ENFORCING INDEPENDENT SPECTRAL DYNAMICS IMPROVES OPTIMIZATION OF CONVOLUTIONAL LAYERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Commonly used optimizers such as Adam implicitly treat each parameter block as a flat vector, ignoring the multilinear structure of the weights in many modern machine learning models. Recent works have shown that exploiting matrix structure can lead to better optimization dynamics. However, these results are largely confined to matrix-valued parameters. The dynamics of layers whose weights are not naturally matrices as convolutional kernels remain poorly understood and standard optimizers do not account for their structure. In this paper, we study the connection between convolutional kernels and circulant matrices and, exploiting their diagonalization by the Fourier transform, we propose a new class of *frequency-decoupled* optimizers with independent spectral evolution. We further analyze existing optimizers within our theoretical framework and propose a method for constructing new decoupled optimizers from the geometry of existing ones. Finally, we show that decoupled optimizers provide other great spectral properties allowing layer compression and show overall better performance on machine learning models utilizing convolutional layers (e.g. CNNs, Neural Operators and Diffusion models) at the cost of computational complexity.

1 INTRODUCTION

Large language models have reached a level of capability that was not anticipated a decade ago (Brown et al., 2020), but their internal computations are only partly understood (Elhage et al., 2021; Olsson et al., 2022). The same architecture is now deployed far beyond language, in vision (Dosovitskiy et al., 2021), structural biology (Jumper et al., 2021), weather forecasting (Bi et al., 2023) and the numerical solution of partial differential equations (Cao, 2021), where a model is trusted only if its internal mechanism can be examined (Rudin, 2019). The component responsible for this success is the attention mechanism (Vaswani et al., 2017), and it is also the component that is hardest to analyze. Its internal structure is intricate (Dong et al., 2021), and the optimization algorithms used to train it act on that structure in ways that are themselves poorly understood (Liu et al., 2020). Recent theoretical work shows that attention subsumes convolution (Cordonnier et al., 2020), and in its most general form both are special cases of a single learnable integral transform (Dhor et al., 2026), a viewpoint also underlying neural operators (Kovachki et al., 2023; Cao, 2021; Guibas et al., 2022). The convolutional layer is therefore a tractable special case of attention, and a theory of its training dynamics is a starting point for a theory of attention.

Layer spectral analysis is one of the common tools for the purpose of analyzing the layer structure and the behavior of a model. Several properties of a network depend on the spectra of its layers, among them the compressibility of large models (Denton et al., 2014; Jaderberg et al., 2014; Idelbayev & Carreira-Perpiñán, 2020), generalization (Bartlett et al., 2017; Neyshabur et al., 2018) and the sharpness of the loss (Yoshida & Miyato, 2017). The spectrum of the trained weights depends on the training process (Martin & Mahoney, 2019; 2021), and therefore on the chosen optimizer (Liu et al., 2025). Indeed, modern optimizers that use the multidimensional structure of the weights (e.g. Muon (Jordan et al., 2024; Liu et al., 2025), Shampoo (Gupta et al., 2018) and SOAP (Vyas et al., 2024)) often outperform optimizers which treat every weight as a vector (e.g. Adam (Kingma & Ba, 2015) and SGD). However, most results on training dynamics concern deep linear networks under strong simplifying assumptions or restrictions on the data (Saxe et al., 2014; Arora et al., 2018;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2019; Gunasekar et al., 2017), while other layer types are studied empirically because a theoretical treatment is difficult. To the best of our knowledge, even for plain CNNs the theory is incomplete, despite their use in modern architectures such as ResNet (He et al., 2016), U-Net (Ronneberger et al., 2015) and ConvNeXt (Liu et al., 2022), in application areas such as computer vision and finance (Zhang et al., 2019), and in recent approaches such as Fourier neural operators (Li et al., 2021) and diffusion models (Ho et al., 2020; Rombach et al., 2022). Fortunately, convolutional layers have several properties that simplify their theoretical treatment considerably, including the treatment of their spectra. Previous work shows that convolutional layers are closely related to circulant matrices (Gray, 2006; Sedghi et al., 2019), which suggests the use of Fourier bases both for analyzing and for optimizing these layers (Mathieu et al., 2014; Sedghi et al., 2019; Trockman & Kolter, 2021).

Contributions. Numerous works suggest optimizing convolutional layers in the Fourier basis (Rippel et al., 2015), but, to the best of our knowledge, none of them provides a thorough theoretical explanation of *why these methods show better results during training or better spectral properties*. In this work we propose a theoretical framework for the optimization of convolutional layers that explains many properties of existing optimizers and provides a way of constructing new optimizers with favourable spectral dynamics. Our contributions are as follows:

- We define a new class of *frequency decoupled* optimizers (Definition 4.1) and show that on a splitting loss these optimizers induce a product of independent per-frequency transition systems (Theorem 4.2). Since the singular values of a circulant matrix are exactly the moduli of its Fourier coefficients, this implies an independent spectral evolution of the isotypic components.
- We construct a theoretical framework by deriving several practical decoupling criteria (Theorem 4.4). We further show that these criteria induce a special structure (LS) on the preconditioner of the optimizer in the Fourier basis.
- We use our framework to analyze common optimizers and show that most of them are not generally decoupled. We then supplement our framework with a method of creating decoupled optimizers given a non-decoupled optimizer and use it to obtain new decoupled optimization methods (Theorems 4.8 and 4.9).
- Finally, we experimentally show that new decoupled optimizers provide better training performance and spectral properties when applied to various deep learning tasks.

We must also note that our framework relies solely on the assumption that the parameter space is a Hilbert space over $\mathbb{C}$ and therefore applies beyond convolutions, for instance, to group convolutions over finite abelian groups and to the spectral layers of Fourier neural operators.

2 Related work

Spectra of convolutional layers. A circulant matrix is diagonalized by the DFT (Gray, 2006). Building on this classical fact, Sedghi et al. (2019) gave an exact characterization of the singular values of a multi-channel convolutional layer as the union of the singular values of its per-frequency symbols – the structural result underlying our framework. The same Fourier representation has subsequently been used to parameterize orthogonal convolutions (Trockman & Kolter, 2021) and to control their spectral norm (Singla & Feizi, 2021). These works exploit frequency structure to manipulate or constrain the *weights*. In contrast, we study whether the *optimization algorithm* that produces these weights respects that structure.

Structure-aware optimizers. Diagonal adaptive methods such as AdaGrad and Adam (Duchi et al., 2011; Kingma & Ba, 2015) operate coordinatewise in a chosen parameterization, whereas second-order and preconditioned methods (Amari, 1998) exploit richer structure, often through Kronecker factorizations of the curvature (Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2024). A complementary line of work replaces Euclidean geometry with spectral geometry. This includes spectral descent (Carlson et al., 2015), the steepest-descent-under-a-norm perspective of Bernstein & Newhouse (2024b;a), Muon (Jordan et al., 2024; Liu et al., 2025), and tensor extensions based on a selected matrix unfolding (Bogachev et al., 2026). These methods are formulated for parameter blocks represented as matrices (or as tensors reduced to matrices through an unfolding). Our

analysis shows that, for convolutional layers, this seemingly innocuous choice of representation can destroy the natural frequency locality of the operator.

Fourier-domain training and dynamics. Fourier structure has long played a role in both the implementation and analysis of neural networks. FFT-based convolution provides an efficient way to train convolutional architectures (Mathieu et al., 2014), while Fourier neural operators parameterize layers directly through their frequency-domain symbols (Li et al., 2021). On the dynamics side, the frequency principle (Rahaman et al., 2019; Xu et al., 2020) documents the empirical tendency of neural networks to learn low-frequency components first. These works exploit, parameterize, or observe frequency structure, but do not characterize when the optimizer itself preserves it.

Related spectral perspectives also arise in the study of optimization dynamics. Exact spectral dynamics are known for deep linear networks (Saxe et al., 2014) and form part of the foundation of the implicit-regularization literature (Gunasekar et al., 2017; Arora et al., 2019). Low-rank adaptation (Hu et al., 2022) modifies training by restricting updates to a low-rank subspace and thereby directly constraining the spectrum of weight updates. Finally, adaptive methods have been observed to underperform SGD on convolutional architectures (Wilson et al., 2017). Our framework provides a complementary structural perspective: rather than asking only how an optimizer changes the spectrum of the weights, we ask which optimizers preserve the intrinsic frequency decomposition of a convolutional layer. Section 4.3 makes this distinction precise.

3 Preliminaries

Notation. Throughout, $\mathcal{H}$ is a finite-dimensional complex Hilbert space of parameters and Γ a finite abelian group acting on it by a unitary representation $\rho\colon \Gamma \to U(\mathcal{H})$. We write $\Omega = \widehat{\Gamma}$ for the character group and χ_ω for the character indexed by ω.

Consider the isotypic decomposition $\mathcal{H} = \bigoplus_\omega \mathcal{H}_\omega$, where

$$\mathcal{H}_\omega = \{v \in \mathcal{H} : \rho(s)v = \chi_\omega(s)\, v \ \ \forall s \in \Gamma\}, \qquad \omega \in \Omega. \tag{1}$$

Then the unitary map $\mathcal{F}$ from the standard (spatial) basis to the basis adapted to (1) (the Fourier basis) is the Fourier transform of the problem. The model case is a layer with kernel $\theta \in \mathbb{C}^{C_{\text{out}} \times C_{\text{in}} \times \Gamma}$ and $\Gamma = \mathbb{Z}_{N_1} \times \cdots \times \mathbb{Z}_{N_d}$ acting by cyclic spatial shifts. Then Ω is the set of Fourier harmonics, $\mathcal{H}_\omega$ consists of the kernels supported on the ω-th harmonic and $q_\omega \equiv \dim_{\mathbb{C}} \mathcal{H}_\omega = C_{\text{out}} C_{\text{in}}$. Additional information about the elements of group theory we use can be found in Appendix B.1.

We also call the $\mathcal{H}_\omega$ *frequency subspaces* and write $\Pi^{\mathcal{H}}_\omega$ for the orthogonal projector onto $\mathcal{H}_\omega$, $g_\omega = \Pi^{\mathcal{H}}_\omega g$ and $H_{\omega\nu} = \Pi^{\mathcal{H}}_\omega H \Pi^{\mathcal{H}}_\nu$ for operators, so that $H = \sum_{\omega,\nu} H_{\omega\nu}$.

Convolution kernels as circulant matrices. Let $\mathcal{C}_\theta$ be the linear operator of circular convolution with the kernel θ, acting from $\mathbb{C}^{C_{\text{in}} \times \Gamma}$ to $\mathbb{C}^{C_{\text{out}} \times \Gamma}$, and let $\mathcal{F}_{\text{in}}, \mathcal{F}_{\text{out}}$ be the unitary spatial DFTs on its input and output. Put $\kappa = \sqrt{|\Gamma|}$. The following two results are due to Sedghi et al. (2019) for $d = 2$, and the extension to a product of cyclic groups is immediate.

Theorem 3.1 (Single-channel convolution; Sedghi et al., 2019)**.** *For $C_{\text{in}} = C_{\text{out}} = 1$ the operator $\mathcal{C}_\theta$ is a d-fold block circulant matrix and is diagonalized by the DFT, $\mathcal{F}\mathcal{C}_\theta\mathcal{F}^* = \kappa\,\mathrm{Diag}(w_\omega)_{\omega\in\Omega}$, where w_ω is the coefficient of θ on the ω-th harmonic. Its singular values are the moduli $\kappa|w_\omega|$.*

Theorem 3.2 (Multi-channel convolution; Sedghi et al., 2019)**.** *For general $C_{\text{in}}, C_{\text{out}}$ the operator $\mathcal{C}_\theta$ is a $C_{\text{out}} \times C_{\text{in}}$ block matrix whose blocks are d-fold block circulant, and, up to a permutation of coordinates,*

$$\mathcal{F}_{\text{out}}\, \mathcal{C}_\theta\, \mathcal{F}_{\text{in}}^* = \bigoplus_{\omega\in\Omega} \kappa\, W_\omega, \qquad W_\omega = \mathrm{mat}(\Pi^{\mathcal{H}}_\omega \theta) \in \mathbb{C}^{C_{\text{out}} \times C_{\text{in}}}. \tag{2}$$

Consequently the singular values of $\mathcal{C}_\theta$ are the union over ω of the singular values of the symbols κW_ω.

The singular spectrum of a convolutional layer therefore splits into the spectra of its $|\Omega|$ symbols, one for each frequency subspace.

Commutation criterion. Both results reflect the fact that circular convolution commutes with spatial shifts. For a general unitary action of Γ, operators with this property are described by the following result.

Theorem 3.3 (Commutation criterion)**.** *For a linear operator $P\colon \mathcal{H} \to \mathcal{H}$ the following are equivalent:* (a) $P(\mathcal{H}_\omega) \subseteq \mathcal{H}_\omega$ *for all ω;* (b) $\mathcal{F}P\mathcal{F}^* = \bigoplus_\omega P_\omega$*;* (c) $[P, \rho(s)] = 0$ *for all $s \in \Gamma$.*

It follows from the isotypic decomposition and Schur's lemma (Serre, 1977, §§2.2, 2.6). For $\Gamma = \mathbb{Z}_N$ it is the characterization of block-circulant matrices as the matrices that commute with the cyclic shift (Davis, 1979, Ch. 3). We denote by $\mathrm{Comm}(\rho(\Gamma))$ the set of operators satisfying these conditions.

Optimizer as a family of maps. For generality, we model optimizers in the following way.

Definition 3.4 (Optimizer)**.** Let θ_t be the parameters, $g_t = \nabla L(\theta_t)$ the supplied gradient of the loss function and h_t the state kept by the optimizer. Then, an *optimizer* is a three-step algorithm: $h_{t+1} = \Phi_t(h_t, g_t)$, $u_t = \Psi_t(h_{t+1}, g_t)$, $\theta_{t+1} = \theta_t - \eta_t u_t$, where $\Phi_t\colon S \times \mathcal{H} \to S$ and $\Psi_t\colon S \times \mathcal{H} \to \mathcal{H}$.

Here we assume that hyperparameters are fixed in advance, and for stochastic gradients statements hold per realization, since decoupling is a property of the maps on arbitrary inputs. Every first-order method we consider can be represented this way by choosing suitable Φ, Ψ maps.

To take a deeper look into the optimizer structure we additionally factor the update into the *preconditioner* P_t and a *carrier* Z_t,

$$\Psi_t(h, g) = P_t(h)\, Z_t(h, g), \tag{3}$$

assuming $P_t(h)$ is linear.

We assume that the state space is indexed by frequencies, $S = \prod_{\omega \in \Omega} S_\omega$ with all S_ω nonempty, Π^S_ω the corresponding projection and $h_\omega = \Pi^S_\omega h$.

4 Main Part

4.1 Frequency decoupling

Using our notation, we can finally define the key concept of our paper.

Definition 4.1 (Frequency decoupling)**.** An optimizer (Φ_t, Ψ_t) is *frequency decoupled* if for every $t \in \mathbb{N}_0$ and every $\omega \in \Omega$ there exist $\Phi_{t,\omega}\colon S_\omega \times \mathcal{H}_\omega \to S_\omega$ and $\Psi_{t,\omega}\colon S_\omega \times \mathcal{H}_\omega \to \mathcal{H}_\omega$ such that for all $(h, g) \in S \times \mathcal{H}$,

$$\big(\Phi_t(h, g)\big)_\omega = \Phi_{t,\omega}(h_\omega, g_\omega), \qquad \big(\Psi_t(h, g)\big)_\omega = \Psi_{t,\omega}(h_\omega, g_\omega).$$

Equivalently, an optimizer is decoupled if the frequencies are independent in the sense that perturbing the other frequencies leaves the ω-th output unchanged. Decoupling can therefore be refuted by a single pair of inputs that agree on frequency ω but disagree after one step. A more formal expression can be found in Appendix C (Theorem C.1).

Note that the definition constrains the *maps* on all inputs, not a single trajectory, and it constrains the optimizer alone. Along a trajectory the recursions then read $h_{t+1,\omega} = \Phi_{t,\omega}(h_{t,\omega}, g_{t,\omega})$ and $u_{t,\omega} = \Psi_{t,\omega}(h_{t+1,\omega}, g_{t,\omega})$, so the algorithm evolves independently on every frequency. Independence of the whole trajectories additionally requires the loss to split. As such we obtain the following theorem.

Theorem 4.2. *Let $L(\theta) = \sum_{\omega \in \Omega} L_\omega(\theta_\omega)$ and let the optimizer be decoupled. Then, there exist maps $T_{s,\omega}\colon S_\omega \times \mathcal{H}_\omega \to S_\omega \times \mathcal{H}_\omega$ with*

$$\begin{pmatrix} h_{t,\omega} \\ \theta_{t,\omega} \end{pmatrix} = \big(T_{t-1,\omega} \circ \cdots \circ T_{0,\omega}\big) \begin{pmatrix} h_{0,\omega} \\ \theta_{0,\omega} \end{pmatrix} \quad (t \geq 1),$$

and the global transitions factor as $T_s = \prod_{\omega \in \Omega} T_{s,\omega}$ and $T_{t-1} \circ \cdots \circ T_0 = \prod_\omega (T_{t-1,\omega} \circ \cdots \circ T_{0,\omega})$.

Now let us discuss the criteria for the frequency decoupling of an optimizer. Recasting Definition 4.1 as three separately checkable conditions, we obtain the following.

Theorem 4.3. *An optimizer is frequency decoupled if and only if for all t, ω*

$$\exists\, \Phi_{t,\omega} :\ \Pi^S_\omega \Phi_t(h, g) = \Phi_{t,\omega}(h_\omega, g_\omega); \tag{$*$}$$

$$g_\omega = \tilde{g}_\omega \ \Rightarrow\ \Pi^{\mathcal{H}}_\omega \Psi_t(h, g) = \Pi^{\mathcal{H}}_\omega \Psi_t(h, \tilde{g}); \tag{$**$}$$

$$h_\omega = \tilde{h}_\omega \ \Rightarrow\ \Pi^{\mathcal{H}}_\omega \Psi_t(h, g) = \Pi^{\mathcal{H}}_\omega \Psi_t(\tilde{h}, g). \tag{$***$}$$

Each condition can be interpreted in the following way: condition ($*$) is locality of the state recursion, ($**$) locality of the step in the gradient, and ($***$) locality of the step in the state. In the following subsections we will show that each non-decoupled method violates exactly one of the three, thus identifying the coupling principle. Proofs of Theorems 4.2 and 4.3 are given in Appendix C.

4.2 Criteria in terms of the preconditioner

Theorem 4.3 is stated for the full update maps. As we will show further (Section 4.3), most modern optimizers couple the frequencies due to the structure of their preconditioner. Thus we state our most powerful criterion.

Theorem 4.4 (Generalized criterion)**.** *Assume that the carrier is local,*

$$Z_t(h, g) = \bigoplus_{\nu \in \Omega} Z_{t,\nu}(h_\nu, g_\nu), \tag{Z}$$

and let $P_{t,\omega\nu}(h) \colon \mathcal{H}_\nu \to \mathcal{H}_\omega$ be the Fourier blocks of $P_t(h)$. For $a \in S_\nu$ set

$$\mathcal{D}_{t,\nu}(a) = \operatorname{span}\big\{ Z_{t,\nu}(a, \xi) - Z_{t,\nu}(a, \widetilde{\xi}) : \xi, \widetilde{\xi} \in \mathcal{H}_\nu \big\}, \qquad b^0_{t,\omega}(h) = \sum_{\nu \in \Omega} P_{t,\omega\nu}(h) Z_{t,\nu}(h_\nu, 0).$$

The optimizer is frequency decoupled if and only if ($*$) holds and, for all t, ω,

$$P_{t,\omega\nu}(h)\, \mathcal{D}_{t,\nu}(h_\nu) = \{0\} \qquad \forall h \in S,\ \forall \nu \neq \omega, \tag{I}$$

and, for all $h, \widetilde{h} \in S$ with $h_\omega = \widetilde{h}_\omega$,

$$\big(P_{t,\omega}(h) - P_{t,\omega}(\widetilde{h})\big)\mathcal{D}_{t,\omega}(h_\omega) = \{0\}, \tag{II}$$

$$b^0_{t,\omega}(h) = b^0_{t,\omega}(\widetilde{h}). \tag{III}$$

Proof sketch. By Theorem 4.3 it suffices to translate conditions ($**$) and ($***$) into conditions on P_t. Writing each carrier block as its base value $Z_{t,\nu}(h_\nu, 0)$ plus the variation $\delta_{t,\nu}(h_\nu, g_\nu) = Z_{t,\nu}(h_\nu, g_\nu) - Z_{t,\nu}(h_\nu, 0)$ gives

$$\Pi^{\mathcal{H}}_\omega \Psi_t(h, g) = b^0_{t,\omega}(h) + \sum_{\nu \in \Omega} P_{t,\omega\nu}(h)\, \delta_{t,\nu}(h_\nu, g_\nu).$$

The ω-block is independent of g_ν, $\nu \neq \omega$, exactly when every term with $\nu \neq \omega$ vanishes, which is (I). Only the term $\nu = \omega$ remains. At $g_\omega = 0$ condition ($***$) reduces to (III), and after subtracting the base part it becomes (II). The full proof is in Appendix D.

Discussion. Conditions (I) and (II) state that the preconditioner is block diagonal in the Fourier basis and that each diagonal block depends only on the state of its own frequency, both required only on the variations of the carrier. Indeed, when the variations of the carrier span the whole frequency subspace $\mathcal{D}_{t,\omega}(h_\omega) = \mathcal{H}_\omega\ \forall\, t, \omega, h_\omega$ conditions (I)–(III) together are equivalent to

$$P_t(h) = \bigoplus_{\omega \in \Omega} P_{t,\omega}(h_\omega). \tag{LS}$$

Condition (III) itself is the decoupling of the "historical momentum" (the step of the optimizer that depends only on the state h_t). Note that it holds automatically if $Z_{t,\nu}(a, 0) = 0$.

Being the most complicated one, Theorem 4.4 is not always the best choice in practice, while for most optimizers it suffices to verify the structure of the preconditioner. In Appendix D we provide more corollaries that are better suited for practical use, two of them being a necessary and a sufficient condition.

4.3 Frequency coupling in existing optimizers

In this subsection, we discuss whether the frequency decoupling holds for popular methods such as SGD, Adam and Muon.

Stochastic gradient descent (SGD). We consider SGD in its most general form $h_{t+1} = m_{t+1} = \beta_t m_t + (1-\beta_t)g_t$, $u_t = a_t m_{t+1} + b_t g_t$. Thus $(a_t, b_t) = (0, 1)$ is plain SGD, $(1, 0)$ is momentum, and $(a_t, b_t) = (\beta_t, 1-\beta_t)$ is the Nesterov variant with normalized momentum. A scalar weight decay preserves decoupling, since $\lambda_t \theta_t$ acts as the scalar λ_t on every $\mathcal{H}_\omega$, whether it is added to the gradient or applied as decoupled weight decay (Loshchilov & Hutter, 2019).

Proposition 4.5. *SGD, SGD with momentum and Accelerated SGD are frequency decoupled.*

Adam. In the coordinate basis Adam (Kingma & Ba, 2015) keeps $h_t = (m_t, v_t)$ with $m_{t+1} = \beta_1 m_t + (1-\beta_1)g_t$, $v_{t+1} = \beta_2 v_t + (1-\beta_2)|g_t| \odot |g_t|$ and, after bias correction, $u_t = \bar{m}_{t+1}/(\sqrt{\bar{v}_{t+1}} + \varepsilon)$ componentwise, i.e. $Z_t(h, g) = \bar{m}$ and $P_t(h) = \mathrm{Diag}((\sqrt{\bar{v}} + \varepsilon)^{-1})$.

Proposition 4.6. *Adam in spatial coordinates is not frequency decoupled.*

Muon with matrix kernels. For a weight presented as a matrix, Muon (Jordan et al., 2024) uses $M_{t+1} = \beta_t M_t + (1-\beta_t)G_t$ and $D_t = \mathrm{msign}(M_{t+1})$, where $\mathrm{msign}(A) = A((A^*A)^{1/2})^+$ is the polar factor of A (Appendix B.2), equal to $U_r V_r^*$ for a compact SVD $A = U_r \Sigma_r V_r^*$.

Proposition 4.7. *Muon on a kernel matrix of size $N_1 \times N_2$ with $N_1 N_2 > 1$ is not decoupled.*

Muon with tensor kernels. A third mechanism arises for tensor weights when unfolding them into matrices. Consider an isometry $\mathcal{R} : \mathbb{C}^{C_{\mathrm{out}} \times C_{\mathrm{in}} \times N} \to \mathbb{C}^{C_{\mathrm{out}} \times (C_{\mathrm{in}} N)}$ and $D_t = \mathcal{R}^{-1}(\mathrm{msign}(\mathcal{R}(M_{t+1})))$. Then, by the unitary equivariance of msign, applying the unitary DFT on the right yields

$$\widehat{D}_\omega = \left[\left(\sum_\nu \widehat{M}_\nu \widehat{M}_\nu^*\right)^{1/2}\right]^+ \widehat{M}_\omega, \tag{4}$$

enabling an explicit dependence on all frequencies through one shared factor.

In general, our framework allows the analysis of any first-order optimizer. Full proofs and an extended list of optimizers are given in Appendix E.

4.4 Per-frequency localization

Block-diagonality of $P(h)$ does not prevent $P_\omega(h)$ from depending on the state of other frequencies, and (LS) also requires localizing the arguments of the operator. In the algorithms of Section 4.3 these arguments are states accumulated on the full gradient, so we re-accumulate them on isolated gradients by running the whole recursion of the method separately for every frequency. First, we consider isometries $E_\omega : \mathbb{C}^{q_\omega} \to \mathcal{H}_\omega \subseteq \mathcal{H}$, so that

$$E_\omega^* E_\nu = \delta_{\omega\nu} I, \qquad \sum_\omega E_\omega E_\omega^* = I, \qquad g_\omega = E_\omega^* g, \qquad g = \sum_\omega E_\omega g_\omega. \tag{5}$$

Informally, E_ω inserts one block in the Fourier basis while also applying the inverse DFT in the spatial basis. We now use these operators to construct a decoupled optimizer out of a given non-decoupled source optimizer.

Localizing the whole recursion. Given a source optimizer (Φ_t, Ψ_t) with state space S^{src}, run an independent replica $\xi_{t,\omega} \in S^{\mathrm{src}}$ per frequency:

$$\xi_{t+1,\omega} = \Phi_t(\xi_{t,\omega}, E_\omega g_{t,\omega}), \qquad u_{t,\omega} = E_\omega^* \Psi_t(\xi_{t+1,\omega}, E_\omega g_{t,\omega}), \qquad \theta_{t+1} = \theta_t - \eta_t \sum_\omega E_\omega u_{t,\omega}, \tag{6}$$

with all replicas initialized at the prescribed source state.

Theorem 4.8. *For any source optimizer, construction (6) is frequency decoupled on the state space $\prod_\omega S^{\mathrm{src}}$.*

Table 1: Localized methods. The full direction can be obtained as $u_t = \sum_\omega E_\omega u_{t,\omega}$. Bars denote bias correction.

Source	Localized method	Local state	Local direction $u_{t,\omega}$
Adam (diagonal)	Fourier-Adam	m_ω, v_ω	$\bar{m}_\omega/(\sqrt{\bar{v}_\omega}+\varepsilon)$
Muon (matrix / standard unfolding)	CommMuon	M_ω	$\mathrm{vec}\big(\mathrm{msign}(M_\omega)\big)$
Muon (general unfolding)	LU-Muon	M_ω	$\mathcal{A}^*_{\omega,\tau}\,\mathrm{msign}(\mathcal{A}_{\omega,\tau}M_\omega)$

Proof sketch. The replica of frequency ω receives only $g_{t,\omega}$ and contributes only to $u_{t,\omega}$. Hence conditions $(*)$–$(***)$ of Theorem 4.3 hold with $\Phi_{t,\omega}(a,b) = \Phi_t(a, E_\omega b)$ and $\Psi_{t,\omega}(a,b) = E^*_\omega \Psi_t(a, E_\omega b)$.

Every frequency is thus trained as the source method would train it if the problem contained this frequency alone. While (6) is already an explicit algorithm, full replicas can be rather expensive. To correct this drawback, we use the concept of a "compact state".

Compact state. For a compact implementation one stores $a_{t,\omega} \in S_\omega$ in place of the replica $\xi_{t,\omega}$ and runs

$$a_{t+1,\omega} = \phi_{t,\omega}(a_{t,\omega}, g_{t,\omega}), \qquad u_{t,\omega} = \psi_{t,\omega}(a_{t+1,\omega}, g_{t,\omega}), \qquad \theta_{t+1} = \theta_t - \eta_t \sum_\omega E_\omega u_{t,\omega}, \tag{7}$$

where $J_\omega \colon S_\omega \to S^{\mathrm{src}}$ parametrizes an invariant family containing the initial state, $J_\omega a_{0,\omega} = h_0^{\mathrm{src}}$ with maps satisfying

$$\Phi_t(J_\omega a, E_\omega b) = J_\omega \phi_{t,\omega}(a,b), \qquad E^*_\omega \Psi_t(J_\omega a, E_\omega b) = \psi_{t,\omega}(a,b) \tag{8}$$

for all $a \in S_\omega$ and $b \in \mathbb{C}^{q_\omega}$.

Theorem 4.9. *Construction* (7) *is frequency decoupled on the state space* $\prod_\omega S_\omega$.

Each frequency thus stores only the degrees of freedom of the source state that it excites by itself. Since the replica of frequency ω accumulates its state from a single harmonic, which the shifts only multiply by a phase, for an equivariant source this state inherits the symmetry of the harmonic and is described by quantities of the size of one frequency block rather than by a full copy of the source state. Full proofs of Theorems 4.8 and 4.9 are given in Appendix F.

Localized optimizers. Applying (6) to the methods of Section 4.3 and compressing the replicas by (7) yields the algorithms of Table 1. For the full derivation see Appendix G. Here we will list how they correct the coupling mechanisms.

- **Fourier-Adam.** Adam couples the frequencies because its preconditioner is diagonal in the spatial rather than the Fourier basis. Fourier-Adam accumulates its second moment $|g_{t,\omega}| \odot |g_{t,\omega}|$ after the frequency has been extracted, which is not the Fourier transform of the spatial second moment, so the preconditioner of each frequency depends on this frequency alone. It coincides with the localization of Adam for $\varepsilon^{\mathrm{src}} = \varepsilon/\sqrt{N}$ and $\eta_t^{\mathrm{src}} = \eta_t/\sqrt{N}$, where $N = |\Gamma|$.
- **CommMuon.** Muon on the standard unfolding couples the frequencies through the shared factor $[(\sum_\nu \widehat{M}_\nu \widehat{M}^*_\nu)^{1/2}]^+$ of (4). In CommMuon this factor reduces to $((M_\omega M^*_\omega)^{1/2})^+$ while the other frequencies vanish, so the polar direction is computed within each block.
- **LU-Muon.** The frequency is isolated before the unfolding of the source is applied, so the polar factor acts on one frequency at a time. Here $\mathcal{A}_{\omega,\tau} = \mathcal{R}_\tau E_\omega$ embeds the block through a candidate unfolding $\mathcal{R}_\tau$ chosen by a *local* rule inherited from the source. The direction admits a positive semidefinite local preconditioner satisfying (LS), and the Newton–Schulz variant remains decoupled.

5 Experiments

We examine a set of architectures (FNO, ViT and DDPM). Each optimizer is evaluated on every architecture using different random seeds, we report the mean ±standard deviation across them. The

hyperparameters is tuned independently for each optimizer–architecture pair via grid search. We report results for the configuration with the best validation performance (experimental protocols can be seen in Appendix A.1). Additional experiments (e.g. ResNet-50, and FNO on physics-informed problem) are considered in Appendix A.2.

5.1 FNO (2D Navier-Stocks)

This section focuses on neural operators for physical systems. We take the two-dimensional incompressible Navier-Stokes problem of PDEBench (Takamoto et al., 2022).

Table 2 reports the one-step training error and the test rollout error over the first 1, 10 and 20 predicted frames, all as relative L_2 error $\|\hat{u} - u\|_2 / \|u\|_2$. We use all Fourier-block singular values, counting conjugate pairs twice. Across the four spectral layers, we average the top-10% shares of squared singular values (C_{10}^{σ}) and frequency-wise weight energy (C_{10}^{ω}), and the largest singular value ($\sigma_{\max}$).

Table 2: PDEBench 2D Navier–Stokes, FNO, one-step and pushforward training: errors and spectra of the spectral layers at the checkpoint selected by the 20-step validation rollout.

	Train	Test rollout L_2 (%)			Spectral layers		
Optimizer	L_2 (%)	1 step	10 steps	20 steps	C_{10}^{σ} (%)	C_{10}^{ω} (%)	$\sigma_{\max}$
One-step training							
Adam	$2.32_{\pm 0.70}$	$5.30_{\pm 1.12}$	$39.8_{\pm 2.5}$	$82.2_{\pm 5.4}$	$63.2_{\pm 0.3}$	$20.8_{\pm 1.4}$	$6.2_{\pm 1.8}$
Fourier-Adam	$2.22_{\pm 0.21}$	$5.10_{\pm 0.31}$	$38.5_{\pm 1.1}$	$81.5_{\pm 2.9}$	$63.1_{\pm 0.9}$	$21.1_{\pm 0.3}$	$5.1_{\pm 0.5}$
Muon	$1.33_{\pm 0.04}$	$5.23_{\pm 0.20}$	$36.7_{\pm 1.4}$	$72.3_{\pm 2.7}$	$55.3_{\pm 0.9}$	$30.1_{\pm 0.3}$	$5.9_{\pm 0.4}$
LU (split, exact)	$1.42_{\pm 0.05}$	$4.78_{\pm 0.20}$	$35.6_{\pm 1.1}$	$71.4_{\pm 0.4}$	$56.9_{\pm 0.6}$	$23.4_{\pm 0.4}$	$4.0_{\pm 0.2}$
LU (split, NS)	$1.54_{\pm 0.28}$	$5.12_{\pm 0.47}$	$37.0_{\pm 1.3}$	$73.9_{\pm 0.8}$	$55.5_{\pm 2.9}$	$22.3_{\pm 1.7}$	$3.6_{\pm 0.5}$
LU (axes)	$1.36_{\pm 0.18}$	$4.48_{\pm 0.32}$	$31.7_{\pm 0.8}$	$63.3_{\pm 0.8}$	$43.4_{\pm 2.2}$	$16.9_{\pm 0.4}$	$3.5_{\pm 0.8}$
Pushforward training							
Adam	$4.83_{\pm 0.26}$	$11.58_{\pm 0.65}$	$32.6_{\pm 1.0}$	$57.2_{\pm 1.5}$	$58.4_{\pm 2.3}$	$18.6_{\pm 0.6}$	$4.2_{\pm 0.2}$
Fourier-Adam	$3.22_{\pm 0.83}$	$9.17_{\pm 0.63}$	$28.2_{\pm 1.4}$	$51.2_{\pm 2.0}$	$56.9_{\pm 0.9}$	$18.3_{\pm 0.9}$	$5.2_{\pm 0.9}$
Muon	$2.65_{\pm 0.34}$	$8.50_{\pm 0.57}$	$29.5_{\pm 0.7}$	$54.6_{\pm 0.9}$	$44.6_{\pm 1.0}$	$24.9_{\pm 1.0}$	$3.1_{\pm 0.1}$
LU (split, exact)	$2.61_{\pm 0.84}$	$8.11_{\pm 1.48}$	$23.0_{\pm 2.2}$	$42.2_{\pm 4.0}$	$50.8_{\pm 0.8}$	$16.9_{\pm 0.0}$	$5.4_{\pm 0.5}$
LU (split, NS)	$2.49_{\pm 0.73}$	$7.95_{\pm 1.23}$	$22.9_{\pm 2.3}$	$41.4_{\pm 3.6}$	$50.0_{\pm 0.6}$	$16.8_{\pm 0.1}$	$5.0_{\pm 0.4}$
LU (axes)	$1.88_{\pm 0.68}$	$6.96_{\pm 1.01}$	$21.6_{\pm 1.7}$	$41.2_{\pm 2.9}$	$47.4_{\pm 4.2}$	$15.3_{\pm 0.3}$	$7.3_{\pm 1.9}$

The localized optimizers outperform their original counterparts. With one-step training LU-Muon (axes) has the lowest test error at every horizon, and in both regimes every seed of it is better at 20 steps than every seed of Muon and of Adam. The split variants improve on Muon after one step and match it at 20 steps. Fourier-Adam matches Adam in the mean, as expected on FNO.

Pushforward training lowers the 20-step error of every optimizer by 18 to 33 points at the price of a higher one-step error, and both localizations beat their sources on every seed: Fourier-Adam is more accurate than Adam at every horizon, and all three LU-Muon variants reach 41–42% against 54.6% for Muon, with LU-Muon (axes) best after 1 and 10 steps.

5.2 ViT

In this section we check whether the effect reaches a transformer through its only convolution. We train ViT-Tiny (Dosovitskiy et al., 2021) on Tiny ImageNet following the protocol of Tensorion (Bogachev et al., 2026). The compared optimizer trains only the convolutional patch embedding, and all other parameters are trained by a shared AdamW (details in Appendix A.1.2).

Table 3 shows that Fourier-Adam improves accuracy over Adam by 2.35 percentage points, despite being applied only to the convolution. LU-Muon (axes), however, falls short of Muon by 1.50 points. Both variants yield more concentrated singular-value spectra and lower entropy ranks, but these shifts are not consistently accompanied by higher accuracy.

Table 3: ViT-Tiny on Tiny ImageNet, validation-selected checkpoint, mean$_{\pm SD}$; train values are running averages over its epoch. Patch-embedding kernel (192×192 unfolding): C_{10}^{σ} and C_{10}^{ω}, energy share of the top 10% of singular values and of 8×8 kernel frequencies; r_H, normalized entropy rank.

	Train	Test		Patch-embedding spectrum		
Optimizer	Loss	Acc. (%)	Loss	C_{10}^{σ} (%)	r_H (%)	C_{10}^{ω} (%)
Adam	$2.665_{\pm 0.162}$	$54.84_{\pm 0.89}$	$3.188_{\pm 0.065}$	$36.8_{\pm 0.7}$	$57.0_{\pm 0.6}$	$16.1_{\pm 0.9}$
Fourier-Adam	$2.422_{\pm 0.147}$	$57.19_{\pm 0.53}$	$3.079_{\pm 0.016}$	$70.5_{\pm 0.8}$	$22.1_{\pm 0.6}$	$22.9_{\pm 0.0}$
Muon	$2.312_{\pm 0.134}$	$59.62_{\pm 0.12}$	$2.939_{\pm 0.020}$	$61.3_{\pm 1.6}$	$30.4_{\pm 1.1}$	$39.6_{\pm 0.7}$
LU-Muon (axes)	$2.393_{\pm 0.131}$	$58.12_{\pm 0.48}$	$3.013_{\pm 0.053}$	$78.4_{\pm 0.6}$	$18.7_{\pm 0.6}$	$22.2_{\pm 0.5}$

5.3 DDPM

Finally, we turn to a generative model. A diffusion model denoises images at every noise level, from coarse layout to fine detail, so the task itself spreads its gradients unevenly over frequencies.

We train a 103.3M DDPM (Ho et al., 2020) on Tiny ImageNet at 64×64 (details in Appendix A.1).

Table 4 reports the denoising error MSE: denoising error ($\times 10^2$), running average over the selected epoch with the training weights (train) and EMA weights on the 10 000 test images (test); FID of 5 000 samples against the 10 000 test images. Kernel spectrum of the EMA weights (unfolding $C_{\text{out}} \times C_{\text{in}} k^2$), median over the 62 convolutions; $\sigma_{\min}$ is the smallest singular value above 10^{-6}.

Table 4: DDPM on Tiny ImageNet.

	MSE ($\times 10^2$)			Kernel spectrum			
Optimizer	Train	Test	FID	$\sigma_{\max}/\sigma_{\min}$	r_s	$\sigma_{\max}$	$\|W\|_*$
Adam	5.29	5.15	31.2	10.1	48.3	5.87	679
Fourier-Adam	5.24	5.10	29.6	8.3	60.3	6.16	675

Fourier-Adam improves every task metric: the denoising error is lower on both the training and the test set, and FID drops from 31.2 to 29.6. The difference is larger in the kernels than in the loss. At almost the same nuclear norm (675 vs. 679), Fourier-Adam spreads the spectrum over more directions: the stable rank grows from 48.3 to 60.3 and $\sigma_{\max}/\sigma_{\min}$ falls from 10.1 to 8.3. Figure 1 compares the radial power spectrum of the samples with that of the test images. Both models generate too little high-frequency power, and the deficit grows with the radius, but beyond the lowest frequencies Fourier-Adam stays closer to the data at every radius. Around radius 40, Adam's samples carry roughly 28% less power than real images and Fourier-Adam's roughly 23%.

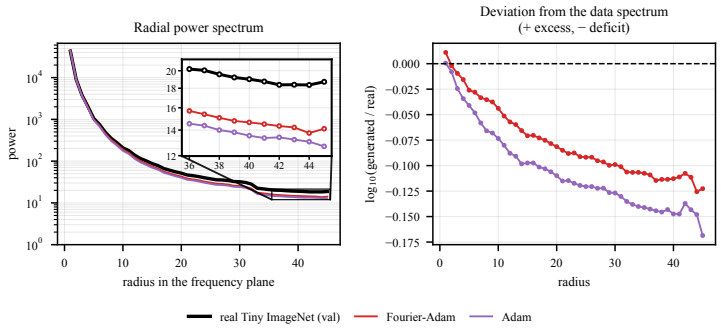

Figure 1: Spectral correlation between generated and reference images

AI USE STATEMENT

In this work, we used generative AI tools (LLMs) to polish the writing, check the text for typos, and compile the textual descriptions of our experimental protocols from our code and configuration files. We did not use generative AI tools to formulate theorem statements, derive proofs, propose hypotheses, design experiments, or interpret results. All proofs, theorem statements, and experimental protocols were developed by the authors without the aid of LLMs. All LLM-compiled protocol descriptions were verified against the actual code and configurations used in our experiments. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All assumptions are stated in Sections 3–4.4 and complete proofs of every claim are given in Appendix. Each proposed method is derived with a proof and specified as explicit pseudocode (Appendix G), including initialization and bias correction, and real-valued parameters are treated in Appendix H. The experimental protocol, datasets, tuning budget and metrics are described in Section 5. Code will be released.

REFERENCES

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning (ICML)*, 2018. arXiv:1802.06509.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1905.13655.

Francesca Arrigo, Michele Benzi, and Caterina Fenu. Computation of generalized matrix functions. *SIAM Journal on Matrix Analysis and Applications*, 37(3):836–860, 2016. arXiv:1512.01446.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.08498.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning, 2024a. arXiv:2410.21265.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology, 2024b. arXiv:2409.20325.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023.

Vladimir Bogachev, Vladimir Aletov, Alexander Molozhavenko, Sergey Kudriashov, and Maxim Rakhuba. Tensorion: A tensor-aware generalization of the muon optimizer, 2026. arXiv:2606.25975.

Nicolas Bourbaki. *Theory of Sets*. Elements of Mathematics. Springer, 2004. Reprint of the 1968 English edition.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2005.14165.

Shuhao Cao. Choose a transformer: Fourier or Galerkin. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2105.14995.

David Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1911.03584.

Philip J. Davis. *Circulant Matrices*. Wiley, 1979.

Emily L. Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. arXiv:1404.0736.

Ashim Dhor, Rasel Mondal, and Pin-Yu Chen. ITNet: A learnable integral transform that subsumes convolution, attention, and recurrence, 2026. arXiv:2606.19538.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.03404.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.11929.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, https://transformer-circuits.pub/2021/framework/index.html, 2021.

Robert M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Adaptive Fourier neural operators: Efficient token mixers for transformers. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2111.13587.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1705.09280.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning (ICML)*, 2018. arXiv:1802.09568.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.03385.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2006.11239.

Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Yerlan Idelbayev and Miguel Á. Carreira-Perpiñán. Low-rank compression of neural nets: Learning the rank of each layer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. In *British Machine Vision Conference (BMVC)*, 2014. arXiv:1405.3866.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6980.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. arXiv:2108.08481.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.08895.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, et al. Muon is scalable for LLM training, 2025. arXiv:2502.16982.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. arXiv:2004.08249.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2201.03545.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1711.05101.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning (ICML)*, 2015. arXiv:1503.05671.

Charles H. Martin and Michael W. Mahoney. Traditional and heavy-tailed self regularization in neural network models. In *International Conference on Machine Learning (ICML)*, 2019. arXiv:1901.08276.

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021. arXiv:1810.01075.

Michaël Mathieu, Mikael Henaff, and Yann LeCun. Fast training of convolutional networks through FFTs. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.5851.

James R. Munkres. *Topology*. Prentice Hall, second edition, 2000.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1707.09564.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, et al. In-context learning and induction heads, 2022. arXiv:2209.11895.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning (ICML)*, 2019. arXiv:1806.08734.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. arXiv:0706.4138.

Oren Rippel, Jasper Snoek, and Ryan P. Adams. Spectral representations for convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. arXiv:1506.03767.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. arXiv:2112.10752.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. arXiv:1505.04597.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6120.

Hanie Sedghi, Vineet Gupta, and Philip M. Long. The singular values of convolutional layers. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1805.10408.

Jean-Pierre Serre. *Linear Representations of Finite Groups*. Springer, 1977.

Sahil Singla and Soheil Feizi. Fantastic four: Differentiable and efficient bounds on singular values of convolution layers. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:1911.10258.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2022.

Audrey Terras. *Fourier Analysis on Finite Groups and Applications*, volume 43 of *London Mathematical Society Student Texts*. Cambridge University Press, 1999.

Asher Trockman and J. Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2104.07167.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.03762.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing shampoo using adam, 2024. arXiv:2409.11321.

Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1705.08292.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020. arXiv:1901.06523.

Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning, 2017. arXiv:1705.10941.

Zihao Zhang, Stefan Zohren, and Stephen Roberts. DeepLOB: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, 67(11):3001–3012, 2019. arXiv:1808.03668.

A Experiment details

A.1 Experimental protocols

A.1.1 FNO (2D Navier–Stokes)

- **Data:** 2D incompressible Navier–Stokes with inhomogeneous forcing from PDEBench (Takamoto et al., 2022) (`NS_Incom`; $\nu = 0.01$ on $[0, 1]^2$, zero velocity outside the domain). Every file holds 4 trajectories of 1,000 frames ($t = 0, 0.005, \ldots, 4.995$) on a 512^2 grid with the velocity (v_x, v_y), a passively advected scalar ("particles") and a forcing field (f_x, f_y) that is constant in time and varies between trajectories. Of the 274 files we drop the 48 whose MD5 checksum occurs twice in the PDEBench file list and split the 33 with the lowest numbers (0–34 without 20 and 21) by file, in order, into 25/4/4 for training/validation/test (100/16/16 trajectories). We keep every 10th frame (100 frames, $\Delta t = 0.05$) and every 8th grid point (64^2, no averaging); the zero-shot test takes every 4th point (128^2). Each of the five channels is normalized by its mean and standard deviation over the training trajectories at 64^2, also at 128^2.
- **Task:** ten consecutive frames of $(v_x, v_y, \text{particles})$ and the two forcing channels (32 input channels) predict the next frame (3 channels). Each epoch draws 64 target frames per training trajectory uniformly, with replacement, from the 90 frames that have ten predecessors (6,400 pairs, 200 steps). The loss is the relative L_2 error of the de-normalized next frame, per channel, averaged over the three channels and the batch.
- **One-step training:** the input is always the true window of the ten preceding frames.
- **Pushforward training** (Brandstetter et al., 2022): for every training pair with target frame t we draw u uniformly from $\{0, \ldots, U_e\}$ with $U_e = \min(e - 1, 3)$ in epoch e (so $U_e = 0, 1, 2, 3, 3, \ldots$), clipped to $u \leq t - 10$, take the true frames $t - u - 10, \ldots, t - u - 1$, roll the model out for u steps without gradient and predict frame t with gradient. No gradient flows through the unrolled steps. The target frames, their order and the initial weights are those of one-step training, and u has its own random stream, so at equal learning rate and seed the two regimes differ only in the input windows.
- **Model:** our implementation of FNO (Li et al., 2021): a lifting 1×1 convolution from the 32 input channels and two grid coordinates to width 16, four spectral layers with modes $|k_1|, |k_2| \leq 12$, each with a pointwise 1×1 path and followed by GELU except the last, and a projection $16 \to 64 \to 3$ with GELU. Every spectral layer stores one real DC block and one complex 16×16 block per conjugate pair of frequencies, 313 blocks per layer; the spectral blocks hold 99.5% of the 642,931 parameters.
- **Optimizers:** the compared optimizer trains the spectral blocks; the lifting, pointwise and projection layers use Adam with learning rate 10^{-3}; no weight decay anywhere. Adam and Fourier-Adam use $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$; Adam keeps separate second moments for the real and imaginary parts of a Fourier coefficient, Fourier-Adam one shared second moment $|g|^2$. Muon orthogonalizes each layer's real parameter unfolded as $C_{\text{out}} \times (\text{rest})$ ($16 \times 10{,}000$) with five float32 Newton–Schulz steps; in the Muon runs only, the gradient norm of all model parameters is clipped at 1. LU-Muon takes a polar factor per Fourier block. The split unfolding factors $C_{\text{out}} = C_{\text{in}} = 4 \cdot 4$ and takes, per block and step, whichever of the three 16×16 unfoldings (the block itself and the two that pair an output with an input factor) has the largest nuclear norm, with the polar factor either exact by float32 SVD (split, exact) or from ten bf16 Newton–Schulz steps (split, NS; the nuclear norm is estimated as $\langle X, Q \rangle$ with Q the Newton–Schulz factor refined by four float32 steps). The axes unfolding takes the block itself, with ten bf16 Newton–Schulz steps. Muon and LU-Muon use momentum 0.95 with Nesterov and no rescaling of the update.

- **Training:** batch 32, 100 epochs, in both regimes; both learning rates follow a cosine decay to zero over the 100 epochs, updated once per epoch.
- **Learning rates:** tuned separately in each regime. Tuning runs cover the first 10 epochs of the 100-epoch schedule with seed 42 and are scored by the mean validation error of the 20-step rollout over epochs 6–10; grids with half-decade steps, extended while the best value lies at an edge. Selected rates (one-step/pushforward): Adam and Fourier-Adam $3 \cdot 10^{-3}$/$3 \cdot 10^{-3}$, Muon $3 \cdot 10^{-2}$/10^{-2}, all LU-Muon variants 10^{-3}/$3 \cdot 10^{-3}$. Main runs use seeds 42, 43 and 44 with the same initial weights for all methods and both regimes.
- **Evaluation:** the model rolls out autoregressively for 90 steps from the first ten frames of every validation or test trajectory; the error over the first h predicted frames is taken over these frames jointly. The checkpoint is selected by the validation error over the first 20 predicted frames, since with one-step training the error of the full 90-step rollout exceeds that of a zero prediction (100%) at every epoch of every run. We report the test error over the first 1, 10 and 20 predicted frames at 64^2; zero-shot at 128^2 the mean errors differ by at most 0.4 points. The training error in the table is the mean one-step loss during the selected epoch; with pushforward training it is averaged over the pairs with $u = 0$.
- **Hardware:** two NVIDIA RTX 2080 Ti.

A.1.2 ViT

- **Data:** Tiny ImageNet (200 classes, 64×64). A fixed random split of the training set gives 90,000 training and 10,000 validation images; the official 10,000-image validation set serves as the test set. Training augmentation: random crops with 4-pixel padding and horizontal flips; per-channel normalization.
- **Model:** ViT-Tiny (Dosovitskiy et al., 2021) in the torchvision implementation, with 8×8 patches (64 tokens), 12 blocks of width 192, 3 heads and MLP width 768, 5.4M parameters. The patch embedding is a convolution $3 \to 192$ with kernel size and stride 8 (36,864 weights, 0.7% of the parameters).
- **Optimizers:** following the protocol of Tensorion (Bogachev et al., 2026), the compared optimizer trains only the patch-embedding kernel, without weight decay. An auxiliary AdamW with learning rate 10^{-3}, $\beta = (0.9, 0.999)$ and weight decay $5 \cdot 10^{-4}$ trains all other parameters. Adam and Fourier-Adam use $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$; Fourier-Adam keeps Adam's moments on the coefficients of the orthonormal DFT of the 8×8 kernel. Muon orthogonalizes the 192×192 kernel unfolding $C_{\text{out}} \times C_{\text{in}} k^2$. LU-Muon (axes) takes the polar factor of the 192×3 block $C_{\text{out}} \times C_{\text{in}}$ at each of the 64 frequencies of the kernel DFT (34 up to conjugation). Both use momentum 0.95 with Nesterov and five Newton–Schulz steps.
- **Training:** cross-entropy loss, batch 128, 400 epochs. Both learning rates follow cosine annealing to zero without warm-up.
- **Learning rates:** tuning runs cover the first 10 epochs of the 400-epoch schedule with seed 42 and are scored by the mean validation loss over their last five epochs; grids with steps of $3\times$, extended while the best value lies at an edge. Selected: Adam $3 \cdot 10^{-4}$, Fourier-Adam 0.1, Muon and LU-Muon 0.03. Main runs use seeds 42, 43 and 44; all methods start from the same initial weights for a given seed.
- **Evaluation:** the checkpoint with the lowest validation loss, tested on the official validation split. Train metrics are running averages over that epoch; spectra are those of the patch-embedding kernel at that checkpoint.
- **Hardware:** NVIDIA H200.

A.1.3 DDPM

- **Data:** Tiny ImageNet at 64×64, without class labels. A fixed random split of the 100,000 training images gives 90,000 images for training and 10,000 held out, of which 1,000 form the validation set; the official 10,000 validation images serve as the test set. Images are scaled to $[-1, 1]$; random horizontal flips are the only augmentation.
- **Model:** DDPM (Ho et al., 2020) with ϵ-prediction, $T = 1000$ steps and the cosine noise schedule (Nichol & Dhariwal, 2021). The U-Net has 103.3M parameters: base width 256, channel multipliers 1, 2, 2, two residual blocks per resolution, self-attention at 16×16,

dropout 0.1. It contains 62 convolutions with 94.3M parameters: 40 of size 3×3 and 22 of size 1×1 (the skip projections of the residual blocks and the query–key–value and output projections of the attention blocks).

- **Optimizers:** the compared optimizer trains the 62 convolution kernels; all other parameters use Adam with learning rate $2 \cdot 10^{-4}$, the same in both runs. Adam and Fourier-Adam use learning rate $2 \cdot 10^{-4}$, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$ and no weight decay; the rate is fixed in advance, not tuned. Fourier-Adam keeps Adam's moments on the coefficients of the unitary DFT of every kernel; on the 22 1×1 kernels, whose DFT is trivial, it coincides with Adam.
- **Training:** 200 epochs with batch 128 (703 steps per epoch, 140,600 steps), each step accumulated over two micro-batches of 64; bf16 autocast in the forward and backward passes with float32 parameters, losses and optimizer states; linear warm-up over 500 steps, cosine decay to zero; EMA of the weights with decay 0.9999; the same initial weights for both optimizers.
- **Evaluation:** the EMA checkpoint with the lowest denoising MSE on the 1,000 validation images (epoch 200 for both optimizers). We report the denoising MSE of the EMA weights on the 10,000 test images and, as the training error, the running mean over the selected epoch with the raw weights. Samples come from the 1000-step ancestral sampler (clipped $\hat{x}_0$, exact posterior variance) with the same noise seeds for both optimizers: FID of 5,000 samples against the test images, and the radial power spectrum of the samples against that of the test images. Reconstructions start from the first eight test images noised to $t = 500$. Kernel spectra are those of the EMA weights (unfolding $C_{\text{out}} \times C_{\text{in}} k^2$).
- **Hardware:** NVIDIA H200.

A.2 Additional results

A.2.1 ResNet-50

ResNet-50. For ImageNet we have a short series with the Adam and Fourier Adam: ResNet-50, 200 epochs, batch 256, bf16, learning rate 10^{-3} with cosine decay, three seeds (Table 5).

Fourier-Adam reaches 72.41% top-1 against 66.68% for Adam, with a lower validation loss (1.163 against 1.363).

Table 5: ResNet-50 on ImageNet, 200 epochs; both optimizers use learning rate 10^{-3} with cosine decay. Top-1 accuracy and loss on the 50,000 official validation images.

	Top-1 (%)				Loss (ep. 200)		Kernel spectrum		
Optimizer	Ep. 1	Ep. 5	Ep. 10	Ep. 200	Train	Val.	C_{10}^{σ} (%)	r_H (%)	$\sigma_{\max}$
Adam	20.31	42.23	51.25	66.68	1.611	1.363	65.0	29.1	2.09
Fourier-Adam	34.46	60.75	67.07	72.41	1.010	1.163	40.3	60.0	18.81

A.2.2 FNO (shallow water)

Motivated by our Navier–Stokes results, we next evaluate on the two-dimensional shallow-water problem from PDEBench (Takamoto et al., 2022), which models a radial dam break. The model maps ten consecutive frames to the next one and is trained on single steps at 64^2. At test time it rolls out the remaining 91 frames autoregressively, feeding its own predictions back as input, both at 64^2 and zero-shot at the native 128^2; we report the relative L_2 error of this rollout.
Table 6 summarizes the results: LU (axes) reaches the lowest one-step and 64^2 rollout errors and is the only optimizer whose validation rollout stays below 6% after the first ten epochs, while at 128^2 Fourier-Adam is slightly ahead.

Table 6: PDEBench shallow water, FNO, at the validation-selected checkpoint: one-step training error and test error of the 91-step rollout, relative L_2, median over three seeds with the minimum–maximum range as a subscript. Max val.: largest validation rollout error over epochs 11–100 and all seeds (> 100: worse than predicting zero). Spectra of the spectral layers: mean$_{\pm \mathrm{SD}}$ over seeds.

	Train	Test rollout L_2		Stability	Spectral layers		
Optimizer	L_2 (%)	64^2 (%)	128^2 (%)	Max val. (%)	C_{10}^{σ} (%)	C_{10}^{ω} (%)	$\sigma_{\max}$
Adam	$0.046_{0.03\text{–}0.21}$	$0.348_{0.31\text{–}1.43}$	$0.664_{0.51\text{–}1.55}$	> 100	$56.1_{\pm 3.8}$	$20.9_{\pm 0.5}$	$11.0_{\pm 1.4}$
Fourier-Adam	$0.046_{0.04\text{–}0.05}$	$0.288_{0.26\text{–}0.38}$	$0.445_{0.41\text{–}0.62}$	> 100	$56.4_{\pm 1.1}$	$21.6_{\pm 0.7}$	$11.0_{\pm 1.2}$
Muon	$0.173_{0.09\text{–}0.36}$	$0.793_{0.47\text{–}2.36}$	$2.374_{0.74\text{–}3.10}$	> 100	$66.1_{\pm 3.4}$	$32.3_{\pm 3.0}$	$64.6_{\pm 27.9}$
LU (axes)	$0.023_{0.02\text{–}0.03}$	$0.256_{0.13\text{–}0.27}$	$0.467_{0.47\text{–}0.78}$	5.4	$47.9_{\pm 0.6}$	$18.7_{\pm 0.3}$	$17.1_{\pm 1.1}$
Shampoo	$0.169_{0.07\text{–}0.40}$	$0.308_{0.21\text{–}1.47}$	$0.701_{0.58\text{–}1.51}$	33.8	$71.8_{\pm 5.0}$	$36.1_{\pm 4.2}$	$22.6_{\pm 12.6}$
Per-freq. Shampoo	$0.042_{0.04\text{–}0.36}$	$1.410_{0.93\text{–}1.54}$	$1.372_{1.21\text{–}1.51}$	10.1	$44.9_{\pm 4.0}$	$22.8_{\pm 3.5}$	$2.7_{\pm 0.6}$

B General Statements

In this section we state mostly well-known general facts that we will use further on.

B.1 Elements of group theory

We use the following general statements, for which we refer to Serre (1977, Ch. 1–2) and Terras (1999, Ch. 1–2).

Proposition B.1 (Characters and isotypic decomposition)**.** *Let Γ be a finite abelian group with a unitary representation ρ on a finite-dimensional complex Hilbert space $\mathcal{H}$. Then,* (i) *the characters form a group $\widehat{\Gamma}$ with $|\widehat{\Gamma}| = |\Gamma|$ and are orthonormal, $|\Gamma|^{-1} \sum_{s\in\Gamma} \overline{\chi_\omega(s)}\chi_\nu(s) = \delta_{\omega\nu}$;* (ii) *$\mathcal{H} = \bigoplus_{\omega\in\Omega} \mathcal{H}_\omega$, the sum being orthogonal, with $\Pi_\omega^{\mathcal{H}} = |\Gamma|^{-1} \sum_{s\in\Gamma} \overline{\chi_\omega(s)}\rho(s)$;* (iii) *for $\Gamma = \mathbb{Z}_n$ acting by $(\rho(s)F)[v] = F[v - s]$ the characters are $\chi_k(s) = e^{-2\pi iks/n}$ and $\mathcal{H}_k$ is spanned by the Fourier harmonic $F_k[v] = n^{-1/2}e^{2\pi ikv/n}$, the characters of a product of cyclic groups being the tensor products of these.*

The following elementary fact, used in the proof of Proposition 4.6, is a special case of Theorem 3.3, because a circulant matrix is constant along its diagonals and a diagonal circulant matrix is therefore scalar.

Lemma B.2. *Let D be diagonal on $\mathbb{C}^{\Gamma\times q}$, $(Dx)[r, a] = d_{r,a}x[r, a]$, and $(\rho(s)x)[r, a] = x[r - s, a]$. Then, $D \in \mathrm{Comm}(\rho(\Gamma))$ iff $d_{r,a}$ does not depend on r for each a.*

Proof. A direct computation gives $(D\rho(s)x)[r, a] = d_{r,a}x[r - s, a]$ and $(\rho(s)Dx)[r, a] = d_{r-s,a}x[r - s, a]$, so $([D, \rho(s)]x)[r, a] = (d_{r,a} - d_{r-s,a})\, x[r - s, a]$. This vanishes for all x and all r, s, a if and only if $d_{r,a} = d_{r-s,a}$ for all r, s, a, i.e. if and only if $d_{r,a}$ is independent of r for each channel a. □

B.2 The polar map

For $A \in \mathbb{C}^{m\times n}$ let $\mathrm{msign}(A) := A((A^*A)^{1/2})^+$ with $\mathrm{msign}(0) = 0$. Here A^* is the Hermitian adjoint and A^+ the Moore–Penrose pseudoinverse. We use the following properties, in which $A = U_r\Sigma_r V_r^*$ is a compact SVD with $r = \mathrm{rank}\, A$.

(i) $\mathrm{msign}(A) = U_rV_r^*$. This is the generalized matrix function of A for $f \equiv 1$ (Arrigo et al., 2016, Def. 2.4, Thm. 3.4). Equivalently $\mathrm{msign}(A) = ((AA^*)^{1/2})^+A$.

(ii) Spectral and nuclear norms are dual: $\max_{\|X\|_2\le 1} \mathrm{Re}\langle A, X\rangle_F = \|A\|_*$, attained at $X^\star = \mathrm{msign}(A)$ (Recht et al., 2010, Prop. 2.1). For complex matrices one uses the Hermitian adjoint and the real part of the inner product. This identifies the Muon direction as a steepest-descent step under the spectral norm (Bernstein & Newhouse, 2024b).

(iii) Unitary equivariance: $\operatorname{msign}(QAR^*) = Q\operatorname{msign}(A)R^*$ for unitary Q, R (Arrigo et al., 2016, Prop. 3.2(ii)). The same identity holds for isometries $Q^*Q = R^*R = I$, since QU_r and RV_r are again isometric singular bases. In that case also $\|QAR^*\|_* = \|A\|_*$.

(iv) Compatibility with direct sums: $\operatorname{msign}\big(\bigoplus_\omega A_\omega\big) = \bigoplus_\omega \operatorname{msign}(A_\omega)$ (Arrigo et al., 2016, Prop. 3.2(iii)).

(v) Compatibility with transposition and conjugation: $\operatorname{msign}(A^T) = \operatorname{msign}(A)^T$ and $\operatorname{msign}(\overline{A}) = \overline{\operatorname{msign}(A)}$, both immediate from the compact SVD.

Newton–Schulz iteration. For computation one may replace $\operatorname{msign}(A)$ by the finite polynomial map $\mathcal{N}_J(A) = Y_J$, where for $A \neq 0$

$$Y_0 = \frac{A}{\|A\|_F + \varepsilon_{\mathrm{NS}}}, \qquad Y_{j+1} = a_j Y_j + b_j (Y_j Y_j^*) Y_j + c_j (Y_j Y_j^*)^2 Y_j, \quad j = 0, \dots, J-1, \tag{9}$$

and $\mathcal{N}_J(0) = 0$. Here J and the real coefficients are fixed and $\varepsilon_{\mathrm{NS}} \geq 0$. Setiing $(a_j, b_j, c_j) = (3/2, -1/2, 0)$ recovers the classical cubic Newton–Schulz iteration (Higham, 2008). Some implementations of Muon may also use quintic polynomials (Jordan et al., 2024).

B.3 Factorization of maps

The following elementary fact is the universal property of the quotient set (Bourbaki, 2004, Ch. II, §6). Its topological version is Munkres (2000, Thm. 22.2).

Lemma B.3 (Factorization through a surjection). *Let $q\colon X \to Z$ be a surjection and $F\colon X \to Y$. Then,*

$$\exists\, \overline{F}\colon Z \to Y :\ F = \overline{F} \circ q \quad \Longleftrightarrow \quad \forall x, \tilde{x} \in X :\ q(x) = q(\tilde{x}) \Rightarrow F(x) = F(\tilde{x}),$$

and in that case $\overline{F}$ is unique.

B.4 Matrix functions and Kronecker products

For a Hermitian matrix C and a function f defined on its spectrum, $f(C)$ is defined through the spectral decomposition of C (Higham, 2008, Ch. 1). We use two standard facts.

Lemma B.4 (Isometric embedding). *Let $E \in \mathbb{C}^{n\times k}$ satisfy $E^*E = I_k$, let $C \in \mathbb{C}^{k\times k}$ be Hermitian, and let f be defined on the spectrum of C and at 0. Then,*

$$f(ECE^*) = Ef(C)E^* + f(0)\,(I_n - EE^*), \qquad \textit{hence} \qquad f(ECE^*)E = Ef(C). \tag{10}$$

Proof. Complete E to a unitary matrix $U = [E\ E_\perp]$. Thus, $ECE^* = U(C \oplus 0)U^*$, and the invariance of matrix functions under unitary similarity together with their compatibility with block-diagonal matrices (Higham, 2008, Thm. 1.13) give $f(ECE^*) = U\big(f(C) \oplus f(0)I\big)U^* = Ef(C)E^* + f(0)E_\perp E_\perp^*$, where $E_\perp E_\perp^* = I_n - EE^*$. Multiplying by E on the right and using $E^*E = I_k$ gives the second identity. □

The following identities are standard (Horn & Johnson, 1991, Ch. 4), and the second one is also used in the analysis of Shampoo (Gupta et al., 2018).

Lemma B.5 (Kronecker products). *Under column-major vectorization $\operatorname{vec}(AXB) = (B^T \otimes A)\operatorname{vec}(X)$ for matrices of compatible sizes. If A and B are positive definite, then $(B^T \otimes A)^p = (B^T)^p \otimes A^p$ for every $p \in \mathbb{R}$.*

C Proofs for Section 4.1

Let $q_\omega(h, g) = (\Pi_\omega^S h, \Pi_\omega^{\mathcal{H}} g)$ be the canonical surjection $S \times \mathcal{H} \to S_\omega \times \mathcal{H}_\omega$ and put $A_{t,\omega} = \Pi_\omega^S \circ \Phi_t$ and $B_{t,\omega} = \Pi_\omega^{\mathcal{H}} \circ \Psi_t$. Consider

$$\exists\, \Phi_{t,\omega}, \Psi_{t,\omega} :\quad A_{t,\omega} = \Phi_{t,\omega} \circ q_\omega, \quad B_{t,\omega} = \Psi_{t,\omega} \circ q_\omega; \tag{$F_{t,\omega}$}$$

$$\forall x, \tilde{x} \in S \times \mathcal{H} :\quad q_\omega(x) = q_\omega(\tilde{x}) \ \Rightarrow\ A_{t,\omega}(x) = A_{t,\omega}(\tilde{x}) \text{ and } B_{t,\omega}(x) = B_{t,\omega}(\tilde{x}). \tag{$I_{t,\omega}$}$$

Condition ($F_{t,\omega}$) is a recast of the definition and requires that the optimizer *computes* the state and the step of frequency ω from the history and gradient of that frequency. On the other hand ($I_{t,\omega}$) says that perturbing the other frequencies does not affect this one. In fact they are equivalent.

Theorem C.1. *For all $t \in \mathbb{N}_0$ and all $\omega \in \Omega$, $(F_{t,\omega}) \iff (I_{t,\omega})$. A decoupled optimizer satisfies $(F_{t,\omega})$ for all t, and the maps $\Phi_{t,\omega}, \Psi_{t,\omega}$ are hence unique.*

Proof. Fix t and ω. The map q_ω is surjective. Given $(a, b) \in S_\omega \times \mathcal{H}_\omega$, choose any h with $h_\omega = a$, which is possible because every S_ν is nonempty, and take $g = b$ regarded as an element of $\mathcal{H} = \bigoplus_\nu \mathcal{H}_\nu$. Lemma B.3 applied to the pair $(A_{t,\omega}, q_\omega)$ shows that the first half of $(F_{t,\omega})$ is equivalent to the first half of $(I_{t,\omega})$, and applied to $(B_{t,\omega}, q_\omega)$ it does so for the second halves. Since Definition 4.1 asserts the existence of the factors in $(F_{t,\omega})$ for every t, a decoupled optimizer satisfies $(F_{t,\omega})$. Finally the uniqueness of the factors follows from the last assertion of Lemma B.3. □

Theorem 4.2. *Let $L(\theta) = \sum_{\omega\in\Omega} L_\omega(\theta_\omega)$ and let the optimizer be decoupled. Then, there exist maps $T_{s,\omega} \colon S_\omega \times \mathcal{H}_\omega \to S_\omega \times \mathcal{H}_\omega$ with*

$$\begin{pmatrix} h_{t,\omega} \\ \theta_{t,\omega} \end{pmatrix} = \left(T_{t-1,\omega} \circ \cdots \circ T_{0,\omega}\right) \begin{pmatrix} h_{0,\omega} \\ \theta_{0,\omega} \end{pmatrix} \quad (t \geq 1),$$

and the global transitions factor as $T_s = \prod_{\omega\in\Omega} T_{s,\omega}$ and $T_{t-1} \circ \cdots \circ T_0 = \prod_\omega (T_{t-1,\omega} \circ \cdots \circ T_{0,\omega})$.

Proof. We will proceed in three steps with this proof. First we will show that the gradient splits across frequencies. Consequently, that the recursions of the optimizer are local. Finally we show that combining these two facts together yields the per-frequency transition.

Step 1. Let $\theta, v \in \mathcal{H}$. Since $L = \sum_\omega L_\omega \circ \Pi^{\mathcal{H}}_\omega$, the differential of L at θ is

$$dL_\theta[v] = \sum_\omega dL_{\omega,\theta_\omega}[v_\omega] = \operatorname{Re} \sum_\omega \left\langle \nabla L_\omega(\theta_\omega), v_\omega \right\rangle.$$

On the other hand, by the definition of the gradient and the orthogonality of the frequency subspaces, $dL_\theta[v] = \operatorname{Re}\langle \nabla L(\theta), v\rangle = \operatorname{Re} \sum_\omega \langle \Pi^{\mathcal{H}}_\omega \nabla L(\theta), v_\omega\rangle$. Comparing the two expressions for v supported on a single $\mathcal{H}_\omega$ yields $\Pi^{\mathcal{H}}_\omega \nabla L(\theta) = \nabla L_\omega(\theta_\omega)$. Thus the ω-block of the gradient depends on θ_ω alone.

Step 2. By Theorem C.1, decoupling provides maps $\Phi_{s,\omega}, \Psi_{s,\omega}$ with $\Pi^S_\omega \circ \Phi_s = \Phi_{s,\omega} \circ q_\omega$ and $\Pi^{\mathcal{H}}_\omega \circ \Psi_s = \Psi_{s,\omega} \circ q_\omega$. Evaluated along the trajectory, these identities read $h_{s+1,\omega} = \Phi_{s,\omega}(h_{s,\omega}, g_{s,\omega})$ and $u_{s,\omega} = \Psi_{s,\omega}(h_{s+1,\omega}, g_{s,\omega})$.

Step 3. Write $g_\omega(\theta_\omega) = \nabla L_\omega(\theta_\omega)$, which by Step 1 equals $g_{s,\omega}$ when $\theta = \theta_s$, and define

$$T_{s,\omega}(h_\omega, \theta_\omega) = \Big(h^+_\omega,\ \theta_\omega - \eta_s \Psi_{s,\omega}\big(h^+_\omega, g_\omega(\theta_\omega)\big)\Big), \qquad h^+_\omega = \Phi_{s,\omega}\big(h_\omega, g_\omega(\theta_\omega)\big).$$

By Step 2 and the update $\theta_{s+1} = \theta_s - \eta_s u_s$ we obtain $(h_{s+1,\omega}, \theta_{s+1,\omega}) = T_{s,\omega}(h_{s,\omega}, \theta_{s,\omega})$, and iterating from $s = 0$ to $t - 1$ proves the first claim. Each $T_{s,\omega}$ acts on its own factor $S_\omega \times \mathcal{H}_\omega$, so the global one-step map is the product $T_s = \prod_\omega T_{s,\omega}$. Since a composition of product maps is the product of the compositions, the last claim follows. □

Theorem 4.3. *An optimizer is frequency decoupled if and only if for all t, ω*

$$\exists\, \Phi_{t,\omega} :\ \Pi^S_\omega \Phi_t(h, g) = \Phi_{t,\omega}(h_\omega, g_\omega); \tag{$*$}$$

$$g_\omega = \tilde{g}_\omega \ \Rightarrow\ \Pi^{\mathcal{H}}_\omega \Psi_t(h, g) = \Pi^{\mathcal{H}}_\omega \Psi_t(h, \tilde{g}); \tag{$**$}$$

$$h_\omega = \tilde{h}_\omega \ \Rightarrow\ \Pi^{\mathcal{H}}_\omega \Psi_t(h, g) = \Pi^{\mathcal{H}}_\omega \Psi_t(\tilde{h}, g). \tag{$***$}$$

Proof. Assume first that the optimizer is decoupled, with $\Phi_{t,\omega}, \Psi_{t,\omega}$ as in Definition 4.1. Condition $(*)$ is the first half of the definition. For $(**)$, let $g_\omega = \tilde{g}_\omega$. Then,

$$\Pi^{\mathcal{H}}_\omega \Psi_t(h, g) = \Psi_{t,\omega}(h_\omega, g_\omega) = \Psi_{t,\omega}(h_\omega, \tilde{g}_\omega) = \Pi^{\mathcal{H}}_\omega \Psi_t(h, \tilde{g}),$$

and $(***)$ follows analogously from $h_\omega = \tilde{h}_\omega$.

Conversely, assume $(*)$–$(***)$. Condition $(*)$ supplies the state maps $\Phi_{t,\omega}$. It remains to construct $\Psi_{t,\omega}$. Given $(a, b) \in S_\omega \times \mathcal{H}_\omega$, choose any (h, g) with $h_\omega = a$ and $g_\omega = b$ and set $\Psi_{t,\omega}(a, b) = \Pi^{\mathcal{H}}_\omega \Psi_t(h, g)$. It is well defined, since if $(\tilde{h}, \tilde{g})$ is another such pair, $(**)$ gives $\Pi^{\mathcal{H}}_\omega \Psi_t(h, g) = \Pi^{\mathcal{H}}_\omega \Psi_t(h, \tilde{g})$ and $(***)$ gives $\Pi^{\mathcal{H}}_\omega \Psi_t(h, \tilde{g}) = \Pi^{\mathcal{H}}_\omega \Psi_t(\tilde{h}, \tilde{g})$. By construction $(\Psi_t(h, g))_\omega = \Psi_{t,\omega}(h_\omega, g_\omega)$ for all (h, g), which corresponds to the second half of Definition 4.1. □

D PROOFS FOR SECTION 4.2

Theorem 4.4 (Generalized criterion). *Assume that the carrier is local,*

$$Z_t(h,g) = \bigoplus_{\nu\in\Omega} Z_{t,\nu}(h_\nu, g_\nu), \tag{Z}$$

and let $P_{t,\omega\nu}(h)\colon \mathcal{H}_\nu \to \mathcal{H}_\omega$ be the Fourier blocks of $P_t(h)$. For $a \in S_\nu$ set

$$\mathcal{D}_{t,\nu}(a) = \operatorname{span}\{Z_{t,\nu}(a,\xi) - Z_{t,\nu}(a,\widetilde{\xi}) : \xi,\widetilde{\xi} \in \mathcal{H}_\nu\}, \qquad b^0_{t,\omega}(h) = \sum_{\nu\in\Omega} P_{t,\omega\nu}(h) Z_{t,\nu}(h_\nu, 0).$$

The optimizer is frequency decoupled if and only if (∗) holds and, for all t, ω,

$$P_{t,\omega\nu}(h)\,\mathcal{D}_{t,\nu}(h_\nu) = \{0\} \qquad \forall h \in S,\ \forall \nu \neq \omega, \tag{I}$$

and, for all $h, \widetilde{h} \in S$ with $h_\omega = \widetilde{h}_\omega$,

$$\big(P_{t,\omega}(h) - P_{t,\omega}(\widetilde{h})\big)\mathcal{D}_{t,\omega}(h_\omega) = \{0\}, \tag{II}$$

$$b^0_{t,\omega}(h) = b^0_{t,\omega}(\widetilde{h}). \tag{III}$$

Proof. By Theorem 4.3 it suffices to show that, under (Z), condition (∗∗) is equivalent to (I), and that, given (I), condition (∗∗∗) is equivalent to (II) together with (III).

Step 1. Put $\delta_{t,\nu}(a,\xi) = Z_{t,\nu}(a,\xi) - Z_{t,\nu}(a,0)$, the variation of the carrier relative to its base value. Therefore, $\delta_{t,\nu}(a,0) = 0$ and $\operatorname{span}\{\delta_{t,\nu}(a,\xi) : \xi \in \mathcal{H}_\nu\} = \mathcal{D}_{t,\nu}(a)$, as each δ is a difference of values of $Z_{t,\nu}(a,\cdot)$, and conversely each such difference equals $\delta_{t,\nu}(a,\xi) - \delta_{t,\nu}(a,\tilde{\xi})$. Writing $Z_{t,\nu}(h_\nu, g_\nu) = Z_{t,\nu}(h_\nu, 0) + \delta_{t,\nu}(h_\nu, g_\nu)$ in (Z) and applying $\Pi^{\mathcal{H}}_\omega P_t(h)$ yields

$$\Pi^{\mathcal{H}}_\omega \Psi_t(h,g) = b^0_{t,\omega}(h) + \sum_{\nu\in\Omega} P_{t,\omega\nu}(h)\,\delta_{t,\nu}(h_\nu, g_\nu)\,: \tag{11}$$

a base part depending on h alone, plus a sum of terms each of which depends on a single g_ν.

Step 2. Fix h. By (11), the ω-block of the update is independent of g_ν for every $\nu \neq \omega$ if and only if each term $P_{t,\omega\nu}(h)\delta_{t,\nu}(h_\nu, g_\nu)$ is constant in g_ν. Since this term vanishes at $g_\nu = 0$, constancy means $P_{t,\omega\nu}(h)\delta_{t,\nu}(h_\nu,\xi) = 0$ for all ξ, i.e. $P_{t,\omega\nu}(h)$ annihilates $\mathcal{D}_{t,\nu}(h_\nu)$. This is (I), while independence of the ω-block from the other blocks of g is exactly (∗∗).

Step 3. Under (I) the sum in (11) reduces to its $\nu = \omega$ term,

$$\Pi^{\mathcal{H}}_\omega \Psi_t(h,g) = b^0_{t,\omega}(h) + P_{t,\omega}(h)\,\delta_{t,\omega}(h_\omega, g_\omega).$$

Let $h_\omega = \widetilde{h}_\omega$. Condition (∗∗∗) asserts that the right-hand side is the same for h and $\widetilde{h}$, for every g_ω. Evaluating at $g_\omega = 0$, where the variation term vanishes, gives $b^0_{t,\omega}(h) = b^0_{t,\omega}(\widetilde{h})$, which is (III). Subtracting this identity leaves $\big(P_{t,\omega}(h) - P_{t,\omega}(\widetilde{h})\big)\delta_{t,\omega}(h_\omega, g_\omega) = 0$ for all g_ω, i.e. the operator annihilates $\mathcal{D}_{t,\omega}(h_\omega)$, which is (II). Conversely, (II) and (III) make both summands agree for h and $\widetilde{h}$, which is (∗∗∗).

Combining the three steps with Theorem 4.3 proves the criterion. □

Although the following theorems can be obtained as corollaries from the theorem above we provide their independent proofs from the first principles for the sake of better understanding and interpretability.

Theorem D.1 (Sufficient condition). *Suppose there are $\Phi_{t,\omega}, P_{t,\omega}, Z_{t,\omega}$ with*

$$\Pi^S_\omega \Phi_t(h,g) = \Phi_{t,\omega}(h_\omega, g_\omega), \tag{∗}$$

$$Z_t(h,g) = \bigoplus_{\nu\in\Omega} Z_{t,\nu}(h_\nu, g_\nu), \tag{Z}$$

$$P_t(h) = \bigoplus_{\omega\in\Omega} P_{t,\omega}(h_\omega). \tag{LS}$$

Then, the optimizer is frequency decoupled.

Proof. Condition (∗) provides the state maps $\Phi_{t,\omega}$ required by Definition 4.1. We construct the update maps. Set $\Psi_{t,\omega}(a,b) = P_{t,\omega}(a)Z_{t,\omega}(a,b)$. Since $P_t(h)$ is the direct sum of the blocks $P_{t,\nu}(h_\nu)$ by (LS), and $Z_t(h,g)$ is the direct sum of the vectors $Z_{t,\nu}(h_\nu, g_\nu)$ by (Z), the ω-component of their product is the product of the ω-components:

$$\Pi^{\mathcal{H}}_{\omega}\Psi_t(h,g) = \Pi^{\mathcal{H}}_{\omega}\big(P_t(h)Z_t(h,g)\big) = P_{t,\omega}(h_\omega)Z_{t,\omega}(h_\omega,g_\omega) = \Psi_{t,\omega}(h_\omega,g_\omega)$$

for all (h,g). Both requirements of Definition 4.1 are therefore satisfied. □

Theorem D.2 (Necessary condition)**.** *Assume* (Z) *and the spanning condition*

$$\operatorname{span}\big\{Z_{t,\omega}(h_\omega,g) - Z_{t,\omega}(h_\omega,\tilde{g}) : g,\tilde{g}\in\mathcal{H}_\omega\big\} = \mathcal{H}_\omega \qquad \forall\,\omega,h_\omega. \tag{S}$$

If the optimizer is decoupled, then (LS) *holds.*

Proof. By Theorem 4.3, decoupling implies (∗∗) and (∗∗∗). We use the former to show that $P_t(h)$ is block diagonal, and the latter to show that its diagonal block depends on h_ω alone. Throughout, $i_\nu\colon \mathcal{H}_\nu \to \mathcal{H}$ denotes the inclusion and $P_{t,\omega\nu}(h) = \Pi^{\mathcal{H}}_{\omega}P_t(h)i_\nu$ the (ω,ν) Fourier block.

Step 1. Fix h and $\nu \neq \omega$, and let $g,\tilde{g}$ agree on every frequency except ν. In particular $g_\omega = \tilde{g}_\omega$, so (∗∗) gives $\Pi^{\mathcal{H}}_{\omega}\Psi_t(h,g) = \Pi^{\mathcal{H}}_{\omega}\Psi_t(h,\tilde{g})$, that is, $\Pi^{\mathcal{H}}_{\omega}P_t(h)\big(Z_t(h,g) - Z_t(h,\tilde{g})\big) = 0$. By (Z) the two carriers differ only in their ν-component,

$$Z_t(h,g) - Z_t(h,\tilde{g}) = i_\nu\big(Z_{t,\nu}(h_\nu,g_\nu) - Z_{t,\nu}(h_\nu,\tilde{g}_\nu)\big),$$

so the previous identity reads $P_{t,\omega\nu}(h)\big(Z_{t,\nu}(h_\nu,g_\nu) - Z_{t,\nu}(h_\nu,\tilde{g}_\nu)\big) = 0$. As $g_\nu,\tilde{g}_\nu$ range over $\mathcal{H}_\nu$, these differences span $\mathcal{H}_\nu$ by (S). A linear operator that vanishes on a spanning set is zero, so $P_{t,\omega\nu}(h) = 0$. Thus $P_t(h)$ is block diagonal.

Step 2. Let $h,\tilde{h}$ satisfy $h_\omega = \tilde{h}_\omega$. By Step 1, $\Pi^{\mathcal{H}}_{\omega}\Psi_t(h,g) = P_{t,\omega}(h)Z_{t,\omega}(h_\omega,g_\omega)$, and the same holds for $\tilde{h}$ with the same carrier, because the carrier depends on h only through $h_\omega = \tilde{h}_\omega$. Condition (∗∗∗) therefore gives $\big(P_{t,\omega}(h) - P_{t,\omega}(\tilde{h})\big)Z_{t,\omega}(h_\omega,g_\omega) = 0$ for every g_ω. Taking differences over $g_\omega,\tilde{g}_\omega$ and invoking (S) once more, the operator $P_{t,\omega}(h) - P_{t,\omega}(\tilde{h})$ vanishes on a spanning set of $\mathcal{H}_\omega$ and is therefore zero. Thus $P_{t,\omega}(h)$ depends on h_ω only, which together with Step 1 is (LS). □

Proposition D.3. (LS) *holds if and only if*

$$[P_t(h),\rho(s)] = 0 \qquad \forall h,s; \tag{a}$$

$$h_\omega = \tilde{h}_\omega \;\Rightarrow\; P_{t,\omega}(h) = P_{t,\omega}(\tilde{h}), \tag{b}$$

where $P_{t,\omega}(h)$ is the diagonal Fourier block of $P_t(h)$.

Proof. Suppose (LS). Thus, $P_t(h)$ is block diagonal in the Fourier basis, so (a) follows from Theorem 3.3 ($b \Rightarrow c$), and the diagonal block equals $P_{t,\omega}(h_\omega)$, which is (b).

Conversely, suppose (a) and (b). By Theorem 3.3 ($c \Rightarrow b$), condition (a) makes $P_t(h)$ block diagonal, $P_t(h) = \bigoplus_\omega P_{t,\omega}(h)$. By (b) the block $P_{t,\omega}(h)$ depends on h only through h_ω, so $\mathbb{P}_{t,\omega}(a) := P_{t,\omega}(h)$ for any h with $h_\omega = a$ is well defined, and $P_t(h) = \bigoplus_\omega \mathbb{P}_{t,\omega}(h_\omega)$ is (LS). □

Corollary D.4 (Preconditioner-level equivalence)**.** *Under* (Z) *and* (S)*, frequency decoupling is equivalent to* (∗) *together with* (LS)*; and* (LS) *holds if and only if* $[P_t(h),\rho(s)] = 0$ *for all* h,s *and the diagonal block* $P_{t,\omega}(h)$ *depends on* h_ω *alone (Proposition D.3).*

Proof. The first assertion combines Theorems D.1 and D.2. The second is Proposition D.3. □

E Proofs for Section 4.3

Proposition 4.5. *SGD, SGD with momentum and Accelerated SGD are frequency decoupled.*

Table 7: Summary of Section 4.3.

Method	Decoupled	Coupling or decoupling mechanism
SGD, momentum, Nesterov	yes	$P_t = I$
Scalar weight decay	yes	absorbed into the gradient
Per-frequency clipping	yes	local scalar factor
Global norm clipping	no	global scalar normalization
Adam (spatial)	no	diagonal in the spatial, not the Fourier, basis (Lemma B.2)
Matrix AdaGrad / full-matrix Adam	no	global scalar normalization
Muon (matrix)	no	global scalar normalization
Muon (tensor, unfolded)	no	unfolding, see (4)

Proof. We verify the hypotheses of Theorem D.1 with $S_\omega = \mathcal{H}_\omega$, i.e. with the local state $h_\omega = m_\omega$. The recursions of SGD in the general form of Section 4.3 act componentwise on the frequency blocks, so

$$\Phi_{t,\omega}(m_\omega, g_\omega) = \beta_t m_\omega + (1-\beta_t) g_\omega, \qquad Z_{t,\omega}(m_\omega, g_\omega) = a_t m_\omega + b_t g_\omega, \qquad P_t = \bigoplus_\omega I_{\mathcal{H}_\omega}$$

satisfy (∗) and (Z) directly, and (LS) holds because P_t is the identity, whose blocks depend on no state at all. Theorem D.1 yields the claim. □

Proposition E.1. *With at least two nonzero frequency subspaces, global norm clipping* $u(g) = \min\{1, c/\|g\|\} g$ *for* $g \neq 0$, $u(0) = 0$, $c > 0$, *is in general not frequency decoupled, whereas* per-frequency clipping $u_\omega = \min\{1, c_\omega/\|g_\omega\|\} g_\omega$ *with fixed* $c_\omega > 0$ *is (Proposition E.2).*

Proof. The strategy is to embed clipping into an auxiliary optimizer whose carrier satisfies the spanning condition (S), so that Theorem D.2 applies and forces the preconditioner to be local. Consequently, the explicit form of the preconditioner contradicts locality.

Step 1. Discard the zero frequency subspaces. For each ω fix a unit vector $e_\omega \in \mathcal{H}_\omega$ and let $V_\omega(\xi) = \xi/\|\xi\|$ for $\xi \neq 0$ and $V_\omega(0) = e_\omega$, so that V_ω maps $\mathcal{H}_\omega$ onto its unit sphere. Consider the algorithm with state $r \in (0,\infty)^\Omega$, input $\xi \in \mathcal{H}$ and

$$\Phi^{\mathrm{aux}}(r, \xi) = r, \qquad Z(r, \xi) = \bigoplus_\omega r_\omega V_\omega(\xi_\omega), \qquad \Psi^{\mathrm{aux}}(r, \xi) = u\big(Z(r, \xi)\big),$$

where u is the clipping map. Since $\|Z(r,\xi)\| = \|r\|_2$, we may write $\Psi^{\mathrm{aux}}(r,\xi) = P(r) Z(r,\xi)$ with $P(r) = \min\{1, c/\|r\|_2\}\, I$.

Step 2. Suppose, for contradiction, that clipping is decoupled. The state map Φ^{aux} is the identity and the carrier Z acts frequency by frequency, so Ψ^{aux} is the composition of a local map with the decoupled map u. The three conditions of Theorem 4.3 are inherited, and the auxiliary algorithm is decoupled as well.

Step 3. The carrier satisfies (Z) by construction. For fixed $r_\omega > 0$ the differences $Z_\omega(r_\omega, \xi) - Z_\omega(r_\omega, \tilde{\xi}) = r_\omega(v - \tilde{v})$, with $v, \tilde{v}$ ranging over the unit sphere of $\mathcal{H}_\omega$, span $\mathcal{H}_\omega$, since for any $x \neq 0$, $x = \frac{\|x\|}{2}\big(v - (-v)\big)$ with $v = x/\|x\|$. This is (S).

Step 4. Theorem D.2 now forces (LS), so the diagonal block $P_\omega(r)$ must depend on r_ω alone. But for $\|r\|_2 > c$ we have $P_\omega(r) = c\big(\sum_\nu r_\nu^2\big)^{-1/2} I_{\mathcal{H}_\omega}$, which changes when some r_ν with $\nu \neq \omega$ is varied at fixed r_ω (possible because $|\Omega| \geq 2$). Hence clipping is not decoupled. □

Proposition E.2. *Per-frequency clipping* $u_\omega = \min\{1, c_\omega/\|g_\omega\|\} g_\omega$, *with fixed* $c_\omega > 0$ *and* $u_\omega = 0$ *when* $g_\omega = 0$, *is frequency decoupled.*

Proof. The algorithm is stateless, so we may take S_ω to be a single point. Conditions (∗) and (∗∗∗) of Theorem 4.3 are then vacuous. The ω-block of the direction, u_ω, is computed from g_ω alone, so it is unchanged when the other blocks of g are modified, which is (∗∗). Theorem 4.3 gives the claim. □

The obstruction in the proof is the dependence of a *shared scalar factor* on the norms of all frequency blocks. We distinguish it into a separate proposition, since it recurs for matrix AdaGrad, for Muon on a kernel matrix and for Muon on an unfolded kernel.

Proposition E.3 (Global scalar normalization). *Let $|\Omega| \geq 2$, $\mathcal{H}_\omega \neq \{0\}$ for all ω, and $a : (0,\infty) \to \mathbb{R}$. The stateless algorithm with direction $T_a(x) = a(\|x\|)x$ for $x \neq 0$ and $T_a(0) = 0$ is frequency decoupled with respect to $\mathcal{H} = \bigoplus_\omega \mathcal{H}_\omega$ if and only if a is constant.*

Proof. Sufficiency. If $a \equiv a_0$, then $T_a(x) = a_0 x$. Taking $S_\omega = \{0\}$, $Z_\omega(x) = x_\omega$ and $P_\omega = a_0 I_{\mathcal{H}_\omega}$, the hypotheses of Theorem D.1 hold trivially, so T_a is decoupled.

Necessity. Suppose T_a is decoupled. We reuse the auxiliary construction of the proof of Proposition E.1. With the same maps V_ω, state $r \in (0,\infty)^\Omega$ and carrier $Z(r,\xi) = \bigoplus_\omega r_\omega V_\omega(\xi_\omega)$, set $\Psi^{\mathrm{aux}}(r,\xi) = T_a(Z(r,\xi)) = P_a(r)Z(r,\xi)$ with $P_a(r) = a(\|r\|_2)\, I$. As before, the auxiliary algorithm is decoupled, and its carrier satisfies (Z) and (S). Corollary D.4 therefore yields (LS), and by Proposition D.3 the diagonal block depends on r_ω alone: there are functions b_ω with

$$a(\|r\|_2) = b_\omega(r_\omega) \qquad \text{for all } \omega \in \Omega \text{ and all } r \in (0,\infty)^\Omega.$$

Choose $\omega \neq \nu$, which is possible because $|\Omega| \geq 2$. Then, $b_\omega(r_\omega) = b_\nu(r_\nu)$ for all $r_\omega, r_\nu > 0$. Varying r_ν at fixed r_ω shows that b_ν is constant, hence so is b_ω, and therefore $r \mapsto a(\|r\|_2)$ is constant on $(0,\infty)^\Omega$. Since $r \mapsto \|r\|_2$ maps $(0,\infty)^\Omega$ onto $(0,\infty)$, the function a is constant. □

Proposition 4.6. *Adam in spatial coordinates is not frequency decoupled.*

Proof. We apply Theorem 4.4 and exhibit a violation of (III).

Step 1. With $h_\omega = (m_\omega, v_\omega)$ the carrier is $Z_{t,\omega}(h_\omega, g_\omega) = \kappa_t m_\omega$, where $\kappa_t = (1-\beta_1^{t+1})^{-1} > 0$. It satisfies (Z) and does not depend on g_ω, so $\mathcal{D}_{t,\nu}(h_\nu) = \{0\}$ for every ν and the base part is

$$b^0_{t,\omega}(m,v) = \kappa_t \sum_{\nu\in\Omega} P_{t,\omega\nu}(v)\, m_\nu, \qquad P_t(v) = \mathrm{Diag}\big((\sqrt{v}+\varepsilon)^{-1}\big).$$

Step 2. Fix t and a positive second moment v such that, for at least one channel a, the map $r \mapsto \bar{v}[r,a]$ is not constant. Subsequently, the diagonal entries of $P_t(v)$ are not constant along r in that channel, so $P_t(v) \notin \mathrm{Comm}(\rho(\Gamma))$ by Lemma B.2, and by Theorem 3.3 it is not block diagonal in the Fourier basis, so there exist $\omega \neq \nu$ and $z \in \mathcal{H}_\nu$ with $P_{t,\omega\nu}(v)z \neq 0$.

Step 3. Let $h = (0, v)$ and $\widetilde{h} = (\widetilde{m}, v)$ with $\widetilde{m}_\nu = z$ and $\widetilde{m}_\mu = 0$ for $\mu \neq \nu$. The two states share the second moment and the ω-component of the first moment (both zero), so $h_\omega = \widetilde{h}_\omega$, but by Step 1,

$$b^0_{t,\omega}(\widetilde{h}) - b^0_{t,\omega}(h) = \kappa_t P_{t,\omega\nu}(v)\, z \neq 0.$$

Condition (III) fails, and Theorem 4.4 shows that Adam is not decoupled. □

Matrix AdaGrad (Duchi et al., 2011) accumulates the full second moment of the vectorized gradient. Adding momentum and allowing both accumulating and exponentially averaged states gives

$$m_{t+1} = \beta_1 m_t + (1-\beta_1)g_t, \qquad H_{t+1} = a_t H_t + b_t g_t g_t^*, \qquad u_t = (\bar{H}_{t+1}^{1/2} + \varepsilon I)^{-1}\bar{m}_{t+1}, \quad (12)$$

with $m_0 = 0$, $H_0 = 0$, bias-corrected $\bar{m}_{t+1}$ and $\bar{H}_{t+1} = H_{t+1}/c_{t+1}$, and pre-specified $a_t \geq 0$, $b_t > 0$, $c_{t+1} > 0$. Taking $a_t = b_t = c_{t+1} = 1$ gives matrix AdaGrad with momentum, and $a_t = \beta_2$, $b_t = 1-\beta_2$, $c_{t+1} = 1-\beta_2^{t+1}$ a full-matrix Adam-like method using $g_t g_t^*$ instead of the diagonal state.

Proposition E.4. *With at least two nonzero frequency subspaces matrix AdaGrad with momentum (12) is not decoupled.*

Proof. It suffices to examine the first step. With $m_0 = 0$ and $H_0 = 0$ the recursions give $m_1 = (1-\beta_1)g$, hence $\bar{m}_1 = g$, and $H_1 = b_0 gg^*$, hence $\bar{H}_1 = (b_0/c_1)\, gg^*$. The positive square root of this rank-one matrix is $\bar{H}_1^{1/2} = \sqrt{b_0/c_1}\, gg^*/\|g\|$, which acts on g as multiplication by $\sqrt{b_0/c_1}\,\|g\|$, so

$$u_0(g) = (\bar{H}_1^{1/2} + \varepsilon I)^{-1}\bar{m}_1 = \frac{g}{\sqrt{b_0/c_1}\,\|g\| + \varepsilon} \quad (g \neq 0), \qquad u_0(0) = 0.$$

Suppose the method were decoupled. By Theorem 4.3, at the fixed initial state $h_0 = 0$ the map $g \mapsto \Psi_0(\Phi_0(h_0, g), g) = u_0(g)$ would be local frequency by frequency. But $u_0 = T_\alpha$ is the global scalar normalization with $\alpha(s) = (\sqrt{b_0/c_1}\, s + \varepsilon)^{-1}$, which is not constant, and by Proposition E.3 such a map is not decoupled when at least two frequency subspaces are nonzero. This contradiction proves the claim. □

Proposition 4.7. *Muon on a kernel matrix of size $N_1 \times N_2$ with $N_1 N_2 > 1$ is not decoupled.*

Proof. Step 1. Let $M \in \mathbb{C}^{N_1 \times N_2}$ be the first moment and F_1, F_2 the unitary DFT matrices of the two spatial axes, so that $\widehat{M} = F_1 M F_2^T$ collects the Fourier coefficients of the momentum. By unitary equivariance of the polar map (Appendix B.2, property (iii)), $\widehat{D} = F_1 \operatorname{msign}(M) F_2^T = \operatorname{msign}(\widehat{M})$. Muon acts on Fourier coefficients by the same rule as on spatial entries, and we work with $\widehat{M}$ throughout. In the notation of Section 4.2 the state is $h = \widehat{M}$, the carrier is $Z_t(h, G) = \widehat{M}$, and $P_t(h)[X] = ((\widehat{M}\widehat{M}^*)^{1/2})^+ X$. The carrier does not depend on the gradient, so $\mathcal{D}_{t,\omega} = \{0\}$ and the base part is $b^0_{t,\omega}(h) = [\operatorname{msign}(\widehat{M})]_\omega$, the ω-th Fourier coefficient of the direction. By Theorem 4.4 it suffices to violate (III), i.e. to exhibit two momenta with the same ω-th coefficient whose directions differ at ω.

Step 2. We may assume $N_1 \geq 2$, since otherwise $N_2 \geq 2$ and we transpose, using $\operatorname{msign}(M^T) = \operatorname{msign}(M)^T$ (Appendix B.2, property (v)). Let e_0 be the first standard basis vector of $\mathbb{R}^{N_2}$ and consider momenta of the form $\widehat{M} = a e_0^T$ with $a \in \mathbb{C}^{N_1}$. They form the direct sum of the frequency subspaces indexed by $(j, 0)$, $j = 0, \ldots, N_1 - 1$. For $a \neq 0$ the matrix $a e_0^T$ has rank one, with compact SVD $(a/\|a\|_2) \cdot \|a\|_2 \cdot e_0^T$, so that

$$\operatorname{msign}(a e_0^T) = \frac{a}{\|a\|_2}\, e_0^T, \qquad b^0_{t,(j,0)}(h) = \frac{a_j}{\|a\|_2}.$$

Thus on this subspace the polar map coincides with the global scalar normalization $T_\alpha(a) = \alpha(\|a\|_2)\, a$ with $\alpha(s) = 1/s$. Concretely, changing a coefficient a_k with $k \neq j$ changes $\|a\|_2$ and hence $b^0_{t,(j,0)}$, while the local state $h_{(j,0)} = a_j$ is unchanged. This violates (III), in agreement with Proposition E.3. □

Derivation of (4). Right multiplication by the unitary DFT matrix turns the unfolding $C_{\text{out}} \times (C_{\text{in}} N)$ into $[\widehat{M}_{\omega_1} \cdots \widehat{M}_{\omega_N}]$. Unitary equivariance and the identity $\operatorname{msign}(A) = ((AA^*)^{1/2})^+ A$ (Appendix B.2, property (i)) give (4). Taking $C_{\text{out}} = 1$ and $N \geq 2$ reduces it to $\widehat{D}_\omega = \widehat{M}_\omega / \sqrt{\sum_\nu \|\widehat{M}_\nu\|_F^2}$, the global scalar normalization excluded by Proposition E.3. □

F Proofs for Section 4.4

Theorem 4.8. *For any source optimizer, construction (6) is frequency decoupled on the state space $\prod_\omega S^{\text{src}}$.*

Proof. We verify the three conditions of Theorem 4.3 for the state space $\prod_\omega S^{\text{src}}$, whose ω-component is the replica ξ_ω. For $a \in S^{\text{src}}$ and $b \in \mathbb{C}^{q_\omega}$ define $\Phi^{\text{loc}}_{t,\omega}(a, b) = \Phi_t(a, E_\omega b)$ and $\Psi^{\text{loc}}_{t,\omega}(a, b) = E_\omega^* \Psi_t(a, E_\omega b)$. By (6) the ω-component of the new state is $\xi_{t+1,\omega} = \Phi^{\text{loc}}_{t,\omega}(\xi_{t,\omega}, g_{t,\omega})$, a function of the ω-components of the state and of the gradient only. This is (∗). Likewise the ω-block of the direction is $u_{t,\omega} = \Psi^{\text{loc}}_{t,\omega}(\xi_{t+1,\omega}, g_{t,\omega})$, which depends neither on g_ν nor on ξ_ν for $\nu \neq \omega$. This gives (∗∗) and (∗∗∗). Theorem 4.3 concludes. No property of the source optimizer has been used. □

Theorem 4.9. *Construction (7) is frequency decoupled on the state space $\prod_\omega S_\omega$.*

Proof. It suffices to show that the directions of construction (7) coincide with those of construction (6). Construction (7) is then construction (6) with the ω-th replica stored as $\xi_{t,\omega} = J_\omega a_{t,\omega}$, and it computes the ω-th state and direction from $a_{t,\omega}$ and $g_{t,\omega}$ alone, so the three conditions of Theorem 4.3 hold exactly as in the proof of Theorem 4.8.

Fix ω and a sequence of gradients. We show by induction on t that $\xi_{t,\omega} = J_\omega a_{t,\omega}$. At $t = 0$ both sides equal h_0^{src}. If $\xi_{t,\omega} = J_\omega a_{t,\omega}$, the first identity in (8) gives $\xi_{t+1,\omega} = \Phi_t(J_\omega a_{t,\omega}, E_\omega g_{t,\omega}) = J_\omega \phi_{t,\omega}(a_{t,\omega}, g_{t,\omega}) = J_\omega a_{t+1,\omega}$, and the second gives $E_\omega^* \Psi_t(\xi_{t+1,\omega}, E_\omega g_{t,\omega}) = \psi_{t,\omega}(a_{t+1,\omega}, g_{t,\omega}) = u_{t,\omega}$. Hence the two constructions produce the same directions. □

G Derivation of the localized methods

Here we derive the methods of Table 1 using Theorem 4.8. For each subsection we first state the source method, compute its replica (6) in closed form and, therefore, read off a per-frequency algorithm. Throughout, $g_{t,\omega} = E_\omega^* g_t$ are the Fourier blocks of the gradient and the direction is assembled as $u_t = \sum_\omega E_\omega u_{t,\omega}$, i.e. by an inverse DFT.

G.1 Derivation scheme

Every subsection below follows the same scheme. It starts from a source optimizer (Φ_t, Ψ_t), whose replica localization (6) is frequency decoupled by Theorem 4.8 but stores a full copy of the source state for every frequency. A proposition then gives the compact form of this replica for a family J_ω, chosen here by inspection (Appendix G.7 shows how to compute it). Step 1 of its proof substitutes $J_\omega a$ and the input $E_\omega b$ into the state recursion of the source, which verifies the first identity in (8) and defines $\phi_{t,\omega}$. Step 2 substitutes them into the direction of the source and defines $\psi_{t,\omega}$, using the functional calculus under an isometric embedding (Lemma B.4) and the identity $\mathrm{msign}(QAV^*) = Q\,\mathrm{msign}(A)V^*$ for isometries Q, V (Appendix B.2, property (iii)). Step 3 concludes that (8) holds, so by the proof of Theorem 4.9 the replica localization coincides with construction (7) with these maps and is frequency decoupled by Theorem 4.9. Further claims of a proposition are proved in additional steps. The method is finally defined by the algorithm that runs this construction.

G.2 Fourier-Adam

Source. Adam in spatial coordinates (Section 4.3) on $\mathcal{H} = \mathbb{C}^{\Gamma \times q}$ with $N = |\Gamma|$, hyperparameters $\beta_1, \beta_2, \varepsilon^{\mathrm{src}}$ and step sizes η_t^{src}. For the spatial DFT, $(E_\omega x)[r] = \varphi_\omega(r)x$ with $|\varphi_\omega(r)| = N^{-1/2}$, so the replica input is $(E_\omega b)[r] = \varphi_\omega(r)b$. We write $\mathbf{1} \in \mathbb{R}^\Gamma$ for the all-ones vector.

Algorithm 1 Fourier-Adam

given step sizes $\eta_t > 0$, $\beta_1, \beta_2 \in [0, 1)$, $\varepsilon > 0$
initialize $t \leftarrow 0$, $\theta_0 \in \mathcal{H}$, $m_{0,\omega} \leftarrow 0$ and $v_{0,\omega} \leftarrow 0$ for all $\omega \in \Omega$
repeat
  $g_t \leftarrow \nabla L(\theta_t)$ ▷ stochastic gradient
  $g_{t,\omega} \leftarrow E_\omega^* g_t$ for all $\omega \in \Omega$ ▷ DFT along the spatial axes
  for each $\omega \in \Omega$ **do**
    $m_{t+1,\omega} \leftarrow \beta_1 m_{t,\omega} + (1 - \beta_1) g_{t,\omega}$
    $v_{t+1,\omega} \leftarrow \beta_2 v_{t,\omega} + (1 - \beta_2)\,|g_{t,\omega}| \odot |g_{t,\omega}|$
    $\bar{m} \leftarrow m_{t+1,\omega}/(1 - \beta_1^{t+1})$, $\bar{v} \leftarrow v_{t+1,\omega}/(1 - \beta_2^{t+1})$
    $u_{t,\omega} \leftarrow \bar{m}/(\sqrt{\bar{v}} + \varepsilon)$ ▷ component-wise in the internal coordinates
  end for
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \sum_\omega E_\omega u_{t,\omega}$ ▷ inverse DFT
  $t \leftarrow t + 1$
until *stopping criterion is met*
return θ_t

Proposition G.1. *The replica localization of Adam admits the compact state* $(m_\omega, v_\omega) \in \mathbb{C}^q \times \mathbb{R}^q_{\geq 0}$ *with*

$$J_\omega(m, v) = \big(E_\omega m,\ N^{-1}\mathbf{1} \otimes v\big), \qquad \textit{i.e.} \qquad m^{(\omega)}[r] = \varphi_\omega(r)m, \quad v^{(\omega)}[r] = N^{-1}v, \tag{13}$$

and is given by $m_{0,\omega} = v_{0,\omega} = 0$ *and*

$$m_{t+1,\omega} = \beta_1 m_{t,\omega} + (1-\beta_1) g_{t,\omega}, \qquad v_{t+1,\omega} = \beta_2 v_{t,\omega} + (1-\beta_2)|g_{t,\omega}| \odot |g_{t,\omega}|,$$
$$u_{t,\omega} = \frac{\sqrt{N}\,\bar{m}_{t+1,\omega}}{\sqrt{\bar{v}_{t+1,\omega}} + \sqrt{N}\,\varepsilon^{\mathrm{src}}},$$

where bars denote the bias corrections of Adam and the operations in $u_{t,\omega}$ *are component-wise. For* $\varepsilon^{\mathrm{src}} = \varepsilon/\sqrt{N}$ *and* $\eta_t^{\mathrm{src}} = \eta_t/\sqrt{N}$ *it is Algorithm 1.*

Proof. Step 1. Substitute $J_\omega(m, v)$ and the input $E_\omega b$ into the recursions of Adam. The first moment is linear in its input, so $\beta_1\varphi_\omega(r)m + (1-\beta_1)\varphi_\omega(r)b = \varphi_\omega(r)\big(\beta_1 m + (1-\beta_1)b\big)$. The second moment receives $|(E_\omega b)[r]| \odot |(E_\omega b)[r]| = N^{-1}|b| \odot |b|$, because $|\varphi_\omega(r)|^2 = N^{-1}$, so $\beta_2 N^{-1} v + (1-\beta_2)N^{-1}|b|\odot|b| = N^{-1}\big(\beta_2 v + (1-\beta_2)|b|\odot|b|\big)$. Hence $\Phi_t(J_\omega(m,v), E_\omega b) = J_\omega \phi_{t,\omega}(m, v, b)$, where $\phi_{t,\omega}$ consists of the two recursions of the statement, and $J_\omega(0,0) = 0$ is the initial state of Adam.

Step 2. The bias corrections act on both sides by the same scalars. At position r the direction of the replica is therefore $\varphi_\omega(r)\,\bar{m}/(N^{-1/2}\sqrt{\bar{v}} + \varepsilon^{\mathrm{src}})$, component-wise in the internal coordinates. Applying E_ω^*, which sums against $\overline{\varphi_\omega(r)}$ over r with $\sum_r |\varphi_\omega(r)|^2 = 1$, gives

$$\psi_{t,\omega}(m, v, b) = \frac{\bar{m}}{N^{-1/2}\sqrt{\bar{v}} + \varepsilon^{\mathrm{src}}} = \frac{\sqrt{N}\,\bar{m}}{\sqrt{\bar{v}} + \sqrt{N}\,\varepsilon^{\mathrm{src}}}.$$

Step 3. Steps 1 and 2 are (8), so the replica localization is construction (7) with these maps (Appendix G.1). For $\varepsilon^{\mathrm{src}} = \varepsilon/\sqrt{N}$ its direction is $\sqrt{N}\,\bar{m}/(\sqrt{\bar{v}}+\varepsilon)$, and with $\eta_t^{\mathrm{src}} = \eta_t/\sqrt{N}$ the update $\theta_{t+1} = \theta_t - \eta_t^{\mathrm{src}} \sum_\omega E_\omega \psi_{t,\omega}$ is that of Algorithm 1. □

G.3 CommAdam and per-frequency matrix AdaGrad

Source. Method (12) of Appendix E, with state (m, H) and direction $(\bar{H}^{1/2} + \varepsilon I)^{-1}\bar{m}$.

Algorithm 2 CommAdam and per-frequency matrix AdaGrad

given step sizes $\eta_t > 0$, $\beta_1 \in [0,1)$, $\varepsilon > 0$ and coefficients $a_t \geq 0$, $b_t > 0$, $c_{t+1} > 0$
  CommAdam: $a_t = \beta_2$, $b_t = 1-\beta_2$, $c_{t+1} = 1 - \beta_2^{t+1}$ with $\beta_2 \in [0,1)$
  per-frequency matrix AdaGrad: $a_t = b_t = c_{t+1} = 1$
initialize $t \leftarrow 0$, $\theta_0 \in \mathcal{H}$, $m_{0,\omega} \leftarrow 0$ and $C_{0,\omega} \leftarrow 0$ for all $\omega \in \Omega$
repeat
  $g_t \leftarrow \nabla L(\theta_t)$ ▷ stochastic gradient
  $g_{t,\omega} \leftarrow E_\omega^* g_t$ for all $\omega \in \Omega$ ▷ DFT
  for each $\omega \in \Omega$ **do**
    $m_{t+1,\omega} \leftarrow \beta_1 m_{t,\omega} + (1-\beta_1) g_{t,\omega}$
    $C_{t+1,\omega} \leftarrow a_t C_{t,\omega} + b_t\, g_{t,\omega} g_{t,\omega}^*$
    $\bar{m} \leftarrow m_{t+1,\omega}/(1-\beta_1^{t+1})$, $\bar{C} \leftarrow C_{t+1,\omega}/c_{t+1}$
    $u_{t,\omega} \leftarrow (\bar{C}^{1/2} + \varepsilon I)^{-1}\bar{m}$ ▷ positive square root
  end for
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \sum_\omega E_\omega u_{t,\omega}$ ▷ inverse DFT
  $t \leftarrow t+1$
until *stopping criterion is met*
return θ_t

Proposition G.2. *The replica localization of* (12) *admits the compact state* $(m_\omega, C_\omega) \in \mathbb{C}^{q_\omega} \times \mathbb{C}^{q_\omega \times q_\omega}$ *with*

$$J_\omega(m, C) = \big(E_\omega m,\ E_\omega C E_\omega^*\big), \tag{14}$$

and is given by $m_{0,\omega} = 0$, $C_{0,\omega} = 0$ *and*

$$m_{t+1,\omega} = \beta_1 m_{t,\omega} + (1-\beta_1) g_{t,\omega}, \qquad C_{t+1,\omega} = a_t C_{t,\omega} + b_t g_{t,\omega} g_{t,\omega}^*,$$
$$u_{t,\omega} = (\bar{C}_{t+1,\omega}^{1/2} + \varepsilon I)^{-1} \bar{m}_{t+1,\omega},$$

with $\bar{m}_{t+1,\omega} = m_{t+1,\omega}/(1-\beta_1^{t+1})$ *and* $\bar{C}_{t+1,\omega} = C_{t+1,\omega}/c_{t+1}$. *Moreover* $C_{t,\omega} \succeq 0$ *for all* t. *This is Algorithm 2.*

Proof. Step 1. The momentum recursion is linear, so $\beta_1 E_\omega m + (1-\beta_1)E_\omega b = E_\omega\big(\beta_1 m + (1-\beta_1)b\big)$. For the state component H, $(E_\omega b)(E_\omega b)^* = E_\omega b b^* E_\omega^*$, hence $a_t E_\omega C E_\omega^* + b_t (E_\omega b)(E_\omega b)^* = E_\omega\big(a_t C + b_t b b^*\big)E_\omega^*$. This is the first identity in (8) with $\phi_{t,\omega}$ given by the recursions of the statement, and $J_\omega(0,0) = 0$. Since $a_t \geq 0$ and $b_t > 0$, the recursion keeps $C_{t,\omega} \succeq 0$.

Step 2. Let $f_\varepsilon(x) = (x^{1/2}+\varepsilon)^{-1}$ on $[0,\infty)$. The source direction at the state $J_\omega(m,C)$ is $f_\varepsilon(E_\omega \bar{C} E_\omega^*) E_\omega \bar{m}$, which equals $E_\omega f_\varepsilon(\bar{C})\bar{m}$ by (10). Applying E_ω^* gives $\psi_{t,\omega} = (\bar{C}^{1/2}+\varepsilon I)^{-1}\bar{m}$.

Step 3. Steps 1 and 2 are (8), so the replica localization is construction (7) with these maps (Appendix G.1), which is Algorithm 2. □

G.4 Per-frequency Shampoo

Source. Matrix Shampoo (Gupta et al., 2018) on the standard unfolding of the kernel, of size $a \times (bN)$, with factors $L \leftarrow \beta_2 L + (1-\beta_2)GG^*$ and $R \leftarrow \beta_2 R + (1-\beta_2)G^*G$, momentum $M \leftarrow \beta_1 M + (1-\beta_1)G$ and direction $(\bar{L}+\delta I)^{-1/4}\bar{M}(\bar{R}+\delta I)^{-1/4}$ with $\delta > 0$. Right multiplication of the unfolding by a unitary matrix F maps $M \mapsto MF$ and $R \mapsto F^*RF$, leaves L unchanged and maps the direction $D \mapsto DF$. We may therefore use Fourier coordinates for the columns, in which the gradient reads $[G_{\omega_1} \ \cdots \ G_{\omega_N}]$ with $G_\omega \in \mathbb{C}^{a\times b}$. Let $V_\omega \colon \mathbb{C}^b \to \mathbb{C}^{bN}$ place a vector into the ω-th block of b coordinates. Then $V_\omega^* V_\omega = I_b$, and for $G \in \mathbb{C}^{a\times b}$ the replica input GV_ω^* has G as its ω-th block of columns and zeros elsewhere.

Algorithm 3 Per-frequency Shampoo

given step sizes $\eta_t > 0$, $\beta_1, \beta_2 \in [0,1)$, $\delta > 0$
  accumulating source: L and R are sums of GG^* and G^*G, without bias correction
initialize $t \leftarrow 0$, $\theta_0 \in \mathcal{H}$, $M_{0,\omega} \leftarrow 0 \in \mathbb{C}^{a\times b}$, $L_{0,\omega} \leftarrow 0 \in \mathbb{C}^{a\times a}$ and $R_{0,\omega} \leftarrow 0 \in \mathbb{C}^{b\times b}$ for all $\omega \in \Omega$
repeat
  $g_t \leftarrow \nabla L(\theta_t)$ ▷ stochastic gradient
  $G_{t,\omega} \leftarrow \mathrm{mat}(E_\omega^* g_t) \in \mathbb{C}^{a\times b}$ for all $\omega \in \Omega$ ▷ DFT
  for each $\omega \in \Omega$ **do**
    $M_{t+1,\omega} \leftarrow \beta_1 M_{t,\omega} + (1-\beta_1)G_{t,\omega}$
    $L_{t+1,\omega} \leftarrow \beta_2 L_{t,\omega} + (1-\beta_2)G_{t,\omega}G_{t,\omega}^*$
    $R_{t+1,\omega} \leftarrow \beta_2 R_{t,\omega} + (1-\beta_2)G_{t,\omega}^* G_{t,\omega}$
    $\bar{M} \leftarrow M_{t+1,\omega}/(1-\beta_1^{t+1})$
    $L_\delta \leftarrow L_{t+1,\omega}/(1-\beta_2^{t+1}) + \delta I_a$, $R_\delta \leftarrow R_{t+1,\omega}/(1-\beta_2^{t+1}) + \delta I_b$
    $D_{t,\omega} \leftarrow L_\delta^{-1/4}\,\bar{M}\,R_\delta^{-1/4}$
  end for
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \sum_\omega E_\omega \,\mathrm{vec}(D_{t,\omega})$ ▷ inverse DFT
  $t \leftarrow t+1$
until *stopping criterion is met*
return θ_t

Proposition G.3. *The replica localization of Shampoo admits the compact state* $(M_\omega, L_\omega, R_\omega)$ *with* $J_\omega(M,L,R) = (MV_\omega^*,\ L,\ V_\omega R V_\omega^*)$ *and is given by*

$$\begin{aligned} L_{t+1,\omega} &= \beta_2 L_{t,\omega} + (1-\beta_2)G_{t,\omega}G_{t,\omega}^*, \quad R_{t+1,\omega} = \beta_2 R_{t,\omega} + (1-\beta_2)G_{t,\omega}^* G_{t,\omega}, \\ M_{t+1,\omega} &= \beta_1 M_{t,\omega} + (1-\beta_1)G_{t,\omega}, \end{aligned} \tag{15}$$

all initialized at zero, and by the direction

$$D_{t,\omega} = L_{\delta,\omega}^{-1/4}\,\bar{M}_{t+1,\omega}\,R_{\delta,\omega}^{-1/4}, \qquad L_{\delta,\omega} = \bar{L}_{t+1,\omega} + \delta I_a, \quad R_{\delta,\omega} = \bar{R}_{t+1,\omega} + \delta I_b, \tag{16}$$

i.e. by Algorithm 3. Under column-major vectorization its local preconditioner is $P_{t,\omega} = (R_{\delta,\omega}^{-1/4})^T \otimes L_{\delta,\omega}^{-1/4} = (R_{\delta,\omega}^T \otimes L_{\delta,\omega})^{-1/4}$.

Proof. *Step 1.* With input GV_ω^* the three recursions of the source give $\beta_1 MV_\omega^* + (1-\beta_1)GV_\omega^* = \big(\beta_1 M + (1-\beta_1)G\big)V_\omega^*$, $\beta_2 L + (1-\beta_2)(GV_\omega^*)(GV_\omega^*)^* = \beta_2 L + (1-\beta_2)GG^*$ and $\beta_2 V_\omega R V_\omega^* + (1-\beta_2)V_\omega G^* G V_\omega^* = V_\omega\big(\beta_2 R + (1-\beta_2)G^*G\big)V_\omega^*$. This is the first identity in (8) with $\phi_{t,\omega}$ given by (15). The bias corrections act on both sides by the same scalars.

Step 2. By (10) with $f(x) = (x+\delta)^{-1/4}$, $(V_\omega \bar{R}_\omega V_\omega^* + \delta I)^{-1/4} V_\omega = V_\omega(\bar{R}_\omega + \delta I_b)^{-1/4}$, and taking adjoints, $V_\omega^*(V_\omega \bar{R}_\omega V_\omega^* + \delta I)^{-1/4} = (\bar{R}_\omega + \delta I_b)^{-1/4} V_\omega^*$. The direction of the replica is therefore $L_{\delta,\omega}^{-1/4}\bar{M}_\omega V_\omega^*(V_\omega \bar{R}_\omega V_\omega^* + \delta I)^{-1/4} = L_{\delta,\omega}^{-1/4}\bar{M}_\omega R_{\delta,\omega}^{-1/4} V_\omega^*$, and extracting the ω-th block of columns gives (16).

Step 3. Steps 1 and 2 are (8), so the replica localization is construction (7) with these maps (Appendix G.1), which is Algorithm 3.

Step 4. The formula for $P_{t,\omega}$ follows from Lemma B.5 with $p = -1/4$. □

G.5 CommMuon

Source. Muon with the exact polar factor, momentum $M \leftarrow \beta_t M + (1-\beta_t)G$ and direction $\mathrm{msign}(M)$, in two settings. The first is the kernel matrix of a two-dimensional single-channel kernel of size $N_1 \times N_2$, and the second is the standard unfolding $a \times (bN)$ of a multi-channel kernel.

Algorithm 4 CommMuon

given step sizes $\eta_t > 0$ and momentum coefficients $\beta_t \in [0,1)$
initialize $t \leftarrow 0$, $\theta_0 \in \mathcal{H}$, $M_{0,\omega} \leftarrow 0 \in \mathbb{C}^{a\times b}$ for all $\omega \in \Omega$
repeat
  $g_t \leftarrow \nabla L(\theta_t)$ ▷ stochastic gradient
  $G_{t,\omega} \leftarrow \mathrm{mat}(E_\omega^* g_t) \in \mathbb{C}^{a\times b}$ for all $\omega \in \Omega$ ▷ DFT
  for each $\omega \in \Omega$ **do**
    $M_{t+1,\omega} \leftarrow \beta_t M_{t,\omega} + (1-\beta_t)G_{t,\omega}$
    $D_{t,\omega} \leftarrow \mathrm{msign}(M_{t+1,\omega})$ ▷ $m/|m|$ if $a = b = 1$
  end for
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \sum_\omega E_\omega \,\mathrm{vec}(D_{t,\omega})$ ▷ inverse DFT
  $t \leftarrow t+1$
until *stopping criterion is met*
return θ_t

Proposition G.4. (i) *For the two-dimensional kernel, let $\omega = (i,j)$ and let $f_i \in \mathbb{C}^{N_1}$, $f_j \in \mathbb{C}^{N_2}$ be the unit Fourier vectors with $E_\omega m = m\, f_i f_j^T$. The replica localization admits the compact state $m_\omega \in \mathbb{C}$ with $J_\omega(m) = m\, f_i f_j^T$ and is given by $m_{t+1,\omega} = \beta_t m_{t,\omega} + (1-\beta_t)g_{t,\omega}$ and $u_{t,\omega} = m_{t+1,\omega}/|m_{t+1,\omega}|$, with $u_{t,\omega} = 0$ if $m_{t+1,\omega} = 0$.* (ii) *For the standard unfolding, in Fourier coordinates for the columns and with V_ω as in the previous subsection, the replica localization admits the compact state $M_\omega \in \mathbb{C}^{a\times b}$ with $J_\omega(M) = MV_\omega^*$ and is given by $M_{t+1,\omega} = \beta_t M_{t,\omega} + (1-\beta_t)G_{t,\omega}$ and $D_{t,\omega} = \mathrm{msign}(M_{t+1,\omega})$. In both cases this is Algorithm 4, case (i) being $a = b = 1$.*

Proof. *Step 1.* In both cases the momentum recursion is linear and the replica input lies in the image of J_ω, so $\beta_t J_\omega(M) + (1-\beta_t)J_\omega(G) = J_\omega\big(\beta_t M + (1-\beta_t)G\big)$. This is the first identity in (8), with zero initial momentum.

Step 2. In case (i), for $m \neq 0$ the rank-one matrix $m f_i f_j^T$ has the compact SVD $f_i\,|m|\,v^*$ with $v = (\overline{m}/|m|)\,\overline{f_j}$, so $\mathrm{msign}(m f_i f_j^T) = f_i v^* = (m/|m|)\, f_i f_j^T = J_\omega(m/|m|)$, and its coefficient on the ω-th harmonic, $m/|m|$, is $\psi_{t,\omega}$. In case (ii), Fourier coordinates for the columns are admissible because $\mathrm{msign}(AF) = \mathrm{msign}(A)F$ for unitary F. By Appendix B.2, property (iii), $\mathrm{msign}(MV_\omega^*) = \mathrm{msign}(M)V_\omega^*$, and extracting the ω-th block of columns gives $\psi_{t,\omega} = \mathrm{msign}(M)$.

Step 3. In both cases Steps 1 and 2 are (8), so the replica localization is construction (7) with these maps (Appendix G.1), which is Algorithm 4. □

G.6 Localized Unfolding Muon

Source. Muon with momentum $M \leftarrow \beta_t M + (1-\beta_t)g$, a finite family of complex-linear isometric unfoldings of the whole parameter, $\mathcal{R}_\tau : \mathcal{H} \to \mathbb{C}^{m_\tau \times n_\tau}$ with $m_\tau n_\tau = \dim_{\mathbb{C}} \mathcal{H}$, $\tau \in \mathcal{T}$, and a selection rule T_t. Its direction is $\mathcal{R}_\tau^{-1}\,\mathrm{msign}(\mathcal{R}_\tau M)$ with $\tau = T_t(M)$. Put

$$\mathcal{A}_{\omega,\tau} = \mathcal{R}_\tau E_\omega, \qquad \mathcal{A}^*_{\omega,\tau} = E^*_\omega \mathcal{R}_\tau^{-1}, \qquad \mathcal{A}^*_{\omega,\tau}\mathcal{A}_{\omega,\tau} = I. \tag{17}$$

The map $\mathcal{A}_{\omega,\tau}$ is an isometric embedding but in general not a bijection onto the whole matrix space, so the inverse step uses the adjoint. In contrast with unfolding inside a Fourier block, the source unfolding may mix spatial with internal coordinates. The frequency is isolated first, the original matrix transform is applied afterwards, and the same output frequency is extracted.

Algorithm 5 Localized Unfolding Muon (LU-Muon)

given step sizes $\eta_t > 0$, momentum coefficients $\beta_t \in [0,1)$, unfoldings $\mathcal{R}_\tau$, $\tau \in \mathcal{T}$, and a selection rule T_t of the source with a tie-breaking rule fixed in advance
initialize $t \leftarrow 0$, $\theta_0 \in \mathcal{H}$, $M_{0,\omega} \leftarrow 0 \in \mathbb{C}^{q_\omega}$ for all $\omega \in \Omega$
repeat
  $g_t \leftarrow \nabla L(\theta_t)$ ▷ stochastic gradient
  $G_{t,\omega} \leftarrow E^*_\omega g_t$ for all $\omega \in \Omega$ ▷ DFT
  for each $\omega \in \Omega$ **do**
    $M_{t+1,\omega} \leftarrow \beta_t M_{t,\omega} + (1-\beta_t) G_{t,\omega}$
    $\tau \leftarrow T_t(E_\omega M_{t+1,\omega})$ ▷ local rule $T_{t,\omega}$
    $B \leftarrow \mathcal{R}_\tau(E_\omega M_{t+1,\omega})$ ▷ $B = \mathcal{A}_{\omega,\tau} M_{t+1,\omega}$
    $D_{t,\omega} \leftarrow E^*_\omega \mathcal{R}_\tau^{-1}\big(\mathrm{msign}(B)\big)$ ▷ $D_{t,\omega} = \mathcal{A}^*_{\omega,\tau}\,\mathrm{msign}(B)$
  end for
  $\theta_{t+1} \leftarrow \theta_t - \eta_t \sum_\omega E_\omega D_{t,\omega}$ ▷ inverse DFT
  $t \leftarrow t+1$
until *stopping criterion is met*
return θ_t

Proposition G.5. *The replica localization of this Muon admits the compact state $M_\omega \in \mathbb{C}^{q_\omega}$ with $J_\omega(M) = E_\omega M$ and is given by $M_{0,\omega} = 0$ and*

$$M_{t+1,\omega} = \beta_t M_{t,\omega} + (1-\beta_t) G_{t,\omega}, \qquad \tau_{t,\omega} = T_{t,\omega}(M_{t+1,\omega}),$$
$$B_{t,\omega} = \mathcal{A}_{\omega,\tau_{t,\omega}} M_{t+1,\omega}, \qquad D_{t,\omega} = \mathcal{A}^*_{\omega,\tau_{t,\omega}}\,\mathrm{msign}(B_{t,\omega}),$$

where $T_{t,\omega}(X) = T_t(E_\omega X)$. This is Algorithm 5. Its direction is $D_{t,\omega} = P_{t,\omega}(M_{t+1,\omega}) M_{t+1,\omega}$ with a positive semidefinite linear operator $P_{t,\omega}$ that depends on the local state only.

Proof. Step 1. The momentum recursion is linear and the input $E_\omega b$ lies in the image of J_ω, so $\Phi_t(E_\omega M, E_\omega b) = E_\omega\big(\beta_t M + (1-\beta_t) b\big)$.

Step 2. In the replica the selected unfolding is $\tau = T_t(E_\omega M) = T_{t,\omega}(M)$, and the source direction is $\mathcal{R}_\tau^{-1}\,\mathrm{msign}(\mathcal{R}_\tau E_\omega M)$. Since $\mathcal{R}_\tau$ is an isometric bijection, $\mathcal{R}_\tau^{-1} = \mathcal{R}_\tau^*$ and $E^*_\omega \mathcal{R}_\tau^{-1} = \mathcal{A}^*_{\omega,\tau}$, so the projected direction is $\psi_{t,\omega} = \mathcal{A}^*_{\omega,\tau}\,\mathrm{msign}(\mathcal{A}_{\omega,\tau} M)$.

Step 3. Steps 1 and 2 are (8), so the replica localization is construction (7) with these maps (Appendix G.1), which is Algorithm 5.

Step 4. Put $B = \mathcal{A}_{\omega,\tau} M$ and $K = ((BB^*)^{1/2})^+ \succeq 0$. By Appendix B.2, property (i), $\mathrm{msign}(B) = KB$, so $D_{t,\omega} = P_{t,\omega}(M) M$ with $P_{t,\omega}(M)[X] = \mathcal{A}^*_{\omega,\tau}\big(K \mathcal{A}_{\omega,\tau} X\big)$. This operator depends on the local state M only, and $\langle X, P_{t,\omega}(M) X\rangle = \mathrm{tr}\big((\mathcal{A}_{\omega,\tau}X)^* K (\mathcal{A}_{\omega,\tau}X)\big) \geq 0$. This exhibits (LS) directly, so LU-Muon also falls under Theorem D.1. □

The selection rule is evaluated on the momentum of one block at a time. For instance, the Tensorion criterion (Bogachev et al., 2026) becomes $T_{t,\omega}(X) \in \arg\max_\tau \|\mathcal{A}_{\omega,\tau} X\|_*$. Choosing a single shared unfolding from the data of *all* frequencies is outside the construction. The method reduces to CommMuon when the unfolding groups whole spatial and whole internal axes (Proposition G.6).

Storing $\mathcal{A}_{\omega,\tau}$ explicitly is unnecessary, since it is a DFT followed by a coordinate permutation, but an arbitrary unfolding may still require polarizing a large matrix for every frequency. The following condition reduces the computation to the small internal block.

Proposition G.6. *Suppose that for every candidate there are isometries $Q_{\omega,\tau}$, $V_{\omega,\tau}$ and an isometric bijection $\mathcal{U}_{\omega,\tau} : \mathbb{C}^{q_\omega} \to \mathbb{C}^{a_{\omega,\tau} \times b_{\omega,\tau}}$ with*

$$\mathcal{A}_{\omega,\tau} X = Q_{\omega,\tau}\, \mathcal{U}_{\omega,\tau}(X)\, V^*_{\omega,\tau}, \qquad Q^*_{\omega,\tau} Q_{\omega,\tau} = V^*_{\omega,\tau} V_{\omega,\tau} = I. \tag{18}$$

Then the LU-Muon direction is computed inside the frequency block, $D_\omega = \mathcal{U}^{-1}_{\omega,\tau} \operatorname{msign}(\mathcal{U}_{\omega,\tau} M_\omega)$. For the canonical unfolding of the internal coordinates this is CommMuon.

Proof. By Appendix B.2, property (iii), $\operatorname{msign}(QAV^*) = Q \operatorname{msign}(A) V^*$ and $\|QAV^*\|_* = \|A\|_*$ for isometries Q, V. Applying this to $\mathcal{A}_{\omega,\tau} M_\omega = Q_{\omega,\tau} \mathcal{U}_{\omega,\tau}(M_\omega) V^*_{\omega,\tau}$ gives $\operatorname{msign}(\mathcal{A}_{\omega,\tau} M_\omega) = Q_{\omega,\tau} \operatorname{msign}(\mathcal{U}_{\omega,\tau} M_\omega) V^*_{\omega,\tau}$. Since the adjoint of (18) is $\mathcal{A}^*_{\omega,\tau} Y = \mathcal{U}^{-1}_{\omega,\tau}\big(Q^*_{\omega,\tau} Y V_{\omega,\tau}\big)$, and $Q^*_{\omega,\tau} Q_{\omega,\tau} = V^*_{\omega,\tau} V_{\omega,\tau} = I$, we obtain $\mathcal{A}^*_{\omega,\tau} \operatorname{msign}(\mathcal{A}_{\omega,\tau} M_\omega) = \mathcal{U}^{-1}_{\omega,\tau} \operatorname{msign}(\mathcal{U}_{\omega,\tau} M_\omega)$. For the canonical $\mathcal{U}_{\omega,\tau}$ this is Algorithm 4. □

For a multidimensional DFT the Fourier vector is a tensor product of one-dimensional Fourier vectors. If the unfolding groups whole spatial and whole internal axes, the spatial factors are absorbed into $Q_{\omega,\tau}$ and $V_{\omega,\tau}$ while the internal axes define $\mathcal{U}_{\omega,\tau}$, so that (18) holds. The equality of nuclear norms in the proof also preserves a maximum-nuclear-norm selection rule.

Finite polar iteration. For computation one may replace msign by the Newton–Schulz map $\mathcal{N}_J$ of (9) (Appendix B.2).

Proposition G.7. *Replacing* msign *by* (9) *in LU-Muon preserves decoupling provided the initial normalization is performed separately for each $B_{t,\omega}$. The same holds for CommMuon.*

Proof. Consider the source method of Proposition G.5 with $\mathcal{N}_J$ in place of msign. Its state recursion is unchanged, so Step 1 of that proof applies. In the replica of frequency ω the matrix passed to $\mathcal{N}_J$ is $\mathcal{R}_\tau(E_\omega M_{t+1,\omega}) = B_{t,\omega}$, so the initial normalization in (9) uses $\|B_{t,\omega}\|_F$ only, and the projected direction is $\mathcal{A}^*_{\omega,\tau} \mathcal{N}_J(B_{t,\omega})$. By Appendix G.1, LU-Muon with $\mathcal{N}_J$ and per-block normalization is the replica localization of this source, and it is decoupled by Theorem 4.9. For both families of Proposition G.4 we have $\|J_\omega(M)\|_F = \|M\|_F$ and $J_\omega(Y) J_\omega(Y)^* J_\omega(Z) = J_\omega(YY^*Z)$, so every iterate of (9) started from $J_\omega(M)$ equals $J_\omega(Y_j)$, where Y_j are the iterates started from M. The same argument therefore applies to CommMuon with the local block $M_{t+1,\omega}$. □

The Nesterov variant is admissible. In the selection rule and in the direction one replaces $M_{t+1,\omega}$ by the local carrier $\beta_t M_{t+1,\omega} + (1-\beta_t) G_{t,\omega}$. For real parameters one takes unfoldings and selection rules compatible with conjugation, $\mathcal{R}_\tau(\overline{X}) = \overline{\mathcal{R}_\tau X}$ and $T_{t,-\omega}(\overline{X}) = T_{t,\omega}(X)$, so that $\mathcal{A}_{-\omega,\tau} \overline{X} = \overline{\mathcal{A}_{\omega,\tau} X}$ and Proposition H.1 applies. For a nuclear-norm selection it suffices to resolve ties in the same order on conjugate frequencies.

G.7 Finding the invariant family

In the derivations above the family J_ω was found by inspection. It can also be computed from the structure of the source method, which turns the derivation into a procedure applicable to other optimizers. The replica input always lies in $\mathcal{H}_\omega$, and the procedure uses two properties of the source.

Symmetry. Suppose that the source is equivariant under Γ and covariant under phases, i.e. there are maps $\tilde{\rho}(s)$ and σ_c on S^{src}, for $s \in \Gamma$ and $c \in \mathbb{C}$ with $|c| = 1$, such that

$$\Phi_t(\tilde{\rho}(s)h, \rho(s)g) = \tilde{\rho}(s)\Phi_t(h,g), \qquad \Phi_t(\sigma_c h, cg) = \sigma_c \Phi_t(h,g), \tag{19}$$

and the initial state is fixed by all of them. For the frequency ω put $\tilde{\rho}_\omega(s) = \sigma_{\overline{\chi_\omega(s)}} \circ \tilde{\rho}(s)$.

Lemma G.8 (Symmetry reduction)**.** *Under* (19) *every replica state $\xi_{t,\omega}$ lies in* $\mathrm{Fix}_\omega = \{h \in S^{\mathrm{src}} : \tilde{\rho}_\omega(s)h = h \ \ \forall s \in \Gamma\}$.

Proof. For $b \in \mathbb{C}^{q_\omega}$ the input satisfies $\rho(s)E_\omega b = \chi_\omega(s)E_\omega b$. Put $c = \overline{\chi_\omega(s)}$. The first identity in (19) and then the second give

$$\tilde{\rho}_\omega(s)\Phi_t(h, E_\omega b) = \sigma_c \Phi_t\big(\tilde{\rho}(s)h, \chi_\omega(s)E_\omega b\big) = \Phi_t\big(\tilde{\rho}_\omega(s)h,\, c\,\chi_\omega(s)E_\omega b\big) = \Phi_t(\tilde{\rho}_\omega(s)h, E_\omega b),$$

because $c\,\chi_\omega(s) = |\chi_\omega(s)|^2 = 1$. Thus $\Phi_t(\cdot, E_\omega b)$ commutes with every $\tilde{\rho}_\omega(s)$ and maps Fix_ω into itself. The initial state lies in Fix_ω, and induction on t gives $\xi_{t,\omega} \in \mathrm{Fix}_\omega$. □

Reachable span. Suppose further that each component of the state follows an affine recursion $h \mapsto \alpha_t h + \zeta_t(g)$ with a scalar α_t, as the exponential averages and accumulations of all methods in this paper do, and that the initial state is zero. Then, by induction on t, the replica state lies in the linear span $\mathcal{K}_\omega$ of the features $\zeta_s(E_\omega b)$ over all s and $b \in \mathbb{C}^{q_\omega}$.

Procedure. (1) Compute Fix_ω from (19) and $\mathcal{K}_\omega$ from the features of the source. (2) Choose J_ω as a linear parametrization of $\mathrm{Fix}_\omega \cap \mathcal{K}_\omega$, which contains every replica state by Lemma G.8 and the reachable-span argument. (3) Obtain $\phi_{t,\omega}$ by writing the state recursion in these coordinates. This is possible because both Fix_ω and $\mathcal{K}_\omega$ are invariant under $h \mapsto \Phi_t(h, E_\omega b)$. (4) Obtain $\psi_{t,\omega} = E_\omega^* \Psi_t(J_\omega a, E_\omega b)$ by substitution, and simplify it with (10) and the isometry identity for msign whenever Ψ_t is built from linear maps, spectral functions of the state and polar factors. Construction (7) with these maps is then the localized algorithm (Appendix G.1).

Table 8 applies the procedure to the methods of this paper. For all of them $\tilde{\rho}(s)$ is the action of the shifts on the respective component of the state, and σ_c multiplies the first moment by c and leaves the second-order components of the state unchanged, since $|cg| \odot |cg| = |g| \odot |g|$ and $(cg)(cg)^* = gg^*$. The symmetry alone does not always suffice. For the matrix state H it only gives $H \in \mathrm{Comm}(\rho(\Gamma))$, i.e. a block-diagonal matrix in the Fourier basis (Theorem 3.3), and the reachable span reduces it to the single block ω. The same holds for the Shampoo factor R. For a stateless source the family is trivial and only step (4) remains. For global norm clipping it gives $E_\omega^* u(E_\omega b) = \min\{1, c/\|b\|\}\, b$, i.e. per-frequency clipping with $c_\omega = c$ (Proposition E.2).

Table 8: The invariant family obtained by the procedure of Appendix G.7. Here $\mathbf{1} \in \mathbb{R}^\Gamma$ is the all-ones vector, and V_ω is the column-block insertion of the Shampoo subsection.

State component	Fix_ω	$\mathcal{K}_\omega$	J_ω
first moment (all methods)	$\mathcal{H}_\omega$	$E_\omega \mathbb{C}^{q_\omega}$	$m \mapsto E_\omega m$
second moment of Adam	shift-invariant vectors	$\mathbf{1} \otimes \mathbb{R}^q$	$v \mapsto N^{-1}\mathbf{1} \otimes v$
matrix state H	$\mathrm{Comm}(\rho(\Gamma))$	$E_\omega \mathbb{C}^{q_\omega \times q_\omega} E_\omega^*$	$C \mapsto E_\omega C E_\omega^*$
Shampoo factor L	$\mathbb{C}^{a\times a}$	$\mathbb{C}^{a\times a}$	$L \mapsto L$
Shampoo factor R	commutant of the column shifts	$V_\omega \mathbb{C}^{b\times b} V_\omega^*$	$R \mapsto V_\omega R V_\omega^*$

Limitations. The replica algorithm (6) is valid for every source, but a small invariant family need not exist. If the source breaks the symmetry (19), for instance Adam with a position-dependent $\beta_2(r)$, then $v^{(\omega)}[r]$ is a separate exponential average at every position, neither Fix_ω nor the reachable span reduces it, and the localized method keeps a full-size second moment for every frequency. Independently of the state, the direction may still require a computation of full size, as for LU-Muon with an arbitrary unfolding, where Proposition G.6 gives a sufficient condition for reducing it to the block.

H On the real parameter case

For real parameters we use the complex continuation of the source recursions and choose $E_{-\omega} = \overline{E_\omega}$ (Section I). The Fourier blocks of a real vector then satisfy $x_{-\omega} = \overline{x_\omega}$, so the real parameter space is the orthogonal direct sum of the real parts of $\mathcal{H}_\omega \oplus \mathcal{H}_{-\omega}$ over the conjugate pairs $\{\omega, -\omega\}$, a self-conjugate frequency forming a pair by itself. On real parameters, frequency decoupling is understood with respect to these pairs.

Proposition H.1. *If the source recursions are compatible with complex conjugation and the initial states are real, then (6) preserves conjugate symmetry and defines a real update. On the real parameter space the independent blocks are the conjugate pairs $\{\omega, -\omega\}$ together with the self-conjugate frequencies.*

Proof. For a real gradient g_t the Fourier blocks satisfy $g_{t,-\omega} = \overline{g_{t,\omega}}$, and by the choice $E_{-\omega} = \overline{E_\omega}$ also $E_{-\omega} g_{t,-\omega} = \overline{E_\omega g_{t,\omega}}$. We show by induction on t that $\xi_{t,-\omega} = \overline{\xi_{t,\omega}}$. At $t = 0$ all replicas start from the same real state. If the identity holds at t, compatibility of Φ_t with conjugation gives

$$\xi_{t+1,-\omega} = \Phi_t\big(\overline{\xi_{t,\omega}}, \overline{E_\omega g_{t,\omega}}\big) = \overline{\Phi_t(\xi_{t,\omega}, E_\omega g_{t,\omega})} = \overline{\xi_{t+1,\omega}}.$$

The same compatibility for Ψ_t then yields $u_{t,-\omega} = \overline{u_{t,\omega}}$, so in $u_t = \sum_\omega E_\omega u_{t,\omega}$ the terms indexed by ω and $-\omega$ are complex conjugates of each other while the self-conjugate terms are real, so u_t is real. Finally, the recursions of the pair $\{\omega, -\omega\}$ involve only $\xi_{\pm\omega}$ and $g_{\pm\omega}$, which is the stated block structure on the real parameter space. □

Proposition H.2. *Frequency decoupling and its violation are preserved for real parameters.* (i) *A frequency decoupled optimizer whose maps are compatible with complex conjugation is decoupled with respect to the conjugate pairs on real parameters. This covers SGD with momentum and Nesterov momentum, per-frequency clipping with $c_{-\omega} = c_\omega$ and the localized methods of Appendix G.* (ii) *If there are at least two nonzero pairs, global norm clipping, the global scalar normalization with a non-constant a, Adam, matrix AdaGrad, Muon on a kernel matrix and Muon on an unfolded kernel are not decoupled with respect to the conjugate pairs on real parameters.*

Proof. Step 1. Let the optimizer be decoupled and compatible with conjugation. By decoupling, the ω-blocks of the new state and of the direction are computed from the ω-blocks of the state and of the gradient, and the same holds for $-\omega$. Compatibility with conjugation keeps real states and gradients real, as in the proof of Proposition H.1. Hence on real parameters every pair is updated from its own components, which is decoupling with respect to the pairs. SGD has real coefficients, per-frequency clipping with $c_{-\omega} = c_\omega$ commutes with conjugation because $\|\overline{g_\omega}\| = \|g_\omega\|$, and for the localized methods the claim is Proposition H.1. Its hypothesis holds for their sources, because conjugation commutes with Hadamard and matrix products, with real functions of Hermitian matrices and with the polar factor (Appendix B.2, property (v)). For LU-Muon it requires unfoldings and selection rules compatible with conjugation (Appendix G).

Step 2. Global norm clipping, the global scalar normalization, the first step of matrix AdaGrad (proof of Proposition E.4) and Muon on an unfolded kernel for momenta supported on a single output channel (as in the derivation of (4)) produce a direction of the form $\alpha(\|x\|)\, x$ with a non-constant α, where x is the gradient or the momentum. Choose s_1, s_2 with $\alpha(s_1) \neq \alpha(s_2)$ and a real x that is nonzero on the pair of ω and has norm less than $\min\{s_1, s_2\}$ there. Completing x on another pair to the total norms s_1 and s_2 gives two real inputs that agree on the pair of ω but whose directions differ on it.

Step 3. For Adam, take $\nu = 0$ in the proof of Proposition 4.6 and a real $z \in \mathcal{H}_0$ that is constant in space and supported on a channel in which $\bar{v}$ is not constant. Then $P_t(v)z$ is real and not constant in space, so $P_{t,\omega 0}(v)z \neq 0$ for some $\omega \neq 0$. The real states $(0, v)$ and (z, v) agree on the pair of ω, whereas by Step 1 of that proof their base parts $b^0_{t,\omega}$ differ by $\kappa_t P_{t,\omega 0}(v)z \neq 0$.

Step 4. For Muon on a kernel matrix, the Fourier coefficients of a real kernel satisfy $\widehat{M}_{-\omega} = \overline{\widehat{M}_\omega}$, and the perturbations used in Step 2 of the proof of Proposition 4.7 must respect this symmetry. Choose a unit vector $v \in \mathbb{C}^{N_1}$ with $v_0 = 0$ and $v_{-j} = \overline{v_j}$, e.g. $v = e_\nu$ for nonzero ν with $\nu = -\nu$ and $v = (e_\nu + e_{-\nu})/\sqrt{2}$ otherwise. For fixed $c > 0$ the family $a(s) = ce_0 + sv$, $s \in \mathbb{R}$, then consists of Fourier coefficients of real kernels. Along it, $\widehat{M}_{(0,0)}(s) = c$ is constant, whereas by Step 2 of that proof, and since $v \perp e_0$, $b^0_{t,(0,0)}(h(s)) = c/\sqrt{c^2 + s^2}$ varies with s. The frequency $(0, 0)$ forms a pair by itself and the perturbation lies on other pairs, so Muon is not decoupled with respect to the pairs either. □

I LIMITATIONS

Real parameters. For real parameters we use the complex continuation of the source recursions (for Adam, the second moment is $|g| \odot |g|$) and choose $E_{-\omega} = \overline{E_\omega}$. If the recursions commute with conjugation and the initial states are real, (6) yields a real update, and the independent blocks are the conjugate pairs $\{\omega, -\omega\}$ and the self-conjugate frequencies (Proposition H.1). Decoupling and its violation are preserved in this setting (Appendix H).

Computational costs. CommMuon and LU-Muon store one momentum block per frequency, exactly as much as Muon, and Fourier-Adam as much as Adam.