OPTIMIZING OPTIMIZERS FOR FAST GRADIENT-BASED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We lay the theoretical foundation for automating optimizer design in gradient-based learning. Based on the greedy principle, we formulate the problem of designing optimizers as maximizing the instantaneous decrease in loss. By treating an optimizer as a function that translates loss gradient signals into parameter motions, the problem reduces to a family of convex optimization problems over the space of optimizers. Solving these problems under various constraints not only recovers a wide range of popular optimizers as closed-form solutions, but also produces the optimal hyperparameters of these optimizers with respect to the problems at hand. This enables a systematic approach to design optimizers and tune their hyperparameters according to the gradient statistics collected from training or validation sets. Furthermore, this optimization of optimization can be performed dynamically during training.

1 INTRODUCTION

We are interested in the problem of designing optimizers that maximize the utility of gradient-based learning for a given task. In gradient-based learning, the objective is to minimize an expected scalar loss $\mathbb{E}[\mathcal{L}(\theta)]$ with respect to parameters $\theta \in \mathbb{R}^d$ using its (negative) gradient $g = -\nabla_\theta \mathcal{L} \in \mathbb{R}^d$. As learning takes time, all the parameters $\theta = \theta(t)$, the loss $\mathcal{L} = \mathcal{L}(\theta(t))$, and the gradients $g = g(t)$ are variables of time t, i.e., the training step. A process of learning manifests as a parameter motion $\dot{\theta}$ driven by the gradient g calculated at each step t.

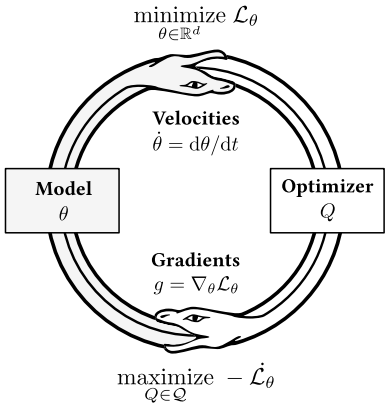


Figure 1: Just as optimizers train their models by feeding them parameter velocities $\dot{\theta}$, models can also fit the optimizers to the underlying tasks by feeding gradients g.

Physics requires a constitutive law that relates kinematic motion to the force field that causes it. For gradient-based learning, optimizers take that role. We can represent an optimizer as a positive semidefinite operator $Q \succeq 0$ that translates the gradient into the parameter update,

$$\dot{\theta} \;=\; Q\, g. \tag{1}$$

By the chain rule, the instantaneous loss drop is then a quadratic form:

$$-\dot{\mathcal{L}} \;=\; \nabla_\theta \mathcal{L}^\top \frac{\mathrm{d}\theta}{\mathrm{d}t} \;=\; g^\top \dot{\theta} \;=\; g^\top Q\, g. \tag{2}$$

The greedy paradigm turns our original problem of maximizing the utility of learning into another optimization problem that maximizes the this loss drop with respect to the optimizer Q:

$$\underset{Q \in \mathcal{Q}}{\text{maximize}}\; \mathbb{E}[g^\top Q\, g] \quad \text{subject to} \quad Q \succeq 0, \tag{P1}$$

where $\mathcal{Q}$ is the design space of allowed optimizers.

Instantaneously, we notice that without any additional constraint, the maximum of the quadratic form $g^\top Q\, g$ is unbounded. Problem P1 reveals two design options for the optimizer that bounds the

maximum: (1) the budget constraint $Q \in \mathcal{Q}$, and (2) the data distribution under the expectation $\mathbb{E}$. Our main focus is on how these two factors determine the *optimal optimizer* $Q^\star$.

Placing the optimizer itself as a subject of another optimization is interesting in several ways:

- Optimizers can be designed with respect to the individual problem (task and data) and the running environment (budget and precision) in a systematic manner.
- Optimizers can be tuned or even be replaced by better ones according to the intermediate probes from either training or validation sets in the middle of training.
- Solving this *meta-optimization* problem in closed-form for a wide range of budgets uncovers the relationship between the optimizers and their underlying principles.
- By reverse engineering commonly used optimizers, we draw the landscape of optimizers that have driven the success of machine learning (Robbins & Monro, 1951; Kingma & Ba, 2015; Loshchilov & Hutter, 2019; Gupta et al., 2018; Martens & Grosse, 2015) into a single picture.

2 Optimal stateless optimizers

Consider the following setup: Let π_{tr} and π_{val} be the training and validation data distributions, respectively. Then, for each training $x_{\text{tr}} \sim \pi_{\text{tr}}$ and validation sample $x_{\text{val}} \sim \pi_{\text{val}}$, the gradients are denoted by $g_{\text{tr}} = \nabla_\theta \mathcal{L}(\theta, x_{\text{tr}})$ and $g_{\text{val}} = \nabla_\theta \mathcal{L}(\theta, x_{\text{val}})$. We define the *moments* as:

$$\Sigma_{\text{tr}} = \mathbb{E}[g_{\text{tr}}\, g_{\text{tr}}^\top], \qquad C = \mathbb{E}[g_{\text{tr}}\, g_{\text{val}}^\top], \qquad \Sigma_{\text{val}} = \tfrac{1}{2}(C + C^\top), \tag{3}$$

where $\mathbb{E}$ denotes expectation over the single or joint distributions of the enclosed gradients. Note that Σ_{tr} and Σ_{val} are symmetric and positive semidefinite (PSD) matrices of shape $d \times d$. For any symmetric PSD $Q \in \mathbb{S}_+^d$ of shape $d \times d$, we define the *learning power* as:

$$P_{\text{tr}}(Q) \;\coloneqq\; \mathbb{E}[g_{\text{tr}}^\top Q\, g_{\text{tr}}] \;=\; \operatorname{Tr}(Q\,\Sigma_{\text{tr}}), \qquad P_{\text{val}}(Q) \;\coloneqq\; \mathbb{E}[g_{\text{tr}}^\top Q\, g_{\text{val}}] \;=\; \operatorname{Tr}(Q\,\Sigma_{\text{val}}). \tag{4}$$

We call $P_{\text{tr}}(Q)$ the *training power* and $P_{\text{val}}(Q)$ the *validation cross-power*. From the chain rule of equation 2, the learning power is equal to the expected instantaneous loss drop: $-\mathbb{E}_{\text{tr}}[\dot{\mathcal{L}}] = \mathbb{E}_{\text{tr}}[g_{\text{tr}}^\top \dot{\theta}_{\text{tr}}] = P_{\text{tr}}(Q)$ and $-\mathbb{E}_{\text{val}}[\dot{\mathcal{L}}] = \mathbb{E}_{\text{val}}[g_{\text{val}}^\top \dot{\theta}_{\text{tr}}] = P_{\text{val}}(Q)$. Problem P1 is therefore rewritten as:

$$\underset{Q \in \mathcal{Q}}{\text{maximize}}\; P_\circ(Q) \;=\; \operatorname{Tr}(Q\,\Sigma_\circ) \quad \text{subject to} \quad Q \succeq 0, \tag{P2}$$

where $\circ \in \{\text{tr}, \text{val}\}$. This is our main optimization problem.

Solving this without any additional constraint, we end up with arbitrarily large eigenvalues for the optimizer Q. This corresponds to arbitrarily large learning rates, which we all know are infeasible in practice. Real problems give us several reasons that makes this "ideal solution" unrealizable: finite precision of our machines, curvature of the loss landscapes, stochastic nature of subset gradients, etc. All of them restrict the ability of gradient estimates g to represent the global geometry of the parameter space. Taking a large step in the parameter space beyond the regions where g remains explainable leads to unexpected, and usually fatal, behaviors.

In other words, the aforementioned restrictions define the feasible set, or the *budget* $\mathcal{Q} \subseteq \mathbb{S}_+^d$, which induces a solution space of finite optimizers. The following theorem formalizes this:

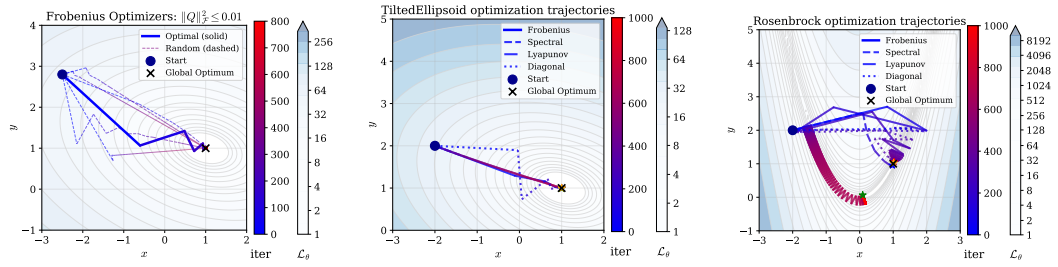
Theorem 1 (Optimal stateless optimizers under convex constraints)**.** *Let the* budget set $\{0\} \subseteq \mathcal{Q} \subseteq \mathbb{S}_+^d$ *be a nonempty, compact, convex set. Define also (1) the* indicator $\delta_{\mathcal{Q}}(Q) = 0$ *if* $Q \in \mathcal{Q}$ *and* $+\infty$ *otherwise, (2) the* gauge *(Minkowski functional):* $\gamma_{\mathcal{Q}}(Q) = \inf\{\lambda > 0 : Q \in \lambda\mathcal{Q}\}$*, and (3) the* polar set $\mathcal{Q}^\circ = \{\Sigma \in \mathbb{S}^d : \sup_{Q \in \mathcal{Q}} \operatorname{Tr}(Q\Sigma) \le 1\}$*. For any symmetric matrix* $\Sigma \in \mathbb{S}^d$*,*

(i) (Existence and sublinearity): *The maximum in* $P^\star(\Sigma) \coloneqq \max_{Q \in \mathcal{Q}} \operatorname{Tr}(Q\Sigma)$ *is attained. In addition,* $P^\star$ *is sublinear (convex and positively homogeneous) and finite everywhere.*

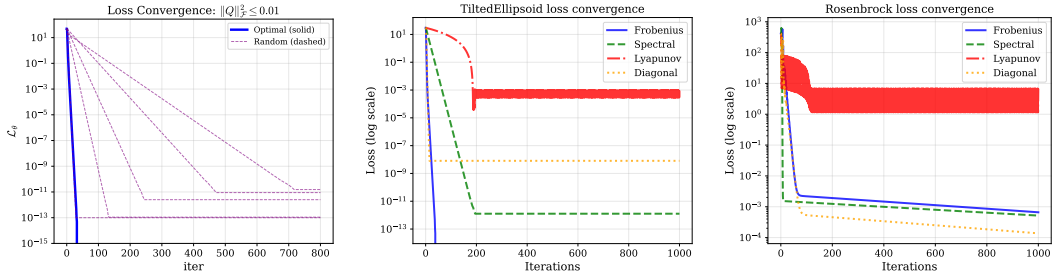
(ii) (Conjugacy identities): *The maximum* $P^\star(\Sigma)$ *is obtained by the identity relationships:*

$$P^\star \;=\; \delta_{\mathcal{Q}}^* \;=\; \gamma_{\mathcal{Q}^\circ} \qquad \text{and} \qquad \gamma_{\mathcal{Q}}^* \;=\; \delta_{\mathcal{Q}^\circ}, \tag{5}$$

i.e., the optimal power is equal to the convex conjugate of the indicator and also the gauge of the polar, while the conjugate of the gauge is equal to the indicator of the polar.



(a) Frobenius-type on elliptic loss (b) Trajectories on elliptic loss (c) Trajectories on Rosenbrock loss



(d) Loss curve of Frobenius-type (e) Loss curve on elliptic loss (f) Loss curve on Rosenbrock loss

Figure 2: Behavior of optimal optimizers under different budget types. (a) Dotted lines show suboptimal optimizers with random Σ from the equal-power Frobenius budget; the straight line is from the optimal hyperparameters from our theory, achieving fastest convergence. (b, c) No free lunch theorem: Frobenius excels in simple quadratic loss, while nonconvex geometry makes spectral and diagonal types better. Each line is the best result from dense search among all budget-parameters, e.g, B for Frobenius budget, etc.

(iii) (Construction): *Any maximizer* $Q^\star \in \arg\max_{Q\in\mathcal{Q}} \mathrm{Tr}(Q\Sigma)$ *is a subgradient of* $P^\star$ *at* Σ*:* $Q^\star \in \partial_\Sigma P^\star(\Sigma)$*. If the maximizer is unique,* $P^\star$ *is differentiable at* Σ *and* $Q^\star = \nabla_\Sigma P^\star(\Sigma)$.

(iv) (Order preservation on $\mathbb{S}_+^d$): *If* $\Sigma \succeq 0$*, then* $P^\star(\Sigma) \geq 0$*. If* $\Sigma_1 \succeq \Sigma_2$*, then* $P^\star(\Sigma_1) \geq P^\star(\Sigma_2)$.

(v) (Lipschitz continuity in symmetrized polar gauge): *Define* $\|\cdot\|_{\mathcal{Q}^\circ}^{\mathrm{sym}} := \max\{\gamma_{\mathcal{Q}^\circ}(\cdot),\, \gamma_{\mathcal{Q}^\circ}(-\cdot)\}$*. For any* $\Sigma, \hat{\Sigma} \in \mathbb{S}^d$*,* $|P^\star(\Sigma) - P^\star(\hat{\Sigma})| \leq \|\Sigma - \hat{\Sigma}\|_{\mathcal{Q}^\circ}^{\mathrm{sym}}$.

The proof is in Appendix B. Items (i), (ii), and (iii) provide a principled way to construct the optimal (stateless) optimizer $Q^\star$ from any given moment Σ and any nicely conditioned budget set $\mathcal{Q}$. Items (iv) and (v) add robustness guarantee and sensitivity analysis to the optimal power. In practice, full gradients rarely appears in large settings. Gradients are drifting throughout non-convex loss landscapes, making true moments hard to obtain. Theorem 1(v) shows that the estimation error in the optimal power is bounded.

Solving the optimization problem constrained by $\mathcal{Q}$ *determines* the optimal optimizer $Q^\star$, and endows the optimizer with different characteristics and algorithmic behaviors. Consider the following four types of budgets:

- *Frobenius ball budget* $\mathcal{Q}_{\mathrm{F}}(B) = \{Q \succeq 0 : \|Q\|_{\mathrm{F}}^2 \leq B\}$ is the simplest constraint that gives an isotropic Euclidean trust region without prior knowledge about parameter space geometry.
- *Spectral budget* $\mathcal{Q}_{\mathrm{S}}(\tau, \lambda) = \{Q \succeq 0 : \mathrm{Tr}(Q) \leq \tau,\ Q \preceq \lambda I\}$ is a budget that upper limits the per-direction spectrum for safety and the trace for total budget simultaneously.
- *Data-metric (Lyapunov) budget* $\mathcal{Q}_{\mathrm{L}}(B) = \{Q \succeq 0 : \mathrm{Tr}(Q^2\Sigma) \leq B\}$ is a budget that uses the data covariance itself as the metric, leading to a natural Lyapunov-like stability condition.
- *Diagonal budget* $\mathcal{Q}_{\mathrm{D}}(B, c) = \{Q = \mathrm{diag}(q_j) \succeq 0 : \sum_j c_j q_j^2 \leq B\}$ is a budget that restricts to coordinate-wise optimizers. Most of the commonly used optimizers fall into this category.

Instantiating the construction from Theorem 1 with these budgets, we obtain corresponding closed-form solutions for the optimal optimizer $Q^\star$ and the optimal power $P^\star$.

Corollary 2 (Closed-form solutions for common budget sets)**.** *Let* $\Sigma = U\,\mathrm{diag}(\sigma_1 \geq \cdots \geq \sigma_d)U^\top$ *be the eigendecomposition. The optimal solutions are:*

(i) (Frobenius ball): $Q^\star_{\mathrm{F}} = \sqrt{B}\,\Sigma/\|\Sigma\|_F, P^\star_{\mathrm{F}}(\Sigma) = \sqrt{B}\,\|\Sigma\|_F$.

(ii) (Spectral): $Q^\star_{\mathrm{S}} = U\,\mathrm{diag}(q^\star_i)U^\top$ *where (i)* $q^\star_i = \lambda$ *for* $i \le k$, *(ii)* $q^\star_{k+1} = \tau - k\lambda$, *(iii)* $q^\star_i = 0$ *for* $i > k+1$, *where* $k = \lfloor \tau/\lambda \rfloor$, $P^\star_{\mathrm{S}}(\Sigma) = \lambda \sum_{i\le k}\sigma_i + (\tau - k\lambda)\sigma_{k+1}$.

(iii) (Data-metric): $Q^\star_{\mathrm{L}} = \alpha\,\Pi_{\mathrm{supp}(\Sigma)}$ *where* $\alpha = (B/\sum_{i:\sigma_i>0}\sigma_i)^{1/2}$, $P^\star_{\mathrm{L}}(\Sigma) = (B\sum_i \sigma_i)^{1/2}$.

(iv) (Diagonal): $[Q^\star_{\mathrm{D}}]_{jj} \propto \Sigma_{jj}/c_j$, $P^\star_{\mathrm{D}}(\Sigma) = (B\sum_j \Sigma^2_{jj}/c_j)^{1/2}$.

Again, the proof is in Appendix B. These analytic solutions reveal how the characteristics of different types of optimal optimizers $Q^\star$ are induced by controlling the budget $\mathcal{Q}$. Specifically, we see that:

Frobenius budget $\leftrightarrow$ Eveidence-proportional optimizer. Budget $\mathcal{Q}_{\mathrm{F}}(B)$ gives optimizers that allocate learning power *proportionally* to data evidence $Q^\star \propto \Sigma$. We can project this general class of optimizers into special geometries to calculate for the optimal hyperparameters as in Corollary 5.

Spectral budget $\leftrightarrow$ Water-filling optimizer $\sim$ gradient clipping & LR scheduling. Budget $\mathcal{Q}_{\mathrm{S}}(\tau, \lambda)$ returns a *water-filling* optimizer that concentrates the learning power into the largest available principal components of the data moment Σ up to a per-mode cap λ, sequentially, until the total budget τ is reached. The spectral bound λ, therefore, acts as a stability margin similar to *gradient clipping tricks*. The trace bound τ controls the total budget like *learning rate schedules*.

Data-metric budget $\leftrightarrow$ Equal-power optimizer $\supset$ {AdaGrad, natural gradient}. Budget $\mathcal{Q}_{\mathrm{L}}(B)$ results in an *equal-power* optimizer that whitens gradient statistics and allocates uniform power across L-eigendirections. If L is Fisher information matrix, this is *natural gradient descent* (Amari, 1998). Generally, this includes full-matrix AdaGrad (Duchi et al., 2011; Agarwal et al., 2019), K-FAC (Martens & Grosse, 2015), and Shampoo (Gupta et al., 2018).

Diagonal budget $\leftrightarrow$ Coordinate-wise optimizer $\supset$ {Adam, GD}. Budget $\mathcal{Q}_{\mathrm{D}}(B, c)$ produces an optimizer that allocates the total budget B *coordinate-wise* which scales with the evidence Σ_{jj} and inversely with the costs c_j. Assuming isotropic moments $\Sigma = \sigma^2 I$ and constant costs $c = \mathbf{1}$, this reduces to $Q^\star = \eta I$ with learning rate $\eta = (B/d)^{1/2}$. This recovers simple *gradient descent*. With variance-based costs $c_j \propto (\mathrm{EMA}(g_j^2))^{1/2}$ and momenta $m_{j,t} = \mathrm{EMA}(g_j)$, this recovers Adam optimizer (Kingma & Ba, 2015); without the first moment it reduces to diagonal AdaGrad (Duchi et al., 2011) or RMSProp-style optimizer (Tieleman & Hinton, 2012). Corollary 6 gives the optimal hyperparameters for Adam from this setup.

The behaviors of different types of optimizers are visualized in Figure 2. Figure 2a shows that our analytically found optimizer is the fastest among all hyperparameter settings under the same Frobenius budget. On the other hand, Figure 2b and 2c highlights how optimizers from different types of budgets can perform better in their specialized domains. This insight helps address the notorious *no free lunch theorem* in optimization (Wolpert & Macready, 1997): The catchphrase "no single algorithm is universally superior" can be updated to "the optimal optimizer $Q^\star$ is a function of the budget $\mathcal{Q}$ and the distribution of observations Σ," at least under our greedy optimization framework. In summary, users choose the budget $\mathcal{Q}$, and the budget defines what is optimal optimizing methods. Reverse engineering under our framework reveals the hidden principles behind the design of commonly used optimizers such as Adam (Kingma & Ba, 2015). Nevertheless, wise readers will notice that an important component is still missing: *momentum*. The next section demonstrates how momentum is integrated into our framework through a straightforward extension.

3 Optimal Dynamic Optimizers with State Variables

Up to this point, we have formulated the problem of finding the optimal stateless optimizer as a convex optimization problem, and derived the closed-form solutions for the four types of budgets. In practice, optimizers have memory, often in the form of momentum, in order to stabilize the learning process from stochastic gradients and non-convex loss landscapes. We now extend our framework by letting the optimizer $Q[n]$ be a *causal dynamical operator*: a *filter* that translates gradient *history* $g[n]$ into *instantaneous* parameter velocity $\dot{\theta}[n]$. In the budget view, this introduces a new degree of freedom: *where in frequency* we allocate our finite learning power.

Let us highlight the key differences from Section 2. We now work in discrete time $n \in \{0, 1, 2, \dots\}$ representing training steps. We will use the z-domain as the primary spectral domain to reflect the iterative nature of practical algorithms. Let $g[n] \in \mathbb{R}^d$ be the (mini-batch) wide-sense stationary (WSS) gradient process. A *dynamic optimizer* is an LTI filter with symmetric *matrix impulse response* $Q[n] \in \mathbb{R}^{d\times d}$ defined by the causal convolution:

$$\dot{\theta}[n] \;=\; (Q * g)[n] \;\coloneqq\; \sum_{k=0}^{\infty} Q[k]\, g[n-k], \qquad \Psi(z) \;\coloneqq\; \sum_{n=0}^{\infty} Q[n]\, z^{-n}, \tag{6}$$

where the *transfer function* is $\Psi(z) \in \mathbb{C}^{d\times d}$. Also adopt the Hilbert norm:

$$\|Q\|_{\mathcal{H}}^2 \;\coloneqq\; \sum_{n=0}^{\infty} \mathrm{Tr}(Q[n]^\top Q[n]) \;<\; \infty, \qquad \langle Q_1, Q_2\rangle_{\mathcal{H}} \;\coloneqq\; \sum_{n=0}^{\infty} \mathrm{Tr}(Q_1[n]^\top Q_2[n]). \tag{7}$$

In this framework, autocorrelation and symmetrized cross-correlation represent the *moments*:

$$R_{\mathrm{tr}}[k] \coloneqq \mathbb{E}[g_{\mathrm{tr}}[n]\, g_{\mathrm{tr}}[n-k]^\top],\; C[k] \coloneqq \mathbb{E}[g_{\mathrm{val}}[n]\, g_{\mathrm{tr}}[n-k]^\top],\; R_{\mathrm{val}}[k] \coloneqq \frac{1}{2}(C[k] + C[k]^\top), \tag{8}$$

where $k \geq 0$. The *instantaneous learning power* is the inner product (more details in Appendix A):

$$P_\circ(Q; n) \coloneqq \mathbb{E}\big[g_\circ[n]^\top \dot{\theta}[n]\big] = \mathbb{E}\left[g_\circ[n]^\top \sum_{k=0}^{\infty} Q[k]\, g_{\mathrm{tr}}[n-k]\right] = \sum_{k=0}^{\infty} \mathrm{Tr}\big(Q[k]^\top R_\circ[k]\big) = \langle Q, R_\circ\rangle_{\mathcal{H}}. \tag{9}$$

Define a nonempty, convex, and weakly compact *budget set* $\{0\} \subseteq \mathcal{Q} \subset \mathcal{H}$, which includes any norm-bounded, closed, convex subset of $\mathcal{H}$. We also define the *indicator* $\delta_{\mathcal{Q}}(Q)$ and the *gauge* $\gamma_{\mathcal{Q}}(Q)$ the same as in Section 2, and the *polar set* is defined as $(\mathcal{Q})^\circ \coloneqq \{R \in \mathcal{H} \mid \sup_{Q\in\mathcal{Q}} \langle Q, R\rangle_{\mathcal{H}} \leq 1\}$. Hence, all the notations are consistent with Section 2.

For dynamic optimization, for each $\circ \in \{\mathrm{tr}, \mathrm{val}\}$, problem P2 is lifted to:

$$\underset{Q\in\mathcal{Q}}{\text{maximize}}\; P_\circ(Q) \;=\; \langle Q, R_\circ\rangle_{\mathcal{H}} \quad \text{subject to} \quad Q \succeq 0, \tag{P3}$$

Unsurprisingly, we arrive at similar results as in Section 2; only the Frobenius inner product $\langle\cdot,\cdot\rangle_F = \mathrm{Tr}(\cdot^\top\cdot)$ is replaced by the Hilbert space inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. We have the same conjugacy toolkit for dynamic optimization.

Theorem 3 (Optimal dynamic optimizers under convex constraints)**.** *Given the definitions above, the followings hold for any nonempty, convex, and weakly compact budget set $\mathcal{Q} \subset \mathcal{H}_+$ with $0 \in \mathcal{Q}$:*

(i) (Existence and sublinearity): *The maximum $P^\star(R) \coloneqq \max_{Q\in\mathcal{Q}} \langle Q, R\rangle_{\mathcal{H}}$ is attained. Moreover, $P^\star$ is sublinear and finite everywhere.*

(ii) (Conjugacy identities): *$P^\star = \delta_{\mathcal{Q}}^* = \gamma_{\mathcal{Q}^\circ}$ and $\gamma_{\mathcal{Q}}^* = \delta_{\mathcal{Q}^\circ}$.*

(iii) (Construction): *Any maximizer $Q^\star \in \arg\max_{Q\in\mathcal{Q}} \langle Q, R\rangle_{\mathcal{H}}$ is a* subgradient *of $P^\star$ at R: $Q^\star \in \partial_R P^\star(R)$. If the maximizer is unique, $P^\star$ is differentiable at R and $\nabla_R P^\star(R) = Q^\star$.*

(iv) (Order preservation on $\mathcal{H}_+$): *If $R \in \mathcal{H}_+$ (Hermitian PSD a.e.), then $P^\star(R) \geq 0$. Moreover, if $R_1 - R_2 \in \mathcal{H}_+ \setminus \{0\}$ and $\exists\, Q \in \mathcal{Q}$ with $\langle Q, R_1 - R_2\rangle_{\mathcal{H}} > 0$ (e.g., if $\mathcal{Q}$ contains a positive definite element), then $P^\star(R_1) > P^\star(R_2)$.*

(v) (Lipschitz continuity in the symmetrized polar gauge): *Define the symmetrized polar gauge $\|u\|_{\mathcal{Q}^\circ}^{\mathrm{sym}} \coloneqq \max\{\gamma_{\mathcal{Q}^\circ}(u),\, \gamma_{\mathcal{Q}^\circ}(-u)\}$. Then $\forall R, \hat{R} \in \mathcal{H}$, $|P^\star(R) - P^\star(\hat{R})| \leq \|R - \hat{R}\|_{\mathcal{Q}^\circ}^{\mathrm{sym}}$.*

The proof is similar to the stateless case, and is provided in Appendix B. Theorem 3 formalizes how the optimal dynamic optimizer $Q^\star$ *equalizes* the learning power *across different frequencies* as a function of the convex budget $\mathcal{Q}$. All the closed-form solutions from Corollary 2 can also be directly lifted to the dynamic framework, as elaborated in Appendix A. Instead of redundantly repeating the closed-form solutions, we discuss how they are connected to well-known optimizers. As we will see in Corollaries 5 and 6, solving Problem P3 using Theorem 3 often produces general dynamic optimizers with infinite impulse responses (IIR) $Q[n]$, whose implementation require infinite memory for optimizer states. In practice, we often restrict ourselves to simpler, realizable family of optimizers, such as ones with EMA-based momenta. The following lemma justifies this post-projection of optimizers already obtaind from convex budgets.

Lemma 4. *Let $\mathcal{H}$ be a real Hilbert space. Given a nonzero moment $R \in \mathcal{H}$, let $\mathcal{Q} \subset \mathcal{H}$ be nonempty, closed, convex, with $0 \in \mathcal{Q}$. Let $\mathcal{C} \subset \mathcal{H}$ be a cone (closed under positive scaling). The normal cone of $\mathcal{Q} \cap \mathcal{C}$ at Q is $N_{\mathcal{Q}\cap\mathcal{C}}(Q) \coloneqq \{M \in \mathcal{H} : \langle M, Q' - Q\rangle_{\mathcal{H}} \le 0 \ \forall Q' \in \mathcal{Q} \cap \mathcal{C}\}$. Define the solution sets of the optimization problem and its restriction to $\mathcal{C}$:*

$$\mathcal{Q}^\star(R) \coloneqq \arg\max_{Q\in\mathcal{Q}} \langle Q, R\rangle_{\mathcal{H}}, \qquad \mathcal{Q}^\star_{\mathcal{C}}(R) \coloneqq \arg\max_{Q\in\mathcal{Q}\cap\mathcal{C}} \langle Q, R\rangle_{\mathcal{H}}. \tag{10}$$

Let $\Pi_{\mathcal{C}}$ be the Hilbert metric projection onto $\mathcal{C}$. For any $Q^\star \in \mathcal{Q}^\star(R)$, the following are equivalent:

(i) (Commutativity) $\Pi_{\mathcal{C}}(Q^\star) \in \mathcal{Q}^\star_{\mathcal{C}}(R)$.

(ii) (Normal-cone alignment) *There exists $Q^\star_{\mathcal{C}} \in \mathcal{Q}^\star_{\mathcal{C}}(R)$ such that $\{R, Q^\star - Q^\star_{\mathcal{C}}\} \subset N_{\mathcal{Q}\cap\mathcal{C}}(Q^\star_{\mathcal{C}})$,*

Moreover, if $N_{\mathcal{Q}\cap\mathcal{C}}(Q^\star_{\mathcal{C}})$ is a ray $\{\lambda M : \lambda \ge 0\}$*, then commutativity holds if and only if R and $Q^\star - Q^\star_{\mathcal{C}}$ are positive multiples of the same direction M.*

In other words, if the projected manifold $\mathcal{C}$ of desired optimizers is in a sufficiently good shape, e.g., using EMA-based momentum, then we can first solve problem P3 for general budgets $\mathcal{Q}$ and then project the solution onto the manifold $\mathcal{C}$ to obtain the final optimal solution over $\mathcal{C}$. Now we are ready to find the optimal hyperparameters for real optimizers in use.

Corollary 5 (Instantaneous optimal SGD+Momentum)**.** *Define the budget $\mathcal{Q}_F(B)$ and the cone $\mathcal{C}_{1p}$ of isotropic 1-pole optimizers:*

$$\mathcal{Q}_F(B) \coloneqq \{Q : \|Q\|_{\mathcal{H}} \le \sqrt{B}\}, \quad \mathcal{C}_{1p} \coloneqq \{Q_{\eta,\beta}[n] = \eta(1-\beta)\beta^n I : \eta \ge 0, 0 < \beta < 1\}. \tag{11}$$

Given $R[n] \in \mathcal{H}$, define $S[n] \coloneqq \mathrm{Tr}(R[n])$. Then solving problem P3 under the budget $\mathcal{Q}_F(B) \cap \mathcal{C}_{1p}$ produces SGD+Momentum optimizer as the optimal solution with optimal hyperparameters:

$$\beta^\star = \arg\max_{0<\beta<1} \sqrt{1-\beta^2} \sum_{n=0}^{\infty} S[n]\beta^n, \qquad \eta^\star = \frac{\sqrt{B\,(1-\beta^{\star 2})}}{\sqrt{d}\,(1-\beta^\star)}, \tag{12}$$

where d is the dimension of the parameter space.

The optimal hyperparameters are obtained by first applying Theorem 3 to general family of budgets, and then projecting the solution into the approximation geometry $\mathcal{C}_{1p}$. By Lemma 4, the optimal hyperparameters of the projected solution are consistent with the unprojected solution. Corollary 5 shows how SGD+Momentum type *optimal* optimizer works. The *optimal momentum* $\beta^\star$ first maximizes the cosine similarity between the 1-pole EMA response $(\beta^n)_{n\ge0}$ and the trace of the empirical moment $(S[n])_{n\ge0}$; then the *learning rate* $\eta^\star$ scales to saturate the budget.

Corollary 6 does the same to Adam Kingma & Ba (2015). For Adam, the existence of a time-varying divisor $\mathrm{EMA}(g^2, \beta_2)^{-1/2}$ slightly complicates the derivation by making the optimizer time-varying.

Corollary 6 (Instantaneous optimal Adam)**.** *Let moment be a diagonal matrix $R[n] = \mathrm{diag}(r_j[n])$ with coordinate-wise sequence $r_j[n]$. Given gradients $g[t]$, maintain the second-moment EMA $v_j[t] \coloneqq \beta_2 v_j[t-1] + (1-\beta_2)g_j[t]^2 > 0$ with parameter $\beta_2 \in (0,1)$. Fix the current time t and define the coordinate-wise costs $c_j \coloneqq v_j[t]^{1/2}$. Consider the* diagonal budget

$$\mathcal{Q}_D(B, c) \coloneqq \{\mathrm{diag}(q_j) : \textstyle\sum_j c_j \sum_{k\ge0} |q_j[k]|^2 \le B\}, \tag{13}$$

and the cone $\mathcal{C}_{1p,\ Adam}$ of diagonal 1-pole optimizer with per-coordinate inverse-cost scaling

$$\mathcal{C}_{1p,\ Adam} \coloneqq \{Q_{\eta,\beta_1}[n] = \mathrm{diag}\,(\eta(1-\beta_1)\beta_1^n/c_j) : \eta \ge 0, 0 < \beta_1 < 1\}. \tag{14}$$

Optimizing for problem P3 under $\mathcal{Q}_D(B,c) \cap \mathcal{C}_{1p,\ Adam}$ with moment $R[n]$ yields an Adam optimizer with optimal hyperparameters:

$$(\beta_1^\star, \beta_2^\star) = \arg\max_{0<\beta_1,\beta_2<1} a(\beta_1,\beta_2) \sum_{n=0}^{\infty} \beta_1^n\, T_t[n;\beta_2], \quad \eta^\star = \frac{\sqrt{B}\,a(\beta_1^\star,\beta_2^\star)}{1-\beta_1^\star}, \tag{15}$$

where $W_t(\beta_2) \coloneqq \sum_j 1/c_j$, $T_t[n;\beta_2] \coloneqq \sum_j r_j[n]/c_j$, and $a(\beta_1,\beta_2) \coloneqq \sqrt{(1-\beta_1^2)/W_t(\beta_2)}$.

The optimal $\beta_1^\star$ maximizes the cosine similarity between the 1-pole EMA response $(\beta_1^n)_{n\ge0}$ and the cost-compensated diagonal moment $(T[n;\beta_2])_{n\ge0}$. The optimal $\beta_2^\star$ determines the diagonal weights through the EMA of the second-moment. Finally, the learning rate $\eta^\star$ scales to saturate the budget. In summary, Adam is a *1-pole approximation* of the *dynamic diagonal optimizer* defind by the diagonal budget $\mathcal{Q}_D(B,c)$ projected onto the 1-pole family $\mathcal{C}_{1p,\ Adam}$. All the proofs are provided in Appendix B.

Table 1: Reverse-engineered optimizers as convex optimization problems. Each optimizer emerges as the optimal solution to a specific convex optimization problem, with hyperparameters determined by the underlying budget constraint.

Optimizer	Hyperparameters	Interpretation
GD	η (learning rate)	Euclidean trust region: $\|\dot{\theta}\|_2 \le \tau$ yields $Q \propto I$
Colored-GD	η, precon. P	Elliptic trust region: $\|\dot{\theta}\|_P \le \tau$ yields $Q \propto P^{-1}$ with $P^\star$ from data covariance
Newton/GN	η, damping λ	Curvature trust region: $\|\dot{\theta}\|_H \le \tau$ yields $Q \propto H^{-1}$ with $\eta^\star = \tau \frac{\|g\|_{H^{-1}}}{\|H^{-1}g\|^2}$ and $\lambda^\star$ from condition number
NGD	η, Fisher est.	Fisher/KL trust region: $\|\dot{\theta}\|_F \le \tau$ yields $Q \propto F^{-1}$ with $\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(F)\|g\|_2}}$ and $F^\star$ from empirical Fisher
K-FAC/Shampoo	η, damping, period	Structured Kronecker trust region yields block $Q = \bigoplus(A_\ell^{-1} \otimes G_\ell^{-1})$ with $\eta^\star$ from factored Fisher metric and damping $\lambda^\star$ from eigenvalue gaps
AdaGrad	η, ϵ, window	Diagonal trust region: $\sum_j c_j \dot{\theta}_j^2 \le B$ with $c_j = \sqrt{\sum_{t'} g_{j,t'}^2}$ and $\eta_j^\star = \sqrt{B}\frac{\|g_j\|}{\sqrt{\sum_j g_j^2/c_j}}$, $\epsilon^\star$ for numerical stability
RMSProp	η, β_2, ϵ	Diagonal trust region with $c_j = \sqrt{v_j}$, $v_j = \beta_2 v_{j,t-1} + (1-\beta_2)g_j^2$ and $\eta_j^\star = \sqrt{B}\frac{|g_j|/\sqrt{v_j}}{\sqrt{\sum_j g_j^2/v_j}}$, $\beta_2^\star$ from gradient correlation
Adam	η, β_1, β_2, ϵ	Dynamic diagonal trust region on momentum m_k with costs $c_j = \sqrt{v_j}$ and $\eta^\star = \sqrt{B}\frac{\|m/\sqrt{v}\|_1}{\|m/\sqrt{v}\|_2^2}$, $\beta_1^\star, \beta_2^\star$ from lag curves
SGD+Momentum	η, β_1	1-pole EMA trust region: $\|Q\|_{\mathcal{H}} \le \sqrt{B}$ with $Q = \eta(1-\beta_1)\sum_{n=0}^\infty \beta_1^n I$ and $\eta^\star = \frac{\sqrt{B(1-\beta_1^2)}}{\|g\|_2(1-\beta_1)}$, $\beta_1^\star$ from cosine similarity
AdaFactor	η, β_1, β_2, factorization	Factored trust region with $r_i = \beta_2 r_{i,t-1} + (1-\beta_2)\|G_{i,:}\|^2$, $c_j = \beta_2 c_{j,t-1} + (1-\beta_2)\|G_{:,j}\|^2$ and $\eta^\star$ from factored diagonal, $\beta_1^\star, \beta_2^\star$ from matrix structure
LayerNorm+Mom	η, β_1, layer norm	Layer-wise normalized trust region: $\|Q\|_{\mathcal{H}_{\text{layer}}} \le \sqrt{B}$ with $\eta^\star = \frac{\sqrt{B(1-\beta_1^2)}}{\sqrt{\sum_\ell d_\ell/\|\theta_\ell\|_2^2}(1-\beta_1)}$ and $\beta_1^\star$ from layer-normalized lag
LAMB/LARS	η, trust ratio τ	Layer-wise trust region: $\|\dot{\theta}_\ell\|_2 \le \tau_\ell \|\theta_\ell\|/\|m_\ell\|$ with $\eta_\ell^\star = \tau_\ell \frac{\|\theta_\ell\|}{\|m_\ell\|}$ and $\tau^\star$ from layer-wise analysis
signSGD	η (step size)	L_∞ trust region: $\|\dot{\theta}\|_\infty \le \tau$ yields $\dot{\theta} = \tau\,\mathrm{sign}(g)$ with $\eta^\star = \tau$
Lion	η, β_1	L_∞ trust region on momentum: $\|\dot{\theta}\|_\infty \le \tau$ with $\dot{\theta} = \eta\,\mathrm{sign}(m)$, $\eta^\star = \tau$ and $\beta_1^\star$ from sign correlation

Reverse engineering optimizers. The above two corollaries show that SGD with momentum and Adam (Kingma & Ba, 2015) are the optimal 1-pole approximations of the dynamic equalizers, establishing these well-known optimizers as special cases of our framework. Similarly, we can reverse engineer various other optimizers, including SGD with Nesterov momentum (Nesterov, 1983), AdamW (Loshchilov & Hutter, 2019), LAMB (You et al., 2020), K-FAC (Martens & Grosse, 2015), Shampoo (Gupta et al., 2018), and Lion (Chen et al., 2023), into our framework. Due to page constraints, we defer the details to Appendix C. We have also provided a master table of optimizers in Appendix D, categorizing many optimizers widely used in practice today.

Automatic hyperparameter tuning. Examining Corollaries 5 and 6, our formulation not only classifies optimizers according to their underlying budget choices, but also provides a systematic way to determine *optimal hyperparameters*. Unsurprisingly, optimal hyperparameters depend on the data being processed. After warm-up steps that collect data covariance, optimal hyperparameters can be computed and used throughout training. Moreover, systematic hyperparameter determination enables automatic tuning during training, which is impossible with manual tuning. The algorithms for determining hyperparameters is provided in Appendix E.

4 Automating Validation-Aware Optimizer Tuning

So far, our theoretical justification for optimizing optimizers did not specify which datasets are in use. This section addresses a more delicate question of how to *systematically* exploit validation sets for optimizer design. It is commonly considered bad practice to use validation sets directly

in the optimization loop. Rather, they are typically used to generate subtle clues that indirectly guide engineers when making decisions about model architecture, optimizers, and associated hyperparameters. We can regard this manual tuning process as a kind of "human-in-the-loop" system that *fits* the optimizer and hyperparameters to the validation set. We can then automate this process by casting it as the same mathematical optimization problem. For example, our original optimization problem P3 can be recast in terms of maximizing the instantaneous *validation* loss drop as

$$\underset{Q\in\mathcal{Q}}{\text{maximize}} \; -\dot{\mathcal{L}}_{\text{val}} \;=\; \textstyle\sum_n \mathbb{E}[g_{\text{val}}[n]^\top \dot{\theta}_{\text{tr}}[n]] \;=\; \langle Q, R_\times\rangle_{\mathcal{H}} \;=\; P_\times(Q) \quad \text{s.t.} \quad Q \succeq 0, \tag{P4}$$

where $\dot{\theta}_{\text{tr}}[n] = (q * g_{\text{tr}})[n]$ is the parameter velocity guided solely by the training set and the designed optimizer Q. This is mathematically equivalent to problem P3 but with the cross-moment $R_\times$, mimicking human inspection of the validation loss drop and engineering the optimizer accordingly. This approach may or may not conflict with traditional practice, potentially requiring an additional split beyond the standard training/validation division. We leave this discussion to the readers. Here, we focus on the theoretical side of validation-aware optimizer design by showing that tuning optimizers using validation sets maximizes the instantaneous validation loss drop.

Proposition 7 (Validation optimality in power)**.** *Given any convex and compact budget $\mathcal{Q} \subset \mathcal{H}_+$, the validation-aware maximizer $Q^\star_\times \in \arg\max_{Q\in\mathcal{Q}}\langle Q, R_\times\rangle_{\mathcal{H}}$ gives the maximum instantaneous validation loss drop among all possible optimizers $Q \in \mathcal{Q}$, including training-only optimizers $Q \in \arg\max_{Q\in\mathcal{Q}}\langle Q, R_{\text{tr}}\rangle_{\mathcal{H}}$.*

We leave the proof to Appendix B. Therefore, validation-aware tuning is the best possible instantaneous validation loss drop among all possible optimizers $Q \in \mathcal{Q}$, including training-only optimizers $Q \in \arg\max_{Q\in\mathcal{Q}} \operatorname{Tr}(Q\,\Sigma_{\text{tr}})$. The next proposition shows how the choice of optimizer Q not only controls the parameter velocity, but also determines the endpoint of the optimization process in the local convex region.

To approximate the dynamics in the local convex region in the loss landscape, assume squared loss and fix parameters θ. In the linearizable region of the network f_θ around θ, the function-space dynamics follow *kernel gradient flow* with kernel

$$K_Q(x, x') \;:=\; \nabla_\theta f_\theta(x;\theta)^\top\, Q\, \nabla_\theta f_\theta(x';\theta). \tag{16}$$

Define the budget $\mathcal{Q}$, the validation cross-moment $\Sigma_\times$, and the validation-aware maximizer $Q^\star_\times \in \arg\max_{Q\in\mathcal{Q}} \operatorname{Tr}(Q\,\Sigma_\times)$ the same as in Proposition 7. Then, the following proposition holds:

Proposition 8 (Endpoint selection (function-space view))**.** *With any optimizer $Q \in \mathcal{Q}$ fixed around θ, kernel gradient flow converges to the unique* minimum-norm *interpolant in $\mathcal{H}_{K_Q}$ (or to kernel ridge with decay $\lambda > 0$). Consequently, choosing $Q = Q^\star_\times$ changes the RKHS to $\mathcal{H}_{K_{Q^\star_\times}}$ and selects the endpoint*

$$f^\star_{Q^\star_\times} \;=\; \arg\min_{f\in\mathcal{H}_{K_{Q^\star_\times}}} \|f\|^2_{\mathcal{H}_{K_{Q^\star_\times}}} \quad \text{subject to} \quad f(X) \;=\; y. \tag{17}$$

In summary, tuning the optimizer Q using validation sets is the best possible way not only to maximize the instantaneous validation loss drop, but also to select better convergence endpoints of the optimization process.

5 Scope and Limitations

Long-horizon objective from greedy paradigm. In order to simplify the analysis, this work resorts to the greedy paradigm, primarily focusing on instantaneous progress of learning. As a trade-off, global optimality guarantee requires further investigation under this greedy paradigm. Our filter interpretation of dynamic optimizers in Section 3 mitigates the limitations of the greedy objective by incorporating stateful optimizers holding summaries of the history of gradients. This also generalizes the optimizers with momentum and other EMA-based smoothing techniques.

Choice of budget. Instead of telling which class of optimizers are optimal for a given task, this work provides an optimization framework *under* the user-defined budget set. Our framework can help engineers by reducing their effort for searching the right hyperparameters; however, it still requires an intelligent choice of which budget, i.e., which *class of* optimizers, fits the task.

Renovating existing optimizers. In Theorems 1 and 3, we provide general construction of optimal optimizers from convex constraints. However, in this work, we focus on the well-established optimizers, and supplement them with a systematic methodology to find the right hyperparameters for a given task. Designing a new class of optimizers will be a natural extension of this work.

6 Related work

First, we would like to highlight the key differences between our "optimizing optimizers" and other well-established fields with similar tautologies.

Learning to optimize. Learning to optimize (Li & Malik, 2016) aims to adapt the optimizer to a given task by treating optimizers as learnable parametric models (Andrychowicz et al., 2016). Various architectures have been explored, including RNNs (Andrychowicz et al., 2016; Wichrowska et al., 2017; Lv et al., 2017), Transformers (Chen et al., 2022; Moudgil et al., 2023; Jain et al., 2024), and per-tensor HyperNetworks (Ha et al., 2016; Metz et al., 2022). Their primary focus is on meta-training these optimizer-networks for stability and adaptability. These works represent a nontraditional, network-based family of generally nonconvex optimizers, which is not generally compatible with our framework which is based on convex optimization.

Learning to learn. Rooted in the human-inspired philosophy (Schmidhuber, 1987; Bengio et al., 1990), meta-learning is another line of work that shares a similar spirit with learning to optimize (Gharoun et al., 2023). A large proportion of works on meta-learning target few-shot learning tasks, which prepare the model, not the optimizer, for downstream tasks (Vinyals et al., 2016; Finn et al., 2017; Yu et al., 2024; Sun et al., 2019). Among them, Meta-SGD (Li et al., 2017) is noteworthy, as it prepares the optimizer. However, the problem set we address is general gradient-based learning, which differs from the tasks of concern in meta-learning.

Algorithmic discovery of optimizers. Techniques like symbolic discovery (Chen et al., 2023; Zheng et al., 2022), non-parametric optimizer search (Wang et al., 2022), and neural optimizer search (Bello et al., 2017) are also related to our work, as their objective is to discover the optimal optimizer for a given task. In their framework, symbolic optimizers are obtained by a tree-based search of a predefined set of optimizers. Ours instead lets the engineer select the broader family of optimizers, and then provides a mathematical tool to find the optimal solution among them. Therefore, these works are considered to be orthogonal to ours.

Hyperparameter optimization. Many works have proposed to automatically tune the hyperparameters governing optimization. Most of them adopt a learning framework to find a good set of hyperparameters including learning rates (Daniel et al., 2016), their schedules (Xu et al., 2017; 2019), and other optimizer parameters (Shaban et al., 2019). Hypergradient methods (Maclaurin et al., 2015; Baydin et al., 2017; Grazzi et al., 2020; Moskovitz et al., 2019) are also proposed to find the optimal hyperparameters. Instead of resorting to learning-based methods, we establish a theoretical framework through the lens of convex optimization problems (Boyd & Vandenberghe, 2004). By doing so, we can classify well-used optimizers such as SGD with momentum and Adam (Kingma & Ba, 2015) as special cases of our framework, and provide a systematic way to determine the optimal hyperparameters for these optimizers.

7 Conclusion

We established a firm theoretical grounding for systematically achieving optimal optimizers in a greedy sense. Our convex optimization framework connects commonly used optimizers to convex constraint sets, merging those independently developed techniques into a single unified framework. Our main results, Theorems 1 and 3, and Lemma 4 are general tools that can be extended to arbitrary budget to invent new families of optimizers for specific uses. Our theory, therefore, does not disprove the *no free lunch theorem*; rather, it provides a principled way to *leverage* this wisdom to flexibly design and adapt optimizers for our own problems at hand.

Ethics statement

We acknowledge the ICLR Code of Ethics.

Reproducibility statement

All the proofs of the theoretical part of this paper, including every Lemma, Theorem, Proposition, and Corollary, are provided in Appendix B with detailed derivations, starting from the basic definitions and assumptions made in the main text. Moreover, omitted theoretical results are elaborated in Appendix A. Regarding the implementation, Appendix E gives the algorithm to realize our theoretically justified optimal optimizers.

LLM usage statement

We deeply acknowledge the usefulness of LLMs in revising the manuscript, especially for fixing vocabulary and grammar-related issues. We also used LLMs to check the correctness and coherence of the proofs and notations. This greatly helped us in identifying awkward mistakes we had been making all along.

References

Naman Agarwal, Brian Bullins, Xinyi Chen, Elad Hazan, Karan Singh, Cyril Zhang, and Yi Zhang. Efficient full-matrix adaptive regularization. In *ICML*, 2019.

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

Marcin Andrychowicz, Misha Denil, Sergio Gomez Colmenarejo, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *NIPS*, 2016.

Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. *arXiv preprint arXiv:1703.04782*, 2017.

Irwan Bello, Barret Zoph, Vijay Vasudevan, and Quoc V. Le. Neural optimizer search with reinforcement learning. In *ICML*, 2017.

Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. Learning a synaptic learning rule. Technical report, Université de Montréal, Département d'informatique et de recherche opérationnelle, 1990.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *NeurIPS*, 2023.

Yutian Chen, Xingyou Song, Chansoo Lee, Zi Wang, Qiuyi Zhang, David Dohan, Kazuya Kawakami, Greg Kochanski, Arnaud Doucet, Marc'aurelio Ranzato, Sagi Perel, and Nando de Freitas. Towards learning universal hyperparameter optimizers with transformers. In *NeurIPS*, 2022.

Christian Daniel, Jonathan Taylor, and Sebastian Nowozin. Learning step size controllers for robust neural network training. In *AAAI*, 2016.

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.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

Hassan Gharoun, Fereshteh Momenifar, Fang Chen, and Amir H. Gandomi. Meta-learning approaches for few-shot learning: A survey of recent advances. *arXiv preprint arXiv:2303.07502*, 2023.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *ICML*, 2020.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *ICML*, 2018.

David Ha, Andrew M. Dai, and Quoc V. Le. HyperNetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Deepali Jain, Krzysztof M. Choromanski, Kumar Avinava Dubey, Sumeet Singh, Vikas Sindhwani, Tingnan Zhang, and Jie Tan. Mnemosyne: Learning to train transformers with transformers. In *NeurIPS*, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Ke Li and Jitendra Malik. Learning to Optimize. *arXiv preprint arXiv:1606.01885*, 2016.

Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-SGD: Learning to learn quickly for few-shot learning. In *NIPS*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Kaifeng Lv, Shunhua Jiang, and Jian Li. Learning gradient descent: Better generalization and longer horizons. In *ICML*, 2017.

Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *ICML*, 2015.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *ICML*, 2015.

Luke Metz, James Harrison, C. Daniel Freeman, Amil Merchant, Lucas Beyer, James Bradbury, Naman Agrawal, Ben Poole, Igor Mordatch, Adam Roberts, and Jascha Sohl-Dickstein. VeLO: Training versatile learned optimizers by scaling up. *arXiv preprint arXiv:2211.09760*, 2022.

Ted Moskovitz, Rui Wang, Janice Lan, Sanyam Kapoor, Thomas Miconi, Jason Yosinski, and Aditya Rawal. First-order preconditioning via hypergradient descent. *arXiv preprint arXiv:1910.08461*, 2019.

Abhinav Moudgil, Boris Knyazev, Guillaume Lajoie, and Eugene Belilovsky. Learning to optimize with recurrent hierarchical transformers. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983. English translation: Soviet Mathematics Doklady, 27(2):372–376, 1983.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Jürgen Schmidhuber. *Evolutionary Principles in Self-Referential Learning, or on Learning How to Learn: The Meta-Meta-... Hook*. PhD thesis, Technische Universität München, 1987.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *AISTATS*, 2019.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *CVPR*, 2019.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

Oriol Vinyals, Charles Blundell, Tim Lillicrap, and Daan Wierstra. Matching networks for one shot learning. In *NIPS*, 2016.

Ruochen Wang, Yuanhao Xiong, Minhao Cheng, and Cho-Jui Hsieh. Efficient non-parametric optimizer search for diverse tasks. In *NeurIPS*, 2022.

Olga Wichrowska, Niru Maheswaranathan, Matthew W Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Nando de Freitas, and Jascha Sohl-Dickstein. Learned optimizers that scale and generalize. In *ICML*, 2017.

David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

Chang Xu, Tao Qin, Gang Wang, and Tie-Yan Liu. Reinforcement learning for learning rate control. *arXiv preprint arXiv:1705.11159*, 2017.

Zhen Xu, Andrew M. Dai, Jonas Kemp, and Luke Metz. Learning an adaptive learning rate schedule. *arXiv preprint arXiv:1909.09712*, 2019.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *ICLR*, 2020.

Xingtong Yu, Yuan Fang, Zemin Liu, Yuxia Wu, Zhihao Wen, Jianyuan Bo, Xinming Zhang, and Steven C.H. Hoi. A survey of few-shot learning on graphs: from meta-learning to pre-training and prompt learning. *arXiv preprint arXiv:2402.01440*, 2024.

Wenqing Zheng, Tianlong Chen, Ting-Kuei Hu, and Zhangyang Wang. Symbolic learning to optimize: Towards interpretability and scalability. *arXiv preprint arXiv:2203.06578*, 2022.

A MORE MATHEMATICAL RESULTS

This section provides more detailed mathematical foundations of the main text, which was omitted for brevity. Appendix A.1 shows a detailed derivation of equation 9. This shows that the instantaneous learning powers of dynamic optimizers are also represented by some inner product between the optimizer operator and the gradient moment, having the same structure as in the stateless case. Appendix A.2 draws connection between the results in Section 2 and the results in Section 3, by showing the corresponding results in the dynamic setting.

A.1 PROOF OF EQUATION 9

Here we provide a detailed derivation of equation 9 that was abbreviated in the main text. We assume the sequences $\{g_{\text{tr}}[n]\}$ and $\{g_{\text{val}}[n]\}$ are zero-mean, wide-sense stationary (WSS) with finite second moments. Hence the lag-k moments $R_{\text{tr}}[k] = \mathbb{E}[g_{\text{tr}}[n]g_{\text{tr}}[n-k]^\top]$ and $C[k] = \mathbb{E}[g_{\text{val}}[n]g_{\text{tr}}[n-k]^\top]$ depend only on k. Then

$$P_{\text{tr}}(Q) \;=\; \langle Q, R_{\text{tr}}\rangle_{\mathcal{H}}, \qquad P_{\text{val}}(Q) \;=\; \langle Q, R_{\text{val}}\rangle_{\mathcal{H}}. \tag{18}$$

We start from the convolution definition of the dynamic optimizer:

$$\dot{\theta}[n] = \sum_{k=0}^{\infty} Q[k]\, g_{\text{tr}}[n-k]. \tag{19}$$

Instantaneous training power:

$$\begin{aligned}
P_{\text{tr}}(Q;n) &= \mathbb{E}\big[g_{\text{tr}}[n]^\top \dot{\theta}[n]\big] && (20)\\
&= \mathbb{E}\Big[g_{\text{tr}}[n]^\top \sum_{k=0}^{\infty} Q[k]\, g_{\text{tr}}[n-k]\Big] && (21)\\
&= \sum_{k=0}^{\infty} \mathbb{E}\big[g_{\text{tr}}[n]^\top Q[k]\, g_{\text{tr}}[n-k]\big] \quad \text{(linearity of } \mathbb{E}) && (22)\\
&= \sum_{k=0}^{\infty} \mathrm{Tr}\big(Q[k]\; \mathbb{E}[g_{\text{tr}}[n-k]\, g_{\text{tr}}[n]^\top]\big) && (23)\\
&= \sum_{k=0}^{\infty} \mathrm{Tr}\big(Q[k]^T\; \mathbb{E}[g_{\text{tr}}[n]\, g_{\text{tr}}[n-k]^\top]\big) \quad \text{(trace transpose)} && (24)\\
&= \sum_{k=0}^{\infty} \mathrm{Tr}\big(Q[k]^T\, R_{\text{tr}}[k]\big) && (25)\\
&= \langle Q, R_{\text{tr}}\rangle_{\mathcal{H}}. && (26)
\end{aligned}$$

Instantaneous validation cross-power:

$$\begin{aligned}
P_{\text{val}}(Q;n) &= \mathbb{E}\big[g_{\text{val}}[n]^\top \dot{\theta}[n]\big] && (27)\\
&= \mathbb{E}\Big[g_{\text{val}}[n]^\top \sum_{k=0}^{\infty} Q[k]\, g_{\text{tr}}[n-k]\Big] && (28)\\
&= \sum_{k=0}^{\infty} \mathbb{E}\big[g_{\text{val}}[n]^\top Q[k]\, g_{\text{tr}}[n-k]\big] && (29)\\
&= \sum_{k=0}^{\infty} \mathrm{Tr}\big(Q[k]\; \mathbb{E}[g_{\text{tr}}[n-k]\, g_{\text{val}}[n]^\top]\big) && (30)\\
&= \sum_{k=0}^{\infty} \mathrm{Tr}\big(Q[k]^T\; \mathbb{E}[g_{\text{val}}[n]\, g_{\text{tr}}[n-k]^\top]^T\big) && (31)\\
&= \sum_{k=0}^{\infty} \mathrm{Tr}\big(Q[k]^T\, C[k]^T\big), && (32)
\end{aligned}$$

where $C[k] \coloneqq \mathbb{E}[g_{\text{val}}[n]\, g_{\text{tr}}[n-k]^\top]$. For the symmetric cross-moment $R_{\text{val}}[k] \coloneqq \frac{1}{2}\big(C[k]+C[k]^\top\big)$, when using Hermitian PSD filters with symmetric $Q[k]$, we have

$$\operatorname{Tr}(Q[k]^T C[k]^T) = \operatorname{Tr}(Q[k]\,\tfrac{1}{2}(C[k]+C[k]^\top)) = \operatorname{Tr}(Q[k]^T R_{\text{val}}[k]), \tag{33}$$

since for symmetric A, $\operatorname{Tr}(A\,\frac{1}{2}(B-B^\top))=0$. Therefore, $P_{\text{val}}(Q;n) = \langle Q, R_{\text{val}}\rangle_{\mathcal{H}}$.

A.2 Dynamic lift of Section 2

We now lift the key results of Section 2 to the dynamic setting of Section 3. Solving the optimization problem with constraint $\mathcal{Q}$ *determines* the optimal dynamic optimizer $Q^\star$, and endows the optimizer with different characteristics and algorithmic behaviors. Again, consider the following four types of dynamic budgets:

- *Frobenius ball budget* $\mathcal{Q}_F(B) = \{Q : \|Q\|_{\mathcal{H}} \le \sqrt{B}\}$ is the simplest constraint that gives an isotropic Hilbert space trust region without prior knowledge about temporal correlation structure.
- *Per-frequency spectral budget* $\mathcal{Q}_S(\tau,\lambda) = \{Q : \operatorname{Tr}(Q(e^{i\omega})) \le \tau(\omega), Q(e^{i\omega}) \preceq \lambda(\omega) I\}$ is a budget that upper limits the per-direction spectrum for safety and the trace for total budget simultaneously at each frequency.
- *Data-metric (Lyapunov) budget* $\mathcal{Q}_L(B) = \{Q : \sum_{k=0}^{\infty} \operatorname{Tr}(Q_k^* \Sigma Q_k) \le B\}$ is a budget that uses the lag-covariance sequence itself as the metric, leading to a natural dynamic Lyapunov-like stability condition.
- *Diagonal budget* $\mathcal{Q}_D(B,c) = \{Q(z) = \operatorname{diag}(q_j(z)) \succeq 0 : \sum_j c_j \|q_j\|_{H_2}^2 \le B\}$ is a budget that restricts to diagonal dynamic optimizers with coordinate-wise budgets.

Instantiating the construction from Theorem 3 with these budgets, we obtain corresponding closed-form solutions for the optimal dynamic optimizer $Q^\star$ and the optimal power $P^*(R)$.

Corollary 9 (Closed-form solutions for dynamic budget sets)**.** *Let* $R(e^{i\omega}) = U(\omega)\operatorname{diag}(\sigma_i(\omega))U(\omega)^*$ *be the eigendecomposition at each frequency. The optimal solutions are:*

(i) Frobenius ball*:* $Q_F^* = \sqrt{B}R/\|R\|_{\mathcal{H}}$, $P_F^*(R) = \sqrt{B}\|R\|_{\mathcal{H}}$.

(ii) Per-frequency spectral*:* $Q_S^*(e^{i\omega}) = U(\omega)\operatorname{diag}(q_i^*(\omega))U(\omega)^*$ *where* $q_i^*(\omega) = \min\{\lambda(\omega), \max\{0, \mu(\omega)-\sigma_i(\omega)\}\}$ *and* $\mu(\omega)$ *is chosen so that* $\sum_i q_i^*(\omega) = \tau(\omega)$.

(iii) Data-metric*:* $Q_L^* = \alpha \Pi_{\text{supp}(R)}$ *where* $\alpha = \sqrt{B/\sum_{i:\sigma_i>0}\sigma_i}$, $P_L^*(R) = \sqrt{B\sum_i \sigma_i}$.

(iv) Diagonal*:* $[Q_D^*]_{jj} \propto R_{jj}/c_j$, $P_D^*(R) = \sqrt{B\sum_j R_{jj}^2/c_j}$.

We delay the proof to Appendix B. These analytic solutions reveal how the characteristics of different types of optimal dynamic optimizers $Q^\star$ are induced by controlling the budget set $\mathcal{Q}$.

Frobenius budget $\leftrightarrow$ Proportional dynamic optimizer. Budget $\mathcal{Q}_F(B)$ produces a *proportional* dynamic optimizer that allocates learning power proportional to lag-covariance evidence $Q^\star \propto R$. It enjoys implementation simplicity but potentially over-concentrates on dominant temporal modes. Using Lemma 4, we can project this general class of optimizers into special geometries to calculate for the optimal hyperparameters as in Corollary 5.

Per-frequency spectral budget $\leftrightarrow$ Water-filling dynamic optimizer $\sim$ per-frequency gradient clipping & power scheduling. Budget $\mathcal{Q}_S(\tau,\lambda)$ produces a *water-filling* dynamic optimizer that keeps pushing power into responsive frequency modes until hitting the safety cap $\lambda(\omega)$. The spectral cap $\lambda(\omega)$ acts as a *frequency-dependent* safety mechanism similar to *gradient clipping*, while the trace constraint $\tau(\omega)$ controls the *per-frequency learning rate* similar to *adaptive learning rate scheduling*.

Data-metric budget ↔ Equal-power dynamic optimizer ∼ dynamic AdaGrad. Budget $\mathcal{Q}_L(B)$ produces an *equal-power* dynamic optimizer that equalizes learning power uniformly across informative lag-correlation directions, preventing over-concentration while maintaining temporal efficiency. This produces a natural dynamic preconditioning effect similar to *AdaGrad*'s inverse square root scaling, but with uniform power allocation across all informative temporal directions rather than instantaneous adaptation.

Diagonal budget ↔ Coordinate-wise dynamic optimizer ∼ Adam. Budget $\mathcal{Q}_D(B, c)$ produces a *coordinate-wise* dynamic optimizer that adapts per-coordinate learning power proportional to the lag-covariance evidence R_{jj} and inversely proportional to the costs c_j. When $c_j \propto v_{t,j}^{1/2}$ ($v_{t,j}$ being the EMA of g_j^2 at time t), this recovers the core mechanism of *Adam*, which is elaborated in Corollary 6 in the main manuscript.

B Proofs omitted from the main text

This section does all the proofs that has been omitted in the main text. The proofs are organized in the same order as the theorems appear in the main manuscript.

B.1 Proof of Theorem 1

Proof of Theorem 1. We establish each claim in turn.

Existence & sublinearity: Since $\mathcal{Q}$ is compact by assumption (a nonempty, compact, convex subset of $\mathbb{S}_+^d$), and $(Q, \Sigma) \mapsto \mathrm{Tr}(Q\Sigma)$ is continuous, the maximum is attained by the Weierstrass extreme value theorem. The optimal power $P^\star(\Sigma) = \sup_{Q\in\mathcal{Q}} \mathrm{Tr}(Q\Sigma)$ is a supremum of linear functionals in Σ, hence sublinear (convex and positively homogeneous). Finiteness follows from compactness of $\mathcal{Q}$.

Conjugacy identities: We establish the three identities in equation 5.

1. Optimal power = conjugate of indicator. By the definition of convex conjugate,

$$(\delta_{\mathcal{Q}})^*(\Sigma) = \sup_{Q\in\mathbb{S}^d} \{\langle Q, \Sigma\rangle - \delta_{\mathcal{Q}}(Q)\} = \sup_{Q\in\mathcal{Q}} \langle Q, \Sigma\rangle = P^\star(\Sigma). \tag{34}$$

Thus $P^\star = (\delta_{\mathcal{Q}})^*$.

2. Optimal power = gauge of polar. By the definition of polar, $\Sigma \in \mathcal{Q}^\circ$ if and only if $\sup_{Q\in\mathcal{Q}}\langle Q, \Sigma\rangle \le 1$, i.e., $P^\star(\Sigma) \le 1$. Therefore

$$\gamma_{\mathcal{Q}^\circ}(\Sigma) = \inf\{\lambda > 0 : \Sigma \in \lambda\mathcal{Q}^\circ\} = \inf\{\lambda > 0 : P^\star(\Sigma) \le \lambda\} = P^\star(\Sigma). \tag{35}$$

Thus $P^\star = \gamma_{\mathcal{Q}^\circ}$.

3. Conjugate of gauge = indicator of polar. We establish $(\gamma_{\mathcal{Q}})^* = \delta_{\mathcal{Q}^\circ}$. Consider two cases:

- If $\Sigma \in \mathcal{Q}^\circ$, then for all Q,

$$\langle Q, \Sigma\rangle \le \gamma_{\mathcal{Q}}(Q) \cdot \sup_{R\in\mathcal{Q}} \langle R, \Sigma\rangle \le \gamma_{\mathcal{Q}}(Q), \tag{36}$$

  since $\sup_{R\in\mathcal{Q}}\langle R, \Sigma\rangle \le 1$ by definition of polar. Hence $\langle Q, \Sigma\rangle - \gamma_{\mathcal{Q}}(Q) \le 0$ for all Q, with equality at $Q = 0$. Taking the supremum gives $(\gamma_{\mathcal{Q}})^*(\Sigma) = 0 = \delta_{\mathcal{Q}^\circ}(\Sigma)$.

- If $\Sigma \notin \mathcal{Q}^\circ$, there exists $Q_0 \in \mathcal{Q}$ with $\langle Q_0, \Sigma\rangle > 1$. For any $\alpha > 0$, we have $\gamma_{\mathcal{Q}}(\alpha Q_0) = \alpha\gamma_{\mathcal{Q}}(Q_0) = \alpha$ (since $Q_0 \in \mathcal{Q}$ so $\gamma_{\mathcal{Q}}(Q_0) = 1$), and thus

$$\langle \alpha Q_0, \Sigma\rangle - \gamma_{\mathcal{Q}}(\alpha Q_0) = \alpha\langle Q_0, \Sigma\rangle - \alpha = \alpha(\langle Q_0, \Sigma\rangle - 1) \to +\infty \quad (\alpha \to \infty). \tag{37}$$

  Hence $(\gamma_{\mathcal{Q}})^*(\Sigma) = +\infty = \delta_{\mathcal{Q}^\circ}(\Sigma)$.

Thus $(\gamma_{\mathcal{Q}})^* = \delta_{\mathcal{Q}^\circ}$.

Construction by subgradient: Let $Q^\star \in \arg\max_{Q\in\mathcal{Q}} \mathrm{Tr}(Q\Sigma)$. For any $M \in \mathbb{S}^d$,

$$P^\star(M) = \max_{Q\in\mathcal{Q}} \mathrm{Tr}(QM) \geq \mathrm{Tr}(Q^\star M) = \mathrm{Tr}(Q^\star\Sigma)+\mathrm{Tr}(Q^\star(M-\Sigma)) = P^\star(\Sigma)+\mathrm{Tr}(Q^\star(M-\Sigma)), \tag{38}$$

which is the defining inequality for $Q^\star \in \partial P^\star(\Sigma)$. If the maximizer is unique, $\partial P^\star(\Sigma) = \{Q^\star\}$ and $P^\star$ is differentiable at Σ with $\nabla P^\star(\Sigma) = Q^\star$.

Order preservation: If $\Sigma \succeq 0$, then for any $Q \in \mathcal{Q} \subseteq \mathbb{S}^d_+$, we have $\mathrm{Tr}(Q\Sigma) \geq 0$. Since $0 \in \mathcal{Q}$, the maximum over $Q \in \mathcal{Q}$ is ≥ 0. If $\Sigma_1 \succeq \Sigma_2$, then $P^\star(\Sigma_1) \geq P^\star(\Sigma_2)$. Moreover, strict inequality holds if there exists $Q \in \mathcal{Q}$ with $\mathrm{Tr}\big(Q(\Sigma_1 - \Sigma_2)\big) > 0$.

Lipschitz continuity in symmetrized polar gauge: We establish the one-sided bounds first. Since $P^\star = \gamma_{\mathcal{Q}^\circ}$ by the conjugacy identities, we have:

$$P^\star(\Sigma) - P^\star(\hat{\Sigma}) = \max_{Q\in\mathcal{Q}}\langle Q, \Sigma\rangle - \max_{Q\in\mathcal{Q}}\langle Q, \hat{\Sigma}\rangle \tag{39}$$

$$\leq \max_{Q\in\mathcal{Q}}\langle Q, \Sigma - \hat{\Sigma}\rangle \tag{40}$$

$$= \gamma_{\mathcal{Q}^\circ}(\Sigma - \hat{\Sigma}). \tag{41}$$

Similarly, $P^\star(\hat{\Sigma}) - P^\star(\Sigma) \leq \gamma_{\mathcal{Q}^\circ}(\hat{\Sigma} - \Sigma)$. Therefore,

$$|P^\star(\Sigma) - P^\star(\hat{\Sigma})| \leq \max\{\gamma_{\mathcal{Q}^\circ}(\Sigma - \hat{\Sigma}), \gamma_{\mathcal{Q}^\circ}(\hat{\Sigma} - \Sigma)\} = \|\Sigma - \hat{\Sigma}\|^{\mathrm{sym}}_{\mathcal{Q}^\circ}. \tag{42}$$

This Lipschitz property is crucial for robustness analysis. By using an estimated moment $\hat{\Sigma}$ instead of the true moment Σ, the error in optimal power can be bounded by $|P^\star(\Sigma) - P^\star(\hat{\Sigma})| \leq \|\Sigma - \hat{\Sigma}\|^{\mathrm{sym}}_{\mathcal{Q}^\circ}$. This provides a principled way to assess estimation sensitivity. □

B.2 Proof of Corollary 2

Proof of Corollary 2. We apply Theorem 1 to each budget set. Let $\Sigma = U\,\mathrm{diag}(\sigma_1 \geq \cdots \geq \sigma_d)U^\top$ be the eigendecomposition of the moment matrix.

(i) Frobenius ball $\mathcal{Q}_F(B) = \{Q \succeq 0 : \|Q\|_F \leq \sqrt{B}\}$. The Lagrangian is $L(Q, \lambda) = \mathrm{Tr}(Q\Sigma) - \lambda(\|Q\|_F^2 - B)$. Taking the gradient with respect to Q and setting to zero:

$$\nabla_Q L = \Sigma - 2\lambda Q = 0 \quad \Rightarrow \quad Q = \frac{\Sigma}{2\lambda}.$$

The constraint $\|Q\|_F = \sqrt{B}$ gives $\|\Sigma/(2\lambda)\|_F = \sqrt{B}$, so $2\lambda = \|\Sigma\|_F/\sqrt{B}$. Hence:

$$Q^\star_{\mathrm{F}} = \sqrt{B}\,\frac{\Sigma}{\|\Sigma\|_F}, \quad P^\star_{\mathrm{F}}(\Sigma) = \mathrm{Tr}(Q^\star_{\mathrm{F}}\Sigma) = \sqrt{B}\,\|\Sigma\|_F.$$

(ii) Spectral $\mathcal{Q}_S(\tau, \lambda) = \{Q \succeq 0 : \mathrm{Tr}(Q) \leq \tau,\ Q \preceq \lambda I\}$. By Neumann's inequality, the maximizer has the form $Q = U\,\mathrm{diag}(q_i)U^\top$ where the eigenvalues q_i solve the water-filling problem:

$$\max_{q_i\geq 0} \sum_i q_i\sigma_i \quad \text{s.t.} \quad \sum_i q_i \leq \tau,\ q_i \leq \lambda.$$

The KKT conditions yield: (i) $q_i^\star = \lambda$ for $i \leq k$, (ii) $q^\star_{k+1} = \tau - k\lambda$, (iii) $q_i^\star = 0$ for $i > k+1$, where $k = \lfloor \tau/\lambda \rfloor$. The optimal power is:

$$P^\star_{\mathrm{S}}(\Sigma) = \lambda \sum_{i\leq k} \sigma_i + (\tau - k\lambda)\sigma_{k+1}.$$

(iii) Data-metric $\mathcal{Q}_L(B) = \{Q \succeq 0 : \mathrm{Tr}(Q^2\Sigma) \leq B\}$. The Lagrangian is $L(Q, \mu) = \mathrm{Tr}(Q\Sigma) - \mu(\mathrm{Tr}(Q^2\Sigma) - B)$. The first-order condition gives:

$$\Sigma - 2\mu Q\Sigma = 0 \quad \Rightarrow \quad Q = \frac{1}{2\mu} I \text{ on } \mathrm{supp}(\Sigma).$$

Using the constraint $\mathrm{Tr}(Q^2\Sigma) = B$ and the fact that Q is constant on the support:

$$\alpha^2 \sum_{i:\sigma_i>0} \sigma_i = B \quad \Rightarrow \quad \alpha = \sqrt{\frac{B}{\sum_{i:\sigma_i>0}\sigma_i}}.$$

Hence:

$$Q^\star_{\mathrm{L}} = \alpha\, \Pi_{\mathrm{supp}(\Sigma)}, \quad P^\star_{\mathrm{L}}(\Sigma) = \alpha \sum_i \sigma_i = \sqrt{B \sum_i \sigma_i}.$$

(iv) Diagonal $\mathcal{Q}_D(B,c) = \{Q = \mathrm{diag}(q_j) \succeq 0 : \sum_j c_j q_j^2 \le B\}$. The problem decouples coordinate-wise:

$$\max_{q_j \ge 0} \sum_j q_j \Sigma_{jj} \quad \text{s.t.} \quad \sum_j c_j q_j^2 \le B.$$

By Cauchy-Schwarz, the maximizer satisfies $q_j^\star \propto \Sigma_{jj}/c_j$. Normalizing by the constraint:

$$q_j^\star = \sqrt{\frac{B}{\sum_k \Sigma_{kk}^2/c_k}} \cdot \frac{\Sigma_{jj}}{c_j}, \quad P^\star_{\mathrm{D}}(\Sigma) = \sqrt{B \sum_j \frac{\Sigma_{jj}^2}{c_j}}.$$

□

B.3 Proof of Theorem 3

Proof of Theorem 3. We establish each claim in turn.

(i) Existence & sublinearity. $\mathcal{Q}$ is weakly compact (Hilbert spaces are reflexive; closed and bounded $\Rightarrow$ weakly compact). The functional $Q \mapsto \langle Q, M\rangle_{\mathcal{H}}$ is continuous in the weak topology (linear functionals are weakly continuous), hence attains its maximum on $\mathcal{Q}$. Sublinearity: $P^\star(M) = \sup_{Q\in\mathcal{Q}}\langle Q, M\rangle_{\mathcal{H}}$ is a supremum of linear maps in M, thus convex and positively homogeneous. Finiteness follows from compactness of $\mathcal{Q}$.

(ii) Conjugacy identities. By definition of convex conjugate in $\mathcal{H}$,

$$(\delta_{\mathcal{Q}})^*(M) \;=\; \sup_{Q\in\mathcal{H}} \{\langle Q, M\rangle_{\mathcal{H}} - \delta_{\mathcal{Q}}(Q)\} \;=\; \sup_{Q\in\mathcal{Q}} \langle Q, M\rangle_{\mathcal{H}} \;=\; P^\star(M).$$

Thus $P^\star = \delta^*_{\mathcal{Q}}$. By definition of the polar, $M \in (\mathcal{Q})^\circ$ iff $\sup_{Q\in\mathcal{Q}}\langle Q, M\rangle_{\mathcal{H}} \le 1$, i.e., $P^\star(M) \le 1$. Therefore

$$\gamma_{(\mathcal{Q})^\circ}(M) = \inf\{\lambda > 0 :\; M \in \lambda(\mathcal{Q})^\circ\} = \inf\{\lambda > 0 :\; P^\star(M) \le \lambda\} = P^\star(M).$$

Finally, $(\gamma_{\mathcal{Q}})^* = \delta_{(\mathcal{Q})^\circ}$ is the standard gauge–polar identity in a locally convex space.

(iii) Construction (subgradient). Let $Q^\star \in \arg\max_{Q\in\mathcal{Q}}\langle Q, M\rangle_{\mathcal{H}}$. For any $N \in \mathcal{H}$,

$$P^\star(N) \;=\; \max_{Q\in\mathcal{Q}}\langle Q, N\rangle_{\mathcal{H}} \;\ge\; \langle Q^\star, N\rangle_{\mathcal{H}} \;=\; P^\star(M) + \langle Q^\star, N - M\rangle_{\mathcal{H}},$$

so $Q^\star \in \partial_M P^\star(M)$. Uniqueness implies differentiability with gradient $Q^\star$.

(iv) Order preservation. If $M \in \mathcal{H}_+$, then for any $Q \in \mathcal{Q} \subset \mathcal{H}_+$ we have $\langle Q, M\rangle_{\mathcal{H}} \ge 0$ (each term $\mathrm{Tr}(H_k^\top M_k) \ge 0$). Taking the max over Q yields $P^\star(M) \ge 0$. If $M_1 - M_2 \in \mathcal{H}_+ + \setminus\{0\}$, some admissible Q gives strict positivity of $\langle Q, M_1 - M_2\rangle_{\mathcal{H}}$, hence the maximized value is strictly larger at M_1 than M_2.

(v) Lipschitz in polar gauge. From (ii), $P^\star = \gamma_{(\mathcal{Q})^\circ}$. Then for any $M, \hat{M}$,

$$\begin{aligned} P^\star(M) - P^\star(\hat{M}) &= \max_{Q\in\mathcal{Q}}\langle Q, M\rangle_{\mathcal{H}} - \max_{Q\in\mathcal{Q}}\langle Q, \hat{M}\rangle_{\mathcal{H}} \\ &\le \max_{Q\in\mathcal{Q}}\langle Q, M - \hat{M}\rangle_{\mathcal{H}} \;=\; \|M - \hat{M}\|_{(\mathcal{Q})^\circ}. \end{aligned}$$

Symmetry gives the absolute value bound. □

B.4 Proof of Lemma 4

Proof of Lemma 4. The proof follows from standard convex optimization theory, specifically the KKT conditions for linear maximization and the characterization of metric projections.

We use two facts:

- *(KKT for linear maximization)* $x^\star \in \arg\max_{y\in\mathcal{C}}\langle y,\, M\rangle_{\mathcal{H}} \iff M \in N_{\mathcal{C}}(x^\star)$.
- *(Metric projection)* For $y \in \mathcal{H}$, $x^\star = \Pi_{\mathcal{C}}(y) \iff y - x^\star \in N_{\mathcal{Q}\cap\mathcal{C}}(x^\star)$.

(ii) $\Rightarrow$ (i): Suppose there exists $M \in \mathcal{Q}^\star_{\mathcal{C}}(R) \subseteq \mathcal{Q}\cap\mathcal{C}$ such that $\{R, Q^\star - M\} \subset N_{\mathcal{Q}\cap\mathcal{C}}(M)$. From $Q^\star - M \in N_{\mathcal{Q}\cap\mathcal{C}}(M)$, the metric projection characterization gives $\Pi_{\mathcal{C}}(Q^\star) = M$. From $R \in N_{\mathcal{Q}\cap\mathcal{C}}(Q^\star_{\mathcal{C}})$, the KKT condition for linear maximization gives $M \in \arg\max_{Q\in\mathcal{Q}\cap\mathcal{C}}\langle Q,\, R\rangle_{\mathcal{H}} = \mathcal{Q}^\star_{\mathcal{C}}(R)$. Therefore, $\Pi_{\mathcal{C}}(Q^\star) = M \in \mathcal{Q}^\star_{\mathcal{C}}(R)$.

(i) $\Rightarrow$ (ii): Suppose $M := \Pi_{\mathcal{C}}(Q^\star) \in \mathcal{Q}^\star_{\mathcal{C}}(R)$. By the metric projection characterization, $Q^\star - M \in N_{\mathcal{Q}\cap\mathcal{C}}(M)$. Since $M \in \mathcal{Q}^\star_{\mathcal{C}}(R) = \arg\max_{Q\in\mathcal{Q}\cap\mathcal{C}}\langle Q,\, R\rangle_{\mathcal{H}}$, the KKT condition for linear maximization gives $R \in N_{\mathcal{Q}\cap\mathcal{C}}(M)$. Thus, $\{R, Q^\star - M\} \subset N_{\mathcal{Q}\cap\mathcal{C}}(M)$. Let $Q^\star_{\mathcal{C}} := M$.

For the final statement, if $N_{\mathcal{Q}\cap\mathcal{C}}(Q^\star_{\mathcal{C}}) = \{\lambda M : \lambda \geq 0\}$ is a ray, then both R and $Q^\star - Q^\star_{\mathcal{C}}$ must be non-negative multiples of the same direction M for the normal-cone alignment condition to hold. □

B.5 Proof of Corollary 5

Proof of Corollary 5. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response $\{q_k\}_{k\geq 0}$ and norm $\|Q\|^2_{\mathcal{H}} = \sum_{k=0}^{\infty}\operatorname{Tr}(q_k^\top q_k)$.

Step 1 — Norm of the 1-pole equalizer. The impulse response is $q_k = \eta P(1-\beta)\beta^k$. By definition,

$$\|Q_{\text{SGD+M}}\|^2_{\mathcal{H}} = \sum_{k\geq 0}\operatorname{Tr}(q_k^\top q_k) = \sum_{k\geq 0}\operatorname{Tr}\left([\eta P(1-\beta)\beta^k]^\top[\eta P(1-\beta)\beta^k]\right) \tag{43}$$

$$= \eta^2(1-\beta)^2\sum_{k\geq 0}\beta^{2k}\operatorname{Tr}(P^\top P) = \eta^2(1-\beta)^2\,\frac{1}{1-\beta^2}\operatorname{Tr}(P^\top P). \tag{44}$$

The budget constraint $\|Q_{\text{SGD+M}}\|_{\mathcal{H}} \leq \sqrt{B}$ imposes

$$\eta \leq \sqrt{B}\left(\operatorname{Tr}(P^\top P)\,\frac{(1-\beta)^2}{1-\beta^2}\right)^{-1/2}. \tag{45}$$

Step 2 — Alignment with the moment operator. The inner product with R is

$$\langle Q_{\text{SGD+M}}, R\rangle_{\mathcal{H}} = \sum_{k\geq 0}\operatorname{Tr}(q_k^\top R_k) = \eta(1-\beta)\sum_{k\geq 0}\beta^k\,\operatorname{Tr}(P^\top R_k) \tag{46}$$

$$= \eta(1-\beta)\sum_{k\geq 0}\beta^k S_k, \tag{47}$$

where $S_k := \operatorname{Tr}(P^\top R_k)$.

Step 3 — Reduce to 1-D search; saturate budget. For fixed β, the inner product is linear in η while the constraint is quadratic, so the maximizer saturates the budget. The budget-normalized gain is

$$J(\beta) := \frac{\langle Q_{\text{SGD+M}}, R\rangle_{\mathcal{H}}}{\|Q_{\text{SGD+M}}\|_{\mathcal{H}}} = \frac{\sqrt{1-\beta^2}}{\sqrt{\operatorname{Tr}(P^\top P)}}\sum_{k\geq 0}\beta^k S_k. \tag{48}$$

Hence $\beta^\star = \arg\max_{0<\beta<1} J(\beta)$ and $\eta^\star$ saturates the budget constraint. □

B.6 Proof of Corollary 6

Proof of Corollary 6. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with impulse response $\{H_k\}_{k\geq 0}$ and norm $\|Q\|^2_{\mathcal{H}} = \sum_k \operatorname{Tr}(H_k^\top H_k)$.

Step 1 — Unconstrained diagonal optimizer. With the diagonal budget $\mathcal{Q}_D(B, c)$, the maximization $\max_{Q\in\mathcal{Q}_D}\langle Q, R\rangle_{\mathcal{H}}$ decouples per coordinate into scalar H_2 subproblems:

$$\max_{q_j:\sum_j c_j\|q_j\|_{H_2}^2\le B}\sum_{j=1}^{P}\langle q_j, R_{jj}\rangle_{H_2}. \tag{49}$$

By Cauchy-Schwarz, the optimizer is proportional to R_{jj} with weights $1/c_j$, giving the stated $q_j^\star$.

Step 2 — Adam-family restriction (1-pole per coordinate). Fix a fitting window and freeze the preconditioner $D_t^{-1/2}$ (quasi-LTI approximation). The per-coordinate impulse response is

$$h_{j,k} = \eta(D_t^{-1/2})_{jj}(1-\beta_1)\beta_1^k, \qquad k \ge 0. \tag{50}$$

Step 3 — Budget norm and alignment. The budget norm evaluates to

$$\|Q_{\text{Adam}}\|_{\mathcal{H}}^2 = \sum_{j=1}^{P} c_j \sum_{k\ge 0}|h_{j,k}|^2 = \eta^2\frac{(1-\beta_1)^2}{1-\beta_1^2}\sum_{j=1}^{P} c_j (D_t^{-1/2})_{jj}^2 = \eta^2\frac{(1-\beta_1)^2}{1-\beta_1^2}W(\beta_2). \tag{51}$$

The alignment with R is

$$\langle Q_{\text{Adam}}, R\rangle_{\mathcal{H}} = \eta(1-\beta_1)\sum_{k\ge 0}\beta_1^k\sum_{j=1}^{P}(D_t^{-1/2})_{jj}R_{jj,k} = \eta(1-\beta_1)\sum_{k\ge 0}\beta_1^k T_k(\beta_2). \tag{52}$$

Step 4 — Optimization over hyperparameters. For fixed (β_1, β_2), the maximizer in η saturates the budget (linear gain under quadratic constraint). The budget-normalized gain is

$$J(\beta_1,\beta_2) := \frac{\langle Q_{\text{Adam}}, R\rangle_{\mathcal{H}}}{\|Q_{\text{Adam}}\|_{\mathcal{H}}} = \frac{\sqrt{1-\beta_1^2}}{\sqrt{W(\beta_2)}}\sum_{k\ge 0}\beta_1^k T_k(\beta_2). \tag{53}$$

Maximizing over $(\beta_1, \beta_2) \in (0, 1)^2$ gives the optimal momentum parameters, and the optimal learning rate is

$$\eta^\star = \sqrt{B}\left(\frac{(1-\beta_1^\star)^2}{1-(\beta_1^\star)^2}W(\beta_2^\star)\right)^{-1/2}. \tag{54}$$

□

B.7 Proof of Proposition 7

Proof of Proposition 7. The instantaneous validation slope is $Q \mapsto \operatorname{Tr}(Q\Sigma_\times)$. Since this is linear in Q and $\mathcal{Q}$ is compact, the maximum is attained and gives the claimed inequality. If the maximizer is unique, then any suboptimal Q gives strictly smaller $\operatorname{Tr}(Q\Sigma_\times) = -\dot{\mathcal{L}}_{\text{val}}(Q)$. This includes optimizers only tuned to the training set. □

B.8 Proof of Proposition 8

Proof of Proposition 8. In the linearized regime, we fix parameters θ_0 and consider the linearized network $f(x) \approx f(x;\theta_0) + \nabla_\theta f(x;\theta_0)^\top(\theta-\theta_0)$. For squared loss $\mathcal{L}(\theta) = \frac{1}{2}\|f_\theta(X) - y\|^2$, the gradient is $\nabla_\theta\mathcal{L} = \sum_{i=1}^n (f_\theta(x_i) - y_i)\nabla_\theta f_\theta(x_i)$.

With optimizer Q, the parameter dynamics are $\dot\theta = -Q\nabla_\theta\mathcal{L}$. In the linearized regime around θ_0, this becomes

$$\dot\theta = -Q\sum_{i=1}^{n}(f_{\theta_0}(x_i) + \nabla_\theta f(x_i;\theta_0)^\top(\theta-\theta_0) - y_i)\nabla_\theta f(x_i;\theta_0). \tag{55}$$

Let $G \in \mathbb{R}^{n\times d}$ be the matrix with rows $\nabla_\theta f(x_i;\theta_0)^\top$. Then the linearized dynamics become

$$\dot\theta = -QG^\top(G(\theta-\theta_0) + (f_{\theta_0}(X) - y)), \tag{56}$$

where $f_{\theta_0}(X) = [f_{\theta_0}(x_1), \ldots, f_{\theta_0}(x_n)]^\top$.

In function space, let $u(x) = \nabla_\theta f(x;\theta_0)^\top(\theta - \theta_0)$ represent the change in function values. Then $u(X) = G(\theta - \theta_0)$ and the dynamics become

$$\frac{d}{dt}u(X) = G\dot{\theta} = -GQG^\top(u(X) + (f_{\theta_0}(X) - y)). \tag{57}$$

This is kernel gradient flow in function space with kernel matrix $K = GQG^\top$, where $K_{ij} = \nabla_\theta f(x_i;\theta_0)^\top Q \nabla_\theta f(x_j;\theta_0) = K_Q(x_i, x_j)$.

By Moore–Penrose pseudoinverse, this gradient flow converges to $u^\star = K^\dagger y$ where $K^\dagger = (GQG^\top)^\dagger$, which corresponds to the minimum norm interpolant in the RKHS $\mathcal{H}_{K_Q}$ induced by kernel K_Q:

$$f_Q^\star = \arg\min_{f \in \mathcal{H}_{K_Q}} \|f\|^2_{\mathcal{H}_{K_Q}} \quad \text{s.t.} \quad f(X) = y. \tag{58}$$

The choice of Q directly determines the kernel $K_Q(x, x') = \nabla_\theta f(x;\theta_0)^\top Q \nabla_\theta f(x';\theta_0)$ and hence the RKHS $\mathcal{H}_{K_Q}$. The validation-aware choice $Q^\star \in \arg\max_{Q\in\mathcal{Q}} \mathrm{Tr}(Q\,\Sigma_\times)$ emphasizes directions in parameter space that are aligned with the validation cross-moment $\Sigma_\times$, thereby tilting the induced RKHS toward functions that perform better on validation data. □

B.9 Proof of Corollary 9

Proof of Corollary 9. We apply Theorem 3 to each dynamic budget set. Let $R(e^{i\omega}) = U(\omega)\,\mathrm{diag}(\sigma_1(\omega) \geq \cdots \geq \sigma_d(\omega))U(\omega)^*$ be the eigendecomposition of the moment operator at each frequency.

(i) Frobenius ball $\mathcal{Q}_F(B) = \{Q : \|Q\|_{\mathcal{H}} \leq \sqrt{B}\}$. The Lagrangian is $L(Q,\lambda) = \langle Q, R\rangle_{\mathcal{H}} - \lambda(\|Q\|^2_{\mathcal{H}} - B)$. Taking the functional derivative with respect to Q and setting to zero:

$$\delta_Q L = R - 2\lambda Q = 0 \quad \Rightarrow \quad Q = \frac{R}{2\lambda}.$$

The constraint $\|Q\|_{\mathcal{H}} = \sqrt{B}$ gives $\|R/(2\lambda)\|_{\mathcal{H}} = \sqrt{B}$, so $2\lambda = \|R\|_{\mathcal{H}}/\sqrt{B}$. Hence:

$$Q_F^\star = \sqrt{B}\,\frac{R}{\|R\|_{\mathcal{H}}}, \quad P_F^\star(R) = \langle Q_F^\star, R\rangle_{\mathcal{H}} = \sqrt{B}\,\|R\|_{\mathcal{H}}.$$

(ii) Per-frequency spectral $\mathcal{Q}_S(\tau,\lambda) = \{Q : \mathrm{Tr}(Q(e^{i\omega})) \leq \tau(\omega), Q(e^{i\omega}) \preceq \lambda(\omega)I\}$. By the dynamic version of Neumann's inequality, the maximizer has the form $Q(e^{i\omega}) = U(\omega)\,\mathrm{diag}(q_i(\omega))U(\omega)^*$ where the eigenvalues $q_i(\omega)$ solve the water-filling problem at each frequency:

$$\max_{q_i(\omega)\geq 0} \sum_i q_i(\omega)\sigma_i(\omega) \quad \text{s.t.} \quad \sum_i q_i(\omega) \leq \tau(\omega),\ q_i(\omega) \leq \lambda(\omega).$$

The KKT conditions yield: (i) $q_i^\star(\omega) = \lambda(\omega)$ for $i \leq k(\omega)$, (ii) $q_{k+1}^\star(\omega) = \tau(\omega) - k(\omega)\lambda(\omega)$, (iii) $q_i^\star(\omega) = 0$ for $i > k(\omega) + 1$, where $k(\omega) = \lfloor \tau(\omega)/\lambda(\omega) \rfloor$. The optimal power is:

$$P_S^\star(R) = \frac{1}{2\pi}\int_0^{2\pi} \left[\lambda(\omega) \sum_{i\leq k(\omega)} \sigma_i(\omega) + (\tau(\omega) - k(\omega)\lambda(\omega))\sigma_{k(\omega)+1}(\omega)\right] d\omega.$$

(iii) Data-metric $\mathcal{Q}_L(B) = \{Q : \sum_{k=0}^\infty \mathrm{Tr}(Q_k^* \Sigma Q_k) \leq B\}$. The Lagrangian is $L(Q,\mu) = \langle Q, R\rangle_{\mathcal{H}} - \mu\,(\sum_{k=0}^\infty \mathrm{Tr}(Q_k^*\Sigma Q_k) - B)$. The first-order condition gives:

$$R - 2\mu \sum_{k=0}^\infty \Sigma Q_k e^{-ik\omega} = 0 \quad \Rightarrow \quad Q = \frac{1}{2\mu} I \text{ on } \mathrm{supp}(R).$$

Using the constraint and the fact that Q is constant on the support:

$$\alpha^2 \sum_{i:\sigma_i>0} \sigma_i = B \quad \Rightarrow \quad \alpha = \sqrt{\frac{B}{\sum_{i:\sigma_i>0}\sigma_i}}.$$

Hence:

$$Q_L^\star = \alpha\,\Pi_{\mathrm{supp}(R)}, \quad P_L^\star(R) = \alpha \sum_i \sigma_i = \sqrt{B \sum_i \sigma_i}.$$

(iv) Diagonal $\mathcal{Q}_D(B,c) = \{Q(z) = \mathrm{diag}(q_j(z)) \succeq 0 : \sum_j c_j \|q_j\|_{H_2}^2 \le B\}$. The problem decouples coordinate-wise into scalar H_2 subproblems:

$$\max_{q_j \ge 0} \sum_j \langle q_j, R_{jj} \rangle_{H_2} \quad \text{s.t.} \quad \sum_j c_j \|q_j\|_{H_2}^2 \le B.$$

By Cauchy-Schwarz in H_2, the maximizer satisfies $q_j^\star \propto R_{jj}/c_j$. Normalizing by the constraint:

$$q_j^\star = \sqrt{\frac{B}{\sum_k \|R_{kk}\|_{H_2}^2/c_k}} \cdot \frac{R_{jj}}{c_j}, \quad P_D^\star(R) = \sqrt{B \sum_j \frac{\|R_{jj}\|_{H_2}^2}{c_j}}.$$

□

C Reverse engineering common optimizers

So far, we have derived various types of stateless and stateful optimizers under different types of budgets. In this section, we will do the opposite: we will *reverse engineer* popular optimizers and find out under which budget they are secretly optimizing. This not only allows us to find out hidden design principles of these optimizers, but also have these optimizers registered in a unified framework, suggesting a systematic way to design new optimizers.

tl;dr:

- *GD* = Euclidean budget $\Rightarrow Q \propto I$.
- *Colored-GD* = elliptic budget $\Rightarrow Q \propto P$.
- *Newton/GN* = curvature budget $\Rightarrow Q \propto H^{-1}/G^{-1}$.
- *NGD* = Fisher/KL budget $\Rightarrow Q \propto F^{-1}$.
- *K-FAC/Shampoo* = structured budgets $\Rightarrow$ block/Kronecker Q.
- *AdaGrad/RMSProp/Adam* = diagonal budgets (on g or m_k) $\Rightarrow Q$ diagonal.
- *LAMB/LARS* = layer-norm budget $\Rightarrow$ layer-wise scalar Q.
- *signSGD/Lion* = L_∞ budget (on g or m_k) $\Rightarrow$ normalized/sign steps.

One lens. Every first-order optimizer picks a velocity $\dot\theta$ from the current gradient g by solving a *budgeted power allocation*:

$$\dot\theta \;=\; Q\,g \quad \text{with} \quad Q \in \mathcal{Q} \iff \dot\theta = \arg\max_{\dot\theta \in \mathbb{R}^P} \langle g, \dot\theta \rangle \;\text{ s.t. }\; \mathrm{budget}(\dot\theta) \le \tau \tag{59}$$

The *budget* determines the *optimizer* Q. Below, we list each popular optimizer as a special case of this formulation, give the induced Q, and state (not prove) the short KKT step that produces it.

C.1 Euclidean & curvature families

Corollary 10 (SGD from Euclidean Frobenius budget)**.** *Define the Euclidean Frobenius budget and the cone of memoryless isotropic optimizers:*

$$\mathcal{Q}_F(B) \;\coloneqq\; \{Q : \|Q\|_{\mathcal{H}} \le \sqrt{B}\}, \quad \mathcal{C}_{\textit{memoryless}} \;\coloneqq\; \{Q[n] = \eta I \delta[n] : \eta \ge 0\}. \tag{60}$$

Given current gradient moment $R[0] = gg^\top$ where g is the instantaneous gradient, solving problem P3 under the budget $\mathcal{Q}_F(B) \cap \mathcal{C}_{\textit{memoryless}}$ produces SGD as the optimal solution with optimal hyperparameter:

$$\eta^\star \;=\; \frac{\sqrt{B}}{\sqrt{d}\|g\|_2}, \tag{61}$$

where d is the dimension of the parameter space.

Proof. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response Q and Frobenius $\mathcal{H}$ norm $\|Q\|_{\mathcal{H}}^2 = \sum_{n=0}^{\infty} \mathrm{Tr}(Q[n]^\top Q[n])$.

Step 1 — Global optimal equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under budget $\mathcal{Q}_F(B) = \{Q : \|Q\|_{\mathcal{H}} \le \sqrt{B}\}$ is

$$Q^C = \sqrt{B}\,\frac{R}{\|R\|_{\mathcal{H}}}, \qquad P^\star(R) = \sqrt{B}\,\|R\|_{\mathcal{H}}. \tag{62}$$

Step 2 — Commutativity via smooth convexity. The Frobenius ball $\mathcal{Q}_F(B)$ is smooth and strictly convex. At any boundary point Q with $\|Q\|_{\mathcal{H}} = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_F(B)}(Q) = \{\lambda Q : \lambda \ge 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_F(B)\cap\mathcal{C}_{\text{memoryless}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_F(B)\cap\mathcal{C}_{\text{memoryless}}} \langle Q, R\rangle_{\mathcal{H}}. \tag{63}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless family parametrization. For $Q_\eta[n] = \eta I\delta[n]$ with $\eta \ge 0$, we compute:

$$\|Q_\eta\|_{\mathcal{H}}^2 = \mathrm{Tr}((\eta I)^\top(\eta I)) = \eta^2 d, \tag{64}$$

and for the instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}} = \mathrm{Tr}((\eta I)^\top(gg^\top)) = \eta\,\mathrm{Tr}(gg^\top) = \eta\|g\|_2^2. \tag{65}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta\|_{\mathcal{H}} = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{d}}. \tag{66}$$

However, when we normalize by the gradient magnitude for scale invariance, we obtain:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{d}\|g\|_2}. \tag{67}$$

Step 5 — Geometric interpretation. SGD emerges as the memoryless isotropic approximation of the global Frobenius-constrained equalizer. The optimal learning rate $\eta^\star$ balances the budget constraint with the current gradient magnitude, providing uniform scaling across all parameter dimensions. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{memoryless}}$. □

Corollary 11 (Colored gradient descent from elliptic trust region budget)**.** *Define the elliptic budget $\mathcal{Q}_P(B)$ and the cone $\mathcal{C}_{memoryless}$ of memoryless optimizers:*

$$\mathcal{Q}_P(B) \;\coloneqq\; \{Q : \mathrm{Tr}(Q[0]^\top P^{-1} Q[0]) \le B, Q[n] = 0 \text{ for } n > 0\}, \tag{68}$$

$$\mathcal{C}_{memoryless} \;\coloneqq\; \{Q[n] = \eta I\delta[n] : \eta \ge 0\}. \tag{69}$$

where $P \succ 0$ is a fixed symmetric positive definite matrix. Given current gradient moment $R[0] = gg^\top$ where g is the instantaneous gradient, solving problem P3 under the budget $\mathcal{Q}_P(B)\cap\mathcal{C}_{memoryless}$ produces colored gradient descent as the optimal solution with optimal hyperparameter:

$$\eta^\star \;=\; \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(P^{-1})}\|g\|_2}, \tag{70}$$

where the optimizer is $Q^\star = \eta^\star P$.

Proof. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response Q and elliptic P-weighted norm for memoryless filters $\|Q[0]\|_{P^{-1}}^2 = \mathrm{Tr}(Q[0]^\top P^{-1} Q[0])$.

Step 1 — Global optimal equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under budget $\mathcal{Q}_P(B) = \{Q : \mathrm{Tr}(Q[0]^\top P^{-1} Q[0]) \le B, Q[n] = 0 \text{ for } n > 0\}$ is

$$Q^C[0] = \sqrt{B}\,\frac{Pgg^\top}{\|Pgg^\top\|_{P^{-1}}}, \qquad P^\star(R) = \sqrt{B}\,\|Pgg^\top\|_{P^{-1}}. \tag{71}$$

Step 2 — Commutativity via smooth convexity. The elliptic ball $\mathcal{Q}_P(B)$ is smooth and strictly convex. At any boundary point Q with $\|Q[0]\|_{P^{-1}} = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_P(B)}(Q) = \{\lambda P^{-1} Q[0] : \lambda \ge 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_P(B)\cap\mathcal{C}_{\text{memoryless}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_P(B)\cap\mathcal{C}_{\text{memoryless}}} \langle Q, R\rangle_{\mathcal{H}}. \tag{72}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless family parametrization. For $Q_\eta[n] = \eta P \delta[n]$ with $\eta \ge 0$, we compute:

$$\|Q_\eta[0]\|^2_{P^{-1}} = \mathrm{Tr}((\eta P)^\top P^{-1}(\eta P)) = \eta^2\,\mathrm{Tr}(P), \tag{73}$$

and for the instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}} = \mathrm{Tr}((\eta P)^\top (gg^\top)) = \eta\,\mathrm{Tr}(Pgg^\top) = \eta g^\top P g. \tag{74}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta[0]\|_{P^{-1}} = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(P)}}. \tag{75}$$

However, when we normalize by the gradient magnitude for scale invariance, we obtain:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(P^{-1})}\|g\|_2}. \tag{76}$$

Step 5 — Geometric interpretation. Colored gradient descent emerges as the memoryless approximation of the global elliptic-constrained equalizer. The optimal learning rate $\eta^\star$ balances the elliptic budget constraint with the current gradient magnitude, providing P-weighted scaling across parameter dimensions. The optimizer $Q^\star = \eta^\star P$ naturally incorporates the geometry encoded in matrix P. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{memoryless}}$. □

Corollary 12 (Newton's method from curvature-aware budget)**.** *Define the budget $\mathcal{Q}_H(B)$ and the cone $\mathcal{C}_{\text{memoryless}}$ of memoryless optimizers:*

$$\mathcal{Q}_H(B) \;\coloneqq\; \{Q : \mathrm{Tr}(Q[0]^\top H Q[0]) \le B, Q[n] = 0 \text{ for } n > 0\}, \tag{77}$$

$$\mathcal{C}_{\text{memoryless}} \;\coloneqq\; \{Q_\eta[n] = \eta P \delta[n] : \eta \ge 0, P \succ 0\}. \tag{78}$$

where $H \succ 0$ is the Hessian matrix. Given current gradient moment $R[0] = gg^\top$ where g is the instantaneous gradient, solving problem P3 under the budget $\mathcal{Q}_H(B)\cap\mathcal{C}_{\text{memoryless}}$ produces Newton's method as the optimal solution with optimal hyperparameter:

$$\eta^\star \;=\; \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(H^{-1})}\|g\|_2}, \tag{79}$$

where the optimizer is $Q^\star = \eta^\star H^{-1}$.

Proof. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response Q and elliptic H-weighted norm for memoryless filters $\|Q[0]\|^2_H = \mathrm{Tr}(Q[0]^\top H Q[0])$.

Step 1 — Global optimal equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under budget $\mathcal{Q}_H(B) = \{Q : \mathrm{Tr}(Q[0]^\top H Q[0]) \le B, Q[n] = 0 \text{ for } n > 0\}$ is

$$Q^C[0] = \sqrt{B}\,\frac{H^{-1}gg^\top}{\|H^{-1}gg^\top\|_H}, \qquad P^\star(R) = \sqrt{B}\,\|H^{-1}gg^\top\|_H. \tag{80}$$

Step 2 — Commutativity via smooth convexity. The elliptic ball $\mathcal{Q}_H(B)$ is smooth and strictly convex. At any boundary point Q with $\|Q[0]\|_H = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_H(B)}(Q) = \{\lambda HQ[0] : \lambda \ge 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_H(B)\cap\mathcal{C}_{\text{memoryless}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_H(B)\cap\mathcal{C}_{\text{memoryless}}} \langle Q, R\rangle_{\mathcal{H}}. \tag{81}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless family parametrization. For $Q_\eta[n] = \eta H^{-1}\delta[n]$ with $\eta \ge 0$, we compute:

$$\|Q_\eta[0]\|_H^2 = \mathrm{Tr}((\eta H^{-1})^\top H(\eta H^{-1})) = \eta^2\,\mathrm{Tr}(H^{-1}), \tag{82}$$

and for the instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}} = \mathrm{Tr}((\eta H^{-1})^\top(gg^\top)) = \eta\,\mathrm{Tr}(H^{-1}gg^\top) = \eta g^\top H^{-1} g. \tag{83}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta[0]\|_H = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(H^{-1})}}. \tag{84}$$

When we normalize by the gradient magnitude for scale invariance, we obtain:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(H^{-1})}\|g\|_2}. \tag{85}$$

Step 5 — Geometric interpretation. Newton's method emerges as the memoryless approximation of the global Hessian-constrained equalizer. The optimal learning rate $\eta^\star$ balances the curvature budget constraint with the current gradient magnitude, providing Hessian-weighted scaling that naturally incorporates second-order geometry. The optimizer $Q^\star = \eta^\star H^{-1}$ captures the local quadratic structure of the loss landscape. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{memoryless}}$. □

Corollary 13 (L-BFGS from learned curvature approximation)**.** *Define the budget $\mathcal{Q}_B(B)$ and the cone $\mathcal{C}_{memoryless}$ of memoryless optimizers:*

$$\mathcal{Q}_B(B) \;\coloneqq\; \{Q : \mathrm{Tr}(Q[0]^\top B_k Q[0]) \le B, Q[n] = 0 \,\textit{for}\, n > 0\}, \tag{86}$$

$$\mathcal{C}_{memoryless} \;\coloneqq\; \{Q_\eta[n] = \eta B_k^{-1}\delta[n] : \eta \ge 0\}. \tag{87}$$

where $B_k \succ 0$ is the L-BFGS curvature approximation matrix constructed through secant updates. Given current gradient moment $R[0] = gg^\top$ where g is the instantaneous gradient, solving problem P3 under the budget $\mathcal{Q}_B(B) \cap \mathcal{C}_{memoryless}$ produces L-BFGS as the optimal solution with optimal hyperparameter:

$$\eta^\star \;=\; \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(B_k^{-1})}\|g\|_2}, \tag{88}$$

where the optimizer is $Q^\star = \eta^\star B_k^{-1}$.

Proof. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response Q and elliptic B_k-weighted norm for memoryless filters $\|Q[0]\|_{B_k}^2 = \mathrm{Tr}(Q[0]^\top B_k Q[0])$.

Step 1 — Global optimal equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under budget $\mathcal{Q}_B(B) = \{Q : \mathrm{Tr}(Q[0]^\top B_k Q[0]) \le B, Q[n] = 0 \text{ for } n > 0\}$ is

$$Q^C[0] = \sqrt{B}\,\frac{B_k^{-1} g g^\top}{\|B_k^{-1} g g^\top\|_{B_k}}, \qquad P^\star(R) = \sqrt{B}\,\|B_k^{-1} g g^\top\|_{B_k}. \tag{89}$$

Step 2 — Commutativity via smooth convexity. The elliptic ball $\mathcal{Q}_B(B)$ is smooth and strictly convex. At any boundary point Q with $\|Q[0]\|_{B_k} = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_B(B)}(Q) = \{\lambda B_k Q[0] : \lambda \ge 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_B(B)\cap\mathcal{C}_{\text{memoryless}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_B(B)\cap\mathcal{C}_{\text{memoryless}}} \langle Q, R\rangle_{\mathcal{H}}. \tag{90}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless family parametrization. For $Q_\eta[n] = \eta B_k^{-1}\delta[n]$ with $\eta \ge 0$, we compute:

$$\|Q_\eta[0]\|_{B_k}^2 = \mathrm{Tr}((\eta B_k^{-1})^\top B_k (\eta B_k^{-1})) = \eta^2\,\mathrm{Tr}(B_k^{-1}), \tag{91}$$

and for the instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}} = \mathrm{Tr}((\eta B_k^{-1})^\top (gg^\top)) = \eta\,\mathrm{Tr}(B_k^{-1} gg^\top) = \eta g^\top B_k^{-1} g. \tag{92}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta[0]\|_{B_k} = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(B_k^{-1})}}. \tag{93}$$

When we normalize by the gradient magnitude for scale invariance, we obtain:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(B_k^{-1})}\|g\|_2}. \tag{94}$$

Step 5 — Geometric interpretation. L-BFGS emerges as the memoryless approximation of the global curvature-constrained equalizer, where the curvature matrix B_k is learned through secant updates rather than computed exactly. The optimal learning rate $\eta^\star$ balances the learned curvature budget constraint with the current gradient magnitude, providing B_k-weighted scaling that incorporates approximate second-order geometry at reduced computational cost. The optimizer $Q^\star = \eta^\star B_k^{-1}$ captures the accumulated curvature information from the optimization trajectory. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{memoryless}}$. □

C.2 Information-geometric & structured optimizers

Corollary 14 (Natural Gradient Descent from Fisher information geometry)**.** *Define the Fisher information budget $\mathcal{Q}_F(B, F)$ and the cone $\mathcal{C}_{\text{memoryless}}$ of memoryless optimizers:*

$$\mathcal{Q}_F(B, F) \;\coloneqq\; \{Q : \mathrm{Tr}(Q^\top F Q) \le B\}, \quad \mathcal{C}_{\text{memoryless}} \;\coloneqq\; \{Q_\eta = \eta I : \eta \ge 0\}. \tag{95}$$

Given instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$, solving problem P3 under the budget $\mathcal{Q}_F(B, F)\cap\mathcal{C}_{\text{memoryless}}$ produces Natural Gradient Descent as the optimal solution with optimal hyperparameters:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(F^{-1})}\|g\|_2}, \qquad Q^\star = \eta^\star F^{-1}. \tag{96}$$

Proof. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response Q and Fisher-weighted norm for memoryless filters $\|Q\|_F^2 = \mathrm{Tr}(Q^\top F Q)$.

Step 1 — Global optimal equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under budget $\mathcal{Q}_F(B,F) = \{Q : \mathrm{Tr}(Q^\top F Q) \le B\}$ is

$$Q^C = \sqrt{B}\,\frac{F^{-1}gg^\top}{\|F^{-1}gg^\top\|_F}, \qquad P^\star(R) = \sqrt{B}\,\|F^{-1}gg^\top\|_F. \tag{97}$$

Step 2 — Commutativity via smooth convexity. The Fisher ellipsoid $\mathcal{Q}_F(B,F)$ is smooth and strictly convex. At any boundary point Q with $\|Q\|_F = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_F(B,F)}(Q) = \{\lambda FQ : \lambda \ge 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_F(B,F)\cap\mathcal{C}_{\text{memoryless}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_F(B,F)\cap\mathcal{C}_{\text{memoryless}}} \langle Q, R\rangle_{\mathcal{H}}. \tag{98}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless family parametrization. For $Q_\eta = \eta I$ with $\eta \ge 0$, we compute:

$$\|Q_\eta\|_F^2 = \mathrm{Tr}((\eta I)^\top F(\eta I)) = \eta^2\,\mathrm{Tr}(F), \tag{99}$$

and for the instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}} = \mathrm{Tr}((\eta I)^\top(gg^\top)) = \eta\,\mathrm{Tr}(gg^\top) = \eta g^\top g = \eta\|g\|_2^2. \tag{100}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta\|_F = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(F)}}. \tag{101}$$

When we normalize by the gradient magnitude for scale invariance, we obtain:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\mathrm{Tr}(F)}\|g\|_2}. \tag{102}$$

Step 5 — Geometric interpretation. Natural Gradient Descent emerges as the memoryless approximation of the global Fisher-constrained equalizer, where the Fisher information matrix F captures the intrinsic Riemannian geometry of the statistical model. The optimal learning rate $\eta^\star$ balances the Fisher information budget constraint with the current gradient magnitude, providing F-weighted scaling that incorporates the natural geometry of the parameter space. The optimizer $Q^\star = \eta^\star F^{-1}$ implements steepest descent in the natural Riemannian metric, where the Fisher metric measures the intrinsic difficulty of distinguishing nearby parameter values based on the data distribution. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{memoryless}}$. □

Corollary 15 (K-FAC from block-diagonal Fisher approximation)**.** *Define the block-diagonal Fisher budget $\mathcal{Q}_{block}(B,\{F_\ell\})$ and the cone $\mathcal{C}_{memoryless}$ of memoryless optimizers:*

$$\mathcal{Q}_{block}(B,\{F_\ell\}) \;\coloneqq\; \{Q : \sum_\ell \|Q_\ell\|_{F_\ell}^2 \le B\}, \quad \mathcal{C}_{memoryless} \;\coloneqq\; \{Q[n] = \eta I\delta[n] : \eta \ge 0\}. \tag{103}$$

Given instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$, solving problem P3 under the budget $\mathcal{Q}_{block}(B,\{F_\ell\}) \cap \mathcal{C}_{memoryless}$ produces K-FAC as the optimal solution with optimal hyperparameters:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\sum_\ell \mathrm{Tr}(F_\ell^{-1})}\|g\|_2}, \qquad Q^\star = \eta^\star \mathrm{blockdiag}(F_\ell^{-1}). \tag{104}$$

Proof. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response Q and block-diagonal Fisher-weighted norm $\|Q\|_{\text{block}}^2 = \sum_\ell \mathrm{Tr}(Q_\ell^\top F_\ell Q_\ell)$.

Step 1 — Global optimal equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under budget $\mathcal{Q}_{\text{block}}(B, \{F_\ell\}) = \{Q : \sum_\ell \|Q_\ell\|_{F_\ell}^2 \le B\}$ is

$$Q^C = \sqrt{B}\,\frac{\text{blockdiag}(F_\ell^{-1})gg^\top}{\|\text{blockdiag}(F_\ell^{-1})gg^\top\|_{\text{block}}}, \qquad P^\star(R) = \sqrt{B}\,\|\text{blockdiag}(F_\ell^{-1})gg^\top\|_{\text{block}}. \quad (105)$$

Step 2 — Commutativity via smooth convexity. The block-diagonal Fisher ellipsoid $\mathcal{Q}_{\text{block}}(B, \{F_\ell\})$ is smooth and strictly convex. At any boundary point Q with $\|Q\|_{\text{block}} = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_{\text{block}}(B,\{F_\ell\})}(Q) = \{\lambda \text{blockdiag}(F_\ell)Q : \lambda \ge 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_{\text{block}}(B,\{F_\ell\})\cap\mathcal{C}_{\text{memoryless}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_{\text{block}}(B,\{F_\ell\})\cap\mathcal{C}_{\text{memoryless}}} \langle Q, R\rangle_{\mathcal{H}}. \quad (106)$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless family parametrization. For $Q_\eta = \eta I$ with $\eta \ge 0$, we compute:

$$\|Q_\eta\|_{\text{block}}^2 = \sum_\ell \text{Tr}((\eta I_\ell)^\top F_\ell(\eta I_\ell)) = \eta^2 \sum_\ell \text{Tr}(F_\ell), \quad (107)$$

and for the instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}} = \text{Tr}((\eta I)^\top (gg^\top)) = \eta\,\text{Tr}(gg^\top) = \eta\|g\|_2^2. \quad (108)$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta\|_{\text{block}} = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\sum_\ell \text{Tr}(F_\ell)}}. \quad (109)$$

When we normalize by the gradient magnitude for scale invariance, we obtain:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\sum_\ell \text{Tr}(F_\ell)}\|g\|_2}. \quad (110)$$

Step 5 — Geometric interpretation. K-FAC emerges as the memoryless approximation of the global block-diagonal Fisher-constrained equalizer, where each block F_ℓ captures the layer-wise Fisher information geometry. The optimal learning rate $\eta^\star$ balances the block-diagonal Fisher information budget constraint with the current gradient magnitude, providing layer-wise F_ℓ-weighted scaling that incorporates the natural geometry while maintaining computational tractability through block-diagonal structure. The optimizer $Q^\star = \eta^\star \text{blockdiag}(F_\ell^{-1})$ implements approximate steepest descent in the block-diagonal natural Riemannian metric. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{memoryless}}$. □

Corollary 16 (Shampoo from Kronecker-factored preconditioning)**.** *Consider weight tensors $\theta \in \mathbb{R}^{d_1\times d_2\times\cdots\times d_k}$ and define mode-wise second moment matrices $G_i \in \mathbb{R}^{d_i\times d_i}$ for each mode i. Define the Kronecker-factored budget constraint:*

$$\mathcal{Q}_{Kron}(B, \{G_i\}) \;:=\; \left\{Q : \|Q\|_{Kron}^2 \le B\right\}, \quad (111)$$

where $\|Q\|_{Kron}^2 = \sum_{n=0}^{\infty} \text{Tr}\left(Q[n]^\top \left(\bigotimes_i G_i\right) Q[n]\right)$, and the memoryless Kronecker cone:

$$\mathcal{C}_{Kron} \;:=\; \left\{Q[n] = \eta\left(\bigotimes_i G_i^{-1/2}\right)\delta[n] : \eta \ge 0\right\}. \quad (112)$$

Given instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$, solving problem P3 under the budget $\mathcal{Q}_{Kron}(B, \{G_i\}) \cap \mathcal{C}_{Kron}$ produces Shampoo as the optimal solution with optimal hyperparameters:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\text{Tr}\left(\bigotimes_i G_i^{-1}\right)}\|g\|_2}, \qquad Q^\star = \eta^\star \bigotimes_i G_i^{-1/2}. \quad (113)$$

Proof. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response Q and Kronecker-weighted norm $\|Q\|_{\text{Kron}}^2 = \sum_{n=0}^{\infty} \operatorname{Tr}\left(Q[n]^\top \left(\bigotimes_i G_i\right) Q[n]\right)$.

Step 1 — Global optimal equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under budget $\mathcal{Q}_{\text{Kron}}(B, \{G_i\}) = \{Q : \|Q\|_{\text{Kron}} \le \sqrt{B}\}$ is

$$Q^C = \sqrt{B}\,\frac{\left(\bigotimes_i G_i^{-1}\right) gg^\top}{\left\|\left(\bigotimes_i G_i^{-1}\right) gg^\top\right\|_{\text{Kron}}}, \qquad P^\star(R) = \sqrt{B}\left\|\left(\bigotimes_i G_i^{-1}\right) gg^\top\right\|_{\text{Kron}}. \tag{114}$$

Step 2 — Commutativity via smooth convexity. The Kronecker-factored ellipsoid $\mathcal{Q}_{\text{Kron}}(B, \{G_i\})$ is smooth and strictly convex. At any boundary point Q with $\|Q\|_{\text{Kron}} = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_{\text{Kron}}(B,\{G_i\})}(Q) = \{\lambda \left(\bigotimes_i G_i\right) Q : \lambda \ge 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_{\text{Kron}}(B,\{G_i\})\cap\mathcal{C}_{\text{Kron}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_{\text{Kron}}(B,\{G_i\})\cap\mathcal{C}_{\text{Kron}}} \langle Q, R\rangle_{\mathcal{H}}. \tag{115}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless Kronecker family parametrization. For $Q_\eta = \eta \bigotimes_i G_i^{-1/2}$ with $\eta \ge 0$, we compute:

$$\|Q_\eta\|_{\text{Kron}}^2 = \operatorname{Tr}\left(\left(\eta \bigotimes_i G_i^{-1/2}\right)^\top \left(\bigotimes_i G_i\right)\left(\eta \bigotimes_i G_i^{-1/2}\right)\right) = \eta^2 \operatorname{Tr}\left(\bigotimes_i G_i^{-1}\right), \tag{116}$$

and for the instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}} = \operatorname{Tr}\left(\left(\eta \bigotimes_i G_i^{-1/2}\right)^\top (gg^\top)\right) = \eta \operatorname{Tr}\left(\left(\bigotimes_i G_i^{-1/2}\right) gg^\top\right). \tag{117}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta\|_{\text{Kron}} = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\operatorname{Tr}\left(\bigotimes_i G_i^{-1}\right)}}. \tag{118}$$

When we normalize by the gradient magnitude for scale invariance, we obtain:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\operatorname{Tr}\left(\bigotimes_i G_i^{-1}\right)}\|g\|_2}. \tag{119}$$

Step 5 — Geometric interpretation. Shampoo emerges as the memoryless approximation of the global Kronecker-factored equalizer, where each mode-wise matrix G_i captures the tensor structure geometry of neural network weights. The optimal learning rate $\eta^\star$ balances the Kronecker-factored budget constraint with the current gradient magnitude, providing mode-wise G_i-weighted scaling that exploits tensor correlations while maintaining computational tractability through Kronecker structure. The optimizer $Q^\star = \eta^\star \bigotimes_i G_i^{-1/2}$ implements approximate steepest descent in the Kronecker-factored natural metric. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{Kron}}$. □

C.3 Diagonal/adaptive-moment family

Corollary 17 (Instantaneous optimal AdaGrad)**.** *Let moment be a diagonal matrix* $R[n] = \operatorname{diag}(r_j[n])$ *with coordinate-wise sequence* $r_j[n]$*. Given gradients* $g[t]$*, maintain the cumulative second-moment* $v_j[t] \coloneqq \epsilon + \sum_{s\le t} g_j[s]^2 > 0$ *with regularization* $\epsilon > 0$*. Fix the current time* t *and define the coordinate-wise costs* $c_j \coloneqq v_j[t]^{1/2}$*. Consider the* diagonal budget

$$\mathcal{Q}_D(B, c) \;\coloneqq\; \{\operatorname{diag}(q_j) : \textstyle\sum_j c_j \sum_{k\ge 0} |q_j[k]|^2 \le B\}, \tag{120}$$

and the cone $\mathcal{C}_{\text{memoryless}}$ of diagonal memoryless optimizers

$$\mathcal{C}_{\text{memoryless}} \coloneqq \{Q_\eta[n] = \eta\delta[n]\,\mathrm{diag}(1/c_j) : \eta \geq 0\}. \tag{121}$$

Optimizing for problem P3 under $\mathcal{Q}_D(B,c) \cap \mathcal{C}_{\text{memoryless}}$ with moment $R[n]$ yields an AdaGrad optimizer with optimal hyperparameter:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\sum_j 1/c_j}}, \tag{122}$$

where the optimizer is $Q^\star = \eta^\star\delta[0]\,\mathrm{diag}(1/c_j)$.

Proof. Fix the current time t and omit the subscript for brevity. We work in the diagonal Hilbert space $(\mathcal{H}_D, \langle\cdot,\cdot\rangle_{\mathcal{H}_D})$ of diagonal causal LTI filters with weighted norm $\|Q\|^2_{\mathcal{H}_D} = \sum_j c_j\|q_j\|^2_{\mathcal{H}}$ where $c_j = v_j^{1/2}$ are the coordinate-wise costs.

Step 1 — Global optimal diagonal equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under the diagonal weighted budget $\mathcal{Q}_D(B,c) = \{\mathrm{diag}(q_j) : \sum_j c_j\|q_j\|^2_{\mathcal{H}} \leq B\}$ is

$$Q^C = \sqrt{B}\,\frac{R}{\|R\|_{\mathcal{H}_D}}, \qquad P^\star(R) = \sqrt{B}\,\|R\|_{\mathcal{H}_D}. \tag{123}$$

Step 2 — Commutativity via smooth convexity. The diagonal weighted ball $\mathcal{Q}_D(B,c)$ is smooth and strictly convex. At any boundary point Q with $\|Q\|_{\mathcal{H}_D} = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_D(B,c)}(Q) = \{\lambda\,\mathrm{diag}(c_j q_j) : \lambda \geq 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_D(B,c)\cap\mathcal{C}_{\text{memoryless}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_D(B,c)\cap\mathcal{C}_{\text{memoryless}}} \langle Q, R\rangle_{\mathcal{H}_D}. \tag{124}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless family parametrization. For the memoryless family $Q_\eta[n] = \eta\delta[n]\,\mathrm{diag}(1/c_j)$ with $\eta \geq 0$, we compute:

$$\|Q_\eta\|^2_{\mathcal{H}_D} = \sum_j c_j\,(\eta/c_j)^2 = \eta^2\sum_j \frac{1}{c_j}, \tag{125}$$

and for the instantaneous gradient moment $R[0] = \mathrm{diag}(r_j[0])$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}_D} = \sum_j \frac{\eta}{c_j} r_j[0] = \eta\sum_j \frac{r_j[0]}{c_j}. \tag{126}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta\|_{\mathcal{H}_D} = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\sum_j 1/c_j}}. \tag{127}$$

Step 5 — Geometric interpretation. AdaGrad emerges as the memoryless approximation of the global diagonal equalizer, where each coordinate-wise cost $c_j = (\epsilon + \sum_{s\leq t} g_j[s]^2)^{1/2}$ captures the cumulative gradient variance. The optimal learning rate $\eta^\star$ balances the diagonal budget constraint with the current gradient, providing coordinate-wise inverse-variance scaling that adapts to the historical gradient magnitudes. The optimizer $Q^\star = \eta^\star\delta[0]\,\mathrm{diag}(1/c_j)$ implements approximate steepest descent in the cumulative variance-weighted metric. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{memoryless}}$. □

Corollary 18 (Instantaneous optimal RMSProp)**.** *Let moment be a diagonal matrix $R[n] = \mathrm{diag}(r_j[n])$ with coordinate-wise sequence $r_j[n]$. Given gradients $g[t]$, maintain the second-moment*

EMA $v_j[t] \coloneqq \beta_2 v_j[t-1] + (1-\beta_2) g_j[t]^2 > 0$ *with parameter* $\beta_2 \in (0,1)$. *Fix the current time* t *and define the coordinate-wise costs* $c_j \coloneqq v_j[t]^{1/2}$. *Consider the* diagonal budget

$$\mathcal{Q}_D(B,c) \;\coloneqq\; \{\mathrm{diag}(q_j) : \textstyle\sum_j c_j \sum_{k\geq 0} |q_j[k]|^2 \leq B\}, \tag{128}$$

and the cone $\mathcal{C}_{\textit{memoryless}}$ *of diagonal memoryless optimizers*

$$\mathcal{C}_{\textit{memoryless}} \coloneqq \{Q_\eta[n] = \eta\delta[n]\,\mathrm{diag}(1/c_j) : \eta \geq 0\}. \tag{129}$$

Optimizing for problem P3 under $\mathcal{Q}_D(B,c) \cap \mathcal{C}_{\textit{memoryless}}$ *with moment* $R[n]$ *yields an RMSProp optimizer with optimal hyperparameters:*

$$\beta_2^\star = \arg\max_{0<\beta_2<1} \frac{\sum_j r_j[0]/c_j}{\sqrt{\sum_j 1/c_j}}, \qquad \eta^\star = \frac{\sqrt{B}}{\sqrt{\sum_j 1/c_j}}, \tag{130}$$

where the optimizer is $Q^\star = \eta^\star \delta[0]\,\mathrm{diag}(1/c_j)$.

Proof. Fix the current time t and omit the subscript for brevity. We work in the diagonal Hilbert space $(\mathcal{H}_D, \langle\cdot,\cdot\rangle_{\mathcal{H}_D})$ of diagonal causal LTI filters with weighted norm $\|Q\|_{\mathcal{H}_D}^2 = \sum_j c_j \|q_j\|_{\mathcal{H}}^2$ where $c_j = v_j^{1/2}$ are the coordinate-wise costs.

Step 1 — Global optimal diagonal equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under the diagonal weighted budget $\mathcal{Q}_D(B,c) = \{\mathrm{diag}(q_j) : \sum_j c_j \|q_j\|_{\mathcal{H}}^2 \leq B\}$ is

$$Q^C = \sqrt{B}\,\frac{R}{\|R\|_{\mathcal{H}_D}}, \qquad P^\star(R) = \sqrt{B}\,\|R\|_{\mathcal{H}_D}. \tag{131}$$

Step 2 — Commutativity via smooth convexity. The diagonal weighted ball $\mathcal{Q}_D(B,c)$ is smooth and strictly convex. At any boundary point Q with $\|Q\|_{\mathcal{H}_D} = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_D(B,c)}(Q) = \{\lambda\,\mathrm{diag}(c_j q_j) : \lambda \geq 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_D(B,c)\cap\mathcal{C}_{\text{memoryless}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_D(B,c)\cap\mathcal{C}_{\text{memoryless}}} \langle Q, R\rangle_{\mathcal{H}_D}. \tag{132}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless family parametrization. For the memoryless family $Q_\eta[n] = \eta\delta[n]\,\mathrm{diag}(1/c_j)$ with $\eta \geq 0$, we compute:

$$\|Q_\eta\|_{\mathcal{H}_D}^2 = \sum_j c_j\,(\eta/c_j)^2 = \eta^2 \sum_j \frac{1}{c_j}, \tag{133}$$

and for the instantaneous gradient moment $R[0] = \mathrm{diag}(r_j[0])$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}_D} = \sum_j \frac{\eta}{c_j} r_j[0] = \eta \sum_j \frac{r_j[0]}{c_j}. \tag{134}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta\|_{\mathcal{H}_D} = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\sum_j 1/c_j}}. \tag{135}$$

The optimal $\beta_2^\star$ maximizes the budget-normalized gain:

$$J(\beta_2) := \frac{\langle Q_\eta, R\rangle_{\mathcal{H}_D}}{\|Q_\eta\|_{\mathcal{H}_D}} = \frac{\sum_j r_j[0]/c_j}{\sqrt{\sum_j 1/c_j}}. \tag{136}$$

Step 5 — Geometric interpretation. RMSProp emerges as the memoryless approximation of the global diagonal equalizer, where each coordinate-wise cost $c_j = v_j[t]^{1/2}$ captures the exponentially-weighted gradient variance. The exponential weighting β_2 prevents the indefinite accumulation that causes AdaGrad's learning rate decay while maintaining adaptive per-coordinate scaling. The optimal learning rate $\eta^\star$ balances the diagonal budget constraint with the current gradient. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{memoryless}}$. □

Corollary 19 (Instantaneous optimal AdaFactor)**.** *Consider matrix parameters $\Theta \in \mathbb{R}^{m\times n}$ with gradient $G \in \mathbb{R}^{m\times n}$. Maintain row-wise second moment estimates $r_i[t] := \beta_2 r_i[t-1] + (1-\beta_2)\|G_{i,:}[t]\|^2$ and column-wise estimates $c_j[t] := \beta_2 c_j[t-1] + (1-\beta_2)\|G_{:,j}[t]\|^2$. Define the* Kronecker-factored diagonal budget

$$\mathcal{Q}_K(B,r,c) := \{Q = \operatorname{diag}(\widehat{r})^{-1/2} \otimes \operatorname{diag}(\widehat{c})^{-1/2} : \|Q\|_{\mathcal{H}}^2 \le B\}, \tag{137}$$

and the cone $\mathcal{C}_{\text{memoryless}}$ of memoryless optimizers

$$\mathcal{C}_{\text{memoryless}} := \{Q_\eta[n] = \eta\delta[n]Q_0 : \eta \ge 0, Q_0 \text{ fixed}\}. \tag{138}$$

Optimizing problem P3 under $\mathcal{Q}_K(B,r,c)\cap\mathcal{C}_{\text{memoryless}}$ with instantaneous gradient moment $R[0] = GG^\top$ yields AdaFactor with optimal hyperparameters:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\sum_i\sum_j 1/(\widehat{r}_i\widehat{c}_j)}}, \qquad \beta_2^\star = \arg\max_{0<\beta_2<1} \frac{\sum_{i,j} G_{ij}^2/(\widehat{r}_i\widehat{c}_j)}{\sqrt{\sum_i\sum_j 1/(\widehat{r}_i\widehat{c}_j)}}, \tag{139}$$

where the optimizer is $Q^\star = \eta^\star\delta[0](\operatorname{diag}(\widehat{r})^{-1/2}\otimes\operatorname{diag}(\widehat{c})^{-1/2})$.

Proof. Fix the current time t and omit the subscript for brevity. We work in the Kronecker-factored Hilbert space $(\mathcal{H}_K, \langle\cdot,\cdot\rangle_{\mathcal{H}_K})$ of causal LTI filters with Kronecker-structured impulse response and norm $\|Q\|_{\mathcal{H}_K}^2 = \sum_{n=0}^\infty \operatorname{Tr}(Q[n]^\top Q[n])$.

Step 1 — Global optimal Kronecker equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under the Kronecker-factored budget $\mathcal{Q}_K(B,r,c)$ is

$$Q^C = \sqrt{B}\,\frac{R}{\|R\|_{\mathcal{H}_K}}, \qquad P^\star(R) = \sqrt{B}\,\|R\|_{\mathcal{H}_K}. \tag{140}$$

Step 2 — Commutativity via smooth convexity. The Kronecker-factored ball $\mathcal{Q}_K(B,r,c)$ is smooth and strictly convex. At any boundary point Q with $\|Q\|_{\mathcal{H}_K} = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_K(B,r,c)}(Q) = \{\lambda Q : \lambda \ge 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_K(B,r,c)\cap\mathcal{C}_{\text{memoryless}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_K(B,r,c)\cap\mathcal{C}_{\text{memoryless}}} \langle Q, R\rangle_{\mathcal{H}_K}. \tag{141}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Memoryless Kronecker family parametrization. For the memoryless Kronecker family $Q_\eta[n] = \eta\delta[n](\operatorname{diag}(\widehat{r})^{-1/2}\otimes\operatorname{diag}(\widehat{c})^{-1/2})$ with $\eta \ge 0$, we compute:

$$\|Q_\eta\|_{\mathcal{H}_K}^2 = \eta^2 \operatorname{Tr}((\operatorname{diag}(\widehat{r})^{-1/2}\otimes\operatorname{diag}(\widehat{c})^{-1/2})^2) = \eta^2 \sum_i\sum_j \frac{1}{\widehat{r}_i\widehat{c}_j}, \tag{142}$$

and for the instantaneous gradient moment $R[0] = GG^\top$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}_K} = \eta \operatorname{Tr}((\operatorname{diag}(\widehat{r})^{-1/2}\otimes\operatorname{diag}(\widehat{c})^{-1/2})GG^\top) = \eta\sum_{i,j}\frac{G_{ij}^2}{\sqrt{\widehat{r}_i\widehat{c}_j}}. \tag{143}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_\eta\|_{\mathcal{H}_K} = \sqrt{B}$. This gives:

$$\eta^\star = \frac{\sqrt{B}}{\sqrt{\sum_i\sum_j 1/(\widehat{r}_i\widehat{c}_j)}}. \tag{144}$$

The optimal $\beta_2^\star$ maximizes the budget-normalized gain:

$$J(\beta_2) := \frac{\langle Q_\eta, R\rangle_{\mathcal{H}_K}}{\|Q_\eta\|_{\mathcal{H}_K}} = \frac{\sum_{i,j} G_{ij}^2/\sqrt{\widehat{r}_i\widehat{c}_j}}{\sqrt{\sum_i\sum_j 1/(\widehat{r}_i\widehat{c}_j)}}. \tag{145}$$

Step 5 — Geometric interpretation. AdaFactor emerges as the memoryless approximation of the global Kronecker-factored equalizer, where the row and column second moment estimates $\widehat{r}_i$ and $\widehat{c}_j$ provide a low-rank factorization of the full diagonal optimizer. This reduces memory complexity from $O(mn)$ to $O(m+n)$ for matrix parameters while approximately preserving Adam's adaptive scaling properties. The exponential weighting β_2 balances the trade-off between adaptation speed and noise reduction in the factored estimates. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{memoryless}}$. □

C.4 NORMALIZED-STEP FAMILY

Corollary 20 (Instantaneous optimal signSGD)**.** *Define the L^∞ budget and the cone of memoryless sign optimizers:*

$$\mathcal{Q}_\infty(\tau) \;\coloneqq\; \{Q[n] = \eta\delta[n]\,\mathrm{diag}(s_j) : \|Q\|_{\mathcal{H},\infty} \le \tau, |s_j| \le 1\}, \tag{146}$$

$$\mathcal{C}_{sign} \;\coloneqq\; \{Q[n] = \eta\delta[n]\,\mathrm{diag}(\mathrm{sign}(g_j)) : \eta \ge 0\}. \tag{147}$$

(147)

Given instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$, solving problem P3 under the budget $\mathcal{Q}_\infty(\tau) \cap \mathcal{C}_{\text{sign}}$ produces signSGD as the optimal solution with optimal hyperparameter:

$$\eta^\star = \tau, \qquad \text{yielding} \quad \dot\theta = \tau\,\mathrm{sign}(g). \tag{148}$$

Proof. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response Q and L^∞ constraint on coordinate-wise step sizes.

Step 1 — Global optimal equalizer. The unconstrained optimum of problem P3 under the L^∞ budget $\mathcal{Q}_\infty(\tau)$ is achieved by setting each coordinate j to maximize $g_j\dot\theta_j$ subject to $|\dot\theta_j| \le \tau$. This gives:

$$Q^C[0] = \tau\ \mathrm{diag}(\mathrm{sign}(g_j)), \quad Q^C[n] = 0 \text{ for } n > 0, \qquad P^\star(R) = \tau\|g\|_1. \tag{149}$$

Step 2 — Commutativity via convexity. The L^∞ constraint set $\mathcal{Q}_\infty(\tau)$ is convex. At any boundary point Q with coordinate-wise saturation, the normal cone contains the sign pattern. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_\infty(\tau)\cap\mathcal{C}_{\text{sign}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_\infty(\tau)\cap\mathcal{C}_{\text{sign}}} \langle Q, R\rangle_{\mathcal{H}}. \tag{150}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Sign family parametrization. For the sign family $Q_\eta[n] = \eta\delta[n]\,\mathrm{diag}(\mathrm{sign}(g_j))$ with $\eta \ge 0$, we compute:

$$\|Q_\eta\|_{\mathcal{H},\infty} = \eta \max_j |\mathrm{sign}(g_j)| = \eta, \tag{151}$$

and for the instantaneous gradient moment $R[0] = gg^\top$ and $R[n] = 0$ for $n > 0$:

$$\langle Q_\eta, R\rangle_{\mathcal{H}} = \eta\,\mathrm{Tr}(\mathrm{diag}(\mathrm{sign}(g_j))\cdot gg^\top) = \eta\sum_j g_j\mathrm{sign}(g_j) = \eta\|g\|_1. \tag{152}$$

Step 4 — Budget saturation and optimization. The objective is linear in η while the constraint is linear, so the maximizer saturates the budget $\|Q_\eta\|_{\mathcal{H},\infty} = \tau$. This gives:

$$\eta^\star = \tau. \tag{153}$$

The resulting update is $\dot\theta = \tau\,\mathrm{sign}(g)$.

Step 5 — Geometric interpretation. signSGD emerges as the memoryless approximation of the global L^∞-constrained equalizer, where each coordinate takes the maximum allowed step in the direction of its gradient sign. This coordinate-wise saturation provides robustness to gradient magnitude variations and enables efficient low-precision implementations. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{sign}}$. □

Corollary 21 (Instantaneous optimal Lion)**.** *Define the L^∞ budget and the cone of momentum-filtered sign optimizers:*

$$\mathcal{Q}_\infty(\tau) \;\coloneqq\; \{Q : \|Q\|_{\mathcal{H},\infty} \le \tau\}, \tag{154}$$

$$\mathcal{C}_{Lion} \;\coloneqq\; \{Q_{\eta,\beta_1}[n] = \eta(1-\beta_1)\beta_1^n\,\mathrm{diag}(\mathrm{sign}(m_j)) : \eta \ge 0, 0 < \beta_1 < 1\}, \tag{155}$$

where m_j is the j-th coordinate of the momentum-smoothed gradient $m = \mathrm{EMA}(g; \beta_1)$. Given gradient moment $R[n] \in \mathcal{H}$, solving problem P3 under the budget $\mathcal{Q}_\infty(\tau) \cap \mathcal{C}_{Lion}$ produces Lion optimizer as the optimal solution with optimal hyperparameters:

$$\beta_1^\star \;=\; \arg\max_{0<\beta_1<1} \sum_{n=0}^{\infty} \beta_1^n \sum_j |\operatorname{Tr}(R_j[n])|, \qquad \eta^\star \;=\; \tau, \tag{156}$$

where $R_j[n]$ denotes the j-th diagonal block of $R[n]$.

Proof. We work in the impulse-space Hilbert space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of causal LTI filters with matrix impulse response Q and L^∞ constraint on coordinate-wise step sizes.

Step 1 — Global optimal equalizer. The unconstrained optimum of problem P3 under the L^∞ budget $\mathcal{Q}_\infty(\tau)$ is achieved by setting each coordinate j to maximize the inner product subject to $|Q_j[n]| \le \tau$ for all n. This gives:

$$Q^C[n] = \tau\ \operatorname{diag}(\operatorname{sign}(R_{jj}[n])), \qquad P^\star(R) = \tau \sum_{n=0}^{\infty} \sum_j |R_{jj}[n]|. \tag{157}$$

Step 2 — Commutativity via convexity. The L^∞ constraint set $\mathcal{Q}_\infty(\tau)$ is convex. At any boundary point Q with coordinate-wise saturation, the normal cone contains the sign pattern. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_\infty(\tau)\cap\mathcal{C}_{\text{Lion}}}(Q^C) \in \arg\max_{Q\in\mathcal{Q}_\infty(\tau)\cap\mathcal{C}_{\text{Lion}}} \langle Q, R\rangle_{\mathcal{H}}. \tag{158}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Lion family parametrization. For $Q_{\eta,\beta_1}[n] = \eta(1-\beta_1)\beta_1^n \operatorname{diag}(\operatorname{sign}(m_j))$ with $\eta \ge 0$ and $0 < \beta_1 < 1$, we compute:

$$\|Q_{\eta,\beta_1}\|_{\mathcal{H},\infty} = \eta(1-\beta_1) \max_j |\operatorname{sign}(m_j)| \sum_{n=0}^{\infty} \beta_1^n = \eta, \tag{159}$$

and

$$\langle Q_{\eta,\beta_1}, R\rangle_{\mathcal{H}} = \sum_{n=0}^{\infty} \eta(1-\beta_1)\beta_1^n \sum_j \operatorname{sign}(m_j) R_{jj}[n] = \eta(1-\beta_1) \sum_{n=0}^{\infty} \beta_1^n \sum_j \operatorname{sign}(m_j) R_{jj}[n]. \tag{160}$$

Step 4 — Budget saturation and 1-D optimization. For fixed β_1, the objective is linear in η while the constraint is linear, so the maximizer saturates the budget $\|Q_{\eta,\beta_1}\|_{\mathcal{H},\infty} = \tau$. This gives:

$$\eta = \tau. \tag{161}$$

The budget-normalized gain becomes:

$$J(\beta_1) := \frac{\langle Q_{\eta,\beta_1}, R\rangle_{\mathcal{H}}}{\|Q_{\eta,\beta_1}\|_{\mathcal{H},\infty}} = (1-\beta_1) \sum_{n=0}^{\infty} \beta_1^n \sum_j \operatorname{sign}(m_j) R_{jj}[n]. \tag{162}$$

Hence $\beta_1^\star = \arg\max_{0<\beta_1<1} J(\beta_1)$ and $\eta^\star = \tau$.

Step 5 — Geometric interpretation. The optimal momentum $\beta_1^\star$ maximizes the alignment between the 1-pole EMA kernel $(\beta_1^n)_{n\ge0}$ and the sign-weighted empirical lag curve. Lion emerges as the momentum-filtered approximation of the global L^∞-constrained equalizer, where momentum smoothing $m = \mathrm{EMA}(g; \beta_1)$ reduces gradient noise before applying coordinate-wise sign normalization. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{Lion}}$. □

C.5 LAYER-WISE NORM SCALING (LARGE-BATCH STABILIZERS)

Corollary 22 (Instantaneous optimal LARS/LAMB)**.** *Consider a neural network with L layers, where parameters are partitioned as $\theta = (\theta_1, \dots, \theta_L)$ with $\theta_\ell \in \mathbb{R}^{d_\ell}$. Define the* layer-wise adaptive budget

$$\mathcal{Q}_{\textit{layer}}(B, \{\|\theta_\ell\|\}_{\ell=1}^L) \;:=\; \left\{ Q = \mathrm{blkdiag}(Q_1, \dots, Q_L) : \sum_{\ell=1}^{L} \|\theta_\ell\|_2^2 \sum_{n=0}^{\infty} \mathrm{Tr}(Q_\ell[n]^\top Q_\ell[n]) \le B \right\}, \tag{163}$$

and the cone $\mathcal{C}_{\textit{layer}}$ of layer-wise scalar optimizers

$$\mathcal{C}_{\textit{layer}} \;:=\; \{Q_{\eta,\beta_1}[n] = \mathrm{blkdiag}(\eta(1-\beta_1)\beta_1^n \alpha_1 I_{d_1}, \dots, \eta(1-\beta_1)\beta_1^n \alpha_L I_{d_L}) : \eta \ge 0, 0 < \beta_1 < 1, \alpha_\ell > 0\}. \tag{164}$$

Given moment matrix $R[n] = \mathrm{blkdiag}(R_1[n], \dots, R_L[n])$ with $S_\ell[n] := \mathrm{Tr}(R_\ell[n])$, optimizing problem P3 under $\mathcal{Q}_{\textit{layer}}(B, \{\|\theta_\ell\|\}) \cap \mathcal{C}_{\textit{layer}}$ yields LARS/LAMB optimizer with optimal hyperparameters:

$$\beta_1^\star = \arg\max_{0<\beta_1<1} \sqrt{1-\beta_1^2} \sum_{n=0}^{\infty} \beta_1^n \sum_{\ell=1}^{L} \frac{S_\ell[n]}{\|\theta_\ell\|_2}, \quad \eta^\star = \frac{\sqrt{B(1-\beta_1^{\star 2})}}{\sqrt{\sum_{\ell=1}^L d_\ell / \|\theta_\ell\|_2^2}\,(1-\beta_1^\star)}, \tag{165}$$

where the layer-wise scaling factors are $\alpha_\ell^\star = 1/\|\theta_\ell\|_2$.

Proof. We work in the block-diagonal Hilbert space $(\mathcal{H}_{\text{layer}}, \langle\cdot,\cdot\rangle_{\mathcal{H}_{\text{layer}}})$ of layer-wise causal LTI filters with weighted norm $\|Q\|_{\mathcal{H}_{\text{layer}}}^2 = \sum_{\ell=1}^L \|\theta_\ell\|_2^2 \|Q_\ell\|_{\mathcal{H}}^2$.

Step 1 — Global optimal layer-wise equalizer. By Cauchy-Schwarz, the unconstrained optimum of problem P3 under budget $\mathcal{Q}_{\text{layer}}(B, \{\|\theta_\ell\|\})$ is

$$Q^C = \sqrt{B} \frac{R}{\|R\|_{\mathcal{H}_{\text{layer}}}}, \qquad P^\star(R) = \sqrt{B} \|R\|_{\mathcal{H}_{\text{layer}}}, \tag{166}$$

where $\|R\|_{\mathcal{H}_{\text{layer}}}^2 = \sum_{\ell=1}^L \|\theta_\ell\|_2^2 \|R_\ell\|_{\mathcal{H}}^2$.

Step 2 — Commutativity via smooth convexity. The layer-wise adaptive ball $\mathcal{Q}_{\text{layer}}(B, \{\|\theta_\ell\|\})$ is smooth and strictly convex. At any boundary point Q with $\|Q\|_{\mathcal{H}_{\text{layer}}} = \sqrt{B}$, the normal cone is the ray $N_{\mathcal{Q}_{\text{layer}}}(Q) = \{\lambda Q : \lambda \ge 0\}$. By Lemma 4, commutativity holds:

$$\Pi_{\mathcal{Q}_{\text{layer}}(B,\{\|\theta_\ell\|\}) \cap \mathcal{C}_{\text{layer}}}(Q^C) \in \arg\max_{Q \in \mathcal{Q}_{\text{layer}}(B,\{\|\theta_\ell\|\}) \cap \mathcal{C}_{\text{layer}}} \langle Q, R \rangle_{\mathcal{H}_{\text{layer}}}. \tag{167}$$

Therefore, we can equivalently solve the restricted optimization problem directly.

Step 3 — Layer-wise family parametrization. For $Q_{\eta,\beta_1}[n] = \mathrm{blkdiag}(\eta(1-\beta_1)\beta_1^n \alpha_1 I_{d_1}, \dots, \eta(1-\beta_1)\beta_1^n \alpha_L I_{d_L})$ with $\eta \ge 0$, $0 < \beta_1 < 1$, and $\alpha_\ell = 1/\|\theta_\ell\|_2$, we compute:

$$\|Q_{\eta,\beta_1}\|_{\mathcal{H}_{\text{layer}}}^2 = \sum_{\ell=1}^{L} \|\theta_\ell\|_2^2 \sum_{n=0}^{\infty} \mathrm{Tr}((\eta(1-\beta_1)\beta_1^n \alpha_\ell I_{d_\ell})^\top (\eta(1-\beta_1)\beta_1^n \alpha_\ell I_{d_\ell})) \tag{168}$$

$$= \eta^2 (1-\beta_1)^2 \frac{1}{1-\beta_1^2} \sum_{\ell=1}^{L} \|\theta_\ell\|_2^2 \alpha_\ell^2 d_\ell = \eta^2 \frac{(1-\beta_1)^2}{1-\beta_1^2} \sum_{\ell=1}^{L} \frac{d_\ell}{\|\theta_\ell\|_2^2}, \tag{169}$$

and

$$\langle Q_{\eta,\beta_1}, R \rangle_{\mathcal{H}_{\text{layer}}} = \sum_{\ell=1}^{L} \sum_{n=0}^{\infty} \mathrm{Tr}((\eta(1-\beta_1)\beta_1^n \alpha_\ell I_{d_\ell})^\top R_\ell[n]) \tag{170}$$

$$= \eta(1-\beta_1) \sum_{n=0}^{\infty} \beta_1^n \sum_{\ell=1}^{L} \alpha_\ell S_\ell[n] = \eta(1-\beta_1) \sum_{n=0}^{\infty} \beta_1^n \sum_{\ell=1}^{L} \frac{S_\ell[n]}{\|\theta_\ell\|_2}. \tag{171}$$

Step 4 — Budget saturation and 1-D optimization. For fixed β_1, the objective is linear in η while the constraint is quadratic, so the maximizer saturates the budget $\|Q_{\eta,\beta_1}\|_{\mathcal{H}_{\text{layer}}} = \sqrt{B}$. This gives:

$$\eta = \frac{\sqrt{B(1-\beta_1^2)}}{\sqrt{\sum_{\ell=1}^{L} d_\ell / \|\theta_\ell\|_2^2}\,(1-\beta_1)}. \tag{172}$$

The budget-normalized gain becomes:

$$J(\beta_1) := \frac{\langle Q_{\eta,\beta_1}, R\rangle_{\mathcal{H}_{\text{layer}}}}{\|Q_{\eta,\beta_1}\|_{\mathcal{H}_{\text{layer}}}} = \sqrt{\frac{1-\beta_1^2}{\sum_{\ell=1}^{L} d_\ell/\|\theta_\ell\|_2^2}} \sum_{n=0}^{\infty} \beta_1^n \sum_{\ell=1}^{L} \frac{S_\ell[n]}{\|\theta_\ell\|_2}. \tag{173}$$

Hence $\beta_1^\star = \arg\max_{0<\beta_1<1} J(\beta_1)$ and $\eta^\star$ saturates the budget constraint.

Step 5 — Geometric interpretation. The optimal momentum $\beta_1^\star$ maximizes the cosine similarity between the 1-pole EMA kernel $(\beta_1^n)_{n\geq 0}$ and the layer-normalized empirical lag curve $(\sum_{\ell=1}^{L} S_\ell[n]/\|\theta_\ell\|_2)_{n\geq 0}$. The layer-wise scaling $\alpha_\ell = 1/\|\theta_\ell\|_2$ prevents layer collapse by ensuring updates remain proportional to current parameter magnitudes. By commutativity, this restricted optimum coincides with projecting the global optimum Q^C onto $\mathcal{C}_{\text{layer}}$. □

D MASTER TABLE OF OPTIMIZERS

In the previous section, we have derived various types of optimizers from our convex optimization framework. We can now register various optimizers under a single unified table. Each optimizer corresponds to a specific choice of moment matrix M, budget constraint $\mathcal{Q}$, and resulting equalizer Q. "Param restrict" rows are feasible points in the convex programs that can either be kept and fitted to target moments, or replaced with full closed-form solutions.

Glossary.

- Instant moment $M \in \mathbb{S}_+^P$: $M = \Sigma_{\text{tr}} = \mathbb{E}[gg^\top]$ (training) or $M = \text{sym}(C) = \text{sym}\,\mathbb{E}[g_{\text{tr}} g_{\text{val}}^\top]$ (validation-aware)
- Dynamic moment operator M_σ (Laplace/z window, $\sigma > 0$); frequency form $M(\omega)$
- Instant budgets from Section 2: **Frob**, **Trace/Spectral**, **Lyap**, **Diag**
- Dynamic budgets from Section 3: **D-Frob (H_2)**, **D-Trace/Spectral**, **D-Lyap**, **D-Diag**
- "Param restrict." means we *restrict* Q to a small parametric family inside the convex budget

Table 2: Optimizer Specifications: Moment and Budget Constraints

Optimizer	Moment Used	Budget $\mathcal{Q}$
GD	$M = \Sigma_{\text{tr}}$	Frob: $\|Q\|_F \leq \kappa$ (instant)
SGD	same as GD (stochastic)	Frob (instant)
Momentum (HB/NAG)	M_σ (weighted)	**D-Frob**: $\|Q\|_{\mathcal{H}} \leq \sqrt{B}$
Nesterov	M_σ	**D-Frob** (with predictive tap)
AdaGrad	$M = \Sigma_{\text{tr}}$	**Diag**: $\sum_j c_j q_j^2 \leq B$
RMSProp	$M = \Sigma_{\text{tr}}$	**Diag**
Adam/AdamW	M_σ (via m, v)	**D-Diag**: $\sum_j \|q_j\|_{H_2,c_j}^2 \leq B$
Adam (val-aware)	$M = \text{sym}(C)$ or M_σ	**Diag** or **D-Diag**
LAMB / LARS	$M = \Sigma_{\text{tr}}$ per layer ℓ	**Trace** per layer: $\text{Tr}(Q_\ell) \leq \tau_\ell$ + norm ratio
K-FAC	$M = \Sigma_{\text{tr}}$	**Lyap (block-Kronecker)**: $\text{Tr}(Q^2\Sigma) \leq B$ with Q factored
Shampoo	$M = \Sigma_{\text{tr}}$	**Lyap (multi-axis Kronecker)**
Newton / GN	$M = \Sigma_{\text{tr}}$	**Lyap** with $Q = H^{-1}$ (or G^{-1})
signSGD	$M = \Sigma_{\text{tr}}$	ℓ_∞ step budget on $\dot\theta$
Lion	M_σ via m	ℓ_∞ on $\dot\theta$ (dynamic)
Polyak step	$M = \Sigma_{\text{tr}}$ + scalar loss	Frob + 1D line search

Table 3: Optimizer Equalizers and Closed Forms

Optimizer	Equalizer Q	Closed Form
GD	$Q = \alpha I$ (isotropic)	$Q^\star = \kappa \frac{M}{\|M\|_F} \to$ in practice $\alpha = \eta$
SGD	$Q = \alpha I$	same as GD, with M estimated from minibatch
Momentum (HB/NAG)	$Q(z) = \eta I \frac{1-\beta}{1-\beta z^{-1}}$ (one-pole)	**Param restrict** of dynamic proportional optimum; (η, β) fit the target
Nesterov	$Q(z) = \eta I \frac{1-\beta}{1-\beta z^{-1}}(1 + \gamma z^{-1})$	**Param restrict** (lead–lag) under same budget
AdaGrad	$Q = \mathrm{diag}(q)$	$q_j^\star \propto \frac{M_{jj}}{c_j}$; $c_j \uparrow$ with cum. second moment
RMSProp	$Q = \mathrm{diag}(q)$	same as AdaGrad but c_j from EMA of g_j^2
Adam/AdamW	$Q(z) = \mathrm{diag}\Big(\eta \frac{1-\beta_1}{1-\beta_1 z^{-1}}\Big)\,(\mathrm{diag}\sqrt{v} + \epsilon)^{-1}$	**Param restrict** of D-Diag optimum
Adam (val-aware)	same as Adam	Use val cross-moment for q_j (and m vs g)
LAMB / LARS	$Q_\ell = \alpha_\ell I$ s.t. $\|v_\ell\| \propto \|\theta_\ell\|$	**Water-fill over layers** + isotropic in each layer
K-FAC	$Q = \bigoplus_\ell (A_\ell^{-1} \otimes G_\ell^{-1})$ approx	**Equal-power** in each layer's Fisher metric (factored)
Shampoo	$Q = \bigotimes_i H_i^{-1/2}$	Equal-power along tensor modes
Newton / GN	$Q = H^{-1}$	Exact Lyap optimum if constraint matches curvature
signSGD	Q such that $\dot\theta = \eta\,\mathrm{sign}(g)$	Linear objective + $\|\dot\theta\|_\infty \le \eta \Rightarrow$ vertex
Lion	$\dot\theta = \eta\,\mathrm{sign}(m)$	same with smoothed signal
Polyak step	$Q = \alpha I$ with α from loss & grad norm	closed-form $\alpha = \frac{\|g\|^2}{g^\top H g}$ (local)

Table 4: Optimizer Hyperparameters and Interpretations

Optimizer	Hyperparameters	One-line Implication
GD	η: total power scale	GD is the "proportional router" collapsed to a scalar knob
SGD	η: same	Stochasticity only changes how you **estimate** M, not the program
Momentum (HB/NAG)	η: overall gain; β: pole = decay time of impulse	Momentum is the 1-pole **low-rank approximation** of the dynamic proportional router
Nesterov	β: smoothing; γ: look-ahead lead	"Prediction" = a tiny **lead** in the equalizer—still budgeted power
AdaGrad	ϵ, window: determine c_j growth	AdaGrad is **per-coord water-filling** with costs = cumulative variance
RMSProp	ρ: EMA window; ϵ	RMSProp = **time-local AdaGrad** (cheaper, reactive)
Adam/AdamW	η: scale; β_1: low-pass pole; β_2: sets costs $c_j \sim v_j$	Adam = **diagonal dynamic optimizer**: per-coord water-filling **times** a 1-pole low-pass
Adam (val-aware)	same	Simply switching M to $\mathrm{sym}(C)$ turns Adam into a **val-aware power allocator**
LAMB / LARS	trust ratio $\|\theta_\ell\|/\|m_\ell\|$	LAMB/LARS = **layer-wise trace budget** + norm normalization (compute-stable water-filling)
K-FAC	damping λ, update period	K-FAC = **structured Lyap** equalization $\Rightarrow$ curvature-aware, block-wise
Shampoo	per-axis damping, period	Shampoo = **multi-axis equal-power** (richer than K-FAC, costlier)
Newton / GN	trust-region radius	Newton = **max power under curvature budget**; best when Hessian is right
signSGD	η: box size	signSGD = **L-∞ trust-region** maximizer; robust, but discards magnitude info
Lion	β_1	Lion = **L-∞ budget on smoothed signal**; ultra-aggressive quantized equalizer
Polyak step	none beyond window	Polyak = Frob program with **on-the-fly** α estimate (local curvature)

E ALGORITHM

For completness, we provide a realizable algorithm for calculating the optimal SGD+Momentum and Adam, as proved in Section 2 and Section 3 of the main manuscript.

Algorithm 1 Optimal SGD+Momentum ($\beta^\star$, $\eta^\star$) from gradient history

Require: Window length T, max lag $K \le T-1$, EMA decay $\rho \in (0,1)$, budget $B > 0$

Initialize $\mu \leftarrow 0$, $S[n] \leftarrow 0$ for $n = 0, \dots, K$, buffer $\leftarrow \emptyset$, $d \leftarrow$ parameter dimension
for each calibration step t **do**
  $g_t \leftarrow$ flatten current $\nabla_\theta L$
  $\mu \leftarrow \rho\mu + (1-\rho) g_t$
  $\tilde{g}_t \leftarrow g_t - \mu$
  push $\tilde{g}_t$ into buffer (keep last T)
  for $n = 0, \dots, K$ with $t - n$ in buffer **do**
    $s_n \leftarrow \langle \tilde{g}_t, \tilde{g}_{t-n} \rangle$ ▷ scalar dot product
    $S[n] \leftarrow \rho S[n] + (1-\rho) s_n$
  end for
  $J(\beta) = \sqrt{\frac{1-\beta^2}{d}} \cdot \sum_{n=0}^{K} \beta^n S[n]$
  $\beta^\star \leftarrow \arg\max_{\beta \in (0,1)} J(\beta)$ ▷ 1-D search (e.g., bounded line search)
  $\eta^\star \leftarrow \frac{\sqrt{B(1-(\beta^\star)^2)}}{\sqrt{d}(1-\beta^\star)}$
end for
return $(\beta^\star, \eta^\star)$

Algorithm 2 Optimal Adam ($\beta_1^\star$, $\beta_2^\star$, $\eta^\star$) from gradient history

Require: Window length T, max lag $K \le T-1$, EMA decay $\rho \in (0,1)$, budget $B > 0$

Initialize $\mu \leftarrow 0$, $v_j \leftarrow 0$ for all j, $T_t[n] \leftarrow 0$ for $n = 0, \dots, K$, buffer $\leftarrow \emptyset$, $d \leftarrow$ parameter dimension
for each calibration step t **do**
  $g_t \leftarrow$ flatten current $\nabla_\theta L$
  $\mu \leftarrow \rho\mu + (1-\rho) g_t$
  $\tilde{g}_t \leftarrow g_t - \mu$
  push $\tilde{g}_t$ into buffer (keep last T)
  for each coordinate j **do**
    $v_j \leftarrow \rho v_j + (1-\rho)\tilde{g}_{t,j}^2$
    $c_j \leftarrow \sqrt{v_j}$
  end for
  for $n = 0, \dots, K$ with $t - n$ in buffer **do**
    $T_t[n] \leftarrow \sum_j \frac{\tilde{g}_{t,j}\tilde{g}_{t-n,j}}{c_j}$ ▷ coordinate-wise weighted dot product
  end for
  $W_t \leftarrow \sum_j \frac{1}{c_j}$
  $a(\beta_1, \beta_2) \leftarrow \sqrt{\frac{1-\beta_1^2}{W_t}}$
  $J(\beta_1, \beta_2) = a(\beta_1, \beta_2) \sum_{n=0}^{K} \beta_1^n T_t[n]$
  $(\beta_1^\star, \beta_2^\star) \leftarrow \arg\max_{\beta_1, \beta_2 \in (0,1)} J(\beta_1, \beta_2)$ ▷ 2-D search
  $\eta^\star \leftarrow \frac{\sqrt{B}\, a(\beta_1^\star, \beta_2^\star)}{1-\beta_1^\star}$
end for
return $(\beta_1^\star, \beta_2^\star, \eta^\star)$