction ψ_θ, the QGT is defined as

$$S_{ij} = \frac{\langle \partial_{\theta_i} \psi_\theta | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle} - \frac{\langle \partial_{\theta_i} \psi_\theta | \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle} \frac{\langle \psi_\theta | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle} = \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x})} \left[\Delta O_i^\dagger(\mathbf{x}) \Delta O_j(\mathbf{x}) \right], \tag{2}$$

with $\Delta O_i(\mathbf{x}) = O_i(\mathbf{x}) - \langle O_i \rangle$ and $O_i(\mathbf{x}) = \partial_{\theta_i} \log \psi_\theta(\mathbf{x})$. Formally, the natural gradient update is $\delta\theta = -\eta \cdot S^{-1} F$, where η is a learning rate and S^{-1} is a pseudo-inverse.

Block-layer QGT. Consider a variational wavefunction $\psi_\theta(\mathbf{x})$, whose parameters θ are partitioned into L disjoint groups corresponding to distinct modules: a patching and embedding layer L_{emb}, transformer encoder layers $\{L_{\text{E}(1)} \dots L_{\text{E}(N)}\}$, and a final output layer L_{out}. Denoting the parameters of layer l as θ_l, we write $\theta = (\theta_{L_{\text{emb}}}, \theta_{L_{\text{E}(1)}}, \dots \theta_{L_{\text{E}(N)}}, \theta_{L_{\text{out}}})$. We construct a block-diagonal approximation of the QGT aligned with the network decomposition

$$S \approx S^{\text{block}} = \text{diag}(S_{L_{\text{emb}}}, S_{L_{\text{E}(1)}}, \cdots, S_{L_{\text{out}}}), \qquad S_l = \langle O_l^\dagger O_l \rangle - \langle O_l^\dagger \rangle \langle O_l \rangle, \tag{3}$$

where O_l is the Jacobian of the log-wavefunction with respect to layer parameters θ_l. Each block S_l is inverted independently, using a uniform diagonal shift λ to stabilize the inversion, $\tilde{S}_l = S_l + \lambda I_l$, and $\delta\theta_l = -\eta \tilde{S}_l^{-1} F_l$. The full parameter update is $\delta\theta = (\delta\theta_{L_{\text{emb}}}, \delta\theta_{L_{\text{E}(1)}}, \dots \delta\theta_{L_{\text{E}(N)}}, \delta\theta_{L_{\text{out}}})$. The block-layer QGT method differs from prior approaches in several ways. (i) Unlike KFAC [30, 31], which factorizes each layer's FIM block as a Kronecker product, we treat each block as a full matrix, preserving all intra-layer correlations. (ii) Compared to MBF [32], which subdivides layers into smaller blocks, our decomposition follows the natural structure of the network modules (i.e., embedding, encoders, output), which has a straightforward extension to deep transformers.

3 Numerical Results

To evaluate the quality of the block-diagonal approximation, we consider metrics defined on the corresponding inverse operators. Let A and B denote the inverses of the block-layer QGT and exact QGT, respectively. Since the QGT defines a metric, approximations may distort geometric orientation (eigenvectors) or spectral scaling (eigenvalues). We therefore use complementary diagnostics probing both alignment and spectral fidelity.

$$\textbf{(a)}\quad \mathrm{F}(A,B)=\frac{\mathrm{Tr}[A^\dagger B]}{\|A\|_F\,\|B\|_F},\qquad \textbf{(b)}\quad \epsilon_F(A,B)=\frac{\|A-B\|_F}{\|B\|_F},$$
$$\textbf{(c)}\quad \kappa(A)=\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)},\qquad \textbf{(d)}\quad r_\lambda(A,B)=\mathrm{corr}\Big(\{\lambda_i^{(A)}\}_\downarrow,\{\lambda_i^{(B)}\}_\downarrow\Big). \tag{4}$$

(a) Frobenius overlap: Cosine of the angle between A and B in matrix space. $\mathrm{F}(A,B)\approx 1$ gives collinear natural-gradient updates for any force F. **(b) Frobenius relative error:** Average relative deviation between A and B over all matrix entries. **(c) Condition number:** Numerical conditioning of the natural gradient. **(d) Eigenvalue spectrum correlation:** Here $\{\lambda_i\}_\downarrow$ are eigenvalues sorted in descending order. Values $r_\lambda\approx 1$ indicate that the approximation preserves the relative scaling of curvature modes, determining the effective step-size weighting in natural-gradient updates.

Heisenberg Chain. As a first benchmark, we consider the spin-1/2 Heisenberg model on a chain of length $L = 16$ with periodic boundary conditions. The Hamiltonian reads $H = \sum_{i=1}^{L} \mathbf{S}_i \cdot \mathbf{S}_{i+1}$, where $\mathbf{S}_i = (\sigma_i^x, \sigma_i^y, \sigma_i^z)/2$ are the spin-1/2 operators at site i. This system is small enough to be exactly solvable while still exhibiting nontrivial correlations, providing a clean benchmark for our block-layer QGT against full SR and exact results. We trained the same Vision-Transformer network [33] for all three settings, using a two-layer encoder. The optimized QGT was stored after 2000 epochs, along with the energy values during optimization and the exact infidelity between the states obtained with the sampled QGT and the exact QGT. Where the infidelity between two quantum states is defined as $\mathcal{I}(|\psi\rangle,|\phi\rangle) = 1 - |\langle\psi|\phi\rangle|^2/(\langle\psi|\psi\rangle\,\langle\phi|\phi\rangle)$, vanishing for identical states and reaching one for orthogonal ones; it is directly related to the Fubini–Study metric [34, 35].

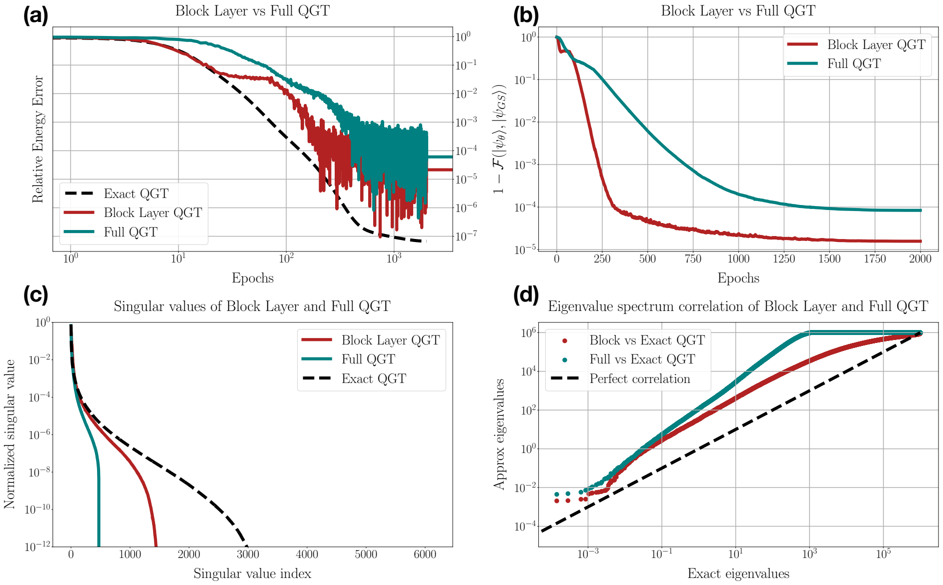

Figure 1: Comparison of block-layer and full QGT optimization for the $L = 16$ Heisenberg chain. (a) Relative energy error; (b) infidelity to the exact ground state; (c) normalized QGT spectra at convergence; (d) correlation between approximate and exact QGT eigenvalues.

Table 1 summarizes a comprehensive comparison between the block-layer and full QGT approximations against the exact QGT computed on the full Hilbert space. All metrics are evaluated on the inverse metric tensor. The results reveal a consistent pattern. Although the full QGT achieves a slightly higher Frobenius overlap with the exact inverse (0.702 vs. 0.656), the block-layer approximation better preserves the eigenvalue spectrum, with a correlation of 0.979 compared to 0.639 for the full QGT. This shows that the block formulation retains the relative weighting of geometric modes—crucial for natural-gradient updates, which depend on the eigenvalue structure of S. Across

Table 1: Diagnostics comparing block-layer QGT and full sampled QGT against the exact QGT

Metric	Block-layer vs Exact	Full vs Exact
Frobenius overlap ↑	0.656	**0.702**
Frobenius relative error ↓	**0.834**	0.864
Condition number ↓	$\mathbf{2.15 \cdot 10^9}$	$2.38 \cdot 10^9$
Eigenvalue spectral correlation ↑	**0.979**	0.639

other metrics, both methods reach similar accuracy, but the block-layer QGT yields smaller Frobenius relative error (0.834 vs. 0.864) and smaller condition number ($2.15 \cdot 10^9$ vs $2.38 \cdot 10^9$).

This structural fidelity translates directly into optimization performance: the accurate eigenvalue ranking of the block-layer QGT produces correctly scaled parameter updates, faster convergence, and lower final infidelity. Furthermore, Table 2 shows faster convergence and reduced simulation time.

Table 2: Summary of measurements for the chain $L = 16$ Heisenberg model.

Method	Energy E	Energy Variance σ^2	Infidelity $\mathcal{I}$	Wall Time
Block-layer QGT	-28.5685(3)	0.008	$1.6 \cdot 10^{-05}$	$\sim$ 25 min
Full QGT	-28.5674(6)	0.026	$8.4 \cdot 10^{-05}$	$\sim$ 1 hour

Frustrated J_1–J_2 model. We consider the spin-1/2 J_1–J_2 Heisenberg model on a square lattice with periodic boundary conditions as a concrete benchmark of our method. The Hamiltonian reads $H = J_1 \sum_{\langle i,j \rangle} \mathbf{S}_i \cdot \mathbf{S}_j + J_2 \sum_{\langle\langle i,j \rangle\rangle} \mathbf{S}_i \cdot \mathbf{S}_j$, where $\langle i, j \rangle$ and $\langle\langle i, j \rangle\rangle$ denote nearest- and next-nearest-neighbor pairs, respectively. When the ratio J_2/J_1 is in the range $\sim 0.4 - 0.6$, the system enters a highly frustrated regime where competing interactions prevent simple magnetic order and give rise to competing states with small energy gaps, making variational optimization challenging. On the 6×6 lattice [Fig. 2(a)], we see that the block-layer QGT converges significantly more rapidly than the full QGT and remains stable even when only a modest number of Monte Carlo samples are available. On the larger 10×10 lattice [Fig. 2(b)], this advantage becomes more pronounced: across all tested sample sizes, block-layer QGT consistently reaches lower relative energy errors, which we evaluate with respect to the best variational energy shown in [9]. This improvement stems from the spectral properties of the QGT: the block approximation suppresses noisy cross-layer correlations, mitigating near-degeneracies and improving conditioning. For comparison, the NTK approach is computationally cheaper and faster at small sample sizes but becomes unstable beyond 2^{14} samples. In contrast, the block-layer QGT remains stable, accurate, and competitively fast, striking a favorable balance between efficiency and robustness.

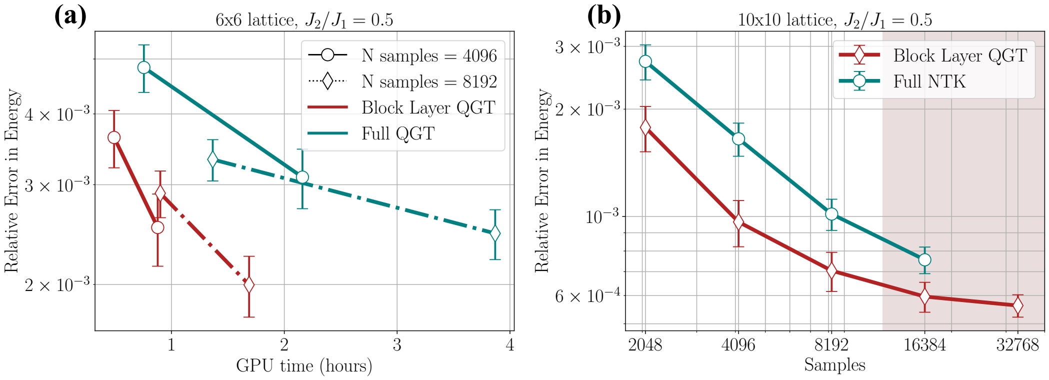

Figure 2: Comparison of block-layer and full QGT optimization in the J_1–J_2 model at $J_2/J_1 = 0.5$. (a): 6×6 lattice—energy error vs. GPU time. (b): 10×10 lattice—energy error vs. samples. Blue (red) shaded background, faster NTK (block-layer QGT) convergence.

4 Conclusions

We observed that a block-layer-diagonal QGT is less rank-deficient than the standard estimator, better capturing its spectrum, while reducing the computational cost at large sample count. This approach provides a scalable alternative to full-matrix natural gradients, achieving faster convergence, lower energies, and more stable training. Beyond VMC, block-structured curvature approximations may improve optimization in other differentiable scientific simulators where full-matrix natural gradients are prohibitive.

Acknowledgments and Disclosure of Funding

Simulations were performed with custom-built extensions to NetKet [36, 37], and at times parallelized with mpi4JAX [38]. This software is built on top of JAX [39] and Flax [40]. We acknowledge insightful discussion with J. Nys. F.V. acknowledges support by the French Agence Nationale de la Recherche through the NDQM project ANR-23-CE30-0018. We acknowledge EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy through grant EHPC-AI-2024A05-006 .

References

[1] Federico Becca and Sandro Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. Cambridge University Press, 2017.

[2] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017. doi: 10.1126/science.aag2302. URL https://www.science.org/doi/abs/10.1126/science.aag2302.

[3] Filippo Vicentini. Machine learning toolbox for quantum many body physics. *Nature Reviews Physics*, 3(3):156–156, January 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00285-7. URL http://dx.doi.org/10.1038/s42254-021-00285-7.

[4] Daniel P. Arovas, Erez Berg, Steven A. Kivelson, and Srinivas Raghu. The hubbard model. *Annual Review of Condensed Matter Physics*, 13(Volume 13, 2022):239–274, 2022. ISSN 1947-5462. doi: https://doi.org/10.1146/annurev-conmatphys-031620-102024. URL https://www.annualreviews.org/content/journals/10.1146/annurev-conmatphys-031620-102024.

[5] Mingpu Qin, Thomas Schäfer, Sabine Andergassen, Philippe Corboz, and Emanuel Gull. The hubbard model: A computational perspective. *Annual Review of Condensed Matter Physics*, 13(1):275–302, March 2022. ISSN 1947-5462. doi: 10.1146/annurev-conmatphys-090921-033948. URL http://dx.doi.org/10.1146/annurev-conmatphys-090921-033948.

[6] Javier Robledo Moreno, Giuseppe Carleo, Antoine Georges, and James Stokes. Fermionic wave functions from neural-network constrained hidden states. *Proceedings of the National Academy of Sciences*, 119(32):e2122059119, 2022. doi: 10.1073/pnas.2122059119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2122059119.

[7] Ao Chen, Zhou-Quan Wan, Anirvan Sengupta, Antoine Georges, and Christopher Roth. Neural network-augmented pfaffian wave-functions for scalable simulations of interacting fermions. 2025. URL https://arxiv.org/abs/2507.10705.

[8] Rajah P. Nutakki, Ahmedeo Shokry, and Filippo Vicentini. Design principles of deep translationally-symmetric neural quantum states for frustrated magnets. 2025. URL https://arxiv.org/abs/2505.03466.

[9] Ao Chen and Markus Heyl. Empowering deep neural quantum states through efficient optimization. *Nature Physics*, 20(9):1476–1481, July 2024. ISSN 1745-2481. doi: 10.1038/s41567-024-02566-1. URL http://dx.doi.org/10.1038/s41567-024-02566-1.

[10] Christopher Roth, Attila Szabó, and Allan H. MacDonald. High-accuracy variational Monte Carlo for frustrated magnets with deep neural networks. *Physical Review B*, 108(5):054410, August 2023. ISSN 2469-9950, 2469-9969. doi: 10.1103/PhysRevB.108.054410.

[11] Markus Schmitt, Marek M. Rams, Jacek Dziarmaga, Markus Heyl, and Wojciech H. Zurek. Quantum phase transition dynamics in the two-dimensional transverse-field ising model. *Science Advances*, 8(37), September 2022. ISSN 2375-2548. doi: 10.1126/sciadv.abl6850. URL http://dx.doi.org/10.1126/sciadv.abl6850.

[12] Luca Gravina, Vincenzo Savona, and Filippo Vicentini. Neural projected quantum dynamics: a systematic study. *Quantum*, 9:1803, July 2025. ISSN 2521-327X. doi: 10.22331/q-2025-07-22-1803. URL http://dx.doi.org/10.22331/q-2025-07-22-1803.

[13] Alessandro Sinibaldi, Douglas Hendry, Filippo Vicentini, and Giuseppe Carleo. Time-dependent neural galerkin method for quantum dynamics, 2024. URL https://arxiv.org/abs/2412.11778.

[14] David Pfau, Simon Axelrod, Halvard Sutterud, Ingrid von Glehn, and James S. Spencer. Accurate computation of quantum excited states with neural networks. *Science*, 385(6711), August 2024. ISSN 1095-9203. doi: 10.1126/science.adn0137. URL http://dx.doi.org/10.1126/science.adn0137.

[15] Adrien Kahn, Luca Gravina, and Filippo Vicentini. Variational subspace methods and application to improving variational monte carlo dynamics. 2025. doi: 10.48550/ARXIV.2507.08930. URL https://arxiv.org/abs/2507.08930.

[16] Douglas Hendry, Alessandro Sinibaldi, and Giuseppe Carleo. Grassmann variational monte carlo with neural wave functions, 2025. URL https://arxiv.org/abs/2507.10287.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

[18] Sandro Sorella. Green function monte carlo with stochastic reconfiguration. *Phys. Rev. Lett.*, 80:4558–4561, May 1998. doi: 10.1103/PhysRevLett.80.4558. URL https://link.aps.org/doi/10.1103/PhysRevLett.80.4558.

[19] Gil Goldshlager, Jiang Hu, and Lin Lin. Fast convergence rates for subsampled natural gradient algorithms on quadratic model problems. 2025. URL https://arxiv.org/abs/2508.21022.

[20] Riccardo Rende, Luciano Loris Viteritti, Lorenzo Bardone, Federico Becca, and Sebastian Goldt. A simple linear algebra identity to optimize large-scale neural network quantum states. *Communications Physics*, 7(1), August 2024. ISSN 2399-3650. doi: 10.1038/s42005-024-01732-4. URL http://dx.doi.org/10.1038/s42005-024-01732-4.

[21] Yi Ren and Donald Goldfarb. Efficient subsampled gauss-newton and natural gradient methods for training neural networks. 2019. URL https://arxiv.org/abs/1906.02353.

[22] Suzanna Becker and Yann Lecun. Improving the convergence of back-propagation learning with second-order methods. 01 1989.

[23] Tom Schaul, Sixin Zhang, and Yann LeCun. No more pesky learning rates. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 343–351, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/schaul13.html.

[24] Nicolas Roux, Pierre-antoine Manzagol, and Yoshua Bengio. Topmoumoute online natural gradient algorithm. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/9f61408e3afb633e50cdf1b20de6f466-Paper.pdf.

[25] Yann Ollivier. Riemannian metrics for neural networks i: feedforward networks. *Information and Inference: A Journal of the IMA*, 4(2):108–153, 03 2015. ISSN 2049-8764. doi: 10.1093/imaiai/iav006. URL https://doi.org/10.1093/imaiai/iav006.

[26] Jannes Nys, Gabriel Pescia, Alessandro Sinibaldi, and Giuseppe Carleo. Ab-initio variational wave functions for the time-dependent many-electron schrödinger equation. *Nature Communications*, 15(1), October 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53672-w. URL http://dx.doi.org/10.1038/s41467-024-53672-w.

[27] Sidhartha Dash, Luca Gravina, Filippo Vicentini, Michel Ferrero, and Antoine Georges. Efficiency of neural quantum states in light of the quantum geometric tensor. *Commun. Phys.*, 8(1), March 2025.

[28] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998. doi: 10.1162/089976698300017746.

[29] Antoine Misery, Luca Gravina, Alessandro Santini, and Filippo Vicentini. Looking elsewhere: improving variational monte carlo gradients by importance sampling. 2025. URL https://arxiv.org/abs/2507.05352.

[30] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/martens15.html.

[31] Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 573–582. JMLR.org, 2016.

[32] Achraf Bahamou, Donald Goldfarb, and Yi Ren. A mini-block fisher method for deep neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2022. URL https://api.semanticscholar.org/CorpusID:253158046.

[33] Luciano Loris Viteritti, Riccardo Rende, and Federico Becca. Transformer variational wave functions for frustrated quantum spin systems. *Phys. Rev. Lett.*, 130:236401, Jun 2023. doi: 10.1103/PhysRevLett.130.236401. URL https://link.aps.org/doi/10.1103/PhysRevLett.130.236401.

[34] Guido Fubini. Sulle metriche definite da una forma hermitiana: nota. *Office graf. C. Ferrari*, 63:502–513, 1904.

[35] Study Eduard. Kürzeste wege im komplexen gebiet. *Mathematische Annalen (in German)*, 60, 1905. doi: 10.1007/bf01457616.

[36] Giuseppe Carleo, Kenny Choo, Damian Hofmann, James ET Smith, Tom Westerhout, Fabien Alet, Emily J Davis, Stavros Efthymiou, Ivan Glasser, Sheng-Hsuan Lin, Marta Mauri, Guglielmo Mazzola, Christian B Pereira, and Filippo Vicentini. Netket: A machine learning toolkit for many-body quantum systems. *SoftwareX*, 10:100311, 2019. doi: 10.1016/j.softx.2019.100311. URL https://www.sciencedirect.com/science/article/pii/S2352711019300974.

[37] Filippo Vicentini, Damian Hofmann, Attila Szabó, Dian Wu, Christopher Roth, Clemens Giuliani, Gabriel Pescia, Jannes Nys, Vladimir Vargas-Calderón, Nikita Astrakhantsev, et al. Netket 3: Machine learning toolbox for many-body quantum systems. *SciPost Physics Codebases*, page 007, 2022.

[38] Dion Häfner and Filippo Vicentini. mpi4jax: Zero-copy mpi communication of jax arrays. *Journal of Open Source Software*, 6(65):3419, 2021. doi: 10.21105/joss.03419. URL https://joss.theoj.org/papers/10.21105/joss.03419.

[39] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: Composable transformations of Python+NumPy programs, 2018.

[40] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2024.