AXIOM: ACTIVATION-CONDITIONED MOMENTUM AS ASSOCIATIVE MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical momentum-based optimizers, such as Stochastic Gradient Descent (SGD) with momentum and Adam accumulate gradient statistics isotropically, treating every direction in parameter space identically and remaining blind to the geometry of the activations that produced each gradient. Motivated by the Nested Learning view of optimization as a system of associative memories, we in this study, reformulate momentum as the solution of a regularized least squares objective that maps an input activation to its local surprise signal, rather than an exponential moving average, yielding a closed form optimizer we call **AXIOM** (Activation-guided eXpressive Iterative Optimization with associative Memory). We derive the resulting update independently through a Sherman–Morrison expansion and through Fenchel–Rockafellar duality, showing that both routes converge on the same rank-one, activation-conditioned correction. We characterize AXIOM as a discrete-time linear time-varying system, prove global uniform exponential stability of its unforced dynamics, and establish an $\mathcal{O}(1/T) + \mathcal{O}(\sigma_g^2)$ nonconvex convergence rate for the forced stochastic system, matching classical first-order guarantees. To test whether activation-aware momentum provides practical benefit under heterogeneous and non-stationary feature geometries, we introduce **Wavelet2VecNet**-a framework that couples pretrained Self-Supervised Learning (SSL)-based contextual embeddings (Wav2Vec2) with a two-dimensional discrete wavelet transform decomposition for compact utterance-level speech representation learning, and evaluate it across *five* speech processing tasks, namely dysarthric speech severity-level assessment, infant cry pathology classification of normal *vs* pathalogical infant cries, environmental acoustic scene and sound classification, and source attribution of deepfake audio. AXIOM matches or exceeds Adam and SGD on four of five tasks, with the largest gains (+7 to +14 accuracy points) on the most acoustically heterogeneous tasks, and no advantage on a near-saturated verification task; specifically, AXIOM improves over Adam by +2.35 points on Dysarthria, +7.48 on Real-DESED, +2.42 on Infant Cry, and +14.43 on UrbanSound8K, while Source Tracing shows a near-tied result. – a pattern consistent with our theoretical account of when activation-conditioned momentum should help. We report ablations across wavelet bases and feature backbones, including a case where AXIOM's advantage does not persist, and discuss the limitations of the present activation-only linear-layer analysis.

1 INTRODUCTION

Progress in deep learning is driven by two complementary forces: increasingly expressive data representations, and the optimization algorithms used to fit them. While representation learning – most visibly through large pretrained foundation models – has advanced rapidly, the optimization process governing how these representations are learned and adapted has changed comparatively little since the introduction of adaptive first-order methods (Kingma & Ba, 2015). Adam and its variants remain the default choice across nearly every modality, especially in case of speech tasks.

A structural property of these optimizers is easy to overlook precisely because it is so deeply embedded in their design: momentum is accumulated *isotropically*. Whether a gradient arrives from a high variance, saturated activation direction or from a rarely visited, low energy one, classical momentum treats it the same way, folding it into a running average governed by a single scalar de-

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

cay rate (or in Adam, a per-coordinate one derived from the gradient history rather than from the current activation). Yet, the activations that produce these gradients are rarely *isotropic*. In speech processing, in particular, the same optimizer is routinely applied across feature distributions with very different statistics – a frozen contextual embeddings of Self-Supervised Learning (SSL)-based encoder, a wavelet decomposed subband representation, a spectrogram, however, there are only a few underived little reason to expect a single isotropic update rule to be equally well suited to all of them (even for State-Of-The-Art (SOTA) networks, such as AASIST (Jung et al., 2022, Singapore), WIN (Shah et al., 2026), Dys. Transformer (Shahamiri et al., 2023)), raising serious concerns about weather to use Adams/SGD for each speech processing task?

This paper asks a narrow, falsifiable question: *If momentum is instead derived as the solution to a regularized optimization problem that explicitly couples the update to the current activation geometry, does this yield an optimizer with strong theoretical guarantees that also helps more when activation statistics are more heterogeneous?* We answer this question in two parts. **First**, following the nested learning framework of optimization algorithms as systems of associative memories (Behrouz et al., 2026, Sydney, Australia), we derive **AXIOM**-a first order optimizer in which the momentum matrix is the unique minimizer of a regularized least squares objective mapping activations to downstream gradients, rather than an exponential moving average. AXIOM is a targeted, hybrid modification rather than a wholesale replacement of existing optimizers: it substitutes only the first moment estimator of an Adam-style optimizer with this associative memory update, while retaining Adam's second moment estimator unchanged (sub-Section 3.4). **Second**, we introduce **Wavelet2VecNet**, a speech representation framework, whose activation statistics vary substantially across tasks by the design, and use it as a testbed to determine, *where* activation-aware momentum help and where it does not?

Contributions of this paper are as follows:

- We derive an update rule of AXIOM from a regularized associative memory objective, and show, via two independent routes (a Sherman–Morrison expansion of the primal problem, and a Fenchel–Rockafellar dual reformulation), that the optimal update rule reduces to a rank-one, error-driven correction along the current activation direction (Section 3). AXIOM replaces only the first moment dynamics of an Adam-style optimizer, retaining its second moment estimator, so the resulting algorithm is a targeted hybridization of the two mechanisms rather than a standalone optimizer built from the scratch.
- We prove global uniform exponential stability of the unforced momentum dynamics and an $\mathcal{O}(1/T) + \mathcal{O}(\sigma_g^2)$ convergence rate for the forced stochastic system under standard nonconvex optimization assumptions, matching classical guarantees while additionally exhibiting activation-conditioned anisotropic filtering that we characterize spectrally (Section 3, full derivations in Appendix C–K).
- We introduce Wavelet2VecNet, coupling frozen Wav2Vec2 contextual embeddings with a *2-D* discrete wavelet transform and a lightweight CNN, and apply it across three distinct speech understanding problems with markedly different activation statistics (Section 4).
- We evaluate AXIOM against Adam and Stochastic Gradient Descent (SGD) across five speech tasks, two additional representation backbones, and four wavelet bases, finding that AXIOM's advantage scales with the *heterogeneity* of the underlying feature distribution, and we report the one setting (source tracing under non-default wavelets), where this advantage does not hold, rather than omitting it (Section 6).

2 Related Work

Adaptive First-Order Optimization. SGD with momentum and Adam-family optimizers (Kingma & Ba, 2015) remain the default choice for training deep networks across tasks such as audio deepfake detection (Shah et al., 2026), dysarthric ASR (Shahamiri et al., 2023), infant cry classification (Ji et al., 2021), and sound event detection (Turpault et al., 2019, New York, USA). These methods adapt learning rates from accumulated gradient statistics independently of the activations that produced them, yielding update geometry that is isotropic (SGD) or axis-aligned (Adam) rather than conditioned on the data manifold. A separate line of work restructures momentum itself: Lion (Chen et al., 2023, New Orleans, Louisiana, USA) discards gradient magnitude in favor of a sign-based

update, and schedule-free methods (Defazio et al., 2024, Vancouver, Canada) reparameterize the momentum sequence to remove explicit learning-rate schedules. These methods change *how* momentum is combined or scheduled; AXIOM instead changes *what* momentum estimates, replacing the exponential moving average with the solution of an activation-conditioned regularized objective – a complementary, not competing, direction.

Nested Learning and Meta-Learned Optimization. Recent work views deep learning architectures and optimizers as nested systems of associative memories, where components such as momentum solve an internal learning problem rather than follow a fixed update rule (Behrouz et al., 2026, Sydney, Australia). AXIOM instantiates this view for first-moment momentum dynamics, primarily for (though not limited to) audio processing tasks.

Associative Memory and Structured Updates. The rank-one, error-driven correction we derive relates to classical associative memory and Hebbian learning (Hebb, 1949), and its anti-Hebbian counterpart (Földiak, 1990). It also echoes the innovation-driven update of recursive estimators such as the Kalman filter (Kalman, 1960) and the predict-and-correct motif of predictive coding (Rao & Ballard, 1999); we treat these as structural analogies rather than mechanistic claims.

Selective State-Space Models. AXIOM's momentum recursion is a discrete-time linear system whose transition operator depends on the current input, structurally similar to selective state-space models such as Mamba (Gu & Dao, 2023, {Last Accessed: 22nd September, 2026}). Unlike Mamba, this time-varying transition is not learned through a gate but follows analytically from the closed-form solution of a convex regularized objective, with no additional parameters.

Curvature- and Geometry-Aware Optimization. Kronecker-factored and second-order methods such as K-FAC (Martens & Grosse, 2015, Lille, France) and Shampoo (Gupta et al., 2018, Vienna, Austria) condition updates on curvature or preconditioning statistics at substantially higher per-step cost than first-order methods. Lightweight curvature-aware optimizers such as Sophia (Liu et al., 2024, Vienna, Austria) narrow this gap via a periodically updated diagonal Hessian proxy, reaching Adam-comparable cost; our cost claim therefore applies specifically to full or Kronecker-factored curvature estimation. AXIOM differs along a separate axis: it conditions on instantaneous *activation* geometry rather than loss curvature, at momentum-level cost, without maintaining any curvature estimate.

Self-Supervised Speech Representations and MultiResolution Analysis. Pretrained encoders such as Wav2Vec2 (Baevski et al., 2020, Vancouver, British Columbia, Canada) provide contextual embeddings that outperform handcrafted acoustic features across many speech tasks. Separately, wavelet decompositions have long supported multi-resolution analysis of non-stationary signals (Mallat, 1989; 1999). Wavelet2VecNet combines these by applying wavelet decomposition directly to contextual embeddings rather than to the raw audio or spectrograms.

Speech Understanding Tasks. We evaluate synthetic speech source attribution using protocols related to ASVspoof (Nautsch et al., 2021) and embedding spaces exhibiting neural collapse (Papyan et al., 2020); dysarthric speech severity assessment, a long-standing clinical problem; infant cry classification; and environmental sound/acoustic scene classification using UrbanSound8K (Salamon et al., 2014, Orlando, Florida, USA) and a DESED-derived acoustic scene corpus (Turpault et al., 2019, New York, USA), together providing a deliberately heterogeneous testbed of activation statistics.

3 AXIOM: Activation-Conditioned Momentum as Associative Memory

3.1 Motivation

Classical momentum can be reformulated as the solution of a regularized optimization problem rather than as an exponential moving average (Behrouz et al., 2026, Sydney, Australia). Given a layer with input activation, $\mathbf{x} \in \mathbb{R}^{d_{\text{in}}}$, downstream gradient (i.e., local surprise signal), $\mathbf{u} \in \mathbb{R}^{d_{\text{out}}}$,

and momentum matrix $\mathbf{M} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, we pose momentum estimation at time $t+1$ as:

$$\mathbf{M}_{t+1} = \arg\min_{\mathbf{M}} J(\mathbf{M}) \triangleq \underbrace{\|\mathbf{M}\mathbf{x}_t - \mathbf{u}_t\|_2^2}_{\text{Association Error}} + \underbrace{\alpha\|\mathbf{M}\|_F^2}_{\text{EWMA-Like Regularizer}} + \underbrace{\frac{1}{2\eta}\|\mathbf{M} - \mathbf{M}_t\|_F^2}_{\text{Retention Term}}, \tag{1}$$

where $\alpha, \eta > 0$ are hyperparameters, $\mathbf{x}_t$ and $\mathbf{u}_t$ denote the activation and downstream gradient, respectively, observed at the current step (we index the objective by the data available at step t, and use $\mathbf{M}_{t+1}$ for its minimizer, consistent with the algorithmic presentation in sub-Section 3.4). Unlike an exponential moving average, which only tracks a running mean of the past gradients, this objective explicitly asks the momentum matrix to *map* the current activation to its associated gradient, casting momentum as an associative memory that stores input–output correspondences rather than a scalar-decayed history.

3.2 Closed-Form Update

Setting $\nabla_{\mathbf{M}} J(\mathbf{M}) = 0$ and applying the Sherman–Morrison identity (Sherman & Morrison, 1950) to avoid an explicit $d_{\text{in}} \times d_{\text{in}}$ matrix inversion (exploiting the rank-one structure of $\mathbf{x}\mathbf{x}^\top$ (STRANG, 1986)) yields the closed form update rule, with $\gamma \triangleq 2\alpha + \frac{1}{\eta}$: (Appendix A)

$$\mathbf{M}_{t+1} = \frac{1}{\gamma}\left(\frac{\mathbf{M}_t}{\eta} + 2\mathbf{g}_t\right)\left(\mathbf{I} - \frac{2\mathbf{x}_t\mathbf{x}_t^\top}{\gamma + 2\mathbf{x}_t^\top\mathbf{x}_t}\right), \qquad \mathbf{g}_t = \mathbf{u}_t\mathbf{x}_t^\top. \tag{2}$$

We independently re-derive this same update via Fenchel–Rockafellar duality (Rockafellar, 1997), reformulating the primal matrix-space problem into a vector dual problem of dimension d_{out} (Appendix C). The dual solution reveals the exact structural form of update rule as a memory retrieval correction: defining the retrieval error $\mathbf{e}_t = \mathbf{u}_t - \bar{\mathbf{M}}_t\mathbf{x}_t$ (with $\bar{\mathbf{M}}_t \triangleq \mathbf{M}_t/\gamma\eta$), the optimal update is exactly given by:

$$\mathbf{M}_{t+1} = \bar{\mathbf{M}}_t + \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}\,\mathbf{e}_t\mathbf{x}_t^\top. \tag{3}$$

Eq. (3) is a rank-one correction: it modifies the memory exclusively along the direction of the current activation, leaves all the orthogonal directions untouched, and vanishes automatically once the stored association is exact ($\mathbf{e}_t = \mathbf{0}$). Because $J(\mathbf{M})$ is strongly convex for $\alpha, \eta > 0$ (Appendix B), this minimizer is unique at every step.

3.3 Stability and Convergence

Writing the momentum recursion in state-space form, $\mathbf{s}_{t+1} = \mathcal{A}_t\mathbf{s}_t + \mathcal{B}_t\mathbf{u}_t$ with $\mathbf{s}_t = \text{vec}(\mathbf{M}_t)$, the transition operator $\mathcal{A}_t = \rho\,(\mathbf{I}_{d_{\text{out}}} \otimes \mathbf{P}_t)$ is symmetric and depends on the current activation through the Householder-like projector, $\mathbf{P}_t = \mathbf{I} - \frac{2}{\gamma+2\|\mathbf{x}_t\|_2^2}\mathbf{x}_t\mathbf{x}_t^\top$, where $\rho = 1/(2\alpha\eta + 1) \in (0, 1)$. This is a discrete-time Linear Time-Varying (LTV) system, whose poles are *anisotropic*, i.e., the eigenvalue (of system matrix) along the active activation direction shrinks as activation energy grows, while orthogonal directions retain a fixed eigenvalue, ρ (Appendix D). This produces an adaptive notch-filtering behavior absent from isotropic momentum, and it underlies our two central theoretical guarantees, described next.

Theorem 1 (Global Uniform Exponential Stability, informal; full statement and proof in Appendix K)**.** *For every activation sequence $\{\mathbf{x}_t\}_{t\geq 0}$ and every $\alpha, \eta > 0$, the unforced AXIOM dynamics $\mathbf{s}_{t+1} = \mathcal{A}_t\mathbf{s}_t$ satisfy $\|\mathbf{s}_t\|_2 \leq \rho^t\|\mathbf{s}_0\|_2$, i.e., the zero-momentum state $\mathbf{s} = 0$ is globally uniformly exponentially stable, independent of the activation sequence encountered during training.*

Theorem 2 (Non-Convex Stationarity Rate, Informal; full statement and proof in Appendix L)**.** *Under standard L-smoothness, unbiased bounded-variance gradient, and bounded-activation assumptions, the forced stochastic system satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla f(W_t)\|_F^2\big] = \mathcal{O}(1/T) + \mathcal{O}(\sigma_g^2),$$

recovering the classical first order non-convex stationarity rate.

Theorem 1 guarantees that AXIOM's momentum state cannot diverge regardless of how the activation distribution evolves during training, while Theorem 2 guarantees the same asymptotic convergence behavior as SGD-type methods. Neither guarantee requires activation statistics to be well-behaved beyond boundedness, which is what motivates testing AXIOM under the heterogeneous conditions in Section 6. We additionally derive, and defer to the Appendix, an information-theoretic account of AXIOM's entropy dynamics (Appendix G),its responsiveness compared with SGD,Adam (Appendix H), its interaction with dropout (Appendix N), and its fixed-point and cross-pattern interference behavior (Appendix O), all of which are consistent with and reinforce the associative memory interpretation above.

3.4 Algorithm and Practical Integration

AXIOM replaces only the first moment estimator; the second moment estimator is retained from Adam, since it addresses a complementary source of difficulty (stochastic gradient variance) that the present objective does not model (Appendix I). Because the update depends on the per-layer activation $\mathbf{x}_t$ and downstream gradient $\mathbf{u}_t$ rather than only the flattened parameter gradient, we integrate AXIOM via forward and backward hooks that capture $\mathbf{x}_t$ and $\mathbf{u}_t$ at every `nn.Linear` and `nn.Conv2d` layer without modifying model code; convolutional activations are unfolded into patch matrices so that the same rank-one update rule applies to both layer types (Section 5).

Algorithm 1 AXIOM Optimizer Step

Require: Learning rate λ, second-moment coefficient β_2, regularization parameters α, η, numerical stability constant ϵ

Require: $m_0 = 0$, $v_0 = 0$

1: **for** $t = 1, 2, \ldots, T$ **do**

2: Compute gradient $g_t = \nabla_\theta \mathcal{L}(\theta_{t-1})$; Obtain per-layer activation x_t and downstream gradient u_t via hooks ▷ $g_t = u_t x_t^\top$ for each layer

3: $\gamma \leftarrow 2\alpha + 1/\eta$ ▷ Effective regularization coefficient

4: $P_t \leftarrow I - \dfrac{2x_t x_t^\top}{\gamma + 2x_t^\top x_t}$ ▷ Activation-conditioned filtering operator

5: $m_t \leftarrow \dfrac{1}{\gamma}\left(\dfrac{m_{t-1}}{\eta} + 2g_t\right) P_t$ ▷ AXIOM associative-memory update

6: $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t \odot g_t$; $\hat{v}_t \leftarrow v_t/(1-\beta_2^t)$ ▷ Adam second-moment estimate

7: $\theta_t \leftarrow \theta_{t-1} - \lambda\, m_t/(\sqrt{\hat{v}_t} + \epsilon)$ ▷ Parameter update

8: **end for**

Because α and η jointly control both stability margin and responsiveness, we tune them online via softplus-reparameterized hypergradient descent (Baydin et al., 2018, Vancouver, BC, Canada) rather than by hand (Appendix M); this treats the fixed-parameter guarantees of Theorems 1–2 as holding locally under a quasi-static approximation, since the hyper-learning rate is chosen several orders of magnitude smaller than the primary learning rate.

4 Wavelet2VecNet: A Testbed for Activation-Aware Optimization

Theorems 1 and 2 hold irrespective of the activation distribution, however, they do not predict *how much* activation-conditioning should help: that depends on how much the activation geometry actually varies across a task's inputs. We therefore, require a representation learning setting in which activation statistics differ substantially both within and across tasks. We introduce Wavelet2VecNet for this purpose.

4.1 Framework Overview

Each utterance is encoded by a frozen, pretrained Wav2Vec2 backbone into contextual embeddings $X = [x_1, \ldots, x_T] \in \mathbb{R}^{T \times 768}$, generated at approximately 20 ms resolution. Utterances are standardized to a maximum duration of four seconds ($T_{\max} = 200$ frames) via symmetric padding,

together with a binary validity mask $M \in \{0,1\}^{200}$ marking genuine speech frames. A single-level *2-D* Discrete Wavelet Transform (2D-DWT) is then applied to the embedding matrix, decomposing it along both the temporal and embedding axes into four subbands $\{LL, LH, HL, HH\}$, each of size 100×384, stacked into $X_{\text{DWT}} \in \mathbb{R}^{4\times100\times384}$. The validity mask is downsampled by matching 1D max-pooling ($200 \to 100$). Applying wavelet decomposition directly to contextual embeddings (rather than to raw audio or spectrograms) exposes the downstream network, and the optimizer, to multiresolution structure already present in the pretrained representation.

4.2 Architecture

A lightweight CNN backbone processes X_{DWT} through three convolutional blocks ($4 \to 32 \to 64 \to 128$ channels, 3×3 kernels, batch normalization, Gaussian Error Linear Unit (GELU) activations, one 2×2 max-pooling operation), with the validity mask propagated and downsampled in lockstep with every spatial pooling step and applied multiplicatively to zero out padded regions. Masked statistical pooling then aggregates valid temporal positions per channel into mean and standard deviation statistics,i.e.,

$$\mu_c = \frac{\sum_t m_t x_{c,t}}{\sum_t m_t}, \qquad \sigma_c = \sqrt{\frac{\sum_t m_t (x_{c,t} - \mu_c)^2}{\sum_t m_t} + \epsilon},$$

yielding a duration independent descriptor, $h_{\text{stat}} \in \mathbb{R}^{256}$. A projection head (FC $256 \to 128$, batch norm, GELU, dropout p=0.3; FC $128 \to 128$) produces an ℓ_2-normalized *128*-D embedding, followed by a task-specific linear classifier. The backbone contains approximately 115,000 trainable parameters, excluding the frozen Wav2Vec2 encoder.

4.3 Tasks

We evaluate on three problems chosen specifically for their differing activation statistics. **Deepfake source tracing** (ASVspoof2019 LA protocol) requires attributing a spoofed utterance to its generating synthesis or voice conversion system. The network is trained as a six-class classifier with an Additive Angular Margin softmax objective, after which the classifier head is discarded and per-source prototype embeddings are constructed for open set, cosine similarity-based verification against synthesis systems *unseen* during training, evaluated via ROC-AUC and Equal Error Rate (EER). Because this task is close to saturated even under baseline optimizers, it serves as a control condition, where little optimization headroom remains. **Dysarthric speech severity-level assessment** is a four-class problem (High/Mid/Low/Very Low intelligibility) with substantial inter-speaker and inter-session acoustic variability. **Infant cry classification** is a binary (Healthy/Pathology) problem on an involuntary, non-linguistic acoustic signal with its own distinct statistics. We additionally evaluate on **UrbanSound8K** (Salamon et al., 2014, Orlando, Florida, USA) and a DESED-derived **acoustic scene classification** corpus (Turpault et al., 2019, New York, USA), both environmental sound domains with substantially more heterogeneous, non-speech activation statistics than the three speech tasks above, providing the widest span of feature geometry in our evaluation.

5 Experimental Setup

Reproducibility. All the experiments are implemented in PyTorch with fixed seeds (seed$= 42$) across Python, NumPy, and PyTorch CPU/GPU backends, deterministic cuDNN convolution algorithms, and automatic mixed precision with dynamic gradient scaling.

Data Pipeline. 2D-DWT subbands are pre-computed offline and cached in memory-mapped arrays, avoiding repeated disk I/O and on-the-fly wavelet decomposition during training; only CNN forward/backward computation occurs inside the training loop.

Optimizer Configuration. We compare AXIOM against Adam and SGD with momentum, each tuned independently. Adam uses a base learning rate of 3×10^{-4}; SGD uses 1×10^{-2}; AXIOM uses $\eta = 3 \times 10^{-4}$ for the momentum objective together with a separate coefficient of 5×10^{-2} governing its retained SGD-style second moment interaction (sub-Section 3.4). All the optimizers otherwise share batch size, number of epochs, and data augmentation settings within each task.

Table 1: Overall performance comparison (db1 wavelet, Wav2Vec2 backbone). Lower EER is better for Source Tracing; higher accuracy is better elsewhere. Best result per task in bold.

Optimizer	**Dysarthria** (Acc ↑)	**Real-DESED** (Acc ↑)	**Infant Cry** (Acc ↑)	**UrbanSound8K** (Acc ↑)	**Source Tracing** (EER ↓)
Adam	91.48	53.92	87.67	57.18	15.61
SGD	89.08	39.24	86.78	48.25	**13.38**
AXIOM	**93.83**	**61.40**	**90.09**	**71.61**	14.65

Table 2: Seen and unseen source-tracing verification accuracy

Optimizer	**Seen Accuracy**	**Unseen Accuracy**
Adam	99.86	84.38
SGD	**99.93**	**86.62**
AXIOM	99.87	84.30

Backbones and Wavelets. In addition to Wav2Vec2, we evaluate the effect of feature backbone choice using HuBERT-Base, XLSR-300M, and Whisper-base encoders, and the effect of wavelet basis choice using the Haar wavelet (*db1*) and higher order (i.e., vanishing movements) Daubechies wavelets (db2, db4, db6), reporting all four for comprehensive evaluation.

6 Results and Discussion

6.1 Overall Performance

AXIOM achieves the best results on four of the five evaluated tasks (Table 1), improving over Adam by $+2.35$ points on Dysarthria, $+7.48$ on Real-DESED, $+2.42$ on Infant Cry, and $+14.43$ on UrbanSound8K. On Source Tracing, SGD obtains the lowest EER, with AXIOM close behind Adam; since Seen/Unseen verification accuracy already exceeds 99.8%/84% for all the three optimizers (Table 2), this task offers little remaining optimization headroom, and the near-tied result is consistent with (rather than a counterexample to) the hypothesis that activation-conditioned momentum helps most when feature geometry is least saturated and most heterogeneous. The clearest support for this pattern is the monotonic relationship between task heterogeneity and gain size: the two environmental-sound tasks (Real-DESED, UrbanSound8K), which involve the most acoustically diverse and non-stationary activation statistics in our evaluation (as in Table 1), show by far the largest improvements.

6.2 Wavelet Basis Ablation

Table 3 reports the full wavelet basis ablation, including two results that qualify the headline claims of Table 1. **First**, all the three optimizers exhibit a sharp, simultaneous drop under *db2* on the Dysarthria task (Adam: $91.48 \rightarrow 66.72$; SGD: $89.09 \rightarrow 57.53$; AXIOM: $93.83 \rightarrow 62.21$). Because this drop is shared across optimizer choice, we attribute it to an interaction between the db2 decomposition and this particular dataset's temporal structure rather than to any weakness specific to AXIOM. **Second**, and more importantly, AXIOM's advantage on source tracing does *not* generalize across wavelet bases: at db2, db4, and db6, AXIOM's EER is worse than both Adam's and SGD's, and worse than AXIOM's own db1 result. We report this result rather than omitting it: AXIOM's benefit appears to be specifically tied to settings with meaningful activation heterogeneity for the optimizer to exploit, and source tracing (already near ceiling performance under every wavelet and optimizer combination) is the task in our evaluation, where that condition is least met.

6.3 Backbone Robustness

To assess whether AXIOM's benefit is tied to a specific representation, we vary the feature backbone (Table 4) and separately, the decomposition basis (Table 5). AXIOM outperforms both baselines on UrbanSound8K across every backbone and every decomposition tested, and on infant cry in five of

Table 3: Effect of wavelet basis on task performance under each optimizer. Dysarthria, Infant Cry, and UrbanSound8K report accuracy (↑); Source tracing reports EER (↓). Best value per row in bold.

Task	Optimizer	db1	db2	db4	db6
Dysarthria (Acc ↑)	Adam	**91.48**	66.72	87.97	90.14
	SGD	89.09	57.53	87.12	**91.48**
	AXIOM	**93.83**	62.21	92.32	93.09
Infant Cry (Acc ↑)	Adam	87.67	**89.65**	87.89	87.67
	SGD	86.78	**87.22**	85.68	86.78
	AXIOM	**90.09**	89.65	88.77	89.21
UrbanSound8K (Acc ↑)	Adam	57.18	56.10	**62.34**	57.47
	SGD	48.25	**50.49**	44.53	47.45
	AXIOM	71.61	70.52	**71.84**	**71.84**
Source Tracing (EER ↓)	Adam	15.61	14.27	**14.20**	14.52
	SGD	13.38	**12.76**	13.88	14.27
	AXIOM	**14.65**	15.45	15.63	15.51

Table 4: Effect of feature extraction backbone on Infant Cry Classification and UrbanSound8K accuracy (↑), db1 wavelet. Best value for each task, and each feature in bold.

	Infant Cry Classification			UrbanSound8K		
Backbone	Adam	SGD	AXIOM	Adam	SGD	AXIOM
HuBERT-Base	**87.67**	83.26	85.02	61.93	52.95	**78.02**
Wav2Vec2	87.67	86.78	**90.09**	57.18	48.25	**71.61**
XLSR-300M	85.68	85.68	**86.34**	37.35	45.39	**68.46**
Whisper-base	87.22	78.19	**87.89**	69.61	54.67	**78.42**

seven settings, with HuBERT-Base as the one backbone, where Adam edges out AXIOM (87.67 *vs.* 85.02). Taken together, the improvements are not an artifact of a single representation choice, though they are also not universal.

6.4 Discussion

Three findings recur across the ablations. **First**, the size of AXIOM's improvement tracks task heterogeneity: gains are largest on the environmental sound tasks, moderate on dysarthria severity-level and infant cry classification, and negligible-to-negative on the near-saturated source tracing task. **Second**, this pattern is not simply "AXIOM wins everywhere" (it is closer to "AXIOM wins, where the theory predicts it should," which we view as a more informative empirical signal than uniform superiority would be). **Third**, no single wavelet basis or backbone is universally optimal for any optimizer, consistent with the well-established observation that feature geometry is task-dependent; AXIOM's activation-conditioned correction inherits and adapts to whatever geometry the chosen representation induces (Appendix E), rather than imposing one of its own.

7 Limitations

AXIOM's update depends explicitly on per-layer input activations, requiring hook-based access to intermediate representations rather than only the flattened parameter gradient used by conventional optimizers; extending the framework to settings, where activations are inaccessible is an open problem. Our theoretical analysis is restricted to the local linear map $\mathbf{u} = \mathbf{M}\mathbf{x}$ and does *not* explicitly model nonlinearities, normalization layers, residual connections, or attention; every affine layer satisfies this formulation locally, however, a full architectural treatment remains future work. We retain Adam's second moment estimator without a joint theoretical treatment of the coupled first and second moment dynamics (Appendix I). Our stability and convergence guarantees are local and rely on standard smoothness, bounded variance, and bounded activation assumptions that real training runs may violate in practice. AXIOM introduces a modest computational overhead from construct-

Table 5: Effect of wavelet basis (as an alternative decomposition, in place of a learned backbone) on Infant Cry and UrbanSound8K accuracy (↑). Best value per row, for various discrete wavelets in bold.

	Infant Cry			UrbanSound8K		
Wavelet	Adam	SGD	AXIOM	Adam	SGD	AXIOM
rbio2.2	87.44	88.77	**90.53**	58.90	50.14	**72.52**
dmey	89.65	85.02	**92.95**	57.07	44.71	**71.55**
db1	87.67	86.78	**90.09**	57.18	48.25	**71.61**
db4	87.89	85.68	**88.77**	62.34	44.53	**71.84**

ing the rank-one projection operator at every step (Appendix M.3), and its softplus-hypergradient hyperparameter adaptation, however it adds a further wall-clock overhead. Finally, our empirical evaluation is restricted to speech and environmental audio tasks with a common wavelet-based front end; we have not yet evaluated AXIOM on vision, language, or large-scale distributed training settings, and the one setting in our own evaluation, where its advantage disappears (source tracing under non-default wavelets) suggests this advantage should not be assumed to transfer automatically to arbitrary domains.

8 Summary and Conclusions

We introduced AXIOM-a first order optimizer that reformulates momentum as the closed form solution of a regularized associative memory objective rather than an exponential moving average, derived it independently via Sherman–Morrison expansion and Fenchel–Rockafellar duality, and proved global uniform exponential stability and a standard non-convex convergence rate for its resulting dynamics. To test whether this activation-conditioned formulation provides practical benefit, we introduced Wavelet2VecNet-a speech representation framework spanning tasks with deliberately heterogeneous activation statistics, and found that AXIOM's empirical advantage over the traditional Adam and SGD scales with that of *heterogeneity*: largest on environmentally diverse acoustic tasks, moderate on clinical and pathological speech tasks, and absent on a near-saturated verification task. We view this as dose-response pattern, rather than uniform superiority, as the paper's central empirical finding, and we report the wavelet condition under which AXIOM's advantage does not hold rather than omit it. Together, the theoretical and empirical results suggest that conditioning momentum on activation geometry is a promising and computationally inexpensive direction for optimization research, while several open problems – operator-valued bias correction, joint first/second-moment analysis, and validation beyond speech and audio remain potential future research direction. Issues while applying pre-conditions on optimizer at each level, leading to solving an optimization problem inside another, which takes large data to converge whithout overfitting remains open research problem.

Reproducibility Statement

Full derivations of the closed form update (sub-Section 3.2), the Fenchel–Rockafellar dual reformulation, the LTV spectral analysis, the Lyapunov stability argument, and the non-convex convergence proof are provided in Appendices C–L. Pseudocode for the full AXIOM update, including hook-based activation capture and the unified matrix representation used for convolutional layers, is given in Algorithm 1 and Appendix M.1. All the experiments use fixed random seeds and deterministic CUDA kernels as described in Section 5; optimizer hyperparameters for Adam, SGD, and AXIOM are reported in Section 5. We commit to releasing code, pre-computed wavelet subband caches, and training configurations upon publication.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:12449–12460, 2020, Vancouver, British Columbia, Canada.

Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. In *International Conference on Learning Representations (ICLR)*, 2018, Vancouver, BC, Canada.

Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. Nested learning: The illusion of deep learning architectures. *Advances in Neural Information Processing Systems (NeurIPS)*, 38:46968–47002, 2026, Sydney, Australia.

Timothy VP Bliss and Sam F Cooke. Long-term potentiation and long-term depression: a clinical perspective. *Clinics*, 66(suppl 1):3–17, 2011.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems (NeurIPS)*, 36:49205–49233, 2023, New Orleans, Louisiana, USA.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ, 2 edition, 2006. ISBN 978-0-471-24195-9.

Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:9974–10007, 2024, Vancouver, Canada.

Peter Földiak. Forming sparse representations by local anti-hebbian learning. *Biological Cybernetics*, 64(2):165–170, 1990.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023, {Last Accessed: 22nd September, 2026}.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning (ICML)*, pp. 1842–1850. PMLR, 2018, Vienna, Austria.

Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. John Wiley & Sons, New York, 1949.

Chunyan Ji, Thosini Bamunu Mudiyanselage, Yutong Gao, and Yi Pan. A review of infant cry analysis and classification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1): 8, 2021.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6367–6371, 2022, Singapore.

RE Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), 3rd *International Conference on Learning Representations, ICLR, San Diego, CA, USA, 2015*, 2015.

Tao Li, Lei Tan, Zhehao Huang, Qinghua Tao, Yipeng Liu, and Xiaolin Huang. Low dimensional trajectory hypothesis is true: DNNs can be trained in tiny subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3411–3420, 2022.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International Conference on Learning Representations (ICLR)*, pp. 1621–1650, 2024, Vienna, Austria.

Stephane G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

Stéphane G. Mallat. *A Wavelet Tour of Signal Processing, 2nd Edition*. Academic Press, 1999. ISBN 978-0-12-466606-1.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning (ICML)*, pp. 2408–2417. PMLR, 2015, Lille, France.

Andreas Nautsch, Xin Wang, Nicholas Evans, Tomi H Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md Sahidullah, Junichi Yamagishi, and Kong Aik Lee. ASVSpoof 2019: Spoofing countermeasures for the detection of synthesized, converted, and replayed speech. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(2):252–265, 2021.

Vardan Papyan, Xiao Y Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 117 (40):24652–24663, 2020.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton University Press, 1997.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 1041–1044, 2014, Orlando, Florida, USA.

Arth J. Shah, Aniket Pandey, and Hemant A. Patil. Winning against audio deepfakes. *IEEE Signal Processing Letters*, 33:2240–2244, 2026. doi: 10.1109/LSP.2026.3695000.

Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3407–3416, 2023.

Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

Dan Simon. *Optimal State Estimation: Kalman, H-infinity, and Nonlinear Approaches*. Wiley, 2006.

G STRANG. Linear algebra and its applications. *Harcourt Brace*, 5, 1986.

Nicolas Turpault, Romain Serizel, Justin Salamon, and Ankit Parag Shah. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE Workshop)*, pp. 253–257. New York University, 2019, New York, USA.

A Derivation of AXIOM

For given optimizer, objective function is:

$$\begin{aligned} m_{t+1} &= \arg\min_M J(M), \\ &= \arg\min_M \left(\|Mx_t - u_t\|_2^2 + \alpha \|M\|_F^2 + \frac{1}{2\eta} \|M - m_t\|_F^2 \right), \end{aligned} \tag{4}$$

where $\|\cdot\|_2$ denotes the standard vector ℓ_2-norm, defined as $\|v\|_2 = \sqrt{\sum_i v_i^2}$, and $\|\cdot\|_F$ denotes the Frobenius norm for matrices, defined as $\|M\|_F = \sqrt{\sum_{i,j} M_{ij}^2} = \sqrt{\mathrm{Tr}(M^\top M)}$.

For finding *critical* points, where the objective is minimal:

$$\nabla_M J(M) = 0, \tag{5}$$

$$\Rightarrow 0 = 2(Mx_t - u_t)x_t^T + 2\alpha M + \frac{1}{\eta}(M - m_t), \tag{6}$$

$$0 = 2\eta M x_t x_t^T - 2\eta u_t x_t^T + 2\eta\alpha M + M - m_t, \tag{7}$$

$$\Rightarrow m_t + 2\eta u_t x_t^T = M\left(2\eta x_t x_t^T + (2\eta\alpha + 1)I\right), \tag{8}$$

$$M = \left(m_t + 2\eta u_t x_t^T\right)\left(2\eta x_t x_t^T + (2\eta\alpha + 1)I\right)^{-1}, \tag{9}$$

$$M = \left(\frac{m_t}{\eta} + 2u_t x_t^T\right)\left(2x_t x_t^T + \left(2\alpha + \frac{1}{\eta}\right)I\right)^{-1}. \tag{10}$$

Let

$$\gamma = 2\alpha + \frac{1}{\eta}.$$

Then:

$$M = \frac{\gamma}{\gamma}\left(\frac{m_t}{\eta} + 2u_t x_t^T\right)\left(2x_t x_t^T + \gamma I\right)^{-1}, \tag{11}$$

$$M = \frac{1}{\gamma}\left(\frac{m_t}{\eta} + 2u_t x_t^T\right)\left(\frac{2}{\gamma}x_t x_t^T + I\right)^{-1}. \tag{12}$$

Since matrix inversion is computationally expensive to compute directly, and $x_t x_t^\top$ is a rank-one outer product (for $x_t \neq 0$) (STRANG, 1986, Appendix D, p. 480), we can apply the Sherman–Morrison lemma Sherman & Morrison (1950) to obtain a closed form expansion of the inverse.

According to the Sherman-Morrison lemma:

$$(B + ab^T)^{-1} = B^{-1} - \frac{B^{-1}ab^T B^{-1}}{1 + b^T B^{-1}a}. \tag{13}$$

Using this formula, we can extend:

$$\left(\frac{2}{\gamma}x_t x_t^T + I\right)^{-1} = I - \frac{2x_t x_t^T}{\gamma + 2x_t^T x_t}. \tag{14}$$

So finally, the optimal matrix M becomes:

$$M = \frac{1}{\gamma}\left(\frac{m_t}{\eta} + 2u_t x_t^T\right)\left(I - \frac{2x_t x_t^T}{\gamma + 2x_t^T x_t}\right). \tag{15}$$

Since gradient $g = u_t x_t^T$, we can write the expression as:

$$M = \frac{1}{\gamma}\left(\frac{m_t}{\eta} + 2g\right)\left(I - \frac{2x_t x_t^T}{\gamma + 2x_t^T x_t}\right). \tag{16}$$

Therefore, the momentum update is given by:

$$m_{t+1} = M. \tag{17}$$

A.1 Other forms of parameterization

Here, the coefficients α and $\frac{1}{2\eta}$ are arbitrary and can be formulated with any form of parameterization with the same expressivity.

We can show that by taking the objective formulation and hyperparameterizing every term, we set:

$$m_{t+1} = \arg\min_M \alpha_1\|Mx_t - u_t\|_2^2 + \alpha_2\|M\|_F^2 + \alpha_3\|M - \alpha_4 m_t\|_F^2. \tag{18}$$

By following the same derivation, we finally obtain the update rule as:

$$m_{t+1} = \left(\left(\frac{\alpha_1}{\alpha_2 + \alpha_3} \right) u_t x_t^T + \left(\frac{\alpha_3 \alpha_4}{\alpha_2 + \alpha_3} \right) m_t \right) \left(I - \frac{\left(\frac{\alpha_1}{\alpha_2 + \alpha_3} \right) x_t x_t^T}{1 + \left(\frac{\alpha_1}{\alpha_2 + \alpha_3} \right) x_t^T x_t} \right). \tag{19}$$

We can show that parameterizing like this is same as our equation where:

$$\frac{\alpha_1}{\alpha_2 + \alpha_3} = \frac{2}{\gamma}, \tag{20}$$

$$\frac{\alpha_3 \alpha_4}{\alpha_2 + \alpha_3} = \frac{1}{\eta \gamma}. \tag{21}$$

We achieve same expressivity by tuning η, γ, so we follow our update rule with η, γ throughout this paper.

B Convexity Analysis

Given our objective:

$$J(M) = \underbrace{\|Mx - u\|_2^2}_{J_1(M)} + \underbrace{\alpha \|M\|_F^2}_{J_2(M)} + \underbrace{\frac{1}{2\eta} \|M - m_t\|_F^2}_{J_3(M)}. \tag{22}$$

To compute derivatives over matrix spaces, we use directional derivatives (i.e. Fréchet derivatives) along an arbitrary perturbation, H.

Computing the first directional derivative (first variation):

$$DJ(M)[H] = \lim_{\epsilon \to 0} \frac{J(M + \epsilon H) - J(M)}{\epsilon}, \tag{23}$$

$$= \frac{d}{d\epsilon} J(M + \epsilon H) \Big|_{\epsilon = 0}. \tag{24}$$

Computing this for each term:

$$J_1(M + \epsilon H) = \langle (Mx - u) + \epsilon Hx, (Mx - u) + \epsilon Hx \rangle, \tag{25}$$

$$= \|Mx - u\|_2^2 + 2\epsilon \langle Mx - u, Hx \rangle + \epsilon^2 \|Hx\|_2^2. \tag{26}$$

Differentiating w.r.t. ϵ at $\epsilon = 0$:

$$DJ_1(M)[H] = 2\langle Mx - u, Hx \rangle, \tag{27}$$

For the second term:

$$J_2(M + \epsilon H) = \alpha \langle M + \epsilon H, M + \epsilon H \rangle_F, \tag{28}$$

$$= \alpha \|M\|_F^2 + 2\alpha\epsilon \langle M, H \rangle_F + \alpha \epsilon^2 \|H\|_F^2. \tag{29}$$

Differentiating w.r.t. ϵ at $\epsilon = 0$:

$$DJ_2(M)[H] = 2\alpha \langle M, H \rangle_F. \tag{30}$$

For the third term:

$$J_3(M + \epsilon H) = \frac{1}{2\eta} \|M - m_t\|_F^2 + \frac{\epsilon}{\eta} \langle M - m_t, H \rangle_F + \frac{\epsilon^2}{2\eta} \|H\|_F^2. \tag{31}$$

Differentiating w.r.t. ϵ at $\epsilon = 0$:

$$DJ_3(M)[H] = \frac{1}{\eta}\langle M - m_t, H\rangle_F. \tag{32}$$

Computing the second directional derivative (Hessian Operator): The Hessian quadratic form is defined as:

$$\nabla^2 J(M)[H,H] = \frac{d^2}{d\epsilon^2} J(M+\epsilon H)\Big|_{\epsilon=0}. \tag{33}$$

Looking back at our expanded quadratic forms in ϵ^2: For J_1, the coefficient of ϵ^2 is $\|Hx\|_2^2$, so its second derivative gives $2\|Hx\|_2^2$.

Summing J_1, J_2, J_3 second derivatives gives the Hessian Operator:

$$\nabla^2 J(M)[H,H] = 2\|Hx\|_2^2 + 2\alpha\|H\|_F^2 + \frac{1}{\eta}\|H\|_F^2, \tag{34}$$

$$\Rightarrow \nabla^2 J(M)[H,H] = 2\|Hx\|_2^2 + \left(2\alpha + \frac{1}{\eta}\right)\|H\|_F^2, \tag{35}$$

$$\Rightarrow \nabla^2 J(M)[H,H] = 2\|Hx\|_2^2 + \gamma\|H\|_F^2. \tag{36}$$

A function J is strongly convex with parameter $\mu > 0$ if its Hessian satisfies:

$$\nabla^2 J(M)[H,H] \geq \mu\|H\|_F^2 \quad \forall H \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}} \tag{37}$$

Evaluating our lower bound:

1. $\|Hx\|_2^2 \geq 0$,
2. Assuming $\alpha > 0$ and $\eta > 0$:

$$\left(2\alpha + \frac{1}{\eta}\right) > 0 \implies \gamma > 0. \tag{38}$$

Therefore, if $\mu = 2\alpha + \frac{1}{\eta}$:

$$\nabla^2 J(M)[H,H] \geq \mu\|H\|_F^2 > 0, \quad \text{for all } H \neq 0. \tag{39}$$

Since we showed that for $\alpha, \eta > 0$, $J(M)$ is strongly convex, it is a convex optimization problem.Therefore, the closed-form update $m_{t+1} = M$ obtained above is the unique global minimizer of the objective.

C Dual Optimization via Fenchel–Rockafellar Duality

Directly solving the primal optimization problem over the matrix space $\mathbb{R}^{d\times n}$ incurs high memory and computational overhead. By appealing to Fenchel-Rockafellar duality, we reformulate the problem into the vector dual space, $\mathbb{R}^d$, proving that the optimal matrix update is a rank-1 perturbation driven by a dual vector, $\mathbf{y}^* \in \mathbb{R}^d$.

C.1 Primal Problem Formulation

Consider the unconstrained regularized objective at time step t:

$$\min_{\mathbf{M}\in\mathbb{R}^{d\times n}} \quad \mathcal{J}(\mathbf{M}) \triangleq \|\mathbf{M}\mathbf{x} - \mathbf{u}\|_2^2 + \alpha\|\mathbf{M}\|_F^2 + \frac{1}{2\eta}\|\mathbf{M} - \mathbf{M}_t\|_F^2. \tag{40}$$

We begin by consolidating the Frobenius norm terms into a single quadratic penalty. Expanding the regularizers:

$$\begin{aligned}
\alpha\|\mathbf{M}\|_F^2 + \frac{1}{2\eta}\|\mathbf{M} - \mathbf{M}_t\|_F^2 &= \left(\alpha + \frac{1}{2\eta}\right)\|\mathbf{M}\|_F^2 - \frac{1}{\eta}\langle \mathbf{M}, \mathbf{M}_t\rangle + \frac{1}{2\eta}\|\mathbf{M}_t\|_F^2, \\
&= \frac{\gamma}{2}\left(\|\mathbf{M}\|_F^2 - 2\left\langle \mathbf{M}, \frac{1}{\gamma\eta}\mathbf{M}_t\right\rangle\right) + \frac{1}{2\eta}\|\mathbf{M}_t\|_F^2, \\
&= \frac{\gamma}{2}\left\|\mathbf{M} - \bar{\mathbf{M}}_t\right\|_F^2 + C_t.
\end{aligned} \tag{41}$$

where the scalar γ, baseline matrix $\bar{\mathbf{M}}_t$, and constant C_t are defined as:

$$\gamma \triangleq 2\alpha + \frac{1}{\eta}, \qquad \bar{\mathbf{M}}_t \triangleq \frac{1}{\gamma\eta}\mathbf{M}_t, \qquad C_t \triangleq \frac{1}{2\eta}\|\mathbf{M}_t\|_F^2 - \frac{\gamma}{2}\|\bar{\mathbf{M}}_t\|_F^2. \tag{42}$$

Dropping C_t ,as it doesnt change optimization decision, the equivalent constrained primal problem with auxiliary residual $\mathbf{r} \in \mathbb{R}^d$ is:

$$\begin{aligned} \min_{\mathbf{M},\mathbf{r}} \quad & f(\mathbf{r}) + g(\mathbf{M}) \\ \text{s.t.} \quad & \mathbf{r} = \mathbf{M}\mathbf{x} - \mathbf{u}, \end{aligned} \tag{43}$$

where $f(\mathbf{r}) \triangleq \|\mathbf{r}\|_2^2$ and $g(\mathbf{M}) \triangleq \frac{\gamma}{2}\|\mathbf{M} - \bar{\mathbf{M}}_t\|_F^2$.

C.2 Fenchel–Rockafellar Dual Derivation

Let $\mathcal{A} : \mathbb{R}^{d\times n} \to \mathbb{R}^d$ be the bounded linear operator $\mathcal{A}(\mathbf{M}) = \mathbf{M}\mathbf{x}$, with its adjoint operator $\mathcal{A}^* : \mathbb{R}^d \to \mathbb{R}^{d\times n}$ given by $\mathcal{A}^*(\mathbf{y}) = \mathbf{y}\mathbf{x}^\top$.

The Fenchel–Rockafellar dual problem is formulated as:

$$\max_{\mathbf{y}\in\mathbb{R}^d} \quad \mathcal{D}(\mathbf{y}) \triangleq -f^*(-\mathbf{y}) - g^*(\mathcal{A}^*\mathbf{y}) - \mathbf{y}^\top\mathbf{u}. \tag{44}$$

1. Fenchel Conjugate of $f(\mathbf{r})$: The convex conjugate $f^*(\mathbf{v}) = \sup_{\mathbf{r}\in\mathbb{R}^d}\{\mathbf{v}^\top\mathbf{r} - \|\mathbf{r}\|_2^2\}$. Setting the gradient to zero yields, $\mathbf{r}^* = \frac{1}{2}\mathbf{v}$. Substituting back:

$$f^*(\mathbf{v}) = \frac{1}{4}\|\mathbf{v}\|_2^2 \implies f^*(-\mathbf{y}) = \frac{1}{4}\|\mathbf{y}\|_2^2. \tag{45}$$

2. Fenchel Conjugate of $g(\mathbf{M})$: The conjugate $g^*(\mathbf{Z}) = \sup_{\mathbf{M}\in\mathbb{R}^{d\times n}}\{\langle\mathbf{Z},\mathbf{M}\rangle - \frac{\gamma}{2}\|\mathbf{M}-\bar{\mathbf{M}}_t\|_F^2\}$. Setting $\nabla_{\mathbf{M}} = \mathbf{Z} - \gamma(\mathbf{M} - \bar{\mathbf{M}}_t) = 0$ yields $\mathbf{M}^* = \bar{\mathbf{M}}_t + \frac{1}{\gamma}\mathbf{Z}$. Substituting back:

$$g^*(\mathbf{Z}) = \langle\mathbf{Z}, \bar{\mathbf{M}}_t\rangle + \frac{1}{2\gamma}\|\mathbf{Z}\|_F^2. \tag{46}$$

Evaluating g^* at $\mathbf{Z} = \mathcal{A}^*\mathbf{y} = \mathbf{y}\mathbf{x}^\top$:

$$g^*(\mathbf{y}\mathbf{x}^\top) = \langle\mathbf{y}\mathbf{x}^\top, \bar{\mathbf{M}}_t\rangle + \frac{1}{2\gamma}\|\mathbf{y}\mathbf{x}^\top\|_F^2 = \mathbf{y}^\top\bar{\mathbf{M}}_t\mathbf{x} + \frac{\|\mathbf{x}\|_2^2}{2\gamma}\|\mathbf{y}\|_2^2. \tag{47}$$

3. Objective Assembly and Strong Duality: Combining the conjugate expressions yields the exact concave dual objective function:

$$\mathcal{D}(\mathbf{y}) = -\frac{1}{4}\|\mathbf{y}\|_2^2 - \frac{\|\mathbf{x}\|_2^2}{2\gamma}\|\mathbf{y}\|_2^2 + \mathbf{y}^\top(\mathbf{u} - \bar{\mathbf{M}}_t\mathbf{x}). \tag{48}$$

To establish strong duality, we verify the Fenchel–Rockafellar qualification condition. In the present finite-dimensional setting, both $f(\mathbf{r}) = \|\mathbf{r}\|_2^2$ and $g(\mathbf{M}) = \frac{\gamma}{2}\|\mathbf{M}-\bar{\mathbf{M}}_t\|_F^2$ are finite-valued quadratic functions defined over their entire ambient vector spaces $\mathbb{R}^d$ and $\mathbb{R}^{d\times n}$, respectively. Hence, they are proper, convex, lower-semicontinuous, and continuous everywhere. In particular,

$$\mathrm{dom}(f) = \mathbb{R}^d, \qquad \mathrm{dom}(g) = \mathbb{R}^{d\times n}, \tag{49}$$

and

$$\mathrm{cont}(f) = \mathbb{R}^d, \qquad \mathrm{cont}(g) = \mathbb{R}^{d\times n}. \tag{50}$$

For the Fenchel–Rockafellar formulation, the primal problem is of the form

$$\inf_{\mathbf{M}} \left\{ g(\mathbf{M}) + f(\mathcal{A}\mathbf{M} - \mathbf{u}) \right\}, \tag{51}$$

where $\mathcal{A}(\mathbf{M}) = \mathbf{M}\mathbf{x}$. Since $\mathrm{dom}(g) = \mathbb{R}^{d\times n}$, we have

$$\mathcal{A}\big(\mathrm{dom}(g)\big) = \mathrm{Range}(\mathcal{A}) \subseteq \mathbb{R}^d. \tag{52}$$

Because f is continuous everywhere,

$$\operatorname{cont}(f) = \mathbb{R}^d. \tag{53}$$

Therefore,

$$\mathcal{A}\big(\operatorname{dom}(g)\big) \cap \operatorname{cont}(f) = \operatorname{Range}(\mathcal{A}) \neq \emptyset, \tag{54}$$

since $\mathbf{0} \in \operatorname{Range}(\mathcal{A})$.

Thus, the Fenchel–Rockafellar qualification condition is satisfied. Consequently, the Fenchel–Rockafellar duality theorem guarantees zero duality gap and attainment of the dual supremum:

$$\inf_{\mathbf{M}\in\mathbb{R}^{d\times n}} \mathcal{J}(\mathbf{M}) = \max_{\mathbf{y}\in\mathbb{R}^d} \mathcal{D}(\mathbf{y}). \tag{55}$$

Furthermore, since $\gamma > 0$, the primal objective is strongly convex and therefore, admits a unique global minimizer, $\mathbf{M}^*$. The dual objective is explicitly given by

$$\mathcal{D}(\mathbf{y}) = -\left(\frac{1}{4} + \frac{\|\mathbf{x}\|_2^2}{2\gamma}\right)\|\mathbf{y}\|_2^2 + \mathbf{y}^\top(\mathbf{u} - \bar{\mathbf{M}}_t\mathbf{x}), \tag{56}$$

whose Hessian is

$$\nabla_{\mathbf{y}}^2\mathcal{D}(\mathbf{y}) = -\left(\frac{1}{2} + \frac{\|\mathbf{x}\|_2^2}{\gamma}\right) I \prec 0. \tag{57}$$

Hence, $\mathcal{D}$ is strictly concave, and the dual optimization problem admits a unique maximizer, $\mathbf{y}^* \in \mathbb{R}^d$.

C.3 Closed-Form Dual Solution & Primal Reconstruction

To find the optimal dual variable $\mathbf{y}^*$, we compute the gradient of $\mathcal{D}(\mathbf{y})$:

$$\nabla_{\mathbf{y}}\mathcal{D}(\mathbf{y}) = -\frac{1}{2}\mathbf{y} - \frac{\|\mathbf{x}\|_2^2}{\gamma}\mathbf{y} + (\mathbf{u} - \bar{\mathbf{M}}_t\mathbf{x}) = 0. \tag{58}$$

Optimal $\mathbf{y}^*$ is:

$$\left(\frac{\gamma + 2\|\mathbf{x}\|_2^2}{2\gamma}\right)\mathbf{y}^* = \mathbf{u} - \bar{\mathbf{M}}_t\mathbf{x} \implies \mathbf{y}^* = \frac{2\gamma}{\gamma + 2\|\mathbf{x}\|_2^2}(\mathbf{u} - \bar{\mathbf{M}}_t\mathbf{x}). \tag{59}$$

By the Fenchel–Rockafellar extremal relation, the optimal primal matrix $\mathbf{M}^*$ is uniquely recovered via $\mathbf{M}^* = \nabla g^*(\mathcal{A}^*\mathbf{y}^*)$:

$$\mathbf{M}^* = \bar{\mathbf{M}}_t + \frac{1}{\gamma}\mathbf{y}^*\mathbf{x}^\top. \tag{60}$$

Substituting $\mathbf{y}^*$ into the reconstruction formula yields the exact rank-1 update:

$$\mathbf{M}^* = \bar{\mathbf{M}}_t + \frac{2}{\gamma + 2\|\mathbf{x}\|_2^2}(\mathbf{u} - \bar{\mathbf{M}}_t\mathbf{x})\mathbf{x}^\top. \tag{61}$$

C.4 Structural Interpretation of the Optimal Memory Update

The Fenchel–Rockafellar reconstruction provides considerably more than an alternative derivation of the optimizer. It reveals the intrinsic structural form of the optimal momentum update:

$$\mathbf{M}_{t+1} = \bar{\mathbf{M}}_t + \frac{2}{\gamma + 2\|\mathbf{x}\|_2^2}\left(\mathbf{u} - \bar{\mathbf{M}}_t\mathbf{x}\right)\mathbf{x}^\top. \tag{62}$$

This expression admits a natural interpretation from the perspective of *associative memory*. The matrix $\mathbf{M}$ may be viewed as a memory operator that stores an association between an activation vector, $\mathbf{x}$ and its desired downstream response, $\mathbf{u}$. The ideal associative memory therefore satisfies:

$$\mathbf{M}\mathbf{x} = \mathbf{u}, \tag{63}$$

meaning that the stored memory perfectly reproduces the desired response, when queried by the corresponding activation.

The quantity

$$\mathbf{e} = \mathbf{u} - \bar{\mathbf{M}}_t \mathbf{x}. \tag{64}$$

is therefore the *memory retrieval error*, representing the discrepancy between the desired association and the response produced by the current memory state. Substituting this residual into the reconstruction formula gives:

$$\mathbf{M}_{t+1} = \bar{\mathbf{M}}_t + \frac{2}{\gamma + 2\|\mathbf{x}\|_2^2} \mathbf{e}\mathbf{x}^\top. \tag{65}$$

Several structural properties follow immediately.

- **Prediction-driven learning.** When the stored memory already satisfies the desired association, i.e.,

  $$\bar{\mathbf{M}}_t \mathbf{x} = \mathbf{u},$$

  the retrieval error vanishes ($\mathbf{e} = \mathbf{0}$), implying

  $$\mathbf{M}_{t+1} = \bar{\mathbf{M}}_t.$$

  Hence, the optimizer performs no unnecessary modification once the correct association has been learned. Learning is driven entirely by prediction error.

- ***Rank-1* memory correction.** The correction term

  $$\mathbf{e}\mathbf{x}^\top$$

  is an outer product and therefore, has rank one (STRANG, 1986, Appendix D, p. 480). Consequently, each optimization step modifies the existing memory through the smallest possible nonzero low-rank perturbation rather than an arbitrary dense matrix update. This structured correction updates the memory only along the direction induced by the current activation, preserving the remaining degrees of freedom of the operator.

- **Localized associative modification.** For every feature direction $\mathbf{z}$ satisfying

  $$\mathbf{x}^\top \mathbf{z} = 0,$$

  the rank-1 correction satisfies

  $$(\mathbf{e}\mathbf{x}^\top)\mathbf{z} = \mathbf{e}(\mathbf{x}^\top \mathbf{z}) = \mathbf{0}.$$

  Therefore, the current optimization step modifies only the subspace associated with the active activation vector while leaving all the orthogonal feature directions unchanged. Each training sample therefore performs a localized memory correction without interfering with unrelated associations.

  This structured correction is conceptually analogous to the low-dimensional adaptation frequently observed in deep learning, where optimization is often dominated by a small number of informative directions despite the high dimensionality of the parameter space.(Li et al. (2022))

From the perspective of nested learning, the optimization process may therefore, be interpreted as the iterative refinement of an associative memory. Rather than directly descending along parameter space, AXIOM repeatedly queries the current memory using the activation $\mathbf{x}$, measures the resulting retrieval error, i.e.,

$$\mathbf{u} - \mathbf{M}\mathbf{x},$$

and applies the unique optimal rank-1 correction that minimizes the local regularized objective. As learning progresses, as the retrieval error approaches zero, the corresponding memory updates naturally vanish. In this view, convergence corresponds to the stabilization of the associative memory, where each stored activation retrieves its desired response with minimal reconstruction error.

Although derived purely from convex optimization principles, the resulting update architecture is remarkably reminiscent of several well-known learning paradigms. The prediction–residual–correction structure closely resembles the computational motif found in hierarchical predictive coding models, in which predictions are generated, residual prediction errors are computed, and these

errors iteratively refine the underlying representation (Rao & Ballard, 1999, pp. 79–80). Likewise, the additive correction through a gain-scaled residual is structurally analogous to recursive state estimators, such as the Kalman filter, where state estimates are recursively refined using newly acquired observations (Kalman, 1960, pp. 40–42). Furthermore, the fact that each optimization step is realized as a rank-one correction naturally echoes observations that learning dynamics in deep neural networks often evolve through low-dimensional update directions rather than unrestricted modifications of the entire parameter space Li et al. (2022). We do not claim equivalence with these frameworks; rather, it is intriguing that such computational principles emerge naturally as a consequence of the optimal solution of the proposed regularized objective.

The localized associative modification is conceptually reminiscent of selective attention mechanisms in biological learning, where synaptic plasticity is preferentially gated toward task-relevant neural representations while unrelated pathways remain largely unaffected. We emphasize that this analogy is purely *structural* rather than mechanistic.

D LTV & Dynamical Systems Analysis of AXIOM

Why AXIOM is a Discrete-Time Linear Time-Varying (LTV) System?

Our momentum update rule is:

$$m_{t+1} = \frac{1}{\gamma}\left(\frac{1}{\eta}m_t + 2g_t\right)P_t, \tag{66}$$

where m_t and g_t denote a row of the memory matrix and its corresponding gradient, respectively, with $m_t, g_t \in \mathbb{R}^{1\times d_{\text{in}}}$, $\gamma = 2\alpha + \frac{1}{\eta}$, and P_t is the symmetric matrix:

$$P_t = I_{d_{\text{in}}} - \frac{2}{\gamma + 2\|x_t\|_2^2}x_t x_t^\top. \tag{67}$$

By considering each output row independently and representing the row vectors as column vectors, we define:

$$\begin{aligned} \text{State vector:} \quad & s_t = m_t^\top \in \mathbb{R}^{d_{\text{in}}}, \\ \text{Control input vector:} \quad & \tilde{g}_t = g_t^\top \in \mathbb{R}^{d_{\text{in}}}. \end{aligned}$$

Taking the transpose of the row-wise momentum update and using the symmetry $P_t^\top = P_t$ gives:

$$\begin{aligned} s_{t+1} &= m_{t+1}^\top, \\ &= \frac{1}{\gamma}P_t^\top\left(\frac{1}{\eta}m_t^\top + 2g_t^\top\right), \\ &= \frac{1}{\gamma\eta}P_t s_t + \frac{2}{\gamma}P_t\tilde{g}_t. \end{aligned} \tag{68}$$

Thus, the momentum dynamics can be written in the standard discrete-time state-space form:

$$s_{t+1} = A_t s_t + B_t\tilde{g}_t, \tag{69}$$

where the transition matrices are:

$$A_t = \rho P_t, \qquad B_t = \frac{2}{\gamma}P_t, \tag{70}$$

with nominal scalar damping factor ρ:

$$\rho = \frac{1}{\gamma\eta} = \frac{1}{2\alpha\eta + 1}, \qquad \rho \in (0, 1). \tag{71}$$

Only the state equation is considered, as the analysis focuses on the internal momentum dynamics. The output equation $y_t = C_t s_t + D_t\tilde{g}_t$ is omitted since the optimizer does not define a separate *observable* output.

Lemma 1 (Symmetry of the Projection Operator)**.** *The matrix*

$$P_t = I_{d_{\text{in}}} - \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}\mathbf{x}_t\mathbf{x}_t^\top \tag{72}$$

is symmetric.

Proof. Since the identity matrix is symmetric and

$$\left(\mathbf{x}_t\mathbf{x}_t^\top\right)^\top = \mathbf{x}_t\mathbf{x}_t^\top, \tag{73}$$

the rank-one matrix $\mathbf{x}_t\mathbf{x}_t^\top$ is symmetric. The scalar coefficient $\frac{2}{\gamma+2\|\mathbf{x}_t\|_2^2}$ is real-valued. Therefore,

$$P_t^\top = I_{d_{\text{in}}} - \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}\mathbf{x}_t\mathbf{x}_t^\top = P_t. \tag{74}$$

□

Interestingly, the resulting state-space model bears structural resemblance to selective state-space models such as Mamba(Gu & Dao (2023, {Last Accessed: 22nd September, 2026})), in which the state transition matrices are conditioned on the current input. Unlike Mamba, however, the time-varying transition in AXIOM is not learned through parameterized gating, but arises analytically from the closed form solution of the underlying regularized optimization problem.

Since the same transition operator P_t acts independently on every row of the memory matrix, the full matrix-valued dynamics decompose into d_{out} identical row-wise systems. To see this explicitly, consider the homogeneous part of the matrix-valued state equation:

$$M_{t+1}^{(h)} = \rho M_t^{(h)} P_t. \tag{75}$$

Writing $M_t \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ row-wise gives

$$M_t = \begin{bmatrix} m_t^{(1)} \\ m_t^{(2)} \\ \vdots \\ m_t^{(d_{\text{out}})} \end{bmatrix}, \qquad m_t^{(i)} \in \mathbb{R}^{1 \times d_{\text{in}}}, \tag{76}$$

and therefore,

$$M_{t+1}^{(h)} = \begin{bmatrix} \rho m_t^{(1)} P_t \\ \rho m_t^{(2)} P_t \\ \vdots \\ \rho m_t^{(d_{\text{out}})} P_t \end{bmatrix}. \tag{77}$$

Thus, all d_{out} rows evolve under the same $d_{\text{in}} \times d_{\text{in}}$ transition matrix, P_t. Representing each row as a column vector, the corresponding state-transition matrix is given by:

$$A_t = \rho P_t^\top = \rho P_t, \tag{78}$$

where the second equality follows from the symmetry of P_t.

Equivalently, the complete matrix-valued dynamics can be represented in vectorized form as:

$$\text{vec}(M_t P_t) = (P_t^\top \otimes I_{d_{\text{out}}})\,\text{vec}(M_t), \tag{79}$$

which gives the full $d_{\text{out}}d_{\text{in}}$-dimensional transition matrix

$$\mathcal{A}_t = \rho(P_t^\top \otimes I_{d_{\text{out}}}) = \rho(P_t \otimes I_{d_{\text{out}}}). \tag{80}$$

The eigenvalues of a Kronecker product are the pairwise products of the eigenvalues of its factors. Since all the eigenvalues of $I_{d_{\text{out}}}$ are equal to one,

$$\text{eig}(\mathcal{A}_t) = \{\rho\lambda_i(P_t)\}, \tag{81}$$

where each eigenvalue of the d_{in}-dimensional row-wise system is repeated d_{out} times. Hence, the row-wise formulation captures all the distinct poles of the full matrix-valued system, while the full formulation introduces only their multiplicities. Consequently, it suffices to analyze a single d_{in}-dimensional row-wise system, when characterizing the distinct poles of AXIOM.

D.1 Why is it Linear Time-Varying (LTV)?

- **Linear:** The state transition depends linearly on the current memory state and input. For each output row,

$$m_{t+1}^{(i)} = \rho m_t^{(i)} P_t + \frac{2}{\gamma} g_t^{(i)} P_t, \tag{82}$$

  so no nonlinear operation is applied to the memory state itself.

- **Time-Varying:** The state-transition matrix

$$A_t = \rho P_t(x_t) \tag{83}$$

  depends explicitly on the current activation x_t, which generally changes at every optimization step, t. Therefore, the transition operator can vary over time.

D.2 Local LTI Analysis & Poles (Eigenvalue Spectral Analysis)

To analyze transient response and stability around a fixed step t, we treat the operator locally as a Linear Time-Invariant (LTI) system with matrix, A_t. The poles of this system are the eigenvalues $\lambda_i(A_t)$ of the state transition matrix, $A_t = \rho P_t$.

Theorem 1 (Pole Locations and Spatial Anisotropy):
At any step t, the eigenvalues $\lambda_i(A_t)$ satisfy:

1. **Along the active activation direction (x_t):**

$$A_t x_t = \rho \left(I - \frac{2 x_t x_t^\top}{\gamma + 2\|x_t\|_2^2} \right) x_t = \rho \left(1 - \frac{2\|x_t\|_2^2}{\gamma + 2\|x_t\|_2^2} \right) x_t, \tag{84}$$

$$\begin{aligned} \implies \lambda_1(A_t) &= \frac{\rho\gamma}{\gamma + 2\|x_t\|_2^2}, \\ &= \frac{1}{(2\alpha\eta + 1)\left(1 + \frac{2}{\gamma}\|x_t\|_2^2\right)}, \\ &= \frac{1}{\eta\left(\gamma + 2\|x_t\|_2^2\right)}. \end{aligned} \tag{85}$$

2. **Along orthogonal invariant directions ($v \perp x_t$):**

$$A_t v = \rho(I - 0)v = \rho v, \tag{86}$$

$$\implies \lambda_i(A_t) = \rho = \frac{1}{2\alpha\eta + 1} = \frac{1}{\eta\gamma}, \quad \forall i \in \{2, \ldots, d_{\text{in}}\}. \tag{87}$$

D.3 Physical & Control-Theoretic Meaning of Anisotropic Pole Placement

- **BIBO Stability:** Since $\alpha > 0$ and $\eta > 0$ (due to strong convexity), all the eigenvalues strictly lie inside the unit circle on the complex z-plane:

$$|\lambda_i(A_t)| < 1 \quad \forall i. \tag{88}$$

  This guarantees local Bounded-Input Bounded-Output (BIBO) stability.

- **Anisotropic Notch Filtering:**
  1. Classical Momentum (SGD/Adam) places poles isotropically ($A_t = \rho I$), treating all directions uniformly.
  2. AXIOM shifts its primary pole λ_1 dynamically based on activation energy, $\|x_t\|_2^2$. As $\|x_t\|_2^2 \to \infty$, $\lambda_1 \to 0$. This acts as an adaptive low-pass notch filter that heavily damps momentum memory along high variance activation directions in order to prevent overshooting, while preserving orthogonal memory directions at $\lambda_i = \rho$, $\forall\, i \in \{2, \ldots, d_{\text{in}}\}$.

We refer to this behavior as *anisotropic notch filtering* because the memory dynamics selectively attenuate directions according to their angular alignment with the current activation. Consider a memory direction v forming an angle θ with x_t. At $\theta = 0^\circ$, the direction is fully aligned with the activation and experiences the strongest attenuation, with effective eigenvalue $\lambda_\parallel$. As θ increases toward 90°, the contribution of the activation-aligned component decreases and the effective retention increases toward $\lambda_\perp$. Sweeping θ from 0° through 360° therefore traces a directional filtering profile in which the activation-aligned region is preferentially suppressed while orthogonal directions are retained more strongly. In this sense, the term 'notch" describes the selective attenuation of a region of the directional state space, while 'anisotropic" emphasizes that this attenuation depends on direction rather than being uniform across all memory components. Since the dependence is through $\cos^2\theta$, the same directional profile repeats every 180°.

E Directional Memory Dynamics and Adaptive Anisotropic Filtering

The previous *duality* analysis established that the optimal momentum update is realized as a *rank-1* associative memory correction,

$$\mathbf{M}_{t+1} = \bar{\mathbf{M}}_t + \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}\left(\mathbf{u}_t - \bar{\mathbf{M}}_t\mathbf{x}_t\right)\mathbf{x}_t^\top,$$

revealing *where* new information is written into memory. The present dynamical systems analysis provides the complementary perspective, characterizing *how* previously stored memory evolves through time. Together, these two viewpoints yield a unified interpretation of AXIOM as an associative memory system that simultaneously performs localized memory writing and direction-dependent memory retention.

The eigenvalue analysis above characterizes the two extremal directions of the state dynamics: the active activation direction $\mathbf{x}_t$ and its *orthogonal complement*. The behavior along an arbitrary direction follows naturally from the spectral decomposition of the state transition matrix.

Let

$$\hat{\mathbf{x}}_t = \frac{\mathbf{x}_t}{\|\mathbf{x}_t\|_2},$$

denote the normalized activation direction, and let $\mathbf{q}$ be any unit vector satisfying

$$\mathbf{q}^\top \hat{\mathbf{x}}_t = 0.$$

Any unit direction in the two-dimensional subspace spanned by $\{\hat{\mathbf{x}}_t, \mathbf{q}\}$ can be expressed as:

$$\mathbf{v} = \cos\theta\hat{\mathbf{x}}_t + \sin\theta\mathbf{q}, \qquad \|\mathbf{v}\|_2 = 1. \tag{89}$$

Since $\|\mathbf{v}\|_2 = 1$, every possible orientation of the state space is represented by a point on the unit hypersphere $S^{d_{\text{in}}-1}$. Consequently, the analysis no longer concerns a single activation direction, but the complete angular distribution of the state dynamics over all possible directions. The *Rayleigh quotient* therefore provides a geometric characterization of how the effective memory retention varies continuously across the hypersphere according to the alignment between the current state direction and the activation vector.

Since both $\hat{\mathbf{x}}_t$ and $\mathbf{q}$ are eigenvectors of the symmetric operator A_t, the effective state transition coefficient along $\mathbf{v}$ is given by the Rayleigh quotient (STRANG, 1986, Sec. 6.4, Eq. (11), pp. 379–380)

$$\lambda_{\text{eff}}(\theta) = \mathbf{v}^\top A_t \mathbf{v} = \lambda_\parallel \cos^2\theta + \lambda_\perp \sin^2\theta, \tag{90}$$

where

$$\lambda_\parallel = \frac{\rho\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2}, \qquad \lambda_\perp = \rho,$$

with $\lambda_\parallel$ representing the eigenvalue of A_t in the one-dimensional eigenspace $\text{span}\{\hat{\mathbf{x}}_t\}$ and $\lambda_\perp$ representing the repeated eigenvalue on the $(d_{\text{in}} - 1)$-dimensional orthogonal subspace

$$\{\mathbf{v} \in \mathbb{R}^{d_{\text{in}}} : \mathbf{v}^\top \hat{\mathbf{x}}_t = 0\}.$$

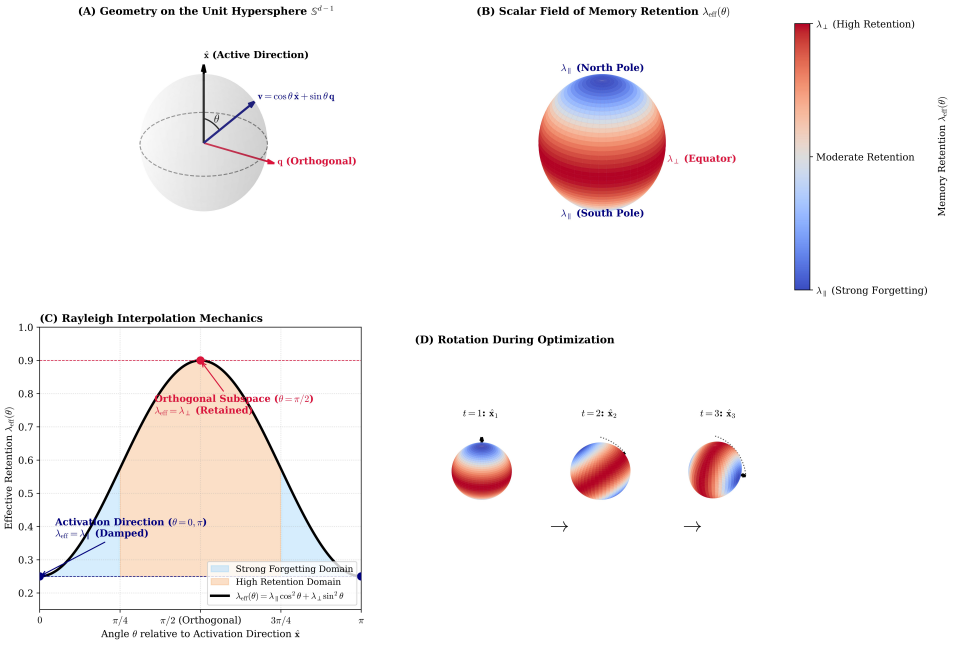

Figure . Directional Memory Dynamics of AXIOM(proposed).
The Rayleigh quotient induces a continuous memory-retention field over the unit hypersphere. The field rotates together with the current activation, producing adaptive anisotropic memory retention.

Figure 1: **Directional Memory Dynamics of AXIOM.** (A) Every unit direction corresponds to a point on the unit hypersphere $S^{d_{\text{in}}-1}$. (B) The Rayleigh quotient defines a continuous memory-retention field over the hypersphere, assigning each direction an effective retention coefficient $\lambda_{\text{eff}}(\theta)$. (C) Memory retention varies smoothly with the angle from the current activation, interpolating continuously between $\lambda_\parallel$ and $\lambda_\perp$. (D) As the activation changes during optimization, the principal damping direction rotates, causing the anisotropic memory-retention field to adapt dynamically to the incoming data.

This expression demonstrates that the state transition dynamics vary continuously with orientation in the parameter space rather than changing abruptly between the *active* and *inactive* directions. Maximum attenuation occurs along the active activation direction ($\theta = 0, \pi$), where $\lambda_{\text{eff}} = \lambda_\parallel$, while minimum attenuation occurs along orthogonal directions ($\theta = \pi/2, 3\pi/2$), where $\lambda_{\text{eff}} = \lambda_\perp$. Intermediate directions exhibit a smooth interpolation between these two limiting cases.

Although the preceding analysis is expressed in terms of a single row of the memory matrix, the same transition operator acts simultaneously on all d_{out} rows of $\mathbf{M}_t$. Thus, the complete matrix-valued state evolves in a $d_{\text{out}}d_{\text{in}}$-dimensional parameter space. To make this structure explicit, consider the homogeneous state dynamics:

$$\mathbf{M}_{t+1}^{(h)} = \rho \mathbf{M}_t^{(h)} P_t. \tag{91}$$

Vectorizing the matrix gives:

$$\text{vec}\left(\mathbf{M}_{t+1}^{(h)}\right) = \rho \left(P_t^\top \otimes I_{d_{\text{out}}}\right) \text{vec}\left(\mathbf{M}_t^{(h)}\right). \tag{92}$$

Since P_t is symmetric, the full $d_{\text{out}}d_{\text{in}}$-dimensional state-transition operator is therefore:

$$\mathcal{A}_t = \rho(P_t \otimes I_{d_{\text{out}}}). \tag{93}$$

The spectrum of $\mathcal{A}_t$ contains the same two distinct eigenvalues as the row-wise system, but with multiplicities induced by the d_{out} independent rows. Specifically, the activation-aligned eigenspace consists of matrix perturbations of the form:

$$\mathbf{H} = \mathbf{a}\hat{\mathbf{x}}_t^\top, \qquad \mathbf{a} \in \mathbb{R}^{d_{\text{out}}}, \tag{94}$$

and has dimension d_{out}, with eigenvalue $\lambda_{\|}$. The activation-orthogonal eigenspace consists of perturbations:

$$\mathbf{H} = \mathbf{a}\mathbf{v}^\top, \qquad \mathbf{a} \in \mathbb{R}^{d_{\text{out}}}, \quad \mathbf{v}^\top \hat{\mathbf{x}}_t = 0, \tag{95}$$

and has dimension $d_{\text{out}}(d_{\text{in}} - 1)$, with eigenvalue $\lambda_{\perp}$. Hence,

$$d_{\text{out}} + d_{\text{out}}(d_{\text{in}} - 1) = d_{\text{out}} d_{\text{in}}, \tag{96}$$

accounting for the complete matrix parameter space. Therefore, the row-wise analysis captures all distinct poles of the full matrix-valued dynamics; the full system simply applies the same directional filtering simultaneously across all output rows.

An equally important interpretation arises from the state-space formulation itself. The transition operator

$$A_t = \rho P_t,$$

may be viewed not only as a stability operator but also as a *memory retention operator*. In the discrete-time state-space dynamics

$$s_{t+1} = A_t s_t + B_t \tilde{g}_t,$$

the term $A_t s_t$ represents the portion of historical momentum that survives into the next optimization step. Consequently, the eigenvalues of A_t directly quantify the fraction of previously accumulated memory retained along each state-space direction.

Viewed from this perspective, the anisotropic pole placement acquires a natural interpretation. Along the active activation direction,

$$\lambda_{\|} < \lambda_{\perp},$$

implying that historical momentum is forgotten more aggressively in precisely those directions currently supported by new evidence. Conversely, orthogonal directions retain a substantially larger fraction of their previous momentum, thereby preserving historical knowledge that is unrelated to the current activation. Rather than uniformly forgetting the past, AXIOM selectively reallocates its memory capacity toward the currently active associative subspace.

The interpolation

$$\lambda_{\text{eff}}(\theta) = \lambda_{\|} \cos^2 \theta + \lambda_{\perp} \sin^2 \theta,$$

further reveals that memory retention varies smoothly with orientation. As the angle gradually rotates away from the active activation direction, progressively more historical momentum is preserved. Thus, memory retention is not binary but continuously distributed across the parameter space according to *geometric alignment* with the current activation.

Geometrically, $\lambda_{\text{eff}}(\theta)$ defines a continuous scalar field on the unit hypersphere, where each direction is assigned its own memory-retention coefficient. Rather than partitioning the parameter space into discrete active and inactive regions, AXIOM continuously modulates memory retention according to angular alignment with the current activation.

A particularly interesting consequence is that the principal damping direction is itself dynamic. Since the activation vector $\mathbf{x}_t$ changes throughout optimization, the eigenspace associated with $\lambda_{\|}$ continuously *rotates*. Consequently, the optimizer does not possess a fixed preferred direction; instead, the direction of strongest forgetting adaptively follows the current data manifold. In this sense, AXIOM may be interpreted as implementing an adaptive anisotropic memory filter, whose retention profile continuously reorients itself according to the incoming data.

When viewed together with the dual optimization analysis, a coherent picture emerges. The Fenchel–Rockafellar formulation showed that new information is incorporated through the unique optimal rank-1 associative correction driven by the current retrieval error. The state-space analysis presented here complements this result by showing that historical information is simultaneously retained in a direction-dependent manner through the transition operator, A_t. Consequently, each optimization step may be interpreted as a coupled process consisting of (i) localized associative memory writing through prediction-error-driven rank-1 updates and (ii) adaptive directional memory retention through anisotropic state evolution. Although derived purely from convex optimization and linear systems analysis, this computational architecture is intriguingly reminiscent of broader principles found in associative memory models, predictive coding, and adaptive state estimation, where learning likewise emerges through the continual balance between correcting prediction error, and preserving useful prior knowledge.

F Local LTI Stability Conditions and Geometric Interpretation

To ensure that the discrete-time state-space system $s_{t+1} = A_t s_t + B_t \tilde{g}_t$ remains well-behaved under online gradient updates, we perform a formal local stability analysis on the localized Linear Time-Invariant (LTI) system operator, $A_t = \rho P_t$.

F.1 Derivation of Parameter Bounds for Local LTI Stability

A discrete-time LTI system $s_{t+1} = A_t s_t + B_t \tilde{g}_t$ is asymptotically stable if and only if all the eigenvalues of the system matrix,A_t satisfy:

$$|\lambda_i(A_t)| < 1, \quad \forall i \in \{1, 2, \ldots, d_{\text{in}}\}. \tag{97}$$

Given the two eigenvalue branches of A_t:

$$\lambda_1(A_t) = \frac{\rho\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2}, \qquad \lambda_i(A_t) = \rho \quad \forall i \in \{2, \ldots, d_{\text{in}}\}. \tag{98}$$

Condition 1: Orthogonal Subspace Bounds (λ_i, $i \in \{2, \ldots, d_{\text{in}}\}$)

$$\left| \frac{1}{2\alpha\eta + 1} \right| < 1. \tag{99}$$

Since $\eta > 0$ and $\alpha > 0$:

$$2\alpha\eta + 1 > 1 \implies 0 < \rho < 1 \iff \alpha \cdot \eta > 0. \tag{100}$$

Condition 2: Active Subspace Bounds (λ_1)

$$\lambda_1(A_t) = \rho \left(\frac{\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2} \right). \tag{101}$$

Given $\gamma = 2\alpha + \frac{1}{\eta}$ with $\alpha > 0$ and $\eta > 0$:

$$\gamma = 2\alpha + \frac{1}{\eta} > 0 \implies \gamma\eta = 2\alpha\eta + 1 > 1. \tag{102}$$

Since $\|\mathbf{x}_t\|_2^2 \geq 0$:

$$0 < \frac{\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2} \leq 1. \tag{103}$$

Multiplying by $\rho \in (0, 1)$:

$$0 < \lambda_1(A_t) \leq \lambda_i(A_t) < 1. \tag{104}$$

Since $\lambda_i(A_t) = \rho$ for all $i \in \{2, \ldots, d_{\text{in}}\}$.

Necessary & Sufficient Local Stability Conditions:

$$\eta > 0, \quad \alpha > 0, \quad \gamma = 2\alpha + \frac{1}{\eta} > 0. \tag{105}$$

F.2 Why Local LTI Stability is Mathematically Required

1. **Suppression of State Explosion:**
   If $|\lambda_i| \geq 1$ for some i, the system is not asymptotically stable, and the corresponding state component does not necessarily decay. Local LTI stability therefore prevents non-decaying or exploding momentum trajectories.
2. **Exponential Noise Decay:**

$$\delta s^{(t+k)} = A_t^k \delta s^{(t)} \implies \|\delta s^{(t+k)}\|_2 \leq \rho^k \|\delta s^{(t)}\|_2, \tag{106}$$

   guaranteeing $\mathcal{O}(\rho^k)$ noise decay.

F.3 Geometric Interpretation in Input Space: Hebbian *vs.* Anti-Hebbian Dynamics

- **Standard Hebbian Update Rule:** $\Delta W \propto y_t x_t^\top$ Hebb (1949) leads to unbounded growth along high variance activation vectors x_t.
- **AXIOM Operator Projection in Input Space:**

$$A_t = \rho P_t = \rho \left(I - \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2} \mathbf{x}_t \mathbf{x}_t^\top \right). \tag{107}$$

- **Decomposition of Momentum State s_t:**

$$s_t = s_t^{\|x_t} + s_t^{\perp x_t}, \tag{108}$$

where:

$$\begin{aligned} \text{Parallel Component:} \quad & s_t^{\|x_t} = \left(\frac{x_t^\top s_t}{\|x_t\|_2^2} \right) x_t, \\ \text{Perpendicular Component:} \quad & s_t^{\perp x_t} = s_t - s_t^{\|x_t}. \end{aligned} \tag{109}$$

- **Action of P_t on Components:**

$$\begin{aligned} P_t s_t^{\perp x_t} &= s_t^{\perp x_t}, \\ P_t s_t^{\|x_t} &= \left(\frac{\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2} \right) s_t^{\|x_t}. \end{aligned} \tag{110}$$

The corrective term

$$-\frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2} \mathbf{x}_t \mathbf{x}_t^\top,$$

acts as an *anti-Hebbian-like projection* (Földiak, 1990, Sec. 6, pp. 166–167), selectively attenuating momentum components aligned with the currently active activation direction. This interpretation is structural rather than mechanistic.

F.4 Neuroscience Analogy: Active *vs.* Passive Memory Consolidation

1. **Active Consolidation (λ_1):**

$$\|\mathbf{x}_t\|_2^2 \to \infty \implies \lambda_1(A_t) = \frac{\rho\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2} \to 0. \tag{111}$$

Although derived entirely from convex optimization and state-space analysis, the resulting dynamics are reminiscent of theories of predictive error correction and activity-dependent plasticity, where recently active pathways undergo more rapid modification than inactive pathways(Bliss & Cooke, 2011, pp. A105) We present this only as an interpretive analogy rather than a mechanistic claim.

2. **Passive Consolidation ($\lambda_i,\ i \in \{2, \ldots, d_{\text{in}}\}$):**

$$\forall v \perp \mathbf{x}_t, \quad \lambda_i(A_t) = \rho = \frac{1}{2\alpha\eta + 1}. \tag{112}$$

Orthogonal pathways undergo steady decay, preserving long-term associative memory in inactive neural circuits.

The *filtering operator* therefore induces two complementary processes within the optimizer. The first is an error-driven rank-one memory insertion established by the dual formulation. The second is an anisotropic forgetting mechanism that selectively suppresses historical momentum aligned with the currently active feature direction while comparatively preserving orthogonal memory components. Consequently, AXIOM continuously reallocates its finite memory capacity toward the most recently activated associative subspaces without globally erasing previously accumulated information.

G Information-Theoretic Perspective: Entropy Dynamics, Directional Mutual Information

In this Section, we construct a formal information-theoretic analysis of the memory matrix dynamics under stochastic gradient forcing. Rather than relying on scalar approximations, we treat the state covariance non-isotropically to capture how the Householder projection operator P_t selectively compresses uncertainty along active feature directions while preserving historical optimization signals across orthogonal subspaces.

G.1 Discrete-Time Gaussian State-Space Model and Covariance Propagation

Let $M_t \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ be the matrix momentum state, and let $s_t \triangleq \text{vec}(M_t) \in \mathbb{R}^d$ denote its vectorization with total state dimension $d = d_{\text{out}} d_{\text{in}}$. Following the standard linear Gaussian state-space formulation commonly adopted in stochastic estimation and Kalman filtering, we model the optimizer state as a multivariate Gaussian process. We model s_t as a multivariate Gaussian stochastic process $s_t \sim \mathcal{N}(\mu_t, \Sigma_t)$, where $\Sigma_t \in \mathbb{R}^{d \times d}$ is the symmetric positive semidefinite state error covariance matrix.

The full stochastic update rule is given by:

$$s_{t+1} = \mathcal{A}_t s_t + \mathcal{B}_t \tilde{g}_t, \tag{113}$$

where:

- $\tilde{g}_t \triangleq \text{vec}(g_t) \sim \mathcal{N}(\bar{g}_t, \Sigma_u)$ is the vectorized stochastic gradient input with covariance $\Sigma_u \in \mathbb{R}^{d \times d}$.
- $\mathcal{A}_t \triangleq I_{d_{\text{out}}} \otimes A_t \in \mathbb{R}^{d \times d}$ is the state transition operator, with $A_t = \rho P_t = \rho \left(I_{d_{\text{in}}} - c_t \mathbf{x}_t \mathbf{x}_t^\top \right)$ and $c_t = \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}$.
- $\mathcal{B}_t \triangleq I_{d_{\text{out}}} \otimes B_t \in \mathbb{R}^{d \times d}$ is the input coupling operator, with $B_t = \frac{2}{\gamma} P_t$.

Under this discrete Gaussian state-space model, the state covariance propagates according to the discrete-time Lyapunov covariance recursion (Simon, 2006, Sec. 4.1, Eq. (4.4), pp. 107–108):

$$\Sigma_{t+1} = \mathcal{A}_t \Sigma_t \mathcal{A}_t^\top + \mathcal{B}_t \Sigma_u \mathcal{B}_t^\top. \tag{114}$$

The covariance propagation equation reveals that uncertainty within AXIOM evolves through two fundamentally different mechanisms. The first term,

$$\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top,$$

transports and reshapes the uncertainty accumulated from previous optimization steps according to the geometry of the state transition operator. Since $\mathcal{A}_t$ possesses anisotropic eigenvalues, previously accumulated uncertainty is attenuated at different rates depending on its alignment with the current activation direction. The second term,

$$\mathcal{B}_t \Sigma_u \mathcal{B}_t^\top,$$

represents the continual injection of new uncertainty arising from stochastic gradient noise. Consequently, each optimization step may be viewed as balancing two competing processes: the adaptive retention of historical uncertainty and the continual incorporation of new stochastic information. This decomposition provides the probabilistic analogue of the deterministic state dynamics developed in the previous section, extending the interpretation of AXIOM from direction-dependent memory evolution to direction-dependent uncertainty evolution. In other words, AXIOM does not merely update its momentum state; it also continuously redistributes the *uncertainty* associated with that state according to the geometry induced by the current activation.

In the absence of stochastic gradient noise ($\Sigma_u = \mathbf{0}$), the covariance recursion reduces to

$$\Sigma_{t+1} = \mathcal{A}_t \Sigma_t \mathcal{A}_t^\top.$$

Since the transition operators satisfy the uniform contraction property $\|\mathcal{A}_t\|_2 \leq \rho < 1$ for every iteration t, repeated covariance propagation yields $\Sigma_t \to \mathbf{0}$. Consequently, the uncertainty associated with the momentum state vanishes asymptotically, and the optimizer approaches a deterministic momentum trajectory.

This covariance evolution forms the basis for the entropy analysis developed next, since for Gaussian state distributions the differential entropy is uniquely determined by the covariance determinant.

G.2 Differential Entropy Balance Equation

The differential entropy $H(s_t)$ of the multivariate Gaussian state $s_t \sim \mathcal{N}(\mu_t, \Sigma_t)$ is (Cover & Thomas, 2006, Ch. 8, Sec. 8.4, pp. 249–250):

$$H(s_t) = \frac{1}{2} \ln \det(\Sigma_t) + \frac{d}{2}(1 + \ln 2\pi). \tag{115}$$

We assume the initial covariance satisfies:

$$\Sigma_0 \succ 0.$$

Since $\rho > 0$ and $\det(P_t) > 0$, the state transition operator $\mathcal{A}_t$ is non-singular for every iteration. Since the state transition operator $\mathcal{A}_t$ is nonsingular for every iteration, positive definiteness is preserved throughout the covariance recursion. Consequently, $\Sigma_t \succ 0$ for all t, ensuring that all inverse and determinant operations appearing below are well-defined.

Lemma 2 (Preservation of Positive-Definite Covariance)**.** *Suppose the initial covariance satisfies*

$$\Sigma_0 \succ 0, \tag{116}$$

and the state transition matrix $\mathcal{A}_t$ is nonsingular for every iteration t. In the absence of stochastic gradient noise, i.e., $\Sigma_u = \mathbf{0}$, the covariance recursion

$$\Sigma_{t+1} = \mathcal{A}_t \Sigma_t \mathcal{A}_t^\top \tag{117}$$

preserves positive definiteness. Consequently,

$$\Sigma_t \succ 0 \qquad \textit{for all } t. \tag{118}$$

Proof. We prove the result by induction. By assumption, $\Sigma_0 \succ 0$. Suppose that $\Sigma_t \succ 0$ for some iteration, t. For any non-zero vector, $z \in \mathbb{R}^d$,

$$z^\top \Sigma_{t+1} z = z^\top \mathcal{A}_t \Sigma_t \mathcal{A}_t^\top z. \tag{119}$$

Since $\mathcal{A}_t$ is non-singular and $z \neq 0$, it follows that $\mathcal{A}_t^\top z \neq 0$. Because $\Sigma_t \succ 0$, we have

$$(\mathcal{A}_t^\top z)^\top \Sigma_t (\mathcal{A}_t^\top z) > 0. \tag{120}$$

Therefore, $z^\top \Sigma_{t+1} z > 0$ for every $z \neq 0$, which implies $\Sigma_{t+1} \succ 0$. By induction, $\Sigma_t \succ 0$ for all t. □

To compute the net entropy change $\Delta H(s_t) = H(s_{t+1}) - H(s_t)$ in the presence of stochastic noise, we decompose Σ_{t+1}:

$$\Sigma_{t+1} = \mathcal{A}_t \Sigma_t \mathcal{A}_t^\top \left(I_d + (\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top)^{-1} \mathcal{B}_t \Sigma_u \mathcal{B}_t^\top \right). \tag{121}$$

Taking the logarithm of determinant on both sides yields the general entropy transition equation:

$$\ln \det(\Sigma_{t+1}) = \ln \det(\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top) + \ln \det \left(I_d + (\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top)^{-1} \mathcal{B}_t \Sigma_u \mathcal{B}_t^\top \right). \tag{122}$$

Applying properties of determinants under Kronecker products, $\det(A \otimes B) = (\det A)^m (\det B)^n$:

$$\det(\mathcal{A}_t) = \det\left(I_{d_{\text{out}}} \otimes \rho P_t\right) = \left(\det(\rho P_t)\right)^{d_{\text{out}}} = \rho^d \left(\det P_t\right)^{d_{\text{out}}}. \tag{123}$$

Using the Matrix Determinant Lemma $\det(I + uv^\top) = 1 + v^\top u$ for the rank-one update $P_t = I_{d_{\text{in}}} - c_t \mathbf{x}_t \mathbf{x}_t^\top$:

$$\det P_t = 1 - c_t \|\mathbf{x}_t\|_2^2 = 1 - \frac{2\|\mathbf{x}_t\|_2^2}{\gamma + 2\|\mathbf{x}_t\|_2^2} = \frac{\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2} = \left(1 + \frac{2}{\gamma}\|\mathbf{x}_t\|_2^2\right)^{-1}. \tag{124}$$

Substituting $\det \mathcal{A}_t = \rho^d \left(1 + \frac{2}{\gamma}\|\mathbf{x}_t\|_2^2\right)^{-d_{\text{out}}}$ into $\frac{1}{2} \ln \det(\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top) = \ln |\det \mathcal{A}_t| + \frac{1}{2} \ln \det \Sigma_t$:

$$\frac{1}{2} \ln \det(\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top) = \frac{1}{2} \ln \det(\Sigma_t) + d \ln \rho - d_{\text{out}} \ln \left(1 + \frac{2}{\gamma}\|\mathbf{x}_t\|_2^2\right). \tag{125}$$

Thus, the exact differential entropy step evolution breaks into an **entropy dissipation** term and a **stochastic injection** term:

$$\Delta H(s_t) = \underbrace{d \ln \rho - d_{\text{out}} \ln \left(1 + \frac{2}{\gamma}\|\mathbf{x}_t\|_2^2\right)}_{\text{Deterministic Entropy Dissipation } \Delta H_{\text{diss}}} + \underbrace{\frac{1}{2} \ln \det \left(I_d + \mathcal{N}_t\right)}_{\text{Stochastic Injection } \Delta H_{\text{inj}}}, \tag{126}$$

where $\mathcal{N}_t \triangleq (\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top)^{-1} \mathcal{B}_t \Sigma_u \mathcal{B}_t^\top \in \mathbb{R}^{d \times d}$ represents the directional Noise-to-Signal matrix operator.

The entropy decomposition derived above naturally separates the effects of the deterministic state dynamics from those of stochastic gradient noise. The stochastic injection term depends on the external gradient covariance and therefore, cannot be characterized independently of the optimization environment. In contrast, the deterministic component depends solely on the intrinsic geometry of the AXIOM state transition operator. The following theorem establishes that this deterministic contribution always contracts the uncertainty of the momentum state and further quantifies how this contraction depends on the activation geometry.

Theorem 3 (Anisotropic Entropy Compression Law)**.** *Suppose* $\mathbf{x}_t \neq \mathbf{0}$ *and* $\rho \in (0, 1)$*. Then the deterministic entropy contribution satisfies*

$$\Delta H_{\text{diss}} < d \ln \rho < 0.$$

Furthermore,

$$\frac{\partial \Delta H_{\text{diss}}}{\partial \|\mathbf{x}_t\|_2^2} = -\frac{2 d_{\text{out}}}{\gamma + 2\|\mathbf{x}_t\|_2^2} < 0,$$

implying that the magnitude of entropy dissipation increases monotonically with the activation energy. Consequently, larger activation magnitudes induce stronger uncertainty compression along the active feature subspace.

Proof. From the entropy decomposition,

$$\Delta H_{\text{diss}} = d \ln \rho - d_{\text{out}} \ln \left(1 + \frac{2}{\gamma}\|\mathbf{x}_t\|_2^2\right).$$

Since

$$\mathbf{x}_t \neq 0,$$

we have

$$\ln \left(1 + \frac{2}{\gamma}\|\mathbf{x}_t\|_2^2\right) > 0.$$

Therefore,

$$\Delta H_{\text{diss}} < d \ln \rho.$$

Because

$$0 < \rho < 1,$$

it follows that

$$d \ln \rho < 0,$$

which proves that the deterministic entropy contribution is strictly negative.

Differentiating with respect to the activation energy gives

$$\frac{\partial \Delta H_{\text{diss}}}{\partial \|\mathbf{x}_t\|_2^2} = -\frac{2d_{\text{out}}}{\gamma + 2\|\mathbf{x}_t\|_2^2} < 0,$$

showing that entropy compression increases monotonically with the activation magnitude. Hence stronger activations produce greater deterministic uncertainty contraction along the active feature direction.

□

G.3 Directional Mutual Information and Anisotropic Filtering

The information provided by the incoming stochastic gradient $\tilde{g}_t$ about the updated state s_{t+1} given the prior state parameterization is quantified by the conditional mutual information $I(s_{t+1}; \tilde{g}_t \mid \mathcal{A}_t s_t)$.

Under Gaussian channel dynamics:

$$\begin{aligned} I(s_{t+1}; \tilde{g}_t \mid \mathcal{A}_t s_t) &= H(s_{t+1} \mid \mathcal{A}_t s_t) - H(s_{t+1} \mid \mathcal{A}_t s_t, \tilde{g}_t) \\ &= \frac{1}{2} \ln \det \left(I_d + \mathcal{B}_t \Sigma_u \mathcal{B}_t^\top (\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top)^{-1} \right). \end{aligned} \tag{127}$$

For analytical tractability, we assume isotropic stochastic gradient noise,

$$\Sigma_u = \sigma_g^2 I_d.$$

Substituting the state transition operators

$$\mathcal{A}_t = I_{d_{\text{out}}} \otimes \rho P_t, \qquad \mathcal{B}_t = I_{d_{\text{out}}} \otimes \frac{2}{\gamma} P_t,$$

into the operator ratio within the determinant yields

$$\begin{aligned} \mathcal{B}_t \Sigma_u \mathcal{B}_t^\top (\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top)^{-1} &= \left(I_{d_{\text{out}}} \otimes \frac{2}{\gamma} P_t \right) (\sigma_g^2 I_d) \left(I_{d_{\text{out}}} \otimes \frac{2}{\gamma} P_t \right) \\ &\quad \times \left[(I_{d_{\text{out}}} \otimes \rho P_t)\, \Sigma_t\, (I_{d_{\text{out}}} \otimes \rho P_t) \right]^{-1}, \\ &= \frac{4\sigma_g^2}{\gamma^2 \rho^2} \left(I_{d_{\text{out}}} \otimes P_t \right) \left(I_{d_{\text{out}}} \otimes P_t \right) \\ &\quad \times \left(I_{d_{\text{out}}} \otimes P_t^{-1} \right) \Sigma_t^{-1} \left(I_{d_{\text{out}}} \otimes P_t^{-1} \right), \\ &= \frac{4\sigma_g^2}{\gamma^2 \rho^2} (I_{d_{\text{out}}} \otimes P_t) \Sigma_t^{-1} (I_{d_{\text{out}}} \otimes P_t^{-1}), \end{aligned} \tag{128}$$

where we have used the identity $(A\Sigma A^\top)^{-1} = A^{-\top} \Sigma^{-1} A^{-1}$ together with the symmetry of the projection operator $P_t = P_t^\top$.

Defining $\mathcal{P}_t \triangleq I_{d_{\text{out}}} \otimes P_t$, the matrix argument inside the determinant can be rewritten as a similarity transformation:

$$I_d + \mathcal{B}_t \Sigma_u \mathcal{B}_t^\top (\mathcal{A}_t \Sigma_t \mathcal{A}_t^\top)^{-1} = \mathcal{P}_t \left(I_d + \frac{4\sigma_g^2}{\gamma^2 \rho^2} \Sigma_t^{-1} \right) \mathcal{P}_t^{-1}.$$

Because similarity transformations preserve the determinant—i.e., $\det(\mathcal{P}_t X \mathcal{P}_t^{-1}) = \det(X)$—the transformation operator $\mathcal{P}_t$ does not alter the spectrum of the product matrix.

Consequently, the conditional mutual information admits the exact expression

$$I(s_{t+1}; \tilde{g}_t \mid \mathcal{A}_t s_t) = \frac{1}{2} \ln \det \left(I_d + \frac{4\sigma_g^2}{\gamma^2 \rho^2} \Sigma_t^{-1} \right). \tag{129}$$

Since the covariance matrix is symmetric positive definite, it admits the eigendecomposition

$$\Sigma_t = U_t \Lambda_t U_t^\top,$$

where

$$\Lambda_t = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d), \qquad \lambda_i > 0.$$

Therefore,

$$\Sigma_t^{-1} = U_t \Lambda_t^{-1} U_t^\top,$$

and the mutual information can be written spectrally as

$$I(s_{t+1}; \tilde{g}_t \mid \mathcal{A}_t s_t) = \frac{1}{2} \sum_{i=1}^{d} \ln \left(1 + \frac{4\sigma_g^2}{\gamma^2 \rho^2 \lambda_i} \right). \tag{130}$$

At first sight this invariance appears surprising, since the optimizer was explicitly constructed to perform anisotropic filtering. However, the invariance is only instantaneous.

The invariance of the determinant under P_t is not saying the geometry is irrelevant. It is saying something subtler: both the propagated state uncertainty and the injected gradient uncertainty are transformed by the same anisotropic operator, inducing a similarity transformation that leaves the instantaneous conditional signal-to-noise ratio spectrum invariant. The directional asymmetry therefore arises not from the current filtering step itself, but from the accumulated covariance Σ_t, whose long-term evolution has already been shaped by repeated anisotropic filtering. In other words, the geometry is not visible in the instantaneous ratio because it has been *absorbed into the state covariance* through the optimizer's history.

Theorem 4 (Spectral Consistency of Anisotropic Covariance Dynamics)**.** *Let*

$$A_t = \rho P_t \in \mathbb{R}^{d_{\mathrm{in}} \times d_{\mathrm{in}}}, \qquad Q_t = \frac{4\sigma_g^2}{\gamma^2} P_t^2 \in \mathbb{R}^{d_{\mathrm{in}} \times d_{\mathrm{in}}},$$

where

$$P_t = I_{d_{\mathrm{in}}} - \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2} \mathbf{x}_t \mathbf{x}_t^\top,$$

is the symmetric operator. Define the corresponding full-system lifted operators acting on the vectorized state space $\mathbb{R}^d$ (with $d = d_{\mathrm{out}} d_{\mathrm{in}}$) by

$$\mathcal{A}_t \triangleq I_{d_{\mathrm{out}}} \otimes A_t \in \mathbb{R}^{d \times d}, \qquad \mathcal{Q}_t \triangleq I_{d_{\mathrm{out}}} \otimes Q_t = \frac{4\sigma_g^2}{\gamma^2} \left(I_{d_{\mathrm{out}}} \otimes P_t^2 \right) \in \mathbb{R}^{d \times d}.$$

Then, both the row-wise operators and the full-system operators commute:

$$A_t Q_t = Q_t A_t \qquad \textit{and} \qquad \mathcal{A}_t \mathcal{Q}_t = \mathcal{Q}_t \mathcal{A}_t.$$

Consequently, both covariance propagation and stochastic uncertainty injection share the same instantaneous spectral geometry induced by the eigenspaces of P_t. In the full vectorized state space, this geometry is inherited through the Kronecker-lifted operators $\mathcal{A}_t$ and $\mathcal{Q}_t$.

Proof. Since $P_t = P_t^\top$, both $A_t = \rho P_t$ and $Q_t = \frac{4\sigma_g^2}{\gamma^2} P_t^2$ are polynomial functions of the same symmetric matrix P_t. Because matrix functions of a common symmetric matrix commute, we have $A_t Q_t = Q_t A_t$.

Applying the properties of the Kronecker product, the full-system operators satisfy:

$$\begin{aligned} \mathcal{A}_t \mathcal{Q}_t &= (I_{d_{\mathrm{out}}} \otimes A_t)(I_{d_{\mathrm{out}}} \otimes Q_t), \\ &= I_{d_{\mathrm{out}}} \otimes (A_t Q_t), \\ &= I_{d_{\mathrm{out}}} \otimes (Q_t A_t), \\ &= (I_{d_{\mathrm{out}}} \otimes Q_t)(I_{d_{\mathrm{out}}} \otimes A_t) = \mathcal{Q}_t \mathcal{A}_t. \end{aligned} \tag{131}$$

To explicitly construct their joint eigenstructure, let $P_t = U_t \Lambda_t U_t^\top$ be the orthogonal eigendecomposition of $P_t \in \mathbb{R}^{d_{\mathrm{in}} \times d_{\mathrm{in}}}$. Then $P_t^2 = U_t \Lambda_t^2 U_t^\top$.

Lifting to the full state space using $\mathcal{U}_t \triangleq I_{d_{\text{out}}} \otimes U_t$, we obtain:

$$\mathcal{A}_t = (I_{d_{\text{out}}} \otimes U_t)(I_{d_{\text{out}}} \otimes \rho\Lambda_t)(I_{d_{\text{out}}} \otimes U_t^\top) = \mathcal{U}_t(I_{d_{\text{out}}} \otimes \rho\Lambda_t)\mathcal{U}_t^\top, \tag{132}$$

$$\mathcal{Q}_t = (I_{d_{\text{out}}} \otimes U_t)\left(I_{d_{\text{out}}} \otimes \frac{4\sigma_g^2}{\gamma^2}\Lambda_t^2\right)(I_{d_{\text{out}}} \otimes U_t^\top) = \mathcal{U}_t\left(I_{d_{\text{out}}} \otimes \frac{4\sigma_g^2}{\gamma^2}\Lambda_t^2\right)\mathcal{U}_t^\top. \tag{133}$$

This demonstrates that both $\mathcal{A}_t$ and $\mathcal{Q}_t$ are diagonalized by the identical orthogonal matrix $\mathcal{U}_t$, proving that the full-system deterministic transition operator and the stochastic noise injection operator share identical eigenvectors, with each row-wise eigenvalue repeated d_{out} times. □

H Responsiveness to Gradients: AXIOM vs. Standard Optimizers

To evaluate how quickly an optimizer incorporates a new mini-batch gradient g_t, we calculate the partial derivative of the next state s_{t+1} with respect to the input gradient $\tilde{g}_t = \text{vec}(g_t)$ (the Gradient Sensitivity Gain Matrix):

$$G_{\text{gain}} \triangleq \frac{\partial s_{t+1}}{\partial \tilde{g}_t}. \tag{134}$$

1. Standard SGD with Momentum (SGDM):

$$s_{t+1} = \beta s_t + \eta\tilde{g}_t \implies G_{\text{gain}}^{\text{SGDM}} = \eta I_d. \tag{135}$$

Behavior: Isotropic, static gain ηI_d. SGDM reacts identically across all directions regardless of feature power $\|\mathbf{x}_t\|_2^2$.

2. Adam / RMSprop:

$$G_{\text{gain}}^{\text{Adam}} \approx \text{diag}\left(\frac{\eta}{\sqrt{v_t} + \epsilon}\right). \tag{136}$$

Behavior: Coordinate-wise adaptive gain. Responsiveness is damped along high-variance historical coordinates, but ignores directional correlations in input activation space ($\mathbf{x}_t$).

3. AXIOM (Proposed):

$$G_{\text{gain}}^{\text{AXIOM}} = \mathcal{B}_t = I_{d_{\text{out}}} \otimes \frac{2}{\gamma}P_t = \frac{2}{\gamma}\left(I_d - \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}(I_{d_{\text{out}}} \otimes \mathbf{x}_t\mathbf{x}_t^\top)\right). \tag{137}$$

Directional Responsiveness Breakdown (Analyzed Per Output Channel):

[leftmargin=*]

1. **Along Parallel Activation Subspace ($g_t \parallel \mathbf{x}_t$):**

$$G_{\text{gain}}^{\text{AXIOM}}\mathbf{x}_t = \frac{2}{\gamma}\left(1 - \frac{2\|\mathbf{x}_t\|_2^2}{\gamma + 2\|\mathbf{x}_t\|_2^2}\right)\mathbf{x}_t = \left(\frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}\right)\mathbf{x}_t. \tag{138}$$

When $\|\mathbf{x}_t\|_2^2 \to \infty$, $G_{\text{gain}} \to 0$. AXIOM automatically reduces its sensitivity along strongly activated directions, with the parallel gain approaching zero as $\|\mathbf{x}_t\|_2^2 \to \infty$.

2. **Along Orthogonal Activation Subspace ($g_t \perp \mathbf{x}_t$):**

$$G_{\text{gain}}^{\text{AXIOM}}v = \frac{2}{\gamma}v. \tag{139}$$

The gain remains unattenuated at $\frac{2}{\gamma}$, maintaining high sensitivity to novelty along directions orthogonal to the current activation.

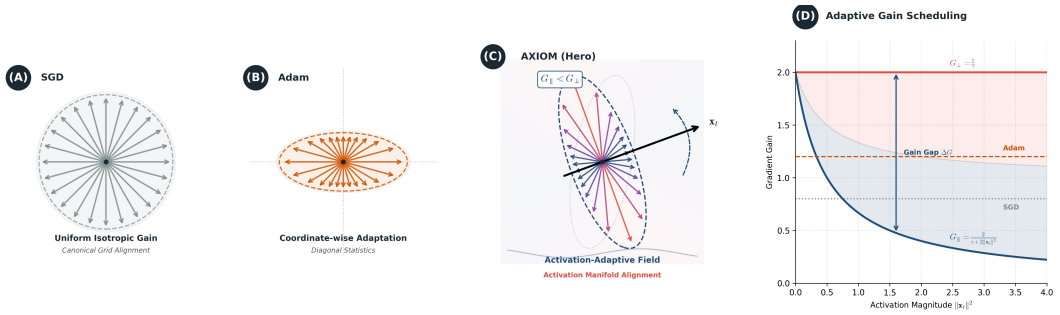

Figure 2: **Comparative Gradient Responsiveness of Different Optimization Algorithms.** (A) SGD applies a uniform isotropic gradient gain, responding equally in every direction of the parameter space. (B) Adam performs coordinate-wise adaptation through accumulated second-moment statistics, producing axis-aligned anisotropic responsiveness that is independent of the current activation geometry. (C) AXIOM performs activation-conditioned anisotropic gain scheduling through the projection operator P_t, suppressing responsiveness along the active activation direction while preserving sensitivity in orthogonal feature directions. Unlike coordinate-aligned methods, the principal responsiveness direction rotates continuously with the incoming activation. (D) Analytical gain comparison illustrating that AXIOM adaptively reduces its parallel gain $G_{\parallel} = 2/(\gamma + 2\|\mathbf{x}\|_2^2)$ as activation magnitude increases, while maintaining a constant orthogonal gain $G_{\perp} = 2/\gamma$, thereby distinguishing between reinforcement of active features and exploration of previously inactive directions.

Classical momentum algorithms employ fixed or coordinate-wise gains that are independent of the current activation manifold. Consequently, they respond equally to all gradient directions or only according to historical coordinate statistics. In contrast, AXIOM couples its responsiveness directly to the geometry of the incoming activation through the projection operator P_t. The optimizer therefore distinguishes between gradients that reinforce an already active feature direction and gradients that introduce information in previously inactive directions. This produces a direction-dependent gain scheduling mechanism that automatically suppresses potentially redundant updates while preserving sensitivity to novel feature directions.

I Why AXIOM Retains Adam's Second-Moment Estimator

The derivation of AXIOM completely replaces the conventional first-moment update by solving a regularized optimization problem over the momentum state. The resulting update

$$\mathbf{M}_{t+1} = \bar{\mathbf{M}}_t + \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2} \left(\mathbf{u}_t - \bar{\mathbf{M}}_t \mathbf{x}_t\right) \mathbf{x}_t^\top \tag{140}$$

is obtained directly as the unique minimizer of the proposed optimization objective and therefore requires no exponential moving-average approximation of the first moment.

However, the optimization objective contains no probabilistic model describing the stochastic variability of the observed gradients. Consequently, while AXIOM determines *how the momentum state should evolve geometrically*, it does not estimate *how reliable the incoming stochastic gradient observation is*. These represent two fundamentally different aspects of stochastic optimization.

The normalization introduced by AXIOM depends exclusively on the current activation energy,

$$\frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}, \tag{141}$$

which arises analytically from the closed-form solution of the regularized optimization problem. This factor measures the geometric leverage of the current activation vector within the linear mapping $\mathbf{M}\mathbf{x}$. Large activation norms correspond to directions in which small changes in the momentum state already produce large changes in the predicted update. Accordingly, AXIOM automatically

reduces the correction magnitude along these high-leverage directions through its optimization-derived normalization.

In contrast, Adam's second-moment estimator Kingma & Ba (2015)

$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2 \tag{142}$$

serves a fundamentally different purpose. Rather than determining the update geometry, it maintains an exponential moving average of the squared stochastic gradients, thereby tracking their second-order statistics across optimization iterations. The corresponding adaptive normalization

$$\frac{1}{\sqrt{\mathbf{v}_t} + \epsilon} \tag{143}$$

reduces the effective learning rate in coordinates exhibiting persistently large gradient fluctuations, thereby improving robustness during stochastic optimization.

Unlike Adam's empirical gradient-statistics normalization, AXIOM's normalization is derived analytically from the optimal solution of the regularized objective. Consequently, the two mechanisms address complementary sources of difficulty. AXIOM adapts updates according to the geometric conditioning induced by the current feature representation, whereas Adam adapts updates according to the observed second-order statistics of the stochastic gradients. One addresses the conditioning of the optimization problem itself; the other addresses the variability of the stochastic optimization process.

Consequently, replacing Adam's second-moment estimator would require introducing an explicit probabilistic model of gradient uncertainty into the optimization objective, which lies outside the scope of the present derivation. For this reason, the proposed optimizer retains Adam's second-moment estimator while replacing only the first-moment dynamics. This separation preserves the optimization-derived momentum update of AXIOM while simultaneously benefiting from variance-adaptive scaling during stochastic optimization.

J Memory–Responsiveness Uncertainty Analysis

In non-stationary stochastic optimization, a fundamental trade-off exists between **historical memory retention** (filtering high-frequency stochastic noise) and **instantaneous gradient responsiveness** (tracking non-stationary gradient signals). In this section, we formalize this interaction as an operator-level analysis, characterizing how AXIOM modulates both historical memory and gradient responsiveness under varying activation geometry.

J.1 Formalization of Memory Retention and Gradient Responsiveness

Consider the stochastic state dynamics

$$s_{t+1} = \mathcal{A}_t s_t + \mathcal{B}_t \tilde{g}_t,$$

where

$$s_t = \mathrm{vec}(M_t), \qquad \tilde{g}_t = \mathrm{vec}(g_t).$$

The transition operator $\mathcal{A}_t$ governs how much historical momentum survives into the next optimization step, whereas $\mathcal{B}_t$ determines the sensitivity of the optimizer to newly observed stochastic gradients. To quantify this trade-off, we introduce the following directional gain functions.

Definition 1 (Directional Memory and Responsiveness Gains)**.** *The memory retention and gradient responsiveness operators are*

$$\mathcal{M}_t := \mathcal{A}_t = I_{d_{\mathrm{out}}} \otimes \rho P_t, \tag{144}$$

$$\mathcal{R}_t := \mathcal{B}_t = I_{d_{\mathrm{out}}} \otimes \frac{2}{\gamma} P_t. \tag{145}$$

Let

$$\hat{x}_t = \frac{x_t}{\|x_t\|_2}$$

denote the normalized activation direction.

The corresponding directional gains along the activation-aligned subspace are defined by

$$m_t(x_t) := \frac{\|\mathcal{M}_t(I_{d_{\text{out}}} \otimes \hat{x}_t)\|}{\|I_{d_{\text{out}}} \otimes \hat{x}_t\|}, \tag{146}$$

$$r_t(x_t) := \frac{\|\mathcal{R}_t(I_{d_{\text{out}}} \otimes \hat{x}_t)\|}{\|I_{d_{\text{out}}} \otimes \hat{x}_t\|}. \tag{147}$$

Since

$$P_t \hat{x}_t = \frac{\gamma}{\gamma + 2\|x_t\|_2^2} \hat{x}_t,$$

these gains simplify to

$$m_t(x_t) = \frac{\rho\gamma}{\gamma + 2\|x_t\|_2^2}, \tag{148}$$

$$r_t(x_t) = \frac{2}{\gamma + 2\|x_t\|_2^2}. \tag{149}$$

J.2 Covariance Evolution of Memory and Gradient Uncertainty

We model the observation through noisy stochastic gradients $\tilde{g}_t = \theta_t + \xi_t$, where $\xi_t \sim \mathcal{N}(0, \sigma_g^2 I_d)$ represents mini-batch gradient variance.

Theorem 5 (Directional Memory–Responsiveness Coupling)**.** *For the stochastic state dynamics*

$$s_{t+1} = \mathcal{A}_t s_t + \mathcal{B}_t \tilde{g}_t,$$

where $\lambda_P = \frac{\gamma}{\gamma + 2\|x_t\|_2^2}$ is the activation-aligned eigenvalue of P_t, assuming the pre-propagation memory-state perturbation has variance σ_s^2 along the activation-aligned subspace and the gradient perturbation has variance σ_g^2, the propagated memory uncertainty and gradient uncertainty variance contributions along the activation-aligned subspace satisfy

$$V_{\text{mem},\|} = (\rho\lambda_P)^2 \sigma_s^2,$$

and

$$V_{\text{grad},\|} = \left(\frac{2}{\gamma}\lambda_P\right)^2 \sigma_g^2.$$

Consequently, their product is:

$$V_{\text{mem},\|} V_{\text{grad},\|} = \frac{4\rho^2\gamma^2\sigma_s^2\sigma_g^2}{(\gamma + 2\|x_t\|_2^2)^4}.$$

Proof. Let the activation-aligned unit direction be

$$\hat{x}_t = \frac{x_t}{\|x_t\|_2},$$

and define its lifted full-state subspace by

$$\mathcal{S}_x = \left\{ a \otimes \hat{x}_t : \ a \in \mathbb{R}^{d_{\text{out}}} \right\}.$$

Since

$$P_t\hat{x}_t = \lambda_P \hat{x}_t, \qquad \lambda_P = \frac{\gamma}{\gamma + 2\|x_t\|_2^2},$$

the memory operator

$$\mathcal{M}_t = \mathcal{A}_t = I_{d_{\text{out}}} \otimes \rho P_t$$

acts on every vector in $\mathcal{S}_x$ with gain

$$G_{\text{mem},\|} = \rho\lambda_P.$$

Therefore, if the memory state has variance σ_s^2 along the activation-aligned direction, linear propagation through $\mathcal{M}_t$ gives

$$V_{\text{mem},\|} = (\rho\lambda_P)^2\sigma_s^2.$$

Similarly, the gradient-responsiveness operator is

$$\mathcal{R}_t = \mathcal{B}_t = I_{d_{\text{out}}} \otimes \frac{2}{\gamma} P_t.$$

Hence its gain along $\mathcal{S}_x$ is

$$G_{\text{grad},\|} = \frac{2}{\gamma}\lambda_P.$$

The stochastic gradient perturbation satisfies $\xi_t \sim \mathcal{N}(0, \sigma_g^2 I_d)$, so its variance along any unit direction, including the activation-aligned direction, is σ_g^2. After propagation through $\mathcal{R}_t$,

$$V_{\text{grad},\|} = \left(\frac{2}{\gamma}\lambda_P\right)^2 \sigma_g^2.$$

Substituting $\lambda_P = \frac{\gamma}{\gamma+2\|x_t\|_2^2}$, we obtain

$$V_{\text{mem},\|} = \frac{\rho^2\gamma^2}{(\gamma + 2\|x_t\|_2^2)^2}\sigma_s^2,$$

and

$$V_{\text{grad},\|} = \frac{4}{(\gamma + 2\|x_t\|_2^2)^2}\sigma_g^2.$$

Their product is therefore

$$V_{\text{mem},\|}V_{\text{grad},\|} = \frac{4\rho^2\gamma^2\sigma_s^2\sigma_g^2}{(\gamma + 2\|x_t\|_2^2)^4}.$$

Thus, increasing the activation magnitude $\|x_t\|_2$ simultaneously reduces the variance carried by the historical memory and the variance transmitted from the instantaneous stochastic gradient along the activation-aligned direction, with $\lambda_P = \frac{\gamma}{\gamma+2\|x_t\|_2^2}$ acting as the common attenuation factor for both terms. □

J.3 Memory–Responsiveness Gain Invariant

Theorem 6 (Activation-Invariant Memory–Responsiveness Ratio)**.** *Let $\eta_{\textit{eff}}(\mathbf{x}_t) \triangleq \frac{r_t(\mathbf{x}_t)}{m_t(\mathbf{x}_t)}$ define the local gain ratio between responsiveness and memory. The ratio is invariant to feature norm:*

$$\eta_{\textit{eff}}(\mathbf{x}_t) = \frac{\frac{2}{\gamma+2\|\mathbf{x}_t\|_2^2}}{\frac{\rho\gamma}{\gamma+2\|\mathbf{x}_t\|_2^2}} = \frac{2}{\rho\gamma} = \textit{Constant}. \tag{150}$$

However, the combined directional gain

$$C(\mathbf{x}_t) \triangleq m_t(\mathbf{x}_t) + \frac{\gamma}{2} r_t(\mathbf{x}_t) = \frac{(1+\rho)\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2} \tag{151}$$

strictly contracts as activation norm $\|\mathbf{x}_t\|_2^2$ increases.

Proof. Substituting the directional gain expressions

$$m_t(x_t) = \frac{\rho\gamma}{\gamma + 2\|x_t\|_2^2}, \qquad r_t(x_t) = \frac{2}{\gamma + 2\|x_t\|_2^2},$$

gives

$$\eta_{\text{eff}}(x_t) = \frac{r_t(x_t)}{m_t(x_t)} = \frac{2}{\rho\gamma},$$

which is independent of the activation norm.

Likewise,

$$m_t(x_t) + \frac{\gamma}{2} r_t(x_t) = \frac{\rho\gamma + \gamma}{\gamma + 2\|x_t\|_2^2} = \frac{(1+\rho)\gamma}{\gamma + 2\|x_t\|_2^2}.$$

Since

$$\frac{\partial C(x_t)}{\partial \|x_t\|_2^2} = -\frac{2(1+\rho)\gamma}{(\gamma + 2\|x_t\|_2^2)^2} < 0,$$

the combined directional gain decreases monotonically with activation energy.

□

Unlike conventional momentum dynamics with separately specified memory and gradient gains, AXIOM does not independently modulate historical memory and gradient responsiveness. Instead, both quantities are modulated by the same activation-dependent operator P_t. As the activation norm increases, both historical memory retention and gradient responsiveness decrease proportionally, preserving a constant relative balance while reducing the overall magnitude of state evolution. Consequently, AXIOM behaves as a direction-dependent gain scheduling mechanism: activation-aligned directions with greater leverage receive proportionally smaller memory retention and gradient responsiveness, while lower-energy activation directions preserve larger gains.

K Lyapunov Stability Analysis of the Unforced LTV System

K.1 Linear Time-Varying (LTV) Autonomous State Dynamics

Consider the unforced (autonomous) momentum dynamics of AXIOM across optimization iterations $t \geq 0$:

$$s_{t+1} = \mathcal{A}_t s_t, \qquad \mathcal{A}_t = I_{d_{\text{out}}} \otimes A_t \in \mathbb{R}^{d \times d}. \tag{152}$$

Here,

$$A_t = \rho P_t, \qquad \rho = \frac{1}{2\alpha\eta + 1} \in (0, 1),$$

and

$$P_t = I_{d_{\text{in}}} - \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2} \mathbf{x}_t \mathbf{x}_t^\top \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}. \tag{153}$$

Since the activation sequence $\{\mathbf{x}_t\}$ changes with every optimization step, the state transition matrix $\mathcal{A}_t$ is inherently time-varying. Classical eigenvalue analysis alone is insufficient to establish global stability across iterations. Instead, stability is analyzed using discrete-time Lyapunov theory for linear time-varying (LTV) systems.

K.2 Construction of the Quadratic Lyapunov Candidate

Consider the quadratic Lyapunov candidate

$$V(s_t) = s_t^\top s_t = \|s_t\|_2^2. \tag{154}$$

Clearly,

$$V(s_t) = \|s_t\|_2^2,$$

which satisfies the standard quadratic bounds

$$c_1\|s_t\|_2^2 \le V(s_t) \le c_2\|s_t\|_2^2$$

with

$$c_1 = c_2 = 1.$$

and therefore V is positive definite and radially unbounded.

K.3 Spectral Properties of the Time-Varying Operator P_t

To evaluate the Lyapunov difference, we first study the spectrum of P_t.

Lemma 3 (Spectral Properties of P_t)**.** *For every activation vector $\mathbf{x}_t \in \mathbb{R}^{d_{\text{in}}}$ and every $\gamma > 0$,*

$$\lambda(P_t) \in \left[\frac{\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2},\, 1\right], \tag{155}$$

and consequently,

$$\|P_t\|_2 = 1. \tag{156}$$

Proof. The matrix P_t is real symmetric and therefore diagonalizable with real eigenvalues.

Since P_t is a rank-one perturbation of the identity, its spectrum can be characterized by considering two invariant subspaces.

(i) Parallel direction

For the activation direction,

$$P_t\mathbf{x}_t = \mathbf{x}_t - \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}\mathbf{x}_t\mathbf{x}_t^\top\mathbf{x}_t \tag{157}$$

$$= \left(1 - \frac{2\|\mathbf{x}_t\|_2^2}{\gamma + 2\|\mathbf{x}_t\|_2^2}\right)\mathbf{x}_t \tag{158}$$

$$= \frac{\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2}\mathbf{x}_t. \tag{159}$$

Hence,

$$\lambda_\parallel = \frac{\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2} \in (0, 1].$$

(ii) Orthogonal subspace

For every vector $v \perp \mathbf{x}_t$,

$$\mathbf{x}_t^\top v = 0,$$

therefore

$$P_t v = v,$$

which gives

$$\lambda_\perp = 1.$$

Thus the eigenvalues of P_t are

$$\left\{1,\ \frac{\gamma}{\gamma + 2\|\mathbf{x}_t\|_2^2}\right\}.$$

Since P_t is symmetric, its induced Euclidean norm equals its largest absolute eigenvalue. Hence,

$$\|P_t\|_2 = \max_i |\lambda_i(P_t)| = 1.$$

□

K.4 Uniform Boundedness of the LTV Operator

Using standard Kronecker-product norm identities,

$$\begin{aligned}
\|\mathcal{A}_t\|_2 &= \|I_{d_{\text{out}}} \otimes A_t\|_2 && (160)\\
&= \|I_{d_{\text{out}}}\|_2 \, \|A_t\|_2 && (161)\\
&= \|A_t\|_2 && (162)\\
&= \|\rho P_t\|_2 && (163)\\
&= \rho \|P_t\|_2 && (164)\\
&= \rho. && (165)
\end{aligned}$$

Since

$$0 < \rho < 1,$$

the state transition operator satisfies

$$\|\mathcal{A}_t\|_2 = \rho < 1, \qquad \forall t.$$

Therefore, the family

$$\{\mathcal{A}_t\}_{t \geq 0}$$

is uniformly contractive.

K.5 Lyapunov Difference

Along the system trajectories,

$$\begin{aligned}
V(s_{t+1}) &= s_{t+1}^\top s_{t+1} && (166)\\
&= s_t^\top \mathcal{A}_t^\top \mathcal{A}_t s_t. && (167)
\end{aligned}$$

Using

$$\|\mathcal{A}_t\|_2 = \rho,$$

we obtain

$$\begin{aligned}
V(s_{t+1}) &= \|\mathcal{A}_t s_t\|_2^2 && (168)\\
&\leq \|\mathcal{A}_t\|_2^2 \|s_t\|_2^2 && (169)\\
&= \rho^2 V(s_t). && (170)
\end{aligned}$$

Therefore,

$$\Delta V(s_t) = V(s_{t+1}) - V(s_t) \leq -(1-\rho^2)V(s_t). \tag{171}$$

Since

$$0 < \rho < 1,$$

the decay constant

$$1 - \rho^2 > 0$$

is independent of both time and the activation sequence. Consequently, unrolling the recursion yields $V(s_t) \leq \rho^{2t} V(s_0)$, which implies

$$\|s_t\|_2 \leq \rho^t \|s_0\|_2.$$

Since the contraction factor $\rho^2 < 1$ is uniform over all t, the unforced LTV system is uniformly exponentially stable.

K.6 Global Uniform Exponential Stability

Theorem 7 (Global Uniform Exponential Stability of AXIOM)**.** *For every activation sequence*

$$\{\mathbf{x}_t\}_{t=0}^{\infty},$$

and every

$$\alpha > 0, \qquad \eta > 0,$$

the origin $s = 0$ of the autonomous LTV system

$$s_{t+1} = \mathcal{A}_t s_t$$

is globally uniformly exponentially stable (GUES).

Furthermore,

$$\|s_t\|_2 \leq \rho^t \|s_0\|_2, \qquad \forall t \geq 0. \tag{172}$$

Proof. From the previous section,

$$V(s_{t+1}) \leq \rho^2 V(s_t).$$

Applying this recursively,

$$V(s_t) \leq (\rho^2)^t V(s_0).$$

Since

$$V(s) = \|s\|_2^2,$$

we obtain

$$\|s_t\|_2^2 \leq \rho^{2t} \|s_0\|_2^2.$$

Taking square roots,

$$\boxed{\|s_t\|_2 \leq \rho^t \|s_0\|_2.}$$

Because

$$0 < \rho < 1,$$

the origin is globally exponentially stable with standard GUES constants $K = 1$ and decay parameter $\lambda = \rho$. Since the uniform contraction bound $\|\mathcal{A}_t\|_2 \leq \rho < 1$ holds uniformly for all optimization iterations and all activation sequences, the convergence is uniform. Hence the origin is globally uniformly exponentially stable (GUES).

□

L Convergence Analysis of the Forced Stochastic System

We analyze the iteration complexity of the non-stationary Linear Time-Varying (LTV) momentum system under stochastic gradient driving forces:

$$s_{t+1} = \mathcal{A}_t s_t + \mathcal{B}_t \tilde{g}_t, \qquad W_{t+1} = W_t - \eta M_{t+1}, \tag{173}$$

where $\tilde{g}_t = \mathrm{vec}(g_t) \in \mathbb{R}^d$, $s_t = \mathrm{vec}(M_t) \in \mathbb{R}^d$, $\eta > 0$ is the learning rate, $\mathcal{A}_t = I_{d_{\text{out}}} \otimes \rho P_t$, $\mathcal{B}_t = I_{d_{\text{out}}} \otimes \frac{2}{\gamma} P_t$, and $P_t = I_{d_{\text{in}}} - \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2} \mathbf{x}_t \mathbf{x}_t^\top$.

L.0.1 Standard Optimization Assumptions

Assumption 1 (L-Smoothness)**.** *The objective function $f : \mathbb{R}^{d_{out} \times d_{in}} \to \mathbb{R}$ is continuously differentiable, bounded below by $f^* > -\infty$, and has L-Lipschitz continuous gradients:*

$$\|\nabla f(W_1) - \nabla f(W_2)\|_F \leq L \|W_1 - W_2\|_F, \quad \forall W_1, W_2 \in \mathbb{R}^{d_{out} \times d_{in}}. \tag{174}$$

Assumption 2 (Unbiased Gradient Estimator & Bounded Variance)**.** *The stochastic mini-batch gradient g_t satisfies:*

$$\mathbb{E}_t[g_t] = \nabla f(W_t), \qquad \mathbb{E}_t\left[\|g_t - \nabla f(W_t)\|_F^2\right] \leq \sigma_g^2, \quad \forall t \geq 0. \tag{175}$$

Assumption 3 (Bounded Activation Norms and Positive Driving Spectrum)**.** *The activation sequences satisfy $0 \leq \|\mathbf{x}_t\|_2^2 \leq X_{\max}^2 < \infty$. Consequently, the symmetric operator P_t satisfies $\mu_P I \preceq P_t \preceq I$, where $\mu_P \triangleq 1 - \frac{2X_{\max}^2}{\gamma + 2X_{\max}^2} = \frac{\gamma}{\gamma + 2X_{\max}^2} > 0$. The driving operator spectrum is strictly bounded:*

$$\mu_B I \preceq \mathcal{B}_t \preceq C_B I, \quad \text{where } \mu_B \triangleq \frac{2\mu_P}{\gamma} = \frac{2}{\gamma + 2X_{\max}^2} > 0, \;\; C_B \triangleq \frac{2}{\gamma}. \tag{176}$$

Assumption 4 (Bounded Gradient)**.** *There exists a constant*

$$G_{\max} < \infty$$

such that

$$\|\nabla f(W_t)\|_F \leq G_{\max}$$

for all iterations.

L.1 Bounded Expected Momentum Norm (Input-to-State Stability)

Lemma 4 (Uniform Expected Momentum Second-Moment Bound)**.** *Under Assumptions 1–4, initialized with $s_0 = 0$, the second moment of the momentum state vector is uniformly bounded for all $t \geq 0$:*

$$\mathbb{E}\left[\|s_t\|_2^2\right] \leq \frac{C_B^2}{(1-\rho)^2} \left(G_{\max}^2 + \sigma_g^2\right), \tag{177}$$

where $G_{\max}$ is the uniform gradient bound from Assumption 4.

Proof. Unrolling $s_t = \sum_{k=0}^{t-1} \left(\prod_{j=k+1}^{t-1} \mathcal{A}_j \right) \mathcal{B}_k \tilde{g}_k$ and noting $\|\mathcal{A}_j\|_2 = \rho \|P_j\|_2 \leq \rho < 1$:

$$\|s_t\|_2 \leq \sum_{k=0}^{t-1} \rho^{t-1-k} \|\mathcal{B}_k\|_2 \|\tilde{g}_k\|_2 \leq C_B \sum_{k=0}^{t-1} \rho^{t-1-k} \|\tilde{g}_k\|_2. \tag{178}$$

Applying Cauchy-Schwarz on the weighted sum with weights $w_k = \rho^{t-1-k}$:

$$\|s_t\|_2^2 \leq C_B^2 \left(\sum_{k=0}^{t-1} \rho^{t-1-k} \right) \left(\sum_{k=0}^{t-1} \rho^{t-1-k} \|\tilde{g}_k\|_2^2 \right) \leq \frac{C_B^2}{1-\rho} \sum_{k=0}^{t-1} \rho^{t-1-k} \|\tilde{g}_k\|_2^2. \tag{179}$$

Expanding $\tilde{g}_k = \mathrm{vec}(\nabla f(W_k)) + \xi_k$, where ξ_k is the zero-mean mini-batch noise vector ($\mathbb{E}_k[\xi_k] = 0$):

$$\begin{aligned} \mathbb{E}\left[\|\tilde{g}_k\|_2^2\right] &= \mathbb{E}\left[\|\nabla f(W_k)\|_F^2\right] + \mathbb{E}\left[\|\xi_k\|_2^2\right] + 2\mathbb{E}\left[\langle \mathrm{vec}(\nabla f(W_k)), \xi_k \rangle\right] \\ &= \mathbb{E}\left[\|\nabla f(W_k)\|_F^2\right] + \mathbb{E}\left[\|\xi_k\|_2^2\right] \leq G_{\max}^2 + \sigma_g^2, \end{aligned} \tag{180}$$

where the inner product cross-term has zero expectation by the conditional unbiasedness of the stochastic gradient ($\mathbb{E}_k[\langle \mathrm{vec}(\nabla f(W_k)), \xi_k \rangle] = 0$). Taking expectations yields:

$$\mathbb{E}\left[\|s_t\|_2^2\right] \leq \frac{C_B^2 (G_{\max}^2 + \sigma_g^2)}{1-\rho} \sum_{k=0}^{t-1} \rho^{t-1-k} \leq \frac{C_B^2 (G_{\max}^2 + \sigma_g^2)}{(1-\rho)^2}. \tag{181}$$

□

L.2 Conditional State Energy Expansion

Before constructing a Lyapunov function, we first characterize the evolution of the momentum-state energy under stochastic gradient excitation.

Recall the stochastic state dynamics

$$s_{t+1} = \mathcal{A}_t s_t + \mathcal{B}_t \tilde{g}_t, \tag{182}$$

where the stochastic gradient is decomposed as

$$\tilde{g}_t = \mathrm{vec}(\nabla f(W_t)) + \xi_t, \tag{183}$$

with

$$\mathbb{E}_t[\xi_t] = 0, \qquad \mathbb{E}_t\left[\|\xi_t\|_2^2\right] \leq \sigma_g^2, \tag{184}$$

and

$$\mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot \mid \mathcal{F}_t]$$

denotes the conditional expectation with respect to the filtration generated by all information available up to iteration t.

Expanding the squared Euclidean norm,

$$\begin{aligned} \|s_{t+1}\|_2^2 &= \|\mathcal{A}_t s_t + \mathcal{B}_t \tilde{g}_t\|_2^2 \\ &= \|\mathcal{A}_t s_t\|_2^2 + 2\left\langle \mathcal{A}_t s_t, \mathcal{B}_t \tilde{g}_t \right\rangle + \|\mathcal{B}_t \tilde{g}_t\|_2^2. \end{aligned} \tag{185}$$

Taking the conditional expectation and using the unbiasedness assumption

$$\mathbb{E}_t[\tilde{g}_t] = \mathrm{vec}(\nabla f(W_t)),$$

yields

$$\begin{aligned} \mathbb{E}_t\left[\|s_{t+1}\|_2^2\right] = \|\mathcal{A}_t s_t\|_2^2 &+ 2\left\langle \mathcal{A}_t s_t, \mathcal{B}_t \,\mathrm{vec}(\nabla f(W_t)) \right\rangle \\ &+ \mathbb{E}_t\left[\|\mathcal{B}_t \tilde{g}_t\|_2^2\right]. \end{aligned} \tag{186}$$

Equation equation 186 provides an exact decomposition of the conditional state energy into three components: (i) propagated historical momentum, (ii) coupling between the previous momentum state and the current gradient, and (iii) gradient-driven energy, including both the mean-gradient and stochastic-noise contributions. In the following subsection, each term is bounded individually to derive a Lyapunov drift inequality for the coupled optimization dynamics.

Lemma 5 (Bound on Propagated Momentum Energy)**.** *The propagated state energy satisfies*

$$\|\mathcal{A}_t s_t\|_2^2 \le \rho^2 \|s_t\|_2^2. \tag{187}$$

Proof. Since

$$\|\mathcal{A}_t\|_2 = \rho,$$

we obtain

$$\|\mathcal{A}_t s_t\|_2 \le \|\mathcal{A}_t\|_2 \|s_t\|_2 = \rho \|s_t\|_2.$$

Squaring both sides yields

$$\|\mathcal{A}_t s_t\|_2^2 \le \rho^2 \|s_t\|_2^2.$$

□

Lemma 6 (Exact Representation of the State–Gradient Coupling)**.** *The deterministic coupling term appearing in the conditional state-energy expansion satisfies*

$$2\left\langle \mathcal{A}_t s_t, \mathcal{B}_t \operatorname{vec}(\nabla f(W_t)) \right\rangle = \frac{4\rho}{\gamma} \left\langle (I_{d_{\text{out}}} \otimes P_t) s_t, (I_{d_{\text{out}}} \otimes P_t) \operatorname{vec}(\nabla f(W_t)) \right\rangle. \tag{188}$$

Proof. Substituting

$$\mathcal{A}_t = \rho (I_{d_{\text{out}}} \otimes P_t),$$

and

$$\mathcal{B}_t = \frac{2}{\gamma} (I_{d_{\text{out}}} \otimes P_t),$$

gives

$$\begin{aligned} &2\left\langle \mathcal{A}_t s_t, \mathcal{B}_t \operatorname{vec}(\nabla f(W_t)) \right\rangle \\ &= 2\left\langle \rho (I_{d_{\text{out}}} \otimes P_t) s_t, \frac{2}{\gamma} (I_{d_{\text{out}}} \otimes P_t) \operatorname{vec}(\nabla f(W_t)) \right\rangle \\ &= \frac{4\rho}{\gamma} \left\langle (I_{d_{\text{out}}} \otimes P_t) s_t, (I_{d_{\text{out}}} \otimes P_t) \operatorname{vec}(\nabla f(W_t)) \right\rangle. \end{aligned} \tag{189}$$

□

Lemma 7 (Young Inequality Bound for the State–Gradient Coupling)**.** *For any parameter $\varepsilon > 0$, the deterministic coupling term satisfies*

$$2\left\langle \mathcal{A}_t s_t, \mathcal{B}_t \operatorname{vec}(\nabla f(W_t)) \right\rangle \le \varepsilon \|\mathcal{A}_t s_t\|_2^2 + \frac{1}{\varepsilon} \|\mathcal{B}_t \operatorname{vec}(\nabla f(W_t))\|_2^2. \tag{190}$$

Furthermore,

$$2\left\langle \mathcal{A}_t s_t, \mathcal{B}_t \operatorname{vec}(\nabla f(W_t)) \right\rangle \le \varepsilon \rho^2 \|s_t\|_2^2 + \frac{C_B^2}{\varepsilon} \|\nabla f(W_t)\|_F^2, \tag{191}$$

where

$$C_B = \frac{2}{\gamma}.$$

Proof. Applying Young's inequality,

$$2a^\top b \le \varepsilon\|a\|^2 + \frac{1}{\varepsilon}\|b\|^2,$$

with

$$a = \mathcal{A}_t s_t, \qquad b = \mathcal{B}_t \operatorname{vec}(\nabla f(W_t)),$$

gives

$$2\left\langle \mathcal{A}_t s_t, \mathcal{B}_t \operatorname{vec}(\nabla f(W_t))\right\rangle \le \varepsilon\|\mathcal{A}_t s_t\|_2^2 + \frac{1}{\varepsilon}\|\mathcal{B}_t \operatorname{vec}(\nabla f(W_t))\|_2^2.$$

Using

$$\|\mathcal{A}_t\|_2 = \rho,$$

and

$$\|\mathcal{B}_t\|_2 \le C_B,$$

we obtain

$$\|\mathcal{A}_t s_t\|_2^2 \le \rho^2\|s_t\|_2^2,$$

and

$$\|\mathcal{B}_t \operatorname{vec}(\nabla f(W_t))\|_2^2 \le C_B^2\|\nabla f(W_t)\|_F^2,$$

which yields the desired result.

□

Lemma 8 (Bound on Stochastic Gradient Energy Injection)**.** *Under Assumption 2 (Unbiased Gradient Estimator and Bounded Variance), the stochastic input satisfies*

$$\mathbb{E}_t\left[\|\mathcal{B}_t\tilde{g}_t\|_2^2\right] \le C_B^2\|\nabla f(W_t)\|_F^2 + C_B^2\sigma_g^2, \tag{192}$$

where

$$C_B = \frac{2}{\gamma}.$$

Proof. Recall that

$$\tilde{g}_t = \operatorname{vec}(\nabla f(W_t)) + \xi_t,$$

where

$$\mathbb{E}_t[\xi_t] = 0, \qquad \mathbb{E}_t\left[\|\xi_t\|_2^2\right] \le \sigma_g^2.$$

Using the induced matrix norm,

$$\begin{aligned}\mathbb{E}_t\left[\|\mathcal{B}_t\tilde{g}_t\|_2^2\right] &= \mathbb{E}_t\left[\tilde{g}_t^\top \mathcal{B}_t^\top \mathcal{B}_t \tilde{g}_t\right] \\ &\le \|\mathcal{B}_t\|_2^2\, \mathbb{E}_t\left[\|\tilde{g}_t\|_2^2\right] \\ &\le C_B^2\, \mathbb{E}_t\left[\|\tilde{g}_t\|_2^2\right]. \end{aligned} \tag{193}$$

Expanding the second moment,

$$\begin{aligned}\mathbb{E}_t\left[\|\tilde{g}_t\|_2^2\right] &= \|\mathrm{vec}(\nabla f(W_t))\|_2^2 + 2\mathbb{E}_t\left[\langle \mathrm{vec}(\nabla f(W_t)), \xi_t\rangle\right] + \mathbb{E}_t\left[\|\xi_t\|_2^2\right] \\ &= \|\nabla f(W_t)\|_F^2 + \mathbb{E}_t\left[\|\xi_t\|_2^2\right] \\ &\le \|\nabla f(W_t)\|_F^2 + \sigma_g^2, \end{aligned} \tag{194}$$

where the cross term has zero conditional expectation because

$$\mathbb{E}_t[\xi_t] = 0.$$

Therefore,

$$\mathbb{E}_t\left[\|\mathcal{B}_t\tilde{g}_t\|_2^2\right] \le C_B^2\left(\|\nabla f(W_t)\|_F^2 + \sigma_g^2\right), \tag{195}$$

which completes the proof.

□

Lemma 9 (Conditional State-Energy Recursion)**.** *Under Assumptions 2–3, the conditional second moment of the momentum state satisfies*

$$\mathbb{E}_t\big[\|s_{t+1}\|_2^2\big] \le (1+\varepsilon)\rho^2\|s_t\|_2^2 + \left(1+\frac{1}{\varepsilon}\right) C_B^2\|\nabla f(W_t)\|_F^2 + C_B^2\sigma_g^2, \tag{196}$$

for every $\varepsilon > 0$, where

$$C_B = \frac{2}{\gamma}.$$

Proof. Starting from the exact conditional energy decomposition established in Equation equation 186,

$$\begin{aligned}\mathbb{E}_t\big[\|s_{t+1}\|_2^2\big] &= \|\mathcal{A}_t s_t\|_2^2 + 2\left\langle \mathcal{A}_t s_t, \mathcal{B}_t\,\mathrm{vec}(\nabla f(W_t))\right\rangle \\ &\quad + \mathbb{E}_t\left[\|\mathcal{B}_t\tilde{g}_t\|_2^2\right]. \end{aligned} \tag{197}$$

Applying Lemma 5,

$$\|\mathcal{A}_t s_t\|_2^2 \le \rho^2\|s_t\|_2^2.$$

Applying Lemma 7,

$$2\left\langle \mathcal{A}_t s_t, \mathcal{B}_t\,\mathrm{vec}(\nabla f(W_t))\right\rangle \le \varepsilon\rho^2\|s_t\|_2^2 + \frac{C_B^2}{\varepsilon}\|\nabla f(W_t)\|_F^2.$$

Finally, applying Lemma 8,

$$\mathbb{E}_t\big[\|\mathcal{B}_t\tilde{g}_t\|_2^2\big] \le C_B^2\|\nabla f(W_t)\|_F^2 + C_B^2\sigma_g^2.$$

Combining the three bounds gives

$$\mathbb{E}_t\left[\|s_{t+1}\|_2^2\right] \le \rho^2\|s_t\|_2^2 + \varepsilon\rho^2\|s_t\|_2^2 + \frac{C_B^2}{\varepsilon}\|\nabla f(W_t)\|_F^2 + C_B^2\|\nabla f(W_t)\|_F^2 + C_B^2\sigma_g^2. \tag{198}$$

Collecting like terms,

$$\boxed{\mathbb{E}_t\left[\|s_{t+1}\|_2^2\right] \le (1+\varepsilon)\rho^2\|s_t\|_2^2 + \left(1+\frac{1}{\varepsilon}\right)C_B^2\|\nabla f(W_t)\|_F^2 + C_B^2\sigma_g^2,}$$

which completes the proof.

□

L.3 Lyapunov Drift Construction

Under Assumptions 1–3, to jointly analyze the optimization objective and the internal momentum dynamics, we introduce the Lyapunov potential

$$\mathcal{L}_t \triangleq f(W_t) + \lambda\|s_t\|_2^2, \tag{199}$$

where $\lambda > 0$ is a scalar weighting parameter that will be chosen later to balance the objective descent and the momentum-state energy.

Taking the conditional expectation,

$$\mathbb{E}_t[\mathcal{L}_{t+1}] = \mathbb{E}_t[f(W_{t+1})] + \lambda\mathbb{E}_t\left[\|s_{t+1}\|_2^2\right]. \tag{200}$$

Applying the smoothness descent inequality,

$$\mathbb{E}_t[f(W_{t+1})] \le f(W_t) - \eta\left\langle\nabla f(W_t), \mathbb{E}_t[M_{t+1}]\right\rangle + \frac{L\eta^2}{2}\mathbb{E}_t\left[\|s_{t+1}\|_2^2\right]. \tag{201}$$

Therefore,

$$\mathbb{E}_t[\mathcal{L}_{t+1}] \le f(W_t) - \eta\left\langle\nabla f(W_t), \mathbb{E}_t[M_{t+1}]\right\rangle + \left(\lambda + \frac{L\eta^2}{2}\right)\mathbb{E}_t\left[\|s_{t+1}\|_2^2\right]. \tag{202}$$

Using

$$\mathbb{E}_t[\tilde{g}_t] = \mathrm{vec}(\nabla f(W_t)),$$

the conditional expectation of the momentum state satisfies

$$\mathbb{E}_t[s_{t+1}] = \mathcal{A}_t s_t + \mathcal{B}_t\,\mathrm{vec}(\nabla f(W_t)).$$

Hence,

$$\begin{aligned}&\left\langle\nabla f(W_t), \mathbb{E}_t[M_{t+1}]\right\rangle \\ &= \mathrm{vec}(\nabla f(W_t))^\top\mathcal{A}_t s_t + \mathrm{vec}(\nabla f(W_t))^\top\mathcal{B}_t\,\mathrm{vec}(\nabla f(W_t)).\end{aligned} \tag{203}$$

Applying Assumption 3,

$$\mathcal{B}_t \succeq \mu_B I,$$

which implies

$$\operatorname{vec}(\nabla f(W_t))^\top \mathcal{B}_t \operatorname{vec}(\nabla f(W_t)) \geq \mu_B \|\nabla f(W_t)\|_F^2. \tag{204}$$

Applying Young's inequality,

$$2a^\top b \leq \delta \|a\|^2 + \frac{1}{\delta}\|b\|^2,$$

with

$$a = \operatorname{vec}(\nabla f(W_t)), \qquad b = \mathcal{A}_t s_t,$$

yields

$$\left|\operatorname{vec}(\nabla f(W_t))^\top \mathcal{A}_t s_t\right| \leq \frac{\delta}{2}\|\nabla f(W_t)\|_F^2 + \frac{1}{2\delta}\|\mathcal{A}_t s_t\|_2^2. \tag{205}$$

Since

$$\|\mathcal{A}_t s_t\|_2^2 \leq \rho^2 \|s_t\|_2^2,$$

we obtain

$$\left|\operatorname{vec}(\nabla f(W_t))^\top \mathcal{A}_t s_t\right| \leq \frac{\delta}{2}\|\nabla f(W_t)\|_F^2 + \frac{\rho^2}{2\delta}\|s_t\|_2^2. \tag{206}$$

Substituting the previous two bounds gives

$$\begin{aligned}\mathbb{E}_t[\mathcal{L}_{t+1}] \leq f(W_t) &- \eta\mu_B\|\nabla f(W_t)\|_F^2 \\ &+ \eta\left(\frac{\delta}{2}\|\nabla f(W_t)\|_F^2 + \frac{\rho^2}{2\delta}\|s_t\|_2^2\right) \\ &+ \left(\lambda + \frac{L\eta^2}{2}\right)\mathbb{E}_t\left[\|s_{t+1}\|_2^2\right].\end{aligned} \tag{207}$$

$$\begin{aligned}\mathbb{E}_t[\mathcal{L}_{t+1}] \leq f(W_t) &- \eta\left(\mu_B - \frac{\delta}{2}\right)\|\nabla f(W_t)\|_F^2 \\ &+ \frac{\eta\rho^2}{2\delta}\|s_t\|_2^2 \\ &+ \left(\lambda + \frac{L\eta^2}{2}\right)\mathbb{E}_t\left[\|s_{t+1}\|_2^2\right].\end{aligned} \tag{208}$$

Substituting the conditional state-energy recursion from Lemma 9 into equation 208 yields

$$\mathbb{E}_t[\mathcal{L}_{t+1}] \le f(W_t) - \eta\left(\mu_B - \frac{\delta}{2}\right)\|\nabla f(W_t)\|_F^2 + \frac{\eta\rho^2}{2\delta}\|s_t\|_2^2 + \left(\lambda + \frac{L\eta^2}{2}\right)\left[(1+\varepsilon)\rho^2\|s_t\|_2^2 + \left(1+\frac{1}{\varepsilon}\right)C_B^2\|\nabla f(W_t)\|_F^2 + C_B^2\sigma_g^2\right]. \tag{209}$$

Expanding the terms gives

$$\mathbb{E}_t[\mathcal{L}_{t+1}] \le f(W_t) - \left[\eta\left(\mu_B - \frac{\delta}{2}\right) - \left(\lambda + \frac{L\eta^2}{2}\right)\left(1+\frac{1}{\varepsilon}\right)C_B^2\right]\|\nabla f(W_t)\|_F^2 + \left[\frac{\eta\rho^2}{2\delta} + \left(\lambda + \frac{L\eta^2}{2}\right)(1+\varepsilon)\rho^2\right]\|s_t\|_2^2 + \left(\lambda + \frac{L\eta^2}{2}\right)C_B^2\sigma_g^2. \tag{210}$$

Recalling that

$$\mathcal{L}_t = f(W_t) + \lambda\|s_t\|_2^2,$$

we obtain

$$\mathbb{E}_t[\mathcal{L}_{t+1}] \le \mathcal{L}_t - \left[\eta\left(\mu_B - \frac{\delta}{2}\right) - \left(\lambda + \frac{L\eta^2}{2}\right)\left(1+\frac{1}{\varepsilon}\right)C_B^2\right]\|\nabla f(W_t)\|_F^2 - \left[\lambda - \frac{\eta\rho^2}{2\delta} - \left(\lambda + \frac{L\eta^2}{2}\right)(1+\varepsilon)\rho^2\right]\|s_t\|_2^2 + \left(\lambda + \frac{L\eta^2}{2}\right)C_B^2\sigma_g^2. \tag{211}$$

Define the coefficients

$$A \triangleq \eta\left(\mu_B - \frac{\delta}{2}\right) - \left(\lambda + \frac{L\eta^2}{2}\right)\left(1+\frac{1}{\varepsilon}\right)C_B^2, \tag{212}$$

$$B \triangleq \lambda - \frac{\eta\rho^2}{2\delta} - (1+\varepsilon)\rho^2\left(\lambda + \frac{L\eta^2}{2}\right). \tag{213}$$

Then equation 211 becomes

$$\mathbb{E}_t[\mathcal{L}_{t+1}] \le \mathcal{L}_t - A\|\nabla f(W_t)\|_F^2 - B\|s_t\|_2^2 + \left(\lambda + \frac{L\eta^2}{2}\right)C_B^2\sigma_g^2. \tag{214}$$

We select

$$\delta = \mu_B,$$

which symmetrically balances the Young inequality and yields

$$\mu_B - \frac{\delta}{2} = \frac{\mu_B}{2}.$$

Consequently,

$$A = \frac{\eta\mu_B}{2} - \left(\lambda + \frac{L\eta^2}{2}\right)\left(1 + \frac{1}{\varepsilon}\right) C_B^2, \tag{215}$$

$$B = \lambda - \frac{\eta\rho^2}{2\mu_B} - (1+\varepsilon)\rho^2\left(\lambda + \frac{L\eta^2}{2}\right). \tag{216}$$

Expanding the coefficient B gives

$$\begin{aligned} B &= \lambda - (1+\varepsilon)\rho^2\lambda - \frac{\eta\rho^2}{2\mu_B} - \frac{(1+\varepsilon)\rho^2 L\eta^2}{2} \\ &= \left(1 - (1+\varepsilon)\rho^2\right)\lambda - \frac{\eta\rho^2}{2\mu_B} - \frac{(1+\varepsilon)\rho^2 L\eta^2}{2}. \end{aligned} \tag{217}$$

Therefore, a necessary condition for the existence of a positive Lyapunov weight satisfying $B > 0$ is

$$1 - (1+\varepsilon)\rho^2 > 0, \tag{218}$$

or equivalently,

$$\boxed{\varepsilon < \frac{1}{\rho^2} - 1.} \tag{219}$$

Under this condition, the coefficient of λ remains strictly positive. Solving the inequality $B > 0$ for λ yields

$$\boxed{\lambda > \frac{\dfrac{\eta\rho^2}{2\mu_B} + \dfrac{(1+\varepsilon)\rho^2 L\eta^2}{2}}{1 - (1+\varepsilon)\rho^2}.} \tag{220}$$

Substituting the expression for A,

$$A = \frac{\eta\mu_B}{2} - \left(\lambda + \frac{L\eta^2}{2}\right)\left(1 + \frac{1}{\varepsilon}\right) C_B^2. \tag{221}$$

Therefore,

$$A > 0$$

holds if and only if

$$\boxed{\lambda < \frac{\eta\mu_B}{2C_B^2\left(1 + \frac{1}{\varepsilon}\right)} - \frac{L\eta^2}{2}.} \tag{222}$$

Combining the lower bound obtained from $B > 0$ with the upper bound required by $A > 0$, a feasible Lyapunov weight exists whenever

$$\boxed{\frac{\frac{\eta\rho^2}{2\mu_B} + \frac{(1+\varepsilon)\rho^2 L\eta^2}{2}}{1-(1+\varepsilon)\rho^2} < \frac{\eta\mu_B}{2C_B^2\left(1+\frac{1}{\varepsilon}\right)} - \frac{L\eta^2}{2}.} \tag{223}$$

Lemma 10 (Existence of an Admissible Lyapunov Weight)**.** *Suppose*

$$0 < \rho < 1,$$

and choose

$$0 < \varepsilon < \frac{1}{\rho^2} - 1.$$

Assume additionally that

$$\boxed{\frac{\mu_B}{C_B^2\left(1+\frac{1}{\varepsilon}\right)} > \frac{\rho^2}{\mu_B\left[1-(1+\varepsilon)\rho^2\right]}.}$$

Then there exists a sufficiently small learning rate

$$\eta^\star > 0$$

such that for every

$$0 < \eta < \eta^\star,$$

the admissible interval

$$(\lambda_{\min}, \lambda_{\max})$$

is non-empty, where

$$\lambda_{\min} = \frac{\frac{\eta\rho^2}{2\mu_B} + \frac{(1+\varepsilon)\rho^2 L\eta^2}{2}}{1-(1+\varepsilon)\rho^2},$$

and

$$\lambda_{\max} = \frac{\eta\mu_B}{2C_B^2\left(1+\frac{1}{\varepsilon}\right)} - \frac{L\eta^2}{2}.$$

Consequently, there exists a positive scalar

$$\lambda \in (\lambda_{\min}, \lambda_{\max})$$

such that both coefficients

$$A > 0, \qquad B > 0$$

hold simultaneously.

Proof. Define the linear and quadratic coefficients

$$a_{\min} = \frac{\rho^2}{2\mu_B\left[1-(1+\varepsilon)\rho^2\right]},$$

$$b_{\min} = \frac{(1+\varepsilon)\rho^2 L}{2\left[1-(1+\varepsilon)\rho^2\right]},$$

and

$$a_{\max} = \frac{\mu_B}{2C_B^2\left(1+\frac{1}{\varepsilon}\right)}.$$

Then

$$\lambda_{\min} = a_{\min}\eta + b_{\min}\eta^2,$$

while

$$\lambda_{\max} = a_{\max}\eta - \frac{L}{2}\eta^2.$$

By the additional assumption,

$$a_{\max} > a_{\min}.$$

Therefore,

$$\lambda_{\max} - \lambda_{\min} = (a_{\max} - a_{\min})\eta - \left(\frac{L}{2} + b_{\min}\right)\eta^2.$$

Hence

$$\lambda_{\max} - \lambda_{\min} = \eta\left[(a_{\max} - a_{\min}) - \left(\frac{L}{2} + b_{\min}\right)\eta\right].$$

Since $a_{\max} - a_{\min} > 0$, there exists

$$\eta^\star > 0$$

such that

$$0 < \eta < \eta^\star \quad \Longrightarrow \quad \lambda_{\max} - \lambda_{\min} > 0.$$

Thus,

$$\lambda_{\min} < \lambda_{\max},$$

so the interval

$$(\lambda_{\min}, \lambda_{\max})$$

is non-empty. Choosing

$$\lambda \in (\lambda_{\min}, \lambda_{\max})$$

guarantees

$$A > 0, \qquad B > 0.$$

□

L.4 Completion of the Lyapunov Drift Argument

By Lemma 9 and the existence result above, there exist positive constants

$$A > 0, \qquad B > 0,$$

such that the master drift inequality becomes

$$\mathbb{E}_t[\mathcal{L}_{t+1}] \leq \mathcal{L}_t - A\|\nabla f(W_t)\|_F^2 - B\|s_t\|_2^2 + C\sigma_g^2, \tag{224}$$

where

$$C = \left(\lambda + \frac{L\eta^2}{2}\right) C_B^2 > 0.$$

Since the state-energy term is non-negative,

$$B\|s_t\|_2^2 \geq 0,$$

discarding it preserves the inequality. Therefore,

$$\mathbb{E}_t[\mathcal{L}_{t+1}] \leq \mathcal{L}_t - A\|\nabla f(W_t)\|_F^2 + C\sigma_g^2. \tag{225}$$

Taking the full expectation and summing equation 225 over $t = 0, \ldots, T-1$ gives

$$\sum_{t=0}^{T-1} \mathbb{E}[\mathcal{L}_{t+1}] \leq \sum_{t=0}^{T-1} \mathbb{E}[\mathcal{L}_t] - A \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(W_t)\|_F^2\right] + TC\sigma_g^2. \tag{226}$$

The Lyapunov terms telescope,

$$\mathbb{E}[\mathcal{L}_T] - \mathcal{L}_0 \leq -A \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(W_t)\|_F^2\right] + TC\sigma_g^2. \tag{227}$$

Rearranging yields

$$A \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(W_t)\|_F^2\right] \leq \mathcal{L}_0 - \mathbb{E}[\mathcal{L}_T] + TC\sigma_g^2. \tag{228}$$

Since

$$\mathcal{L}_t = f(W_t) + \lambda\|s_t\|_2^2,$$

with $\lambda > 0$, it follows that

$$\mathcal{L}_t \geq f(W_t) \geq f^,$$

where f denotes the infimum of the objective function.

Consequently,

$$\mathbb{E}[\mathcal{L}_T] \geq f^.$$

Substituting this bound gives

$$A\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_F^2\right] \le \mathcal{L}_0 - f^{+TC\sigma_g^2}. \quad (229)$$

Dividing both sides by AT yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_F^2\right] \le \frac{\mathcal{L}_0 - f}{AT} + \frac{C}{A}\sigma_g^2. \quad (230)$$

Therefore,

$$\boxed{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_F^2\right] = \mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}(\sigma_g^2).} \quad (231)$$

In particular, in the deterministic setting ($\sigma_g^2 = 0$),

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(W_t)\|_F^2\right] = \mathcal{O}\left(\frac{1}{T}\right),$$

recovering the classical first-order stationarity rate for smooth non-convex optimization.

M Hyperparameter Selection: Adaptive Tuning via Softplus Hypergradients

M.1 Alternative Tuning Paradigms and Empirical Pitfalls

Selecting suitable values for the decay rate parameter $\alpha > 0$ and learning rate $\eta > 0$ is critical, as they dictate both the spectral dissipation factor $\rho = \frac{1}{2\alpha\eta+1} \in (0,1)$ and the overall optimizer stability. During initial development, we explored several intuitive adaptive heuristics:

- **Reliability-Based Heuristics:** Scaling α dynamically based on gradient direction variance (e.g., directional signal-to-noise ratio).
- **Energy-Based Dissipation:** Matching ρ_t to kinetic energy fluctuations $\|M_t\|_F^2$.
- **LTV Pole Placement Paradigm:** Dynamically shifting eigenvalues to target specific decay trajectories based on local Hessian estimates.

But none produced promising results

M.2 Hypergradient Descent with Softplus Enclosure

To overcome these failure modes, we adopt a continuous, gradient-based hyperparameter optimization scheme based on *Hypergradient Descent* ((Baydin et al., 2018, Vancouver, BC, Canada)), enforcing strict positivity via a smooth Softplus parametrization.

Let α and η be reparametrized through unconstrained real-valued latent hyperparameters $\theta_\alpha, \theta_\eta \in \mathbb{R}$:

$$\alpha(\theta_\alpha) = \frac{1}{\beta}\ln\left(1 + e^{\beta\theta_\alpha}\right), \qquad \eta(\theta_\eta) = \frac{1}{\beta}\ln\left(1 + e^{\beta\theta_\eta}\right), \quad (232)$$

where $\beta > 0$ (default $\beta = 1$) controls the smoothness of the Softplus transformation, strictly guaranteeing $\alpha > 0$ and $\eta > 0$ for all real latent values.

The parameters θ_α and θ_η are updated online along the negative hypergradient of the primary loss function $\mathcal{L}_{t+1} = f(W_{t+1})$ with respect to the latent states:

$$\theta_{\alpha,t+1} = \theta_{\alpha,t} - \mu_\theta\frac{\partial\mathcal{L}_{t+1}}{\partial\theta_\alpha}, \qquad \theta_{\eta,t+1} = \theta_{\eta,t} - \mu_\theta\frac{\partial\mathcal{L}_{t+1}}{\partial\theta_\eta}, \quad (233)$$

where $\mu_\theta > 0$ represents the hyper-learning rate.

Applying the chain rule through the update step $W_{t+1} = W_t - \eta M_{t+1}$:

$$\frac{\partial \mathcal{L}_{t+1}}{\partial \theta_\alpha} = \left\langle \nabla f(W_{t+1}),\, \frac{\partial W_{t+1}}{\partial \alpha} \right\rangle \cdot \sigma(\beta\theta_\alpha), \tag{234}$$

$$\frac{\partial \mathcal{L}_{t+1}}{\partial \theta_\eta} = \left\langle \nabla f(W_{t+1}),\, \frac{\partial W_{t+1}}{\partial \eta} \right\rangle \cdot \sigma(\beta\theta_\eta), \tag{235}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the standard logistic sigmoid function arising from the derivative of the Softplus transformation $\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{softplus}(z) = \sigma(z)$, and $\nabla f(W_{t+1})$ is approximated using the consecutive mini-batch gradient g_{t+1}.

M.3 Computational Complexity and Limitations

Time and Memory Complexity: Evaluating the hypergradients requires differentiating through the optimizer update and evaluating the scalar elementwise Softplus/sigmoid mappings. When the rank-one structure of the AXIOM update is exploited, the additional operations associated with the activation-conditioned state update scale linearly with the number of layer parameters.

Optimizer Step	Asymptotic Time	Additional Memory
Standard Adam / SGDM	$\mathcal{O}(d)$	$\mathcal{O}(d)$
AXIOM (Fixed α, η)	$\mathcal{O}(d)$	$\mathcal{O}(d_{\text{in}})$
AXIOM (Softplus Hypergradient)	$\mathcal{O}(d)$ plus hypergradient cost	Implementation-dependent

Table 6: Asymptotic computational and memory characteristics, with $d = d_{\text{out}} d_{\text{in}}$. The AXIOM update exploits the rank-one structure of the activation-conditioned operator rather than explicitly forming the $d_{\text{in}} \times d_{\text{in}}$ matrix.

Primary Limitation (Hypergradient Overhead): While Softplus hypergradient adaptation reduces the need for manual tuning of α and η, it introduces additional computational and implementation complexity:

1. **Differentiation Through the Optimizer:** Computing the hypergradients requires differentiating through the optimizer update, introducing additional computation and bookkeeping compared with fixed-parameter optimization.
2. **Hyper-Hyperparameter Dependence:** Although α and η are adapted automatically, the hypergradient procedure introduces additional scalar hyperparameters, including the hyper-learning rate μ_θ and the Softplus parameter β. Their selection therefore remains a practical consideration.

The theoretical development in the preceding sections, including the Lyapunov stability analysis, Global Uniform Exponential Stability (GUES), and the non-convex convergence analysis, is carried out under the assumption of fixed optimizer parameters (α, η, γ). In the practical implementation of AXIOM, these parameters are updated online using Hypergradient Descent. Consequently, the resulting optimizer becomes a parameter-varying discrete-time LTV system.

When the hyperparameter updates are sufficiently slow relative to the momentum dynamics, the parameters (α, η, γ) may be regarded as approximately constant over a short optimization horizon. Under this frozen-parameter (quasi-static) assumption, the local system dynamics are described by the theoretical model developed above.

Under this quasi-static interpretation, the previously established fixed-parameter stability and convergence results provide a local reference for the adaptive dynamics, while the adaptive optimizer can be viewed as a slowly varying sequence of frozen LTV systems. A complete analysis of the fully coupled parameter-varying system, in which both the optimizer state and the hyperparameters evolve simultaneously, is beyond the scope of this work and is left for future investigation.

N Discussion: Interaction Between AXIOM and Dropout

Although AXIOM is derived independently of neural network regularization techniques, its operator formulation provides an interesting perspective on the effect of dropout. Since the optimizer is driven directly by the activation vector, randomly masking neurons modifies the geometry of the state-space operator itself. This section analyzes how dropout changes the memory dynamics of AXIOM and concludes with an interpretive analogy to distributed biological computation.

N.1 Dropout-Induced Modification of the State Operator

Let the dropout mask at iteration t be

$$D_t = \mathrm{diag}(m_1, m_2, \ldots, m_{d_{\mathrm{in}}}),$$

where

$$m_i \in \{0, 1\}, \qquad \mathbb{P}(m_i = 1) = 1 - p.$$

The activation vector becomes

$$\tilde{\mathbf{x}}_t = D_t \mathbf{x}_t.$$

Consequently, the AXIOM update becomes

$$\mathbf{M}_{t+1} = \bar{\mathbf{M}}_t + \frac{2}{\gamma + 2\|\tilde{\mathbf{x}}_t\|_2^2} \left(\mathbf{u}_t - \bar{\mathbf{M}}_t \tilde{\mathbf{x}}_t\right) \tilde{\mathbf{x}}_t^\top. \tag{236}$$

Therefore, dropout influences AXIOM solely through the modified activation geometry. The optimizer itself is unchanged; instead, the state-transition operator is constructed from the surviving activation coordinates.

N.2 Effect on the State-Transition Operator

The corresponding operator acting on the state becomes

$$P_t^{\mathrm{drop}} = I - \frac{2}{\gamma + 2\|\tilde{\mathbf{x}}_t\|_2^2} \tilde{\mathbf{x}}_t \tilde{\mathbf{x}}_t^\top. \tag{237}$$

Unlike the original operator, this transformation acts only on the randomly selected active neurons. Consequently, dropout changes the activation-aligned direction and therefore the geometry of the state-transition operator rather than merely scaling the optimization step. The optimization geometry therefore adapts automatically to the stochastic feature subspace induced by the dropout mask.

N.3 Gradient Responsiveness Under Dropout

Since

$$\|\tilde{\mathbf{x}}_t\|_2^2 \le \|\mathbf{x}_t\|_2^2,$$

it follows that

$$\frac{2}{\gamma + 2\|\tilde{\mathbf{x}}_t\|_2^2} \ge \frac{2}{\gamma + 2\|\mathbf{x}_t\|_2^2}. \tag{238}$$

Hence, AXIOM therefore applies a larger correction gain along the masked activation direction when dropout reduces the activation energy. Intuitively, because fewer neurons participate in the current representation, AXIOM automatically allocates a larger correction to the remaining active feature directions.

N.4 Memory Retention Under Dropout

The activation-aligned eigenvalue of the state transition operator becomes

$$\lambda_{\parallel}^{\text{drop}} = \frac{\rho\gamma}{\gamma + 2\|\tilde{\mathbf{x}}_t\|_2^2}. \tag{239}$$

Since

$$\|\tilde{\mathbf{x}}_t\|_2^2 \le \|\mathbf{x}_t\|_2^2,$$

we obtain

$$\lambda_{\parallel}^{\text{drop}} \ge \lambda_{\parallel}. \tag{240}$$

Therefore, dropout increases the memory-retention eigenvalue along the activation-aligned direction, causing historical momentum to decay more slowly than in the fully activated network.

N.5 Entropy Dynamics Under Dropout

Replacing the activation vector by the masked activation modifies the deterministic entropy dissipation term as

$$\Delta H_{\text{diss}}^{\text{drop}} = d \ln \rho - d_{\text{out}} \ln \left(1 + \frac{2}{\gamma}\|\tilde{\mathbf{x}}_t\|_2^2\right). \tag{241}$$

Since

$$\|\tilde{\mathbf{x}}_t\|_2^2 \le \|\mathbf{x}_t\|_2^2,$$

the logarithmic compression term decreases, implying weaker deterministic entropy dissipation. Consequently, dropout preserves a larger amount of uncertainty within the momentum state than the fully activated network.

N.6 Expected Activation Energy Under Dropout

Assuming independent Bernoulli dropout,

$$\mathbb{E}[D_t] = (1-p)I,$$

and, for a fixed activation vector $\mathbf{x}_t$,

$$\mathbb{E}\left[\|\tilde{\mathbf{x}}_t\|_2^2\right] = (1-p)\|\mathbf{x}_t\|_2^2. \tag{242}$$

Thus, dropout reduces the expected activation energy by the factor $1-p$. Since the activation-aligned correction gain and memory retention depend nonlinearly on $\|\tilde{\mathbf{x}}_t\|_2^2$, their expectations cannot in general be obtained by simply substituting the expected activation energy into their respective expressions. The exact expected gains depend on the distribution of the dropout mask.

N.7 Interpretive Analogy to Distributed Biological Computation

Although AXIOM is derived entirely from convex optimization and state-space analysis, its dynamics admit an interpretive analogy to distributed biological computation. In this analogy, partial activation corresponds to computation being carried by a subset of available pathways, while the remaining active pathways continue to participate in the computation. Similarly, dropout randomly removes a subset of activation coordinates, while AXIOM recomputes its activation-conditioned

state-transition operator and correction gain using the resulting masked activation vector. The optimizer therefore continues to update through the surviving activation directions without modifying its underlying optimization rule.

This comparison is intended solely as an interpretive analogy rather than a mechanistic model of neural compensation.

O Fixed-Point and Repeated-Activation Analysis of AXIOM

The main body of this report analyzes the transient behavior, stability, and convergence properties of AXIOM. In this appendix, we examine the asymptotic behavior of the optimizer under repeated presentation of the same activation pattern. Although this analysis is not required for the derivation of AXIOM, it provides additional insight into the associative-memory interpretation of the proposed update rule.

O.1 Fixed-Point Characterization

Consider the AXIOM momentum update

$$\mathbf{M}_{t+1} = \bar{\mathbf{M}}_t + \frac{2}{\gamma + 2\|\mathbf{x}\|_2^2} \left(\mathbf{u} - \bar{\mathbf{M}}_t \mathbf{x}\right) \mathbf{x}^\top, \tag{243}$$

where the same activation vector $\mathbf{x}$ and desired correction $\mathbf{u}$ are repeatedly observed.

A fixed point satisfies

$$\mathbf{M}_{t+1} = \bar{\mathbf{M}}_t = \mathbf{M}^\star.$$

Substituting into the update equation gives

$$\mathbf{0} = \frac{2}{\gamma + 2\|\mathbf{x}\|_2^2} \left(\mathbf{u} - \mathbf{M}^\star \mathbf{x}\right) \mathbf{x}^\top. \tag{244}$$

Since

$$\gamma > 0,$$

and

$$\mathbf{x} \neq \mathbf{0},$$

the scalar coefficient is strictly positive, yielding the equilibrium condition

$$\boxed{\mathbf{M}^\star \mathbf{x} = \mathbf{u}.} \tag{245}$$

Thus, the equilibrium momentum matrix maps the repeatedly presented activation vector exactly to the desired output.

O.2 Interpretation as an Associative Memory

The fixed-point equation

$$\mathbf{M}^\star \mathbf{x} = \mathbf{u}$$

admits a natural associative-memory interpretation. Rather than explicitly storing gradients, AXIOM continuously adjusts the momentum matrix until the current activation vector is mapped to the desired update direction. In this sense, the momentum matrix acts as an associative mapping between feature representations and optimization corrections.

Unlike conventional exponential moving-average momentum, the equilibrium of AXIOM is therefore characterized by an input–output association rather than by accumulation of historical gradients alone.

O.3 Repeated Presentation of the Same Pattern

Suppose the same pair

$$(\mathbf{x}, \mathbf{u})$$

is presented repeatedly.

Define the prediction error

$$\mathbf{e}_t = \mathbf{u} - \mathbf{M}_t\mathbf{x}. \tag{246}$$

Recalling $\bar{\mathbf{M}}_t = \rho\mathbf{M}_t$, multiplying the update equation on the right by $\mathbf{x}$ yields

$$\begin{aligned}\mathbf{M}_{t+1}\mathbf{x} &= \bar{\mathbf{M}}_t\mathbf{x} + \frac{2\|\mathbf{x}\|_2^2}{\gamma + 2\|\mathbf{x}\|_2^2}\left(\mathbf{u} - \bar{\mathbf{M}}_t\mathbf{x}\right) \\ &= \frac{\gamma}{\gamma + 2\|\mathbf{x}\|_2^2}\bar{\mathbf{M}}_t\mathbf{x} + \frac{2\|\mathbf{x}\|_2^2}{\gamma + 2\|\mathbf{x}\|_2^2}\mathbf{u} \\ &= \lambda_{\|}\mathbf{M}_t\mathbf{x} + (1 - \lambda_{\|})\mathbf{u}, \end{aligned} \tag{247}$$

where

$$\lambda_{\|} = \frac{\rho\gamma}{\gamma + 2\|\mathbf{x}\|_2^2}, \qquad 0 < \lambda_{\|} < 1.$$

Consequently, the error vector evolves according to

$$\begin{aligned}\mathbf{e}_{t+1} &= \mathbf{u} - \mathbf{M}_{t+1}\mathbf{x} \\ &= \mathbf{u} - \left[\lambda_{\|}\mathbf{M}_t\mathbf{x} + (1 - \lambda_{\|})\mathbf{u}\right] \\ &= \lambda_{\|}(\mathbf{u} - \mathbf{M}_t\mathbf{x}) \\ &= \lambda_{\|}\mathbf{e}_t, \end{aligned} \tag{248}$$

which yields the closed-form solution

$$\boxed{\mathbf{e}_t = \lambda_{\|}^t\mathbf{e}_0.} \tag{249}$$

Since

$$0 < \lambda_{\|} < 1,$$

the prediction error converges exponentially to zero,

$$\lim_{t\to\infty} \mathbf{e}_t = \mathbf{0},$$

implying

$$\lim_{t\to\infty} \mathbf{M}_t\mathbf{x} = \mathbf{u}.$$

O.4 Discussion

This analysis complements the stability results presented in the main text. The Lyapunov analysis establishes that the momentum dynamics remain globally stable under arbitrary activation sequences, whereas the present analysis considers the special case of repeated presentation of a single activation pattern. Under this setting, AXIOM converges exponentially toward a matrix that exactly reproduces the desired association for that activation vector. This behavior further supports the interpretation of AXIOM as a structured associative memory, where optimization proceeds by progressively refining an input-dependent mapping rather than solely averaging past gradients.

P Interference Analysis Between Different Activation Patterns

The previous appendix analyzed repeated presentation of the same activation vector. We now examine how learning one activation pattern influences the response to another pattern. This provides additional insight into the associative-memory behavior of AXIOM.

P.1 Influence of One Update on Another Activation

Consider an update generated by the activation pair

$$(\mathbf{x}_i, \mathbf{u}_i),$$

producing the rank-one correction

$$\Delta \mathbf{M}_i = \frac{2}{\gamma + 2\|\mathbf{x}_i\|_2^2} (\mathbf{u}_i - \mathbf{M}\mathbf{x}_i)\, \mathbf{x}_i^\top. \tag{250}$$

Suppose that, immediately afterwards, another activation vector $\mathbf{x}_j$ is presented.

The change induced on the predicted response for the second activation is

$$\begin{aligned} \Delta \mathbf{y}_j &= \Delta \mathbf{M}_i \mathbf{x}_j \\ &= \frac{2}{\gamma + 2\|\mathbf{x}_i\|_2^2} (\mathbf{u}_i - \mathbf{M}\mathbf{x}_i)\, \mathbf{x}_i^\top \mathbf{x}_j. \end{aligned} \tag{251}$$

Therefore,

$$\boxed{\Delta \mathbf{y}_j \propto \mathbf{x}_i^\top \mathbf{x}_j.} \tag{252}$$

The interaction between two activation patterns is therefore governed entirely by their feature-space inner product.

P.2 Orthogonal Activations

Proposition 1. *If*

$$\mathbf{x}_i^\top \mathbf{x}_j = 0,$$

then learning the activation $\mathbf{x}_i$ produces no first-order interference on the prediction associated with $\mathbf{x}_j$.

Proof. Substituting

$$\mathbf{x}_i^\top \mathbf{x}_j = 0$$

into the previous expression gives

$$\Delta \mathbf{y}_j = \mathbf{0}.$$

Hence the update generated by $\mathbf{x}_i$ does not alter the prediction for the orthogonal activation $\mathbf{x}_j$.

□

P.3 Correlated Activations

When

$$\mathbf{x}_i^\top \mathbf{x}_j \neq 0,$$

the previous update influences the prediction of the new activation according to

$$\|\Delta \mathbf{y}_j\|_2 = \frac{2|\mathbf{x}_i^\top \mathbf{x}_j|}{\gamma + 2\|\mathbf{x}_i\|_2^2} \|\mathbf{u}_i - \mathbf{M}\mathbf{x}_i\|_2. \tag{253}$$

Hence the degree of interference increases continuously with feature similarity. Activations occupying nearby directions in feature space therefore share optimization information, whereas geometrically separated activations evolve almost independently.

P.4 Geometric Interpretation

The above analysis reveals that AXIOM naturally partitions the feature space according to activation geometry. Since every optimization step produces only a rank-one correction, each update modifies the momentum state exclusively along the direction of the current activation vector. Subsequent activations are influenced only to the extent that they overlap with this direction through the inner product $\mathbf{x}_i^\top \mathbf{x}_j$.

Consequently, For a given update state, the immediate cross-pattern interference is governed by activation overlap rather than temporal ordering alone. Orthogonal activations remain decoupled, whereas correlated activations share information through overlapping projection directions.

P.5 Discussion

This result provides an additional interpretation of AXIOM as an associative optimization mechanism. Rather than globally modifying every parameter direction after each gradient update, AXIOM performs localized rank-one corrections whose influence naturally decays with feature dissimilarity. As a result, optimization updates are intrinsically organized by the geometry of the activation space, suggesting a potential mechanism for reduced interference between geometrically dissimilar feature representations.