Continual Learning as a Geometric Flow

T.Y. Tsui

University of Pennsylvania.

Contributing authors: tytsui@seas.upenn.com;

Abstract

Real-world learning happens in time-ordered, non-IID data streams without task boundaries. Yet purely gradient descent is structurally mismatched to this regime: with persistent change and finite capacity, it fixes an early local geometry that constrains later updates, so adaptation proceeds mainly by overwriting formed structure. Rather than adding auxiliary mechanisms to mitigate this issue, we argue that continual learning should be treated as the general learning regime, with empirical risk minimization only as a stationary limit. Based on this view, we propose a coupled learning dynamics that avoids geometric lock-in by introducing an online EMA K-FAC metric that carries persistent information and induces conceptors. These conceptors shape the metric to keep reusable directions accessible while damping drift, while a circulation that conserves unused-capacity of the conceptors drives continuous reallocation toward reuse under change. Under true streaming protocols with no task boundaries, methods without external replay largely break down, whereas our approach can learn effectively. This work offers a new perspective that treats continual learning as the general learning regime and provides a learning dynamic aligned with that regime.

Keywords: Continual Learning

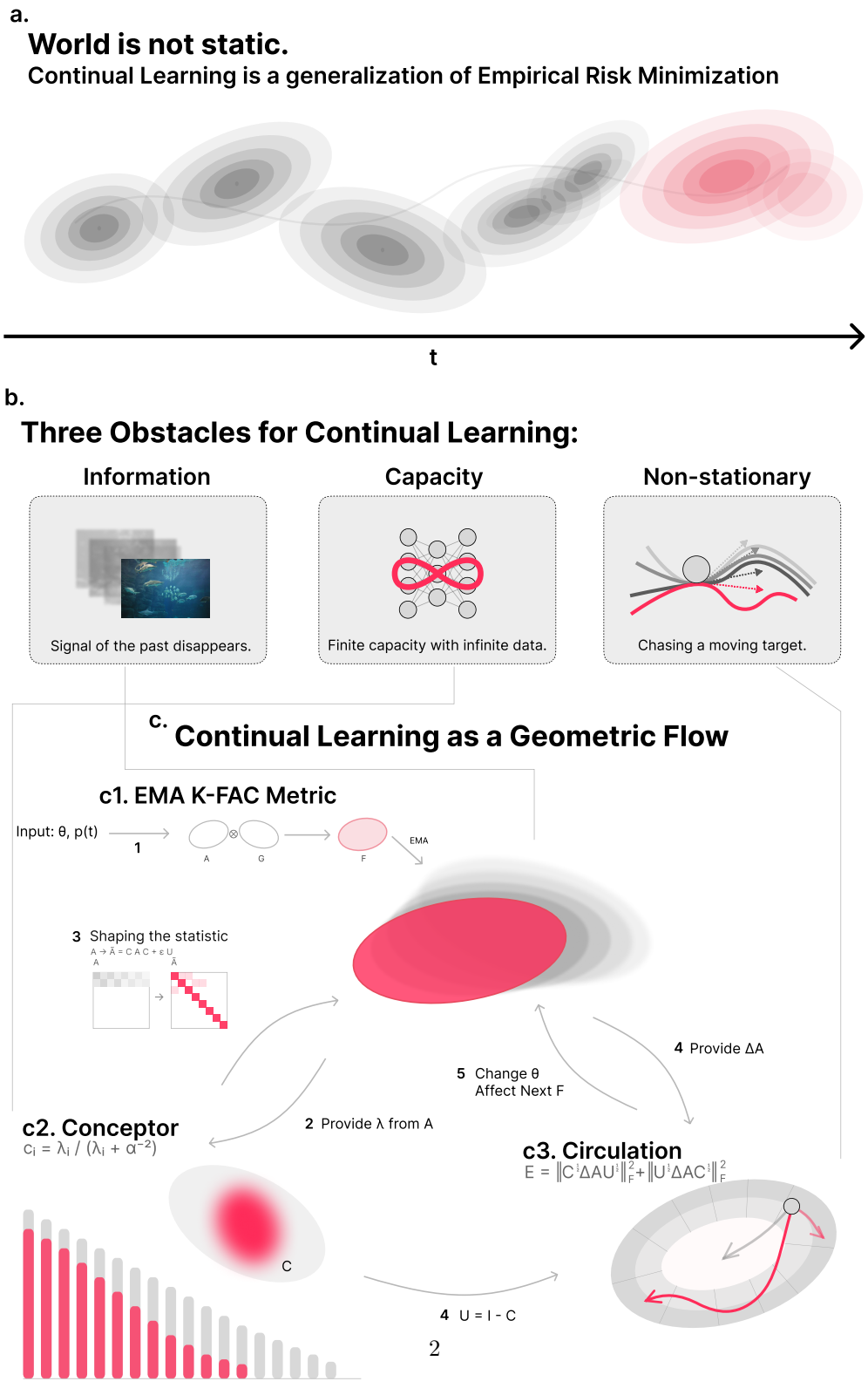

Fig. 1 Overview. Continual learning is the general learning regime because real data streams evolve over time and are non IID rather than stationary i.i.d. **a:** A non stationary stream with a drifting data distribution. **b:** We target three obstacles of continual learning: the information obstacle where past training signal vanishes, the capacity obstacle where finite parameters must absorb infinite data stream, and the non stationary obstacle where the learning target shifts over time. **c:** We treat learning as a coupled evolution of parameters and their online geometry in such regime. c1: EMA K-FAC maintains a streaming estimate of the Fisher geometry as a block structured metric. c2: Conceptors split the running activation statistics into persistent and weakly supported directions, producing a complement subspace that tracks transient usage. c3: A leakage potential measures instantaneous activation energy in the complement subspace and defines a conservative drive. A circulation term rotates the Fisher preconditioned leakage gradient through a skew generator, producing a tangent update that redistributes representation while preserving leakage to first order.

Continual learning (CL) asks how a learner can remain competent as its input distribution changes over time (Wang et al. 2024; Gunasekara et al. 2023). In the real world, inputs arrive as a non-IID stream $x_t \sim \mathcal{P}_t$ with $\dot{\mathcal{P}}_t \neq 0$. Learning is therefore best cast as a controlled dynamical system whose objective is to sustain competence under drift subject to finite compute and memory. Continual learning arises across many domains: financial markets, recommender systems, robotics, model fine-tuning, and biological learning all fit this setting, where the stream is time-ordered and the underlying distribution evolves.

Classical ERM arises only in the stationary limit: when $\mathcal{P}_t \equiv \mathcal{P}$ over the update horizon and the objective is time-invariant, Φ reduces to descent on the fixed risk $L(\theta) = \mathbb{E}_{x\sim\mathcal{P}}[\ell(\theta; x)]$. That stationarity is, at best, an approximation, valid only when drift is negligible relative to adaptation timescales. One can enlarge datasets and use replay buffers to reduce temporal correlations so that a non-IID, non-stationary problem more closely resembles IID training; this remains an approximation to stationarity, not its recovery. A common consequence of this simplification is out-of-distribution mismatch. Modern methods mitigate it with regularization, and it is often treated as an unavoidable cost.

However, non-IID structure is not a failure mode to be simplified away but a natural regime to exploit, as it reveals temporal and causal regularities that enable abstraction, reuse, and continual adaptation. The world is never static, so learning must be inherently adaptive.

Therefore, rather than viewing continual learning as a sub-problem of modern machine learning, we argue that it is the general case and that offline optimization is the stationary limit. As a result, an effective learning dynamic should be compatible with this world view.

Three obstacles to continual learning.

Viewed as a long-horizon dynamical problem, continual learning faces three distinct obstacles.

First is an *information obstacle*: when past training signals vanish from the stream, the corresponding supervision is no longer available unless the learner maintains memory, replay, or an explicitly recoverable sufficient statistic. Regularization provides only a lossy summary, and isolation can reduce interference via routing, but neither is equivalent to keeping the original signal accessible.

Second is a *capacity obstacle*: a system of finite complexity must absorb an effectively unbounded history, which forces compression, abstraction, and selective forgetting. What should be preserved as reusable structure and what should be discarded as transient detail is ultimately a question of representation and objective, not merely an optimization trick.

Third is a *non-stationarity obstacle*: in an ongoing stream, the goal is not to converge and stop, but to continually track a moving target under drift. Standard optimization dynamics tends to drive parameters toward a local optimum of the current distribution, even when that optimum will soon become obsolete; as a result, the instantaneous

gradient can be misaligned with long-horizon performance. Many CL methods compensate for this equilibrium bias rather than replacing the underlying dynamics, leaving a persistent structural tension between stability and plasticity.

Continual learning is currently framed as an auxiliary.

Despite major progress, the dominant framing remains optimization-centric: we assume a loss, apply gradient descent on the current data, and then add auxiliary mechanisms to avoid drifting too far from the past minimum. This has produced three canonical families: regularization(Kirkpatrick et al. 2017; Zenke et al. 2017; Aljundi et al. 2018; Li and Hoiem 2016), replay(Rebuffi et al. 2017; Lopez-Paz and Ranzato 2017; Buzzega et al. 2020; Bellitto et al. 2024), and parameter isolation(Mallya and Lazebnik 2018; Mallya et al. 2018; Serra et al. 2018; Gao et al. 2024). Together they largely define the modern CL toolbox. A detailed summary of prior work can be found in Appendix A. However, they all suffer from their structural limit, making current methods rely on episodic benchmarks, where the stream is simplified into task boundaries and the learner is evaluated in a setting that can mask the core difficulty of true online adaptation.

A structural limitation: single-point commitment is unavoidable in pure optimization.

Under purely gradient-descent dynamics, updates are dominated by monotone descent on the instantaneous objective, so indefinite non-stationarity forces a tradeoff between overwriting past structure and constraining motion into rigidity. This limitation is inseparable from the capacity and non-stationarity obstacle. In a stream, the learner must reallocate finite degrees of freedom, and there typically exist many parameterizations with similar short-horizon loss but different long-horizon reuse profiles. Descent does not search over these near-equivalent configurations: it concentrates the trajectory into a single basin, fixing a single local geometry from which all subsequent gradients are evaluated. As a result, both interference and capacity allocation become consequences of the particular basin reached early, rather than quantities the system can continuously redistribute under drift.

Geometry as the state.

In this paper, we treat continual learning as controlled dynamics in an information space, where the evolving state is an online estimate of Fisher geometry maintained through exponential moving averages of Kronecker-factored statistics. Instead of relying on explicit replay to keep old signals alive, the geometry itself carries persistent information about sensitivity and can be updated continuously from incoming data without storing exemplars. Crucially, the same geometry also defines a split in representation space between directions that are repeatedly supported by the stream and directions that remain weakly supported and brittle. We then feed this split back into the geometry construction so that shared directions remain accessible while transient drift directions are continuously damped. In this way, the metric is not just a preconditioner; it is also a compact, continuously refreshed memory of what the stream persistently supports.

Capacity functional and circulation.

Purely gradient descent tends to lock the learner into a single local configuration, making capacity allocation a one-shot byproduct of the early trajectory. We make this pressure explicit by measuring how much the current batch uses directions identified as nonreusable by the persistent-versus-transient split introduced by the geometry. To turn capacity allocation from a deterministic early outcome into a quantity that can be continually rebalanced, the dynamics needs a component that promotes lateral motion across near-equivalent configurations under drift, rather than only descending along the instantaneous objective. We therefore introduce a circulation term that is computed in the same evolving Fisher geometry and is constructed to preserve the stress energy from the time-varying geometry. This circulation component complements the descent step by enabling controlled exploration that keeps capacity reallocation active over time.

Coupled update under streaming.

Under drift, the update rule must recalibrate to the current stream while avoiding premature lock-in. At each step, the incoming batch refreshes the online geometry, the persistent-versus-transient split is recomputed from the same statistics, and the geometry is reshaped accordingly. Both the gradient descent direction and the circulation term are then computed coherently with respect to this same updated geometry. Because they share a single evolving metric, the circulation remains well-scaled relative to the current notion of stress energy, and the gradient descent remains measured in a coordinate system that reflects the present stream.

Results

We evaluate continual learning on CIFAR-10 and CLEAR-100 under a *true* non-episodic data stream: training examples arrive sequentially with no explicit task boundaries and no task identifier, and the learner is updated online throughout the stream. We consider three stream variants:

- **S1 (prior drift).** Class sampling frequencies vary continuously over time and may revisit earlier modes, inducing a drifting label prior without any episode segmentation.
- **S2 (prior and context drift).** In addition to drifting class frequencies as in S1, the input distribution changes over time via temporally varying augmentations (e.g., rotation, blur, jitter, noise), creating recurring covariate shift under a continuous stream.
- **S3 (prior and label drift).** The stream combines drifting class frequencies with time-varying label semantics, i.e., the mapping from data to observed labels changes over time.

Importantly, the stream does not enforce uniform class exposure: some classes may appear much more or much less frequently than others depending on the stream realization. As a result, overall accuracy is entangled with class appearance probability and is not directly comparable to standard episodic protocols or offline ERM. We therefore

report complementary metrics that reflect stream dynamics: online and rolling accuracy (Fig. 2), stability–plasticity trade-offs for EMA K-FAC, shared-embedding UMAP projections of learned representations to assess feature organization and the effect of conceptors (Fig. 4), and circulation diagnostics (ORI rate and time-warpness). Additional experiments, including learning with self-supervised features on CLEAR-100, are reported in Sec. 3.

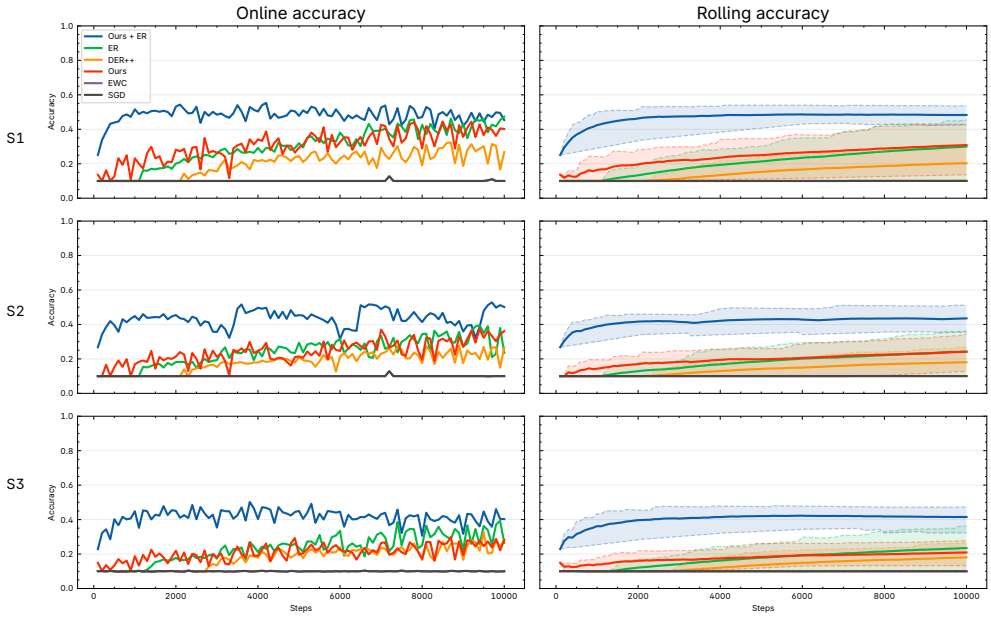

Fig. 2 Test accuracy on the full CIFAR-10 test set measured after each online update. Left: instantaneous online accuracy. Right: rolling accuracy computed as a sliding-window average over recent updates. Across all three streaming protocols, our method learns effectively without a replay buffer, while buffer-free baselines show no learning activity. When an external replay buffer is available, Ours+ER achieves consistently higher accuracy and faster improvement than standard replay-based baselines (ER, DER++). Shaded bands indicate variability across runs.

Figure 2 reports test accuracy on the full CIFAR-10 test set after each online update, along with a rolling (sliding-window) average that suppresses high-frequency fluctuations. Across S1–S3, buffer-free baselines such as SGD and EWC show almost zero improvement under the continuous stream. Replay-based baselines (ER, DER++) improve robustness by reintroducing past signal, but require external memory. In contrast, our method learns effectively without replay across all three streams. When replay is available, Ours+ER achieves the strongest performance in every stream, with particularly large gains in early and mid-stream stages.

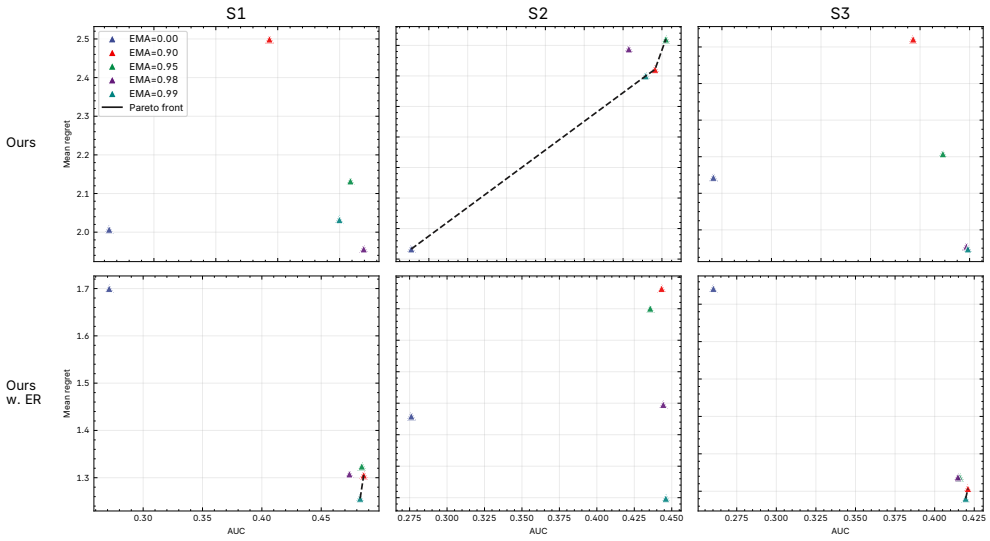

Fig. 3 Stability vs. plasticity across EMA settings for K-FAC. Each point corresponds to an EMA coefficient used to smooth the K-FAC curvature statistics. *Stability* is measured by the area under the curve (AUC) of the **online test accuracy** trajectory, i.e., the time-average of test accuracy evaluated after each online update over the full stream. *Plasticity* is measured by the **mean online regret** over the stream, computed at each step as the gap between the learner's rolling-window loss and an oracle *opt-loss* baseline on the same window, and then averaged across time (lower is better).

To quantify the stability–plasticity trade-off induced by curvature smoothing, we sweep the exponential moving-average (EMA) coefficient used in the K-FAC statistics and summarize outcomes in a two-objective space (Fig. 3). We use the AUC of online accuracy over time as a metric for *stability* and the mean regret over time as a metric for *plasticity* (lower is better). Across all three stream protocols, EMA$=0$ yields the poorest configurations, and is dominated by EMA-smoothed settings. With EMA enabled, the non-dominated set forms a stream-dependent Pareto frontier: S2 exhibits the clearest trade-off between higher stability and lower regret, whereas S1 and S3 admit a broader range of near-optimal EMA values. When an external replay buffer is available, the frontier shifts toward more favorable outcomes and becomes less sensitive to the exact EMA choice, indicating improved robustness to this hyperparameter.

Fig. 4 Feature geometry over time under Stream S1. UMAP projection of penultimate-layer features for a fixed set of CIFAR-10 test images at multiple stream steps. A single UMAP embedding is fit once using features pooled across all methods and time points within Stream S1, then reused for every panel so coordinates are comparable. Rows show methods and columns show stream steps; points are colored by ground-truth class. Several baselines contract onto thin manifolds and intermittently fragment into small disconnected islands as the stream progresses, indicating reduced feature diversity and less stable organization under drift. Our method maintains a broader, more coherently structured feature space over time, and adding replay further stabilizes the global geometry. For comparision with other stream protocols, see results in 3.2

To probe representation quality beyond accuracy, we visualize feature embeddings over time using a shared UMAP coordinate system (Fig. 4). Underperforming baselines often compress into thin manifolds or collapse into tight regions as the stream progresses, with limited class separation. In contrast, our method maintains a broader and more coherently organized feature space over time, and adding replay further stabilizes the global geometry. The ablation in Fig. 4 also isolates the role of conceptors: removing conceptors yields a near-collapse in representation, consistent with the drop in robustness observed in the no-replay setting.

Discussion

These results emphasize that continual learning under a *true* non-episodic stream differs qualitatively from episodic evaluation. When class exposure is non-uniform and drift is continuous, performance is inherently time-dependent: the central question is whether a method remains functional as the stream shifts, revisits, and occasionally changes abruptly, not merely whether it produces a strong terminal model. This motivates treating evaluation as a dynamical problem and prioritizing metrics that capture degradation, recovery, and stability–plasticity trade-offs.

The accuracy trajectories illustrate this dynamical view. Instantaneous accuracy reflects the immediate consequence of each update, while rolling accuracy exposes sustained failures and recovery that are otherwise obscured by noise. Replay improves robustness by preserving training signal, but it externalizes the information obstacle into a memory budget.

The EMA sweep clarifies a core dynamical issue in stream continual learning: second-order updates must trade off variance reduction against responsiveness to a moving objective. Interpreted as temporal regularization of the Fisher geometry, EMA turns a high-variance instantaneous curvature estimate into a more reliable control signal for online preconditioning. The fact that EMA$=0$ is consistently dominated across streams is therefore not a minor tuning artifact but evidence that unsmoothed curvature is too noisy and time-correlated to support stable long-horizon adaptation.

The stream-dependent shape of the Pareto frontier further distinguishes when stability–plasticity tension is intrinsic versus when it is largely avoidable. Under covariate drift (S2), the geometry itself changes rapidly, so stronger smoothing improves stability but introduces lag, producing a sharper trade-off with plasticity. Under S1 and S3, the broader near-optimal region suggests that once temporal averaging is present, performance is comparatively robust to the precise smoothing level.

Finally, replay shifts the frontier toward the favorable region and reduces sensitivity to EMA, consistent with complementarity between memory and the update rule. Replay preserves past training signal, while EMA-stabilized curvature determines how efficiently new signal is integrated, so improved updates can increase replay *utility* rather than merely reducing replay *necessity*.

The representation diagnostics provide a mechanistic lens on these behaviors. In the shared UMAP space, collapsing or highly fragmented feature geometry aligns with brittleness under drift and revisitation, since reusable structure is not retained in a stable form. By contrast, the broader, persistent organization maintained by our

method is consistent with robustness under continual change, indicating that stability in stream learning is tightly linked to preserving an expressive but well-structured feature space. The conceptor ablation strengthens this interpretation by showing that removing conceptors induces representational collapse, highlighting their role in preserving reusable structure in the absence of stored data.

Overall, the coupling between performance curves and feature geometry argues for evaluation protocols that explicitly reflect non-episodic stream realities. Metrics that quantify sustained degradation, recovery speed, and stability–plasticity trade-offs are essential complements to final accuracy in this regime, and representation diagnostics can serve as early indicators of failure modes such as progressive collapse.

1 Methods

1.1 Problem setup: continual learning as the general regime

We study learning under a time-indexed non-IID stream $(x_t, y_t) \sim \mathcal{P}_t$ with drift $\dot{\mathcal{P}}_t \neq 0$. The learner updates parameters $\theta_t \in \Theta \subset \mathbb{R}^n$ online,

$$\theta_{t+1} = \Phi(\theta_t; x_t, y_t, \mathcal{S}_t), \tag{1}$$

where $\mathcal{S}_t$ denotes internal state including online statistics and compact memory.

Offline empirical risk minimization is recovered under stationarity. If $\mathcal{P}_t \equiv \mathcal{P}$ is stationary over the adaptation horizon and the update rule is chosen as stochastic descent on

$$L(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\, \ell(\theta; x, y), \tag{2}$$

the learning dynamics reduces to standard offline training.

We state the inclusion relation through a time-scale ratio. Let T_{drift} be a characteristic time such that $\mathcal{P}_t$ changes by $O(1)$ over T_{drift}. Let T_{learn} be a characteristic time such that θ changes by $O(1)$ over T_{learn}. Define $\varepsilon \coloneqq T_{\text{learn}}/T_{\text{drift}}$. The regime $\varepsilon \to 0$ is quasi-stationary and yields ERM-like behavior. The regime $\varepsilon = O(1)$ is continual learning. Thus offline i.i.d. learning is a singular limit of continual learning as $\varepsilon \to 0$.

1.2 System Overview

Our learning dynamic is defined as

$$\dot{\theta} = -\operatorname{grad}_{g_t} L_t(\theta) + \kappa\, \Omega^{z}_{g_t}(\theta)\, \operatorname{grad}_{g_t} E_t(\theta), \tag{3}$$

where $L_t(\theta)$ is the instantaneous loss on the current stream, $E_t(\theta)$ is the kinematic energy of the activation-side metric residual, and g_t is an online Fisher geometry estimated from the stream. The Riemannian gradient under g_t is defined as

$$\operatorname{grad}_{g_t} f(\theta) \coloneqq g_t(\theta)^{-1} \nabla_\theta f(\theta), \qquad \langle u, v\rangle_{g_t} \coloneqq u^\top g_t(\theta) v, \tag{4}$$

and $\kappa \geq 0$ is a constant that scales the circulation amplitude in parameter space. The operator $\Omega^{z(t)}_{g_t}(\theta)$ is g_t-skew (Axiom 11) and is parameterized by an internal state

$z(t)$ (Sec. 1.4); thus $\langle v, \Omega^z_{g_t}(\theta)v\rangle_{g_t} = 0$ and the circulation preserves E_t to first order (Theorem 4).

1.2.1 Online geometry and memory

We approximate g_t by Kronecker-factored blocks $\{F_{\ell,t}\}_\ell$ with EMA-updated factors $A_{\ell,t}$ and $G_{\ell,t}$ (Sec. 1.3.2):

$$F_{\ell,t} \approx G_{\ell,t} \otimes A_{\ell,t}, \qquad A_{\ell,t} = \beta A_{\ell,t-1} + (1-\beta)\widehat{A}_{\ell,t}, \qquad G_{\ell,t} = \beta G_{\ell,t-1} + (1-\beta)\widehat{G}_{\ell,t}, \tag{5}$$

where $\widehat{A}_{\ell,t}$ and $\widehat{G}_{\ell,t}$ are current-minibatch estimates, and $\beta \in (0,1)$ is the EMA coefficient. From $A_{\ell,t}$ we compute conceptors and complements,

$$C_{\ell,t} = A_{\ell,t}(A_{\ell,t} + \alpha^{-2} I)^{-1}, \qquad U_{\ell,t} = I - C_{\ell,t}, \tag{6}$$

and feed the split back into the geometry by shaping the activation factor,

$$\widetilde{A}_{\ell,t} = C_{\ell,t} A_{\ell,t} C_{\ell,t} + \varepsilon_A U_{\ell,t}, \qquad \widetilde{F}_{\ell,t} \approx G_{\ell,t} \otimes \widetilde{A}_{\ell,t}, \tag{7}$$

with aperture $\alpha > 0$ and floor $\varepsilon_A > 0$. Damped inverses are computed as in (22).

1.2.2 Energy and circulation

Here we define the kinematic energy of the online metric as a channel-coupled quadratic form of the activation-covariance residual. Let $\widehat{A}_\ell(\theta)$ denote the instantaneous minibatch activation covariance at parameters θ, and define the residual

$$\Delta A_{\ell,t}(\theta) \coloneqq \widehat{A}_\ell(\theta) - A_{\ell,t}. \tag{8}$$

We then define

$$E_t(\theta) = \sum_\ell \Big[\mathrm{tr}\big(\Delta A_{\ell,t}(\theta)\, C_{\ell,t}\, \Delta A_{\ell,t}(\theta)\, U_{\ell,t}\big) + \mathrm{tr}\big(\Delta A_{\ell,t}(\theta)\, U_{\ell,t}\, \Delta A_{\ell,t}(\theta)\, C_{\ell,t}\big)\Big], \tag{9}$$

where $C_{\ell,t}$ and $U_{\ell,t}$ are treated as fixed within step t.

The circulation term is evaluated in the same evolving geometry and is implemented by mapping covectors to whitened coordinates, applying a skew action, and mapping back to parameter space (Algorithm 1). The skew generator $z(t)$ is updated by an endogenous chaotic map (Algorithm 2) with internal gain ω_z and internal step size Δt_z.

1.2.3 Learning dynamics discretization

We discretize (28) by Lie–Trotter splitting with step size Δt:

$$\theta_{t+\frac{1}{2}} = \theta_t - \Delta t\ \mathrm{grad}_{g_t} L_t(\theta_t), \tag{10}$$

$$\theta_{t+1} = \theta_{t+\frac{1}{2}} + \Delta t\, \kappa\, \Omega^z_{g_t}(\theta_{t+\frac{1}{2}})\, \mathrm{grad}_{g_t} E_t(\theta_{t+\frac{1}{2}}), \tag{11}$$

Algorithm 1 Learning Dynamic (one online update)

Input: Parameters θ; minibatch (x, y); EMA factors $\{A_\ell, G_\ell\}_\ell$; internal state z.
Input: Hyperparameters: EMA rate β, damping λ, conceptor (α, ε_A), step size Δt, circulation amplitude κ, max-update bound τ.

Update geometry (online Fisher / K-FAC + conceptor shaping)

Compute minibatch statistics $\{\widehat{A}_\ell, \widehat{G}_\ell\}_\ell$ and update EMA factors using (21)
for all layers ℓ **do**
  $C_\ell \leftarrow A_\ell(A_\ell + \alpha^{-2} I)^{-1}$; $U_\ell \leftarrow I - C_\ell$ ▷ Eq. (23)
  $\widetilde{A}_\ell \leftarrow C_\ell A_\ell C_\ell + \varepsilon_A U_\ell$ ▷ Eq. (24)
end for
Define shaped metric blocks $\widetilde{F}_\ell \approx G_\ell \otimes \widetilde{A}_\ell$ with damping λ

Whiten / Dewhiten in the shaped geometry

Let $\widetilde{F}$ denote the block-diagonal K-FAC approximation built from $\{\widetilde{F}_\ell\}_\ell$.
Define $\mathrm{Whiten}(g; \widetilde{F}, \lambda) := \widetilde{F}^{-1/2} g$ (covector whitening),
  and $\mathrm{Dewhiten}(u; \widetilde{F}, \lambda) := \widetilde{F}^{-1/2} u$ (map back to a parameter-space vector).

(1) Gradient descent in the shaped geometry

$L \leftarrow \ell(\theta; x, y)$
$g_L \leftarrow \nabla_\theta L$
$\Delta\theta_{\mathrm{gd}} \leftarrow -\Delta t\, \widetilde{F}^{-1} g_L$ ▷ Eq. (17) with $\widetilde{F}$
$\Delta\theta_{\mathrm{gd}} \leftarrow \mathrm{GlobalRescale}(\Delta\theta_{\mathrm{gd}}; \tau)$
$\theta \leftarrow \theta + \Delta\theta_{\mathrm{gd}}$

(2) Circulation from kinematic energy

Compute instantaneous activation covariances $\{\widehat{A}_\ell(\theta; x)\}_\ell$ on the minibatch
for all layers ℓ **do**
  $\Delta A_\ell \leftarrow \widehat{A}_\ell - A_\ell$ ▷ Residual, Eq. (8)
end for
$E \leftarrow \sum_\ell \left(\mathrm{tr}(\Delta A_\ell C_\ell \Delta A_\ell U_\ell) + \mathrm{tr}(\Delta A_\ell U_\ell \Delta A_\ell C_\ell)\right)$ ▷ Eq. (25), C_ℓ, U_ℓ fixed within step
$g_E \leftarrow \nabla_\theta E$
$\xi \leftarrow \mathrm{Whiten}(g_E; \widetilde{F}, \lambda)$ ▷ $\xi = \widetilde{F}^{-1/2} g_I$
$z \leftarrow \mathrm{ChaosStep}(z)$ ▷ Alg. 2
$u \leftarrow \mathrm{Skew}(z)\, \xi$ ▷ $u = \Omega^z \xi$ in whitened coordinates
$\Delta\theta_{\mathrm{circ}} \leftarrow +\Delta t\, \kappa\, \mathrm{Dewhiten}(u; \widetilde{F}, \lambda)$
$\Delta\theta_{\mathrm{circ}} \leftarrow \mathrm{GlobalRescale}(\Delta\theta_{\mathrm{circ}}; \tau)$
$\theta \leftarrow \theta + \Delta\theta_{\mathrm{circ}}$
return θ, updated $\{A_\ell, G_\ell\}_\ell$, updated z

All K-FAC inverses use damped factors (22). We apply a global update rescale that bounds the maximum absolute parameter update and prevents rare spikes in minibatch statistics from producing unstable steps.

1.3 Derivation from first principles

The derivation follows a forced chain. Axiom 1 asserts that, near previously good solutions, the relevant update-induced damage is second-order. Axiom 2 fixes "smallness" to be local distributional change measured by predictive KL. Together they imply that the step-controlling quadratic form is the Hessian of local KL, which uniquely equals the Fisher information metric (Theorem 1, using Lemma 1). With this metric, Axiom 3 yields natural-gradient descent (Theorem 2). To realize this online in deep networks with layer-local computation, we approximate the Fisher blocks by Kronecker factors, leading to EMA K-FAC (Theorem 3). Finally, under indefinite non-stationarity and finite capacity, we formalize an instantaneous kinematic energy functional E_t and require a conservative redistribution term that preserves E_t to first order; this forces the circulation form in (28) (Theorem 4). An explicit parametrization of the skew generator ("endogenous chaos") is given separately in Sec. 1.4.

1.3.1 Second-order dominance and metric necessity

Axiom 1 (Quadratic damage near past optima)**.** *For a smooth loss $L(\theta)$, when θ is close to a minimizer of a past regime, the first-order term is small and the leading-order change induced by an update $\delta\theta$ is governed by the quadratic term in the local expansion.*

By Taylor expansion,

$$L(\theta + \delta\theta) = L(\theta) + \nabla L(\theta)^\top \delta\theta + \tfrac{1}{2}\,\delta\theta^\top H(\theta)\,\delta\theta + o(\|\delta\theta\|^2), \tag{12}$$

where $H(\theta) = \nabla^2 L(\theta)$. Under Axiom 1, any step-control principle that ignores a quadratic form risks uncontrolled damage near past optima.

Axiom 2 (Small step means small distributional change)**.** *A learning step is small when it induces small change in the model distribution, measured locally by Kullback–Leibler divergence in the predictive family.*

Lemma 1 (Local KL induces Fisher metric)**.** *For a probabilistic model $q_\theta(y|x)$ and a data distribution $\mathcal{P}_t$, the local expansion of KL satisfies*

$$\mathrm{KL}(q_{\theta+\delta\theta} \,\|\, q_\theta) = \tfrac{1}{2}\,\delta\theta^\top g_t(\theta)\,\delta\theta + o(\|\delta\theta\|^2), \tag{13}$$

with Fisher information metric

$$g_t(\theta) \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{P}_t}\Big[\nabla_\theta \log q_\theta(y|x)\,\nabla_\theta \log q_\theta(y|x)^\top\Big]. \tag{14}$$

Proof Axiom 2 fixes local KL as the operational step-size notion in the predictive family. Fix x and expand $\mathrm{KL}(q_{\theta+\delta\theta}(\cdot|x) \,\|\, q_\theta(\cdot|x))$ around $\delta\theta = 0$. Let $s_\theta(y|x) \coloneqq \nabla_\theta \log q_\theta(y|x)$. The first-order term vanishes since $\mathbb{E}_{y\sim q_\theta(\cdot|x)}[s_\theta(y|x)] = 0$. The Hessian at $\delta\theta = 0$ equals $\mathbb{E}_{y\sim q_\theta(\cdot|x)}[s_\theta(y|x)s_\theta(y|x)^\top]$. Taking expectation over $(x, y) \sim \mathcal{P}_t$ yields (14), and the second-order expansion yields (13). □

Theorem 1 (Fisher necessity from second-order damage and local KL)**.** *Assume Axiom 1 and Axiom 2. If step size is controlled by a second-order quadratic form consistent with local predictive KL, then the controlling metric is uniquely*

$$g_t(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{P}_t}\Big[\nabla_\theta \log q_\theta(y|x)\,\nabla_\theta \log q_\theta(y|x)^\top\Big], \tag{15}$$

i.e., the Fisher information metric.

Proof Axiom 1 requires that, in the regime relevant to preserving previously good solutions, update control is governed by a quadratic form $\frac{1}{2}\,\delta\theta^\top M_t(\theta)\,\delta\theta$ to leading order. Axiom 2 specifies that the operational notion of "small" update is local predictive KL, so the quadratic form must match the second-order expansion of $\mathrm{KL}(q_{\theta+\delta\theta}\|q_\theta)$ at $\delta\theta = 0$. By Lemma 1, this

expansion is $\mathrm{KL}(q_{\theta+\delta\theta}\|q_\theta) = \frac{1}{2}\,\delta\theta^\top g_t(\theta)\,\delta\theta + o(\|\delta\theta\|^2)$. Therefore $M_t(\theta)$ must equal $g_t(\theta)$, and the metric is uniquely the Fisher information. □

Axiom 3 (Steepest local loss decrease under a KL trust region)**.** *At each step t, the descent component achieves maximal first-order decrease of the instantaneous loss L_t subject to a local KL budget.*

A minimal local model is

$$\min_{\delta\theta}\ \nabla L_t(\theta)^\top \delta\theta \quad \text{subject to} \quad \tfrac{1}{2}\,\delta\theta^\top g_t(\theta)\,\delta\theta \le \eta. \tag{16}$$

Theorem 2 (Natural-gradient descent direction)**.** *Under Axiom 3 and Theorem 1, the descent direction is*

$$\dot\theta_{\mathrm{gd}} = -\operatorname{grad}_{g_t} L_t(\theta), \qquad \operatorname{grad}_{g_t} L_t(\theta) \coloneqq g_t(\theta)^{-1}\nabla L_t(\theta). \tag{17}$$

Proof By Theorem 1, the KL trust region induces the quadratic constraint with matrix $g_t(\theta)$. Form the Lagrangian $\mathcal{L}(\delta\theta,\mu) = \nabla L_t(\theta)^\top\delta\theta + \mu(\frac{1}{2}\,\delta\theta^\top g_t(\theta)\delta\theta - \eta)$, $\mu \ge 0$. Stationarity gives $\nabla L_t(\theta) + \mu g_t(\theta)\delta\theta = 0$, hence $\delta\theta \propto -g_t(\theta)^{-1}\nabla L_t(\theta)$. Therefore the steepest feasible decrease is achieved along $-\operatorname{grad}_{g_t} L_t(\theta)$. □

1.3.2 Tractable online Fisher approximation and EMA K-FAC

Theorem 2 provides the *form* of the descent direction in terms of the Fisher metric g_t. To make Axiom 1 actionable in deep networks, we must maintain and invert an online approximation to this *second-order* object.

Axiom 4 (Tractable online second-order metric)**.** *The learner maintains an online, invertible approximation $\widetilde{g}_t(\theta)$ to the Fisher metric $g_t(\theta)$ that: (i) updates from current-minibatch statistics without storing exemplars, (ii) is representable and invertible with layer-local memory and compute, and (iii) is block-diagonal across layers, $\widetilde{g}_t = \mathrm{blkdiag}(\widetilde{F}_{1,t},\dots,\widetilde{F}_{L,t})$.*

Lemma 2 (Layerwise Fisher block as a Kronecker second moment)**.** *Consider a layer ℓ with pre-activation $s_\ell = W_\ell a_{\ell-1}$ (bias handled by the usual augmentation). Let $\delta_\ell \coloneqq \nabla_{s_\ell}\log q_\theta(y|x)$ denote the backpropagated score with respect to s_ℓ. Then the Fisher block for $\mathrm{vec}(W_\ell)$ is*

$$F_{\ell,t} = \mathbb{E}_{(x,y)\sim\mathcal{P}_t}\left[\left(a_{\ell-1}a_{\ell-1}^\top\right)\otimes\left(\delta_\ell\delta_\ell^\top\right)\right]. \tag{18}$$

Proof By definition of Fisher (Lemma 1), a layerwise block equals the second moment of the layerwise score. For the layer parameters, $\nabla_{W_\ell}\log q_\theta(y|x) = \delta_\ell a_{\ell-1}^\top$. Using $\mathrm{vec}(uv^\top) = v\otimes u$ yields $\nabla_{\mathrm{vec}(W_\ell)}\log q_\theta(y|x) = a_{\ell-1}\otimes\delta_\ell$. Therefore

$$F_{\ell,t} = \mathbb{E}\Big[(a_{\ell-1}\otimes\delta_\ell)(a_{\ell-1}\otimes\delta_\ell)^\top\Big] = \mathbb{E}\Big[(a_{\ell-1}a_{\ell-1}^\top)\otimes(\delta_\ell\delta_\ell^\top)\Big],$$

which is (18). □

Axiom 5 (K-FAC factorization)**.** *For each layer ℓ, we approximate the mixed second moment in (18) by*

$$\mathbb{E}\big[(a_{\ell-1}a_{\ell-1}^\top) \otimes (\delta_\ell \delta_\ell^\top)\big] \approx \mathbb{E}[a_{\ell-1}a_{\ell-1}^\top] \otimes \mathbb{E}[\delta_\ell \delta_\ell^\top]. \tag{19}$$

Theorem 3 (K-FAC block form of the online Fisher geometry)**.** *Under Axioms 1 and 4, Lemma 2, and Axiom 5, a tractable layerwise approximation to g_t is given by K-FAC blocks*

$$F_{\ell,t} \approx G_{\ell,t} \otimes A_{\ell,t}, \qquad A_{\ell,t} \coloneqq \mathbb{E}[a_{\ell-1}a_{\ell-1}^\top], \quad G_{\ell,t} \coloneqq \mathbb{E}[\delta_\ell \delta_\ell^\top], \tag{20}$$

where expectations are taken under $\mathcal{P}_t$.

Proof Axiom 1 forces the use of a second-order metric, and Axiom 4 forces a layerwise, invertible representation of that metric. By Lemma 2, the exact layerwise Fisher block equals the mixed second moment $\mathbb{E}[(a_{\ell-1}a_{\ell-1}^\top) \otimes (\delta_\ell \delta_\ell^\top)]$. Applying Axiom 5 yields $F_{\ell,t} \approx \mathbb{E}[a_{\ell-1}a_{\ell-1}^\top] \otimes \mathbb{E}[\delta_\ell \delta_\ell^\top]$. Defining these marginals as $A_{\ell,t}$ and $G_{\ell,t}$ gives (20). □

Instantaneous minibatch estimates are $\widehat{A}_{\ell,t}$ and $\widehat{G}_{\ell,t}$. We update the factors by exponential weighting,

$$A_{\ell,t} = \beta A_{\ell,t-1} + (1-\beta)\widehat{A}_{\ell,t}, \qquad G_{\ell,t} = \beta G_{\ell,t-1} + (1-\beta)\widehat{G}_{\ell,t}, \tag{21}$$

and compute damped inverses with Tikhonov regularization,

$$A_{\ell,t}^{-1} \approx (A_{\ell,t} + \lambda I)^{-1}, \qquad G_{\ell,t}^{-1} \approx (G_{\ell,t} + \lambda I)^{-1}. \tag{22}$$

1.3.3 Past information constraints and compact memory in geometry

Axiom 6 (Recoverable past information constraint)**.** *A continual learner must retain a recoverable statistic of past sensitivity that persists beyond the lifespan of any single minibatch, without storing exemplars.*

Axiom 7 (Persistent-subspace separation)**.** *The memory statistic must define a bounded operator that separates persistent directions from weakly supported directions in representation space.*

Let $A_{\ell,t} \succeq 0$ be an EMA activation covariance for layer ℓ. A bounded shrinkage operator satisfying Axiom 7 is the conceptor

$$C_{\ell,t} = A_{\ell,t}\big(A_{\ell,t} + \alpha^{-2} I\big)^{-1}, \qquad U_{\ell,t} = I - C_{\ell,t}, \tag{23}$$

with aperture $\alpha > 0$.

Axiom 8 (Geometry-coupled memory)**.** *The persistent-subspace split must enter the learning dynamics through the same geometry that defines gradient descent.*

We implement Axiom 8 by shaping the activation factor as

$$\widetilde{A}_{\ell,t} = C_{\ell,t}\, A_{\ell,t}\, C_{\ell,t} + \varepsilon_A\, U_{\ell,t}, \tag{24}$$

with $\varepsilon_A > 0$.

1.3.4 Capacity pressure, kinematic energy, and circulation

Axiom 9 (Finite capacity induces capacity pressure)**.** *Under indefinite non-stationarity, finite model capacity induces a pressure to allocate degrees of freedom toward reusable structure and away from brittle transient directions.*

Let $\widehat{A}_\ell(\theta)$ denote the instantaneous activation covariance on the current minibatch under parameters θ.

Definition 1 (Kinematic energy functional)**.** *At step t we define the instantaneous functional*

$$E_t(\theta) = \sum_\ell \Big[\mathrm{tr}\big(\Delta A_{\ell,t}(\theta)\, C_{\ell,t}\, \Delta A_{\ell,t}(\theta)\, U_{\ell,t}\big) + \mathrm{tr}\big(\Delta A_{\ell,t}(\theta)\, U_{\ell,t}\, \Delta A_{\ell,t}(\theta)\, C_{\ell,t}\big)\Big], \tag{25}$$

where $C_{\ell,t}$ and $U_{\ell,t}$ are treated as fixed within the step and $\Delta A_{\ell,t}(\theta) \coloneqq \widehat{A}_\ell(\theta) - A_{\ell,t}$.

Axiom 10 (Conservative redistribution under capacity pressure)**.** *The learning dynamics contains a circulation component driven by the capacity pressure that preserves the instantaneous functional E_t to first order and is linear in its Riemannian gradient.*

Axiom 11 (g-skew generator)**.** *There exists an operator $\Omega^z_{g_t}(\theta)$, parameterized by an internal state $z(t)$, such that for all $u, v \in T_\theta\Theta$,*

$$\langle u, \Omega^z_{g_t}(\theta) v\rangle_{g_t} = -\langle \Omega^z_{g_t}(\theta) u, v\rangle_{g_t}. \tag{26}$$

Lemma 3 (Canonical g-skew form)**.** *Axiom 11 holds if and only if*

$$\Omega^z_{g_t}(\theta) = g_t(\theta)^{-1}\Omega^z(\theta), \qquad \Omega^z(\theta)^\top = -\Omega^z(\theta). \tag{27}$$

Proof Assume Axiom 11 and define $\Omega^z(\theta) \coloneqq g_t(\theta)\Omega^z_{g_t}(\theta)$. Then for all u, v, $u^\top\Omega^z(\theta)v = \langle u, \Omega^z_{g_t}(\theta)v\rangle_{g_t} = -\langle \Omega^z_{g_t}(\theta)u, v\rangle_{g_t} = -u^\top\Omega^z(\theta)^\top v$, hence $\Omega^z(\theta)^\top = -\Omega^z(\theta)$ and $\Omega^z_{g_t} = g_t^{-1}\Omega^z$. Conversely, if $\Omega^z_{g_t} = g_t^{-1}\Omega^z$ with $\Omega^{z\top} = -\Omega^z$, then $\langle u, \Omega^z_{g_t} v\rangle_{g_t} = u^\top\Omega^z v = -(\Omega^z u)^\top v = -\langle \Omega^z_{g_t} u, v\rangle_{g_t}$, which is (26). □

Theorem 4 (Learning dynamics from axioms)**.** *Under Theorem 2 and Axioms 10–11, the learning dynamics on* Θ *takes the form*

$$\dot{\theta} = -\operatorname{grad}_{g_t} L_t(\theta) + \kappa\, \Omega^z_{g_t}(\theta)\, \operatorname{grad}_{g_t} E_t(\theta), \tag{28}$$

with constant amplitude $\kappa \geq 0$*. Moreover, for frozen* t *(treating* $E_t(\cdot)$ *as the instantaneous functional at step* t*), the circulation term preserves* E_t *to first order:*

$$\frac{d}{dt} E_t(\theta) = \left\langle \operatorname{grad}_{g_t} E_t(\theta), \dot{\theta} \right\rangle_{g_t} = -\left\langle \operatorname{grad}_{g_t} E_t(\theta), \operatorname{grad}_{g_t} L_t(\theta) \right\rangle_{g_t}. \tag{29}$$

Proof By Theorem 2, the descent component under the KL trust region is $-\operatorname{grad}_{g_t} L_t(\theta)$. Axiom 10 requires an additional component that is linear in $\operatorname{grad}_{g_t} E_t(\theta)$, and Axiom 11 requires that this component be generated by a g_t-skew operator; together these imply the form $\kappa\, \Omega^z_{g_t}(\theta) \operatorname{grad}_{g_t} E_t(\theta)$, yielding (28).

For evolution at frozen t, differentiate $E_t(\theta)$ along the θ-flow: $\frac{d}{dt} E_t(\theta) = \langle \operatorname{grad}_{g_t} E_t(\theta), \dot{\theta} \rangle_{g_t}$. Substitute (28) to obtain two terms. The circulation term vanishes because for any v, $\langle v, \Omega^z_{g_t}(\theta) v \rangle_{g_t} = -\langle \Omega^z_{g_t}(\theta) v, v \rangle_{g_t}$ by Axiom 11, hence it equals its own negative and must be zero. This yields (29). □

1.4 Endogenous Chaos

We instantiate $\Omega^{z(t)}$ via an endogenous chaotic engine parameterized by an internal gain ω_z.

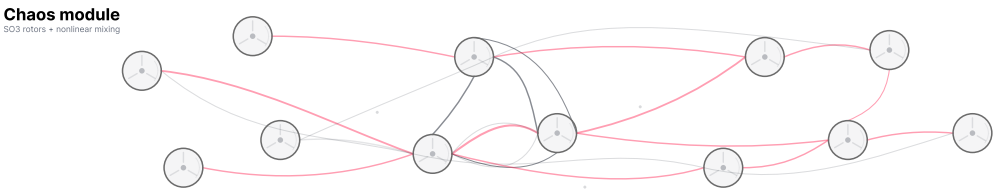

Fig. 5 A visualization of our chaos module.

Definition 2. (Blockwise Cayley update). *Let* $h(z) = \tanh(VRz) + \varepsilon$ *be a vector field with parameters* V, R, ε*. The internal state* $z \in \mathbb{R}^{3B}$ *is partitioned into* B *blocks* $z_b \in \mathbb{R}^3$*, each updated by a reversible rotation*

$$z_b^+ = Q_b(z)^\top z_b, \qquad Q_b(z) = \left(I - \tfrac{s}{2}\widehat{h_b(z)}\right)^{-1}\left(I + \tfrac{s}{2}\widehat{h_b(z)}\right) \in \mathrm{SO}(3), \tag{30}$$

where $s = \Delta t_z\, \omega_z$ *and* $\widehat{h_b(z)} \in \mathfrak{so}(3)$ *is the hat map of* $h_b(z)$*.*

The skew-symmetric field is then $\Omega^z(\theta) = \operatorname{diag}\left(\widehat{h_1(z)}, \ldots, \widehat{h_B(z)}\right)$, where each block couples to a subspace of θ-dynamics.

Lemma 4. (Casimir invariance). *The dynamics in (30) are reversible and preserve the Euclidean norm $C_b(z) := \|z_b\|^2$ of each block b. The invariant manifold is $(\mathbb{S}^2)^B$.*

Proof Since $Q_b(z) \in \mathrm{SO}(3)$, it is orthogonal with determinant $+1$, hence $\|z_b^+\| = \|Q_b^\top z_b\| = \|z_b\|$. □

Algorithm 2 ChaosStep (internal skew-generator update)

Input: State z partitioned into blocks $\{z_b\}_{b=1}^B$, $z_b \in \mathbb{R}^3$.
Input: Step size Δt_z and internal gain ω_z.
for $b = 1$ to B **do**
  Compute endogenous field $h_b \leftarrow h_b(z)$
  Form the 3×3 skew matrix $\widehat{h}_b$ from h_b
  $Q_b \leftarrow (I - \frac{1}{2}(\Delta t_z \omega_z)\widehat{h}_b)^{-1}(I + \frac{1}{2}(\Delta t_z \omega_z)\widehat{h}_b)$
  $z_b \leftarrow Q_b^\top z_b$
  $z_b \leftarrow z_b/(\|z_b\| + 10^{-12})$ ▷ Casimir / norm preservation
end for
return updated z

1.5 Degeneracy case in offline ERM: reduction to SGD

We show that standard offline optimization is a degeneracy of (28) under stationarity. The reduction follows from (i) stationarity, which freezes the online geometry, and (ii) stationarity of the activation medium, which yields zero kinematic energy.

1.5.1 Axioms for the stationary regime

Axiom 12 (Stationary stream)**.** *The data stream is stationary over the adaptation horizon: $\mathcal{P}_t \equiv \mathcal{P}$.*

Axiom 13 (EMA convergence of online Fisher factors)**.** *Under Axiom 12, the exponentially weighted K-FAC factors converge (in expectation) to time-invariant limits:*

$$A_{\ell,t} \to A_\ell, \qquad G_{\ell,t} \to G_\ell. \tag{31}$$

Consequently, any deterministic function of these factors (e.g., conceptors $C_{\ell,t}$, complements $U_{\ell,t}$, and shaped factors $\widetilde{A}_{\ell,t}$) also converges to time-invariant limits.

Axiom 14 (Vanishing kinematic energy under a stationary medium)**.** *In the stationary regime with a stationary activation medium, the residual vanishes and the induced kinematic energy is zero:*

$$\Delta A_\ell(\theta) := \widehat{A}_\ell(\theta) - A_\ell = 0 \quad \Longrightarrow \quad I(\theta) = 0, \tag{32}$$

where A_ℓ is the stationary limit of $A_{\ell,t}$ and I is the time-invariant functional defined below.

1.5.2 Frozen geometry and frozen energy functional

Lemma 5 (Time-invariant geometry)**.** *Under Axioms 12–13, the online Fisher geometry becomes time-invariant. In particular, the (shaped) block-diagonal approximation $\widetilde{g}_t$ converges to a fixed metric g.*

Proof Axiom 13 gives convergence of the K-FAC factors and hence of the blockwise Fisher approximation (including any shaping that is a deterministic function of the factors). Therefore $\widetilde{g}_t \to g$ for some fixed g. □

Lemma 6 (Time-invariant kinematic energy functional)**.** *Under Axioms 12–13, the kinematic energy becomes time-invariant:*

$$E(\theta) = \sum_{\ell} \Big[\mathrm{tr}\big(\Delta A_\ell(\theta)\, C_\ell\, \Delta A_\ell(\theta)\, U_\ell\big) + \mathrm{tr}\big(\Delta A_\ell(\theta)\, U_\ell\, \Delta A_\ell(\theta)\, C_\ell\big) \Big], \tag{33}$$

where C_ℓ and U_ℓ are the stationary limits of $C_{\ell,t}$ and $U_{\ell,t}$.

Proof By Axiom 13, $A_{\ell,t} \to A_\ell$ and hence $C_{\ell,t} \to C_\ell$ and $U_{\ell,t} \to U_\ell$. Substituting these stationary limits into Definition 1 yields the time-invariant functional (33). □

1.5.3 Reduction of the dynamics

Theorem 5 (Degeneracy to natural-gradient descent)**.** *Assume Axioms 12–14. Then the learning dynamics (28) degenerates to natural-gradient flow under the fixed metric g:*

$$\dot{\theta} = -\operatorname{grad}_g L(\theta), \tag{34}$$

where $L(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\, \ell(\theta; x, y)$.

Proof By Lemma 5 and Lemma 6, the stationary regime induces a fixed geometry g and a fixed functional I. In (28), the conservative component is $\kappa\, \Omega_g^z(\theta)\, \operatorname{grad}_g I(\theta)$. By Axiom 14, $I(\theta) = 0$ (hence $\operatorname{grad}_g I(\theta) = 0$), so this term vanishes, and the remaining dynamics is $\dot{\theta} = -\operatorname{grad}_g L(\theta)$. □

Lemma 7 (Critical points are unchanged under a fixed SPD metric)**.** *If g is symmetric positive definite, then*

$$\operatorname{grad}_g L(\theta) = 0 \iff \nabla L(\theta) = 0. \tag{35}$$

Proof By definition, $\operatorname{grad}_g L = g^{-1}\nabla L$. Since g^{-1} is invertible (SPD), $g^{-1}\nabla L = 0$ if and only if $\nabla L = 0$. □

1.5.4 Reduction to SGD in Euclidean specialization

In a Euclidean specialization (or locally whitened coordinates) where $g \approx I$, (34) reduces to standard gradient flow $\dot{\theta} = -\nabla L(\theta)$. Its stochastic discretization under minibatch sampling yields SGD.

2 Datastream Protocol

2.1 CIFAR-10

2.2 CLEAR-100

3 Ablation Studies

3.1 CLEAR-100

3.2 Additional results of Learned Features

Fig. 6 U-Map of learned features under stream s2 protocal.

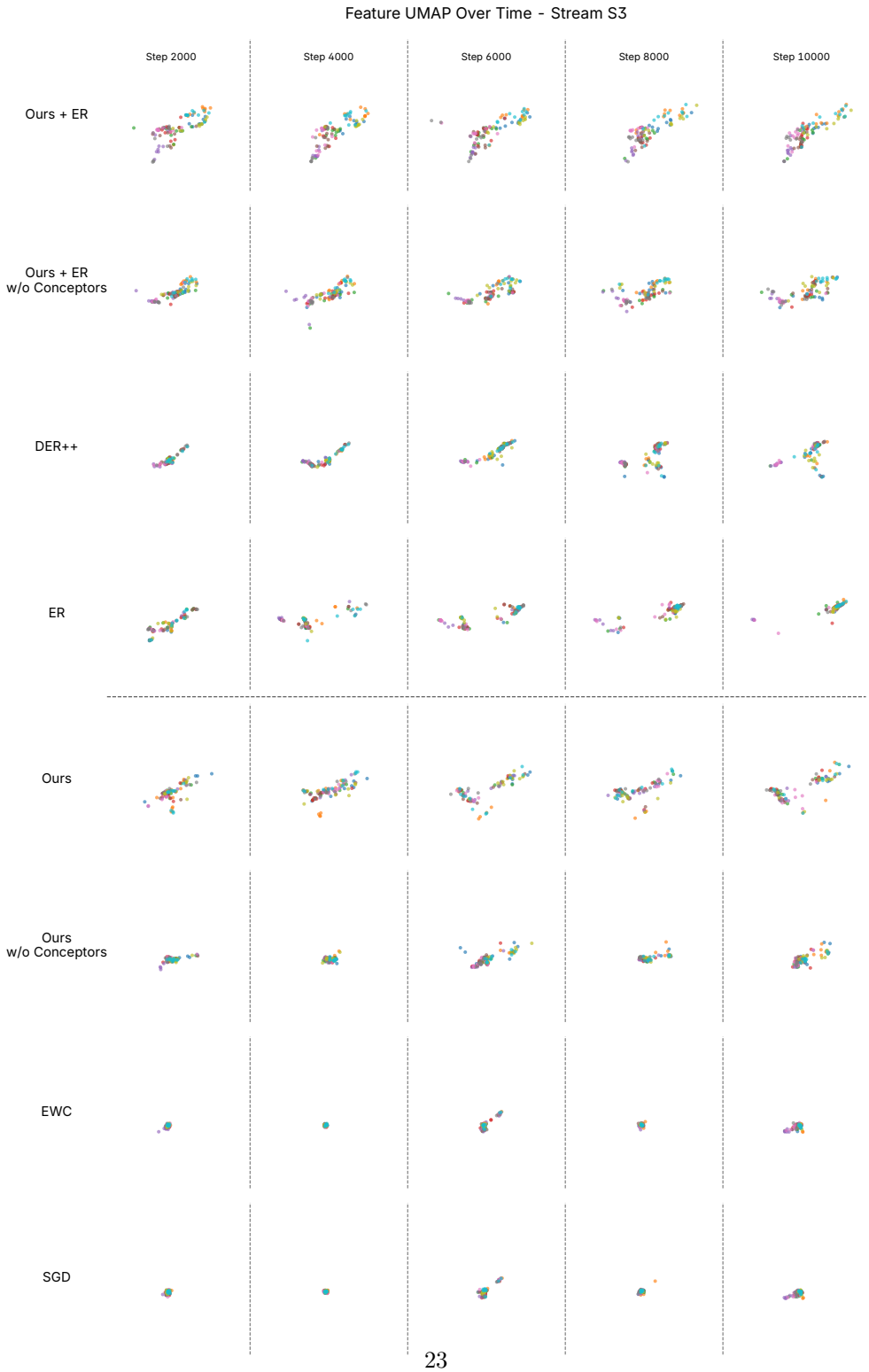

Fig. 7 U-Map of learned features under stream s3 protocal.

Method	Effective rank		Cosine sim		Top1 frac		Tail mass 10		Scatter ratio	
	Mean	Final	Mean	Final	Mean	Final	Mean	Final	Mean	Final
Stream S1										
SGD	0.084	0.039	0.971	0.911	0.583	0.778	0.062	0.016	0.261	0.214
EWC	0.084	0.040	0.972	0.914	0.583	0.774	0.061	0.016	0.264	0.218
Ours	0.098	0.115	0.484	0.457	0.469	0.379	0.058	0.081	0.587	0.550
ER	0.041	0.028	0.733	0.720	0.698	0.871	0.003	0.001	1.071	1.022
DER++	0.048	0.064	0.691	0.616	0.677	0.607	0.007	0.011	0.591	0.501
Ours+ER	0.110	0.121	0.397	0.417	0.396	0.366	0.039	0.054	0.681	0.524
Stream S2										
SGD	0.086	0.067	0.974	0.920	0.564	0.549	0.063	0.030	0.290	0.226
EWC	0.087	0.067	0.974	0.918	0.560	0.558	0.063	0.031	0.289	0.226
Ours	0.108	0.104	0.497	0.482	0.448	0.470	0.057	0.057	0.335	0.382
ER	0.035	0.027	0.751	0.809	0.760	0.870	0.003	0.001	0.892	0.666
DER++	0.040	0.052	0.706	0.584	0.745	0.648	0.006	0.004	0.571	0.235
Ours+ER	0.105	0.101	0.412	0.417	0.408	0.420	0.038	0.035	0.599	0.556
Stream S3										
SGD	0.014	0.004	0.835	0.755	0.565	0.837	0.090	0.005	0.387	0.169
EWC	0.014	0.003	0.834	0.747	0.566	0.866	0.090	0.004	0.386	0.136
Ours	0.022	0.017	0.510	0.493	0.343	0.451	0.127	0.107	0.290	0.242
ER	0.007	0.006	0.720	0.657	0.595	0.731	0.016	0.010	1.004	0.995
DER++	0.008	0.009	0.667	0.642	0.638	0.578	0.024	0.019	0.592	0.453
Ours+ER	0.019	0.030	0.416	0.401	0.358	0.219	0.098	0.174	0.605	0.480

Table 1 Streaming regimes (S1–S3) with geometry/diversity metrics. Shaded rows highlight our methods.

Appendix A Prior Work Expanded

A.1 Regularization protects parameters, but can become rigidity.

Regularization-based CL constrains updates so that parameters deemed important to the past remain unchanged (Kirkpatrick et al. 2017; Zenke et al. 2017; Aljundi et al. 2018; Li and Hoiem 2016). While this can slow destructive drift, it often turns continual learning into a stability–plasticity tradeoff that is resolved by freezing. The learner becomes increasingly rigid, absorbs new information only weakly, and may collapse to low-diversity codes under long-horizon non-stationarity. In short, regularization can reduce damage by suppressing motion, but it does not provide a principled way to keep moving without breaking structure.

A.2 Replay preserves samples, but does not fix destructive dynamics.

Replay-based CL stabilizes training by reintroducing past samples (Rebuffi et al. 2017; Lopez-Paz and Ranzato 2017; Buzzega et al. 2020; Bellitto et al. 2024). This can work well, yet it scales by accumulation and bookkeeping: what to store, how to prioritize, and how to compress. It also relies on the same underlying descent dynamics that caused the damage in the first place. Replay can patch forgetting by revisiting old data, but it does not by itself guarantee that the parameter trajectory remains non-destructive under indefinite non-stationarity, especially when the stream is long, correlated, and continuously shifting.

A.3 Parameter isolation avoids interference by construction, but often sidesteps the lifelong setting.

A third line of work mitigates forgetting by allocating task-specific subnetworks or adaptation paths, such as learned masks, hard gating, or per-task adapters (Mallya and Lazebnik 2018; Mallya et al. 2018; Serra et al. 2018; Gao et al. 2024). These methods typically rely on explicit task boundaries, task identifiers, or routing heuristics, and they trade the original problem for capacity growth and bookkeeping over task-specific solutions. In effect, they reduce forgetting by avoiding shared representation learning, rather than by enabling a single learner to remain plastic while preserving internal structure under an open-ended stream.

A.4 Benchmarks often sanitize the stream and sometimes redefine the problem.

A separate but decisive issue is evaluation. Many CL benchmarks operationalize "continual" via i.i.d. datasets partitioned into tasks, classes, or episodes (Krizhevsky 2009; Lin et al. 2021), turning a genuinely open-ended stream into a curated sequence of finite blocks. This practice quietly imports assumptions that are rarely true in the wild. First, episodicity assumes distribution shifts arrive as clean task boundaries and training is implicitly reset at each boundary (Koh et al. 2022). Second, i.i.d. splits masquerade as time, as shuffling or class partitioning creates an artificial notion of non-stationarity that is statistically convenient but behaviorally unfaithful (Cossu et al. 2022). Third, a single-pass premise assumes that once a sample is seen it never reappears, effectively making "remembering" synonymous with memorizing a vanishing set rather than maintaining stable structure under recurring regimes (Hemati et al. 2025). As a result, methods can look successful while optimizing for benchmark-specific artifacts, such as buffer accounting, task identifiers, or boundary-aware heuristics, rather than for the core requirement of continual learning: remaining plastic without destructively rewriting internal organization under a genuinely ongoing stream.

Appendix B Information Stress Energy Tensor

We can interpret the kinematic residual energy as an information geometry stress energy. This appendix defines the *stress energy* tensor family used in our circulation objective. For each layer, we maintain EMA factors (A, G) and compute instantaneous factors $(\hat{A}, \hat{G})$ from the current batch (or live-captured activations/gradients). The *curvature residuals* are

$$\Delta A := \hat{A} - A, \qquad \Delta G := \hat{G} - G, \tag{B1}$$

and all channelized tensors below are built from these residuals. We introduce conceptors on both Kronecker factors. On the activation side,

$$P_0 := C_A, \qquad P_1 := U_A := I - C_A, \tag{B2}$$

and on the gradient side,

$$Q_0 := C_G, \qquad Q_1 := U_G := I - C_G. \tag{B3}$$

To measure channel coupling without explicitly forming square roots in implementation, we use the identity $\|X^{\frac{1}{2}} M Y^{\frac{1}{2}}\|_F^2 = \mathrm{tr}(M^\top X M Y)$ for symmetric X, Y; nevertheless, for clarity we present the energy in the symmetric "half-power" form. Define the channelized residual blocks

$$\Delta A^{rs}_{\frac{1}{2}} := P_r^{\frac{1}{2}} \, (\Delta A) \, P_s^{\frac{1}{2}}, \qquad \Delta G^{pq}_{\frac{1}{2}} := Q_p^{\frac{1}{2}} \, (\Delta G) \, Q_q^{\frac{1}{2}}, \qquad p, q, r, s \in \{0, 1\}. \tag{B4}$$

These blocks enumerate all $2^4 = 16$ (C/U)-combinations across both sides. We then define a *gauge-indexed* family of residual channel tensors in the Kronecker-product parameter space by

$$\mathcal{K}^{pq,rs} := \Delta G^{pq}_{\frac{1}{2}} \otimes \Delta A^{rs}_{\frac{1}{2}}, \qquad p, q, r, s \in \{0, 1\}, \tag{B5}$$

and the corresponding *kinematic energy tensor entries* as Frobenius norms

$$\mathcal{E}^{pq,rs} := \left\|\mathcal{K}^{pq,rs}\right\|_F^2 = \left\|\Delta G^{pq}_{\frac{1}{2}}\right\|_F^2 \left\|\Delta A^{rs}_{\frac{1}{2}}\right\|_F^2, \tag{B6}$$

where the last equality uses $\|X \otimes Y\|_F = \|X\|_F \|Y\|_F$. The full residual kinematic energy for a layer can be taken as the sum over selected channels (e.g., coupling-only channels), or as the sum over all channels, depending on which leakage mode is enforced:

$$\mathcal{E}_{\text{all}} := \sum_{p,q \in \{0,1\}} \sum_{r,s \in \{0,1\}} \mathcal{E}^{pq,rs}. \tag{B7}$$

In particular, the *coupling energy* on the activation side that we use in the main text is the restriction to the off-diagonal $(r, s) \in \{(0, 1), (1, 0)\}$ channels,

$$\mathcal{E}^A_{\text{couple}} := \left\|C_A^{\frac{1}{2}} \, \Delta A \, U_A^{\frac{1}{2}}\right\|_F^2 + \left\|U_A^{\frac{1}{2}} \, \Delta A \, C_A^{\frac{1}{2}}\right\|_F^2, \tag{B8}$$

and analogously one may define $\mathcal{E}^{G}_{\text{couple}}$ by replacing $(C_A, U_A, \Delta A)$ with $(C_G, U_G, \Delta G)$.

For an explicit all channels view, we assemble the 2×2 channel block matrices of half-powered residuals:

$$\mathbf{\Delta A}_{\frac{1}{2},\text{ch}} := \begin{bmatrix} \Delta A^{00}_{\frac{1}{2}} & \Delta A^{01}_{\frac{1}{2}} \\ \Delta A^{10}_{\frac{1}{2}} & \Delta A^{11}_{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} C_A^{\frac{1}{2}} \Delta A\, C_A^{\frac{1}{2}} & C_A^{\frac{1}{2}} \Delta A\, U_A^{\frac{1}{2}} \\ U_A^{\frac{1}{2}} \Delta A\, C_A^{\frac{1}{2}} & U_A^{\frac{1}{2}} \Delta A\, U_A^{\frac{1}{2}} \end{bmatrix}, \tag{B9}$$

$$\mathbf{\Delta G}_{\frac{1}{2},\text{ch}} := \begin{bmatrix} \Delta G^{00}_{\frac{1}{2}} & \Delta G^{01}_{\frac{1}{2}} \\ \Delta G^{10}_{\frac{1}{2}} & \Delta G^{11}_{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} C_G^{\frac{1}{2}} \Delta G\, C_G^{\frac{1}{2}} & C_G^{\frac{1}{2}} \Delta G\, U_G^{\frac{1}{2}} \\ U_G^{\frac{1}{2}} \Delta G\, C_G^{\frac{1}{2}} & U_G^{\frac{1}{2}} \Delta G\, U_G^{\frac{1}{2}} \end{bmatrix}. \tag{B10}$$

The corresponding grand 4×4 block tensor over channel indices is the Kronecker product

$$\mathbf{K}_{\text{ch}} \;:=\; \mathbf{\Delta G}_{\frac{1}{2},\text{ch}} \otimes \mathbf{\Delta A}_{\frac{1}{2},\text{ch}}, \tag{B11}$$

whose (p, r)-by-(q, s) block is exactly $\mathcal{K}^{pq,rs} = \Delta G^{pq}_{\frac{1}{2}} \otimes \Delta A^{rs}_{\frac{1}{2}}$. Expanding this explicitly (ordering $(p, r) \in \{(0,0), (0,1), (1,0), (1,1)\}$ for block rows and (q, s) similarly for block columns) yields

$$\mathbf{K}_{\text{ch}} = \begin{bmatrix} \Delta G^{00}_{\frac{1}{2}} \otimes \Delta A^{00}_{\frac{1}{2}} & \Delta G^{00}_{\frac{1}{2}} \otimes \Delta A^{01}_{\frac{1}{2}} & \Delta G^{01}_{\frac{1}{2}} \otimes \Delta A^{00}_{\frac{1}{2}} & \Delta G^{01}_{\frac{1}{2}} \otimes \Delta A^{01}_{\frac{1}{2}} \\ \Delta G^{00}_{\frac{1}{2}} \otimes \Delta A^{10}_{\frac{1}{2}} & \Delta G^{00}_{\frac{1}{2}} \otimes \Delta A^{11}_{\frac{1}{2}} & \Delta G^{01}_{\frac{1}{2}} \otimes \Delta A^{10}_{\frac{1}{2}} & \Delta G^{01}_{\frac{1}{2}} \otimes \Delta A^{11}_{\frac{1}{2}} \\ \Delta G^{10}_{\frac{1}{2}} \otimes \Delta A^{00}_{\frac{1}{2}} & \Delta G^{10}_{\frac{1}{2}} \otimes \Delta A^{01}_{\frac{1}{2}} & \Delta G^{11}_{\frac{1}{2}} \otimes \Delta A^{00}_{\frac{1}{2}} & \Delta G^{11}_{\frac{1}{2}} \otimes \Delta A^{01}_{\frac{1}{2}} \\ \Delta G^{10}_{\frac{1}{2}} \otimes \Delta A^{10}_{\frac{1}{2}} & \Delta G^{10}_{\frac{1}{2}} \otimes \Delta A^{11}_{\frac{1}{2}} & \Delta G^{11}_{\frac{1}{2}} \otimes \Delta A^{10}_{\frac{1}{2}} & \Delta G^{11}_{\frac{1}{2}} \otimes \Delta A^{11}_{\frac{1}{2}} \end{bmatrix}. \tag{B12}$$

Finally, the same channelization can be written as a single indexed sum by using E_{pq} as the 2×2 matrix units (a single 1 at entry (p, q)):

$$\mathbf{K}_{\text{ch}} = \sum_{p,q \in \{0,1\}} \sum_{r,s \in \{0,1\}} \left(E_{pq} \otimes E_{rs}\right) \left(\Delta G^{pq}_{\frac{1}{2}} \otimes \Delta A^{rs}_{\frac{1}{2}}\right), \quad \mathbf{E}_{\text{ch}} = \sum_{p,q} \sum_{r,s} \left(E_{pq} \otimes E_{rs}\right) \mathcal{E}^{pq,rs}. \tag{B13}$$

Here $\mathbf{E}_{\text{ch}}$ denotes the corresponding 4×4 matrix of scalar channel energies whose entries are $\mathcal{E}^{pq,rs} = \|\Delta G^{pq}_{\frac{1}{2}} \otimes \Delta A^{rs}_{\frac{1}{2}}\|_F^2$.

References

Aljundi R, Babiloni F, Elhoseiny M, et al (2018) Memory aware synapses: Learning what (not) to forget. In: European Conference on Computer Vision (ECCV), pp 144–161, https://doi.org/10.1007/978-3-030-01219-9_9

Bellitto G, Palazzo S, Spampinato C, et al (2024) Saliency-driven experience replay for continual learning. In: Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 9 - 15, 2024

Buzzega P, Boschini M, Porrello A, et al (2020) Dark experience for general continual learning: a strong, simple baseline. In: Advances in Neural Information Processing Systems (NeurIPS)

Cossu A, Graffieti G, Pellegrini L, et al (2022) Is class-incremental enough for continual learning? Frontiers in Artificial Intelligence https://doi.org/10.3389/frai.2022.829842

Gao X, Dong S, He Y, et al (2024) Beyond prompt learning: Continual adapter for efficient rehearsal-free continual learning. In: European Conference on Computer Vision (ECCV), pp 89–106

Gunasekara N, Pfahringer B, Gomes HM, et al (2023) Survey on online streaming continual learning. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI 2023). ijcai.org, pp 6628–6637, https://doi.org/10.24963/IJCAI.2023/743

Hemati H, Pellegrini L, Duan X, et al (2025) Continual learning in the presence of repetition. Neural Networks 183:106920. https://doi.org/10.1016/j.neunet.2024.106920

Kirkpatrick J, Pascanu R, Rabinowitz N, et al (2017) Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences 114(13):3521–3526. https://doi.org/10.1073/pnas.1611835114

Koh H, Kim D, Ha JW, et al (2022) Online continual learning on class incremental blurry task configuration with anytime inference. In: International Conference on Learning Representations (ICLR)

Krizhevsky A (2009) Learning multiple layers of features from tiny images. Tech. Rep. TR-2009, University of Toronto

Li Z, Hoiem D (2016) Learning without forgetting. In: European Conference on Computer Vision (ECCV), pp 614–629, https://doi.org/10.1007/978-3-319-46493-0_37

Lin Z, Shi J, Pathak D, et al (2021) The CLEAR benchmark: Continual LEArning on real-world imagery. In: NeurIPS Datasets and Benchmarks Track

Lopez-Paz D, Ranzato M (2017) Gradient episodic memory for continual learning. In: Advances in Neural Information Processing Systems (NeurIPS)

Mallya A, Lazebnik S (2018) Packnet: Adding multiple tasks to a single network by iterative pruning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Mallya A, Davis D, Lazebnik S (2018) Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In: European Conference on Computer Vision (ECCV)

Rebuffi S, Kolesnikov A, Sperl G, et al (2017) icarl: Incremental classifier and representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

Serra J, Suris D, Miron M, et al (2018) Overcoming catastrophic forgetting with hard attention to the task. In: International Conference on Machine Learning (ICML)

Wang L, Zhang X, Su H, et al (2024) A comprehensive survey of continual learning: Theory, method and application. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(8):5362–5383. https://doi.org/10.1109/TPAMI.2024.3367329

Zenke F, Poole B, Ganguli S (2017) Continual learning through synaptic intelligence. In: Precup D, Teh YW (eds) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, Proceedings of Machine Learning Research, vol 70. PMLR, pp 3987–3995