IMPROVED ELASTIC WEIGHT CONSOLIDATION AS AN OPTIMIZATION CONSTRAINT FOR CONTINUAL LEARNING

Davide Wiest
Axym Labs
TU Darmstadt
Darmstadt, Germany
dw@axym.org

June 17, 2026

ABSTRACT

Elastic Weight Consolidation (EWC) is widely used in continual learning, yet its practical form relies on the empirical Fisher matrix, whose justification beyond probabilistic classification remains incomplete. The empirical Fisher differs mathematically from the true Fisher, aligns with it only under specific assumptions, and is nevertheless used much more broadly in practice. We address this gap by deriving EWC again from first principles. Starting from the continual-learning problem itself, we formulate old-task solution preservation as a constrained optimization problem (CLOC) and develop PL-CLOC, a population, low-order variant for deriving scalable mechanisms. Within this framework, preserving the old scalar loss value recovers exactly the empirical Fisher matrix (EF) used in EWC. This places EWC on direct optimization-theoretic footing without reliance on the true Fisher matrix. Preserving outputs up to motion along the old-task loss level set recovers the improved empirical Fisher (IEF) and leads to IEWC, a generalized and normalized sibling of EWC. EF implicitly scales local old-task sensitivity directions by the output's gradient magnitude, whereas IEF does not. We identify EWC-DR's vanishing-importance pathology as a special case of this gradient-scale effect on well-fitted data, and provide the improved empirical Fisher as the corresponding general solution. More broadly, the derivation separates the outer EWC mechanism from output geometry, enabling geometry-aware continual learning within the same regularization template. Empirically, IEWC improves old-distribution retention over EF-EWC across classification, regression, diffusion, and segmentation, and a sliced-Wasserstein output geometry gives selective stability gains in diffusion. Together, these results motivate IEWC as a theoretical and practical successor.

Keywords continual learning · elastic weight consolidation · empirical Fisher · constrained optimization

1 Continual Learning as Optimization Constraint

Consider the problem of continual learning, where data from distributions (tasks, environments) are presented sequentially [1–4]. Concretely, we first train a model on distribution A and obtain parameters θ_A^*, and only then can we optimize with respect to a new distribution B. If we plainly pick a minimal loss/cost $\min_\theta L_B(\theta)$ as our objective, then the optimization does not have to respect performance on A, which usually degrades it too much to be practical. Hence, we have to guide or constrain optimization in a way. A natural fit for this problem is the framework of constrained optimization, with solution preservation being the constraint in question. Define both a quantity $Q : \Theta \to \mathcal{H}$ that is relevant to this preservation and a constraint set $\mathcal{C} \subset \mathcal{H}$ that implies preservation is met. Before exploring concrete choices in later sections, we continue on the common thread. These two objects let us define a new objective:

$$\min_{\theta \in \Theta} L_B(\theta) \qquad \text{subject to} \qquad q(\theta) \in \mathcal{C}$$

This gives a mathematical formulation of preservation, and the next section extends it to become more practical.

2 Population and Local Approximations

To obtain a differentiable objective suitable for gradient-based optimization, we apply the standard Moreau-Yosida regularization of the constraint indicator, which for closed convex $\mathcal{C}$ yields the squared-distance penalty [5]:

$$\min_\theta J(\theta) = \min_\theta L_B(\theta) + \frac{\lambda}{2}\text{dist}(q(\theta), \mathcal{C})^2 = \min_\theta L_B(\theta) + \frac{\lambda}{2}\inf_{u\in\mathcal{C}} \text{dist}(q(\theta), u)^2.$$

We call this the **Continual Learning as Optimization Constraint (CLOC)** formulation. GEM and A-GEM are similar in spirit in that they formulate preservation through constraints on old examples [6, 7]. CLOC turns the old-task constraint into a proximal penalty built after task-A training. In a population setting, e.g. when we have a dataset, we can translate this into a population risk objective by wrapping it in an expectation over the data and using sample-based analogues ℓ_B of L_B and q of Q, then taking the empirical average. Define the unscaled proximal loss as $p(a, \theta) = \text{dist}(q(a, \theta), \mathcal{C})^2$.

$$\min_\theta \mathbb{E}_{b\sim\mathcal{D}_B}[\ell_B(b, \theta)] + \frac{\lambda}{2}\mathbb{E}_{a\sim\mathcal{D}_A}[p(a, \theta)]$$

$$\approx \min_\theta \frac{1}{N_B}\sum_{i=1}^{N_B}\ell_B(b_i, \theta) + \frac{\lambda}{2N_A}\sum_{i=1}^{N_A} p(a_i, \theta)$$

The two-distribution presentation is only for notational simplicity; additional old distributions can be handled by adding analogous constraints. We refer to this population version as **population CLOC (P-CLOC)**. One efficiency consideration is the penalty p: Ideally, we let the constraint contribute to each gradient step with its gradient. However, recomputing the full old-distribution penalty p and its associated factors many times during optimization on distribution B is usually prohibitively expensive, since it depends on the dataset D_A associated with A. Additionally, reusing dataset D_A later on does not respect the continual learning setting. For this reason, replace the exact penalty by a local k-th order Taylor-polynomial approximation around the old solution $\theta^\star_A$:

$$p(a, \theta^\star_A + \delta) \approx \sum_{j=0}^{k}\frac{1}{j!}\nabla^j_\theta p(a, \theta^\star_A)\left[\delta^{\otimes j}\right], \qquad \delta = \theta - \theta^\star_A.$$

This turns the computation of p into a repeated evaluation of precomputed derivative tensors and bounds memory requirements to $O(d^k)$ for a d-dimensional parameterization. Since memory scales exponentially with k, the standard scalable surrogate is a low-order Taylor approximation. We refer to this local approximation as **low-order CLOC (L-CLOC)**, and to its population version as **PL-CLOC**. It remains to specify the preserved quantity q and constraint set $\mathcal{C}$. Notice that the gradient of p is dependent on only one sample at once, which allows us to analyze p and its gradients on a per-sample basis.

3 Old-Task Loss Preservation

Since ℓ_A is our best practical surrogate on performance on distribution A, a natural choice is to use ℓ_A itself to define preservation. We say that the model f_θ preserved the solution if it preserved the old loss value for distribution A. This leads to:

$$q(a, \theta) = \ell_A(f_\theta(x)) - \ell_A(f_{\theta^\star_A}(x)).$$

Since we want $q(a, \theta)$ to be zero, the constraint set's definition follows as

$$\mathcal{C} = \{0\} \subset \mathbb{R}.$$

Thus the hard constraint is

$$q(a, \theta) \in \mathcal{C},$$

and the corresponding unscaled proximal loss is

$$p(a, \theta) = \text{dist}(q(a, \theta), \mathcal{C})^2 = q(a, \theta)^2.$$

To obtain a low-order Taylor approximation, use the perturbation around the old solution $\delta = \theta - \theta^\star_A$ in the first-order expansion of q:

$$q(a, \theta^\star_A + \delta) \approx g^\top\delta,$$

where

$$g = \nabla_\theta \ell_A(f_{\theta^\star_A}(a)).$$

Equivalently, writing

$$z_A = f_{\theta_A^\star}(a) \qquad r_i = \nabla_z \ell_A(z)\big|_{z=f_{\theta_A^\star}(a)}, \qquad J_i = \nabla_\theta f_{\theta_A^\star}(a),$$

the chain rule gives

$$g_i = J_i^\top r_i.$$

Substituting this expansion into the preservation penalty gives

$$\frac{\lambda}{2N_A}\sum_{i=1}^{N_A} q(a_i,\theta)^2 \;\approx\; \frac{\lambda}{2N_A}\sum_{i=1}^{N_A}(g_i^\top \delta)^2.$$

Collecting the gradient outer products yields

$$\frac{\lambda}{2N_A}\sum_{i=1}^{N_A}(g_i^\top \delta)^2 = \frac{\lambda}{2}\,\delta^\top\left(\frac{1}{N_A}\sum_{i=1}^{N_A} g_i g_i^\top\right)\delta = \frac{\lambda}{2}\,\delta^\top M_1 \delta.$$

Therefore the resulting matrix is

$$M_1 = \frac{1}{N_A}\sum_{i=1}^{N_A} g_i g_i^\top = \frac{1}{N_A}\sum_{i=1}^{N_A} J_i^\top r_i r_i^\top J_i.$$

We call this an importance matrix as it encodes how important it is to preserve a direction in parameter space.

4 Empirical Fisher Regularization

This is exactly the empirical Fisher matrix (EF) used in Elastic Weight Consolidation (EWC) [8], which we will call EF-EWC from now on:

$$M_1 = F_{\mathrm{EF}}.$$

Hence EF-EWC approximates a constrained optimization problem with a smooth proximal penalty, preserving the old scalar loss value sample by sample. This places EF-EWC within the broader family of parameter-importance regularizers [9, 10] and Bayesian or online continual-learning approximations [11–13].

The original EF-EWC derivation uses properties specific to probabilistic classification models. After training on task A, one considers the posterior over parameters

$$p(\theta \mid \mathcal{D}_A) \propto p(\mathcal{D}_A \mid \theta)\, p(\theta),$$

and uses it as a prior when learning task B. Expanding the negative log-posterior around the old solution $\theta_A^\star$ yields a local quadratic approximation

$$-\log p(\theta \mid \mathcal{D}_A) \approx -\log p(\theta_A^\star \mid \mathcal{D}_A) + \frac{1}{2}(\theta-\theta_A^\star)^\top H_A(\theta_A^\star)(\theta-\theta_A^\star),$$

where $H_A(\theta_A^\star)$ is the Hessian of the negative log-posterior at the old solution. Under the usual assumptions for log-likelihood models, and neglecting the prior curvature term or absorbing it into the approximation, this Hessian can be replaced by the Fisher information matrix

$$F(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_A}\left[\nabla_\theta \log p_\theta(y \mid x)\, \nabla_\theta \log p_\theta(y \mid x)^\top\right].$$

For negative log-likelihood losses, this Fisher matrix coincides with the curvature of the expected loss under standard regularity conditions, and therefore matches the corresponding Hessian-based local quadratic model near an optimum, as in the natural-gradient and approximate-curvature literature [14–16]. Since classification models explicitly define a predictive distribution $p_\theta(y \mid x)$ over labels, the true Fisher can be computed as population object by sampling from said distribution. In practice, however, EWC replaces it by empirical approximations computed from the old dataset, which leads to the empirical Fisher form used above. However, this framework leaves open a justification of the empirical Fisher outside of negative log-likelihood losses. This is an open problem because EWC is used for different tasks and objectives as well, for example for diffusion [17]. Furthermore, Kunstner et al. show that the empirical Fisher is a different mathematical object than the true Fisher and approximates the true Fisher well only in specific settings [18]. We re-derive EF-EWC from scratch using PL-CLOC to resolve these two issues, basing the usage of said matrix on solid theoretical grounding without reliance on the true Fisher matrix.

5 Level-Set Preservation in Output Space

Our first mechanism is derived from preservation of the loss value. An extension of the same preservation constraint is to impose it through the old-task loss level set in output space. More concretely, we preserve the output per-sample up to motions that stay on this set. This output-space perspective connects to distillation and functional regularization approaches to continual learning [19–22], which let the model match old predictions or predictive functions. This gives a geometric formulation of the same underlying constraint: instead of penalizing raw change in the scalar loss coordinate, we penalize displacement away from the local constraint set in output space. Let $a \in D_A$ be an old sample, and write $z_A = f_{\theta_A^\star}(x)$. The exact old-loss preservation set in output coordinates is

$$\mathcal{C} = \{z : \ell_A(z) = \ell_A(z_A)\}.$$

Thus the preserved quantity is the model output itself:

$$q(a, \theta) = f_\theta(x).$$

The hard constraint is

$$q(a, \theta) \in \mathcal{C},$$

and the corresponding unscaled proximal loss is

$$p(a, \theta) = \mathrm{dist}_G\left(q(a, \theta), \mathcal{C}\right)^2,$$

where $G \succ 0$ is the chosen local output metric.

This exact penalty is generally not tractable. To obtain a practical form, use the standard first-order local model of the constraint set near the old output z_A, namely its tangent plane. Let

$$r = \nabla_z \ell_A(z)\big|_{z=z_A}.$$

For a small output displacement

$$\Delta z = f_\theta(x) - f_{\theta_A^\star}(x),$$

the first-order old-loss change is

$$\ell_A(z_A + \Delta z) - \ell_A(z_A) \approx r^\top \Delta z.$$

Hence the local constraint set is

$$T = \{\Delta z : r^\top \Delta z = 0\}.$$

Replacing the exact level set by this tangent hyperplane gives the local proximal loss

$$p(a, \theta) \approx \mathrm{dist}_G(\Delta z, T)^2.$$

For the metric $|v|_A^2 = v^\top G v$, the squared distance to the tangent hyperplane follows from projecting Δz onto the gradient:

$$\mathrm{dist}_G(\Delta z, T)^2 = \frac{(r^\top \Delta z)^2}{r^\top G^{-1} r}.$$

Linearizing the model output around $\theta_A^\star$, with

$$\delta = \theta - \theta_A^\star, \qquad J = \nabla_\theta f_{\theta_A^\star}(x),$$

we have

$$\Delta z \approx J\delta.$$

Substituting this into the local regularization term gives

$$\frac{\lambda}{2} p(a, \theta_A^\star + \delta) \approx \frac{\lambda}{2} \frac{(r^\top J\delta)^2}{r^\top G^{-1} r}.$$

Equivalently,

$$\frac{\lambda}{2} p(a, \theta_A^\star + \delta) \approx \frac{\lambda}{2} \delta^\top \left(J^\top \frac{r r^\top}{r^\top G^{-1} r} J \right) \delta.$$

Taking the empirical average over old-distribution samples $a_1, \ldots, a_N$ yields an importance matrix of the form

$$M_{2,A} = \frac{1}{N} \sum_{i=1}^{N} J_i^\top \frac{r_i r_i^\top}{r_i^\top G_i^{-1} r_i} J_i.$$

In the Euclidean-output case $G_i = I$, this reduces to

$$M_2 = \frac{1}{N} \sum_{i=1}^{N} J_i^\top \frac{r_i r_i^\top}{|r_i|^2} J_i.$$

6 Improved Empirical Fisher Regularization

This recovers the improved empirical Fisher matrix introduced by Wu et al. [23]:

$$M_2 = F_{\mathrm{IEF}}.$$

Hence the IEF is the first-order local approximation of a constrained optimization problem that protects old behavior geometrically: instead of preserving the scalar loss value, it preserves the output up to motion along the loss level set. The original IEF paper derives the same normalization from approximate natural-gradient descent. In that setting, the empirical Fisher reweights sample contributions by squared output-gradient magnitude, which can make it a poor curvature surrogate; IEF removes this scale factor and retains the normalized output-loss direction. In our derivation, the same normalization arises because the penalty is defined through distance to the old-loss level set rather than through raw change in the scalar loss coordinate.

7 Relation to Gauss-Newton

Consider the scalar least-squares setting in Euclidean geometry. Suppose the model output is scalar, such that

$$\ell_n(z) = \frac{1}{2}(z - y_n)^2.$$

Then

$$r_n = z_n^A - y_n$$

is scalar, and for $r_n \neq 0$ we have

$$\frac{r_n^2}{r_n^2} = 1.$$

Therefore the Euclidean IEF contribution reduces to

$$J_n^\top J_n,$$

and averaging gives

$$M_{\mathrm{IEF}} = \frac{1}{N}\sum_{n=1}^{N} J_n^\top J_n.$$

This is exactly the Gauss–Newton or output-Jacobian form for squared-error regression [15, 24, 25], the standard practical preconditioning matrix for second-order optimization. This matches our geometric intuition: Loss preservation is governed by how parameters move the output, which is encoded by the Jacobian J, not by the current output gradient magnitude. This equivalence was made explicit in [23]. In fact, the IEF was motivated as an approximate Gauss–Newton algorithm.

8 Normalization and Sample Weighting

To compare both matrices in a common notation, let $G_n \succ 0$ denote the chosen output metric for sample n. Although F_{EF} itself does not depend on G_n, its summands can be rewritten relative to said output metric.

Define

$$\hat{r}_n^G = \frac{r_n}{\sqrt{r_n^\top G_n^{-1} r_n}}, \qquad a_n^G = J_n^\top \hat{r}_n^G, \qquad w_n^G = r_n^\top G_n^{-1} r_n.$$

Then

$$M_{\mathrm{IEF},G} = \frac{1}{N}\sum_{n=1}^{N} a_n^G (a_n^G)^\top, \qquad F_{\mathrm{EF}} = \frac{1}{N}\sum_{n=1}^{N} w_n^G\, a_n^G (a_n^G)^\top.$$

If we interpret both matrices as empirical approximations of local old-task sensitivity directions, represented by $a_n^G(a_n^G)^\top$, then they use the same directions but differ in how strongly each sample contributes. $M_{\mathrm{IEF},G}$ keeps the original empirical weighting of the old-task samples, while F_{EF} multiplies each sample's contribution by the metric-dependent squared magnitude

$$w_n^G = r_n^\top G_n^{-1} r_n.$$

We will analyze this difference more closely below and see that this particular weighting is problematic in general. The next section treats the important special case of probabilistic classification models.

9 Gradient-Scale Effects in Classification

The output-gradient weighting in EF has an important classification-specific symptom. Consider one data sample. For cross-entropy with softmax probabilities p_n and one-hot label y_n, the output-loss gradient with respect to logits is

$$r_n = p_n - y_n.$$

For a confidently correct prediction with class c and small ε,

$$p_{n,c} = 1 - \varepsilon,$$

we have

$$\|r_n\|_2^2 = \|p_n - y_n\|_2^2 \in \Theta(\varepsilon^2).$$

Thus the EF contribution of this sample becomes small as the model becomes more confident and correct. This is coherent with the way we derived the EF: the scalar first-order loss change becomes small. In continual learning, this is undesirable: The samples that the old model classifies best may receive very low importance. EWC-DR identifies this failure mode in classification and proposes a classification-specific repair through a logit-reversal operation, literally flipping the signs of the model's logits [26]. In the present framework, their observation can be read as a symptom of the weighting issue identified in Section 8: F_{EF} importance depends on output-gradient scale. This evidently does not occur for M_{IEF}, and we make this point precise in the next section. Hence, the normalization analysis yields a formal problem identification and its generalizable solution, the Improved Empirical Fisher matrix. Additionally, this analysis allows us to make our statement about weighting more precise: EF changes the old-task weighting toward high-loss samples and away from confidently fitted ones.

10 Behavior Near Well-Fitted Solutions

The loss scale dependence of F_{EF} is further accentuated by the fact that we calculate F_{EF} *after* training on A, where we expect the model to have fitted the data well. In fact, this approximate optimality of $\theta_A^\star$ is a necessary assumption in the derivation of EF-EWC [8] to relate the true Fisher to the Hessian. This also matches modern practice, where highly expressive models are often trained to fit the data very well, frequently up to interpolation [27]. This justifies theoretical treatment outside of the classification case.

To make the comparison precise, consider the per-sample summands

$$S_n^{\mathrm{EF}} = J_n^\top r_n r_n^\top J_n, \qquad S_n^{\mathrm{IEF},G} = J_n^\top \frac{r_n r_n^\top}{r_n^\top G_n^{-1} r_n} J_n,$$

where $G_n \succ 0$ is the chosen output metric. We assume that G_n is uniformly positive definite and bounded, and that J_n is bounded under the operator norm. More precisely, assume there exists a constant $C_J < \infty$ such that

$$\|J_n^\top v\|_2 \le C_J \|v\|_2$$

for the relevant directions, and constants $0 < m_G \le M_G < \infty$ such that

$$m_G I \preceq G_n \preceq M_G I.$$

Using the same G_n-normalized direction as above,

$$\hat{r}_n^G = \frac{r_n}{\sqrt{r_n^\top G_n^{-1} r_n}}, \qquad a_n^G = J_n^\top \hat{r}_n^G, \qquad w_n^G = r_n^\top G_n^{-1} r_n, \qquad b_n^G = \|J_n^\top \hat{r}_n^G\|_2^2,$$

This gives

$$S_n^{\mathrm{IEF},G} = a_n^G (a_n^G)^\top, \qquad S_n^{\mathrm{EF}} = w_n^G \, a_n^G (a_n^G)^\top.$$

For the spectral norm,

$$\|S_n^{\mathrm{EF}}\|_2 = w_n^G b_n^G, \qquad \|S_n^{\mathrm{IEF},G}\|_2 = b_n^G.$$

Since

$$\|\hat{r}_n^G\|_2^2 = \frac{\|r_n\|_2^2}{r_n^\top G_n^{-1} r_n},$$

the metric bounds give

$$m_G \le \|\hat{r}_n^G\|_2^2 \le M_G.$$

Thus the bounded-Jacobian assumption gives

$$b_n^G \le C_J^2 M_G,$$

so

$$\|S_n^{\mathrm{EF}}\|_2 \in O(w_n^G), \qquad \|S_n^{\mathrm{IEF},G}\|_2 \in O(1).$$

The sharper lower-bound statement requires a nondegeneracy condition in the normalized output-gradient direction. If there exists $c_J > 0$ such that

$$\|J_n^\top \hat{r}_n^G\|_2 \ge c_J$$

for the relevant samples, then

$$\|S_n^{\mathrm{EF}}\|_2 \in \Theta(w_n^G), \qquad \|S_n^{\mathrm{IEF},G}\|_2 \in \Theta(1).$$

We now interpret this for practical situations. Near the end of training on distribution A, we expect many old samples to have small gradients. The norm analysis above then shows that EF contains the additional multiplicative factor $w_n^G = r_n^\top G_n^{-1} r_n$, whereas IEF removes that factor. Remaining low-magnitude summands of the IEF come from the model sensitivity term b_n^G, not from the scale of the output-loss gradient itself. In the classification setting with $G_n = I$, a confidently correct prediction with $p_{n,c} = 1 - \varepsilon$ yields

$$\|r_n\|_2^2 = \|p_n - y_n\|_2^2 \in \Theta(\varepsilon^2).$$

Therefore

$$\|S_n^{\mathrm{EF}}\|_2 = w_n^I b_n^I \in \Theta(\varepsilon^2)\, b_n^I, \qquad \|S_n^{\mathrm{IEF},I}\|_2 = b_n^I.$$

Under the same nondegeneracy condition on $J_n^\top \hat{r}_n^I$, this gives $\|S_n^{\mathrm{EF}}\|_2 \in \Theta(\varepsilon^2)$ and $\|S_n^{\mathrm{IEF},I}\|_2 \in \Theta(1)$.

The singular case $r_n = 0$ remains and must be handled explicitly. As done in [23], one practical solution is to add a damping factor τ to the quotient, yielding

$$M_{\mathrm{IEF},G} = \frac{1}{N} \sum_{n=1}^{N} \frac{g_n g_n^\top}{r_n^\top G_n^{-1} r_n + \tau}, \qquad g_n = J_n^\top r_n.$$

11 Output-Space Geometry

The original EWC derivation is tied to the true Fisher matrix, and therefore to the local KL geometry of probabilistic models [8]. From that perspective, one might conclude that EWC is theoretically tied to KL-type distances and therefore mismatched with settings that require different metrics. This concern is especially relevant in modern generative modeling, where Wasserstein-type structure appears explicitly in parts of the flow-matching and optimal-transport literature [28, 29]. However, our derivation of the empirical Fisher starting from **PL-CLOC** does not require compatibility with KL geometry. In fact, our derivation of F_{EF} sidesteps this problem by not relying on any output-space distance metric at all. This matches empirical results, where the empirical Fisher is useful in diffusion settings too [17]. In contrast, the preceding derivation and resulting form of the IEF is compatible with arbitrary output metrics. Concretely, the output-level derivation only requires a local metric $G_n \succ 0$ on output displacements. In practice, G_n should be read as a quadratic surrogate of the output-space distance one wants to preserve local to the point $f_{\theta^\star}(a_n)$. For the mathematical relation of M_{IEF} to G_n-based distance, see Section 5. The canonical Euclidean choice $G_n = I$ recovers the algebraic form from [23]. For regression or NLL-based classification setups, this may already be the right local model. Other settings warrant other choices: if the outputs lie on a probability simplex, one may instead choose a local statistical geometry. If the outputs are elements of a distribution, a Wasserstein-type local geometry may be more appropriate; our experiments use a sliced-Wasserstein surrogate for this purpose [30].

12 Additional Theoretical Properties

We move further analysis of M_{IEF} into the appendix. Namely, we treat invariance to monotone relabelings of the loss and better hyperparameter interpretability in Appendix A, and analyze rank stability of the two matrices subject to a simple contamination model in Appendix B.

13 Improved Elastic Weight Consolidation

After the theoretical analysis of M_{IEF}, we turn to the remaining design choices needed for an implementation. First, in Section 10 we already identified the singular case $r_n = 0$ and introduced damping with parameter τ as the practical

remedy. Second, using a full matrix is usually prohibitively expensive in memory and computation, so an approximation is necessary. [8] suggested a diagonal approximation, which we adopt, while noting the empirical evidence from diffusion continual learning that indicates that diagonal Fisher approximations can underperform in diffusion settings [17]. In addition, the contamination model of Appendix B suggests that the leading eigenspace of M_{IEF} can be more stable than that of F_{EF} in settings with atypical samples or samples that the model did not learn. Section 8 identified the empirical Fisher's implicit sample-weighting profile. Since alternative aggregation weights can be useful in some continual-learning settings, we include an explicit weighting function; possible weighting mechanisms appear in FROMP and GSS [22, 31]. Therefore, let

$$h : \mathcal{D}_A \times \Theta \to \mathbb{R}_{\geq 0}, \qquad a \mapsto h(a, \theta),$$

be the weighting function that assigns each sample its corresponding sample importance coefficient. To keep λ as the global regularization strength, normalize the coefficients as

$$\bar{h}(a_i, \theta) = \frac{h(a_i, \theta)}{\frac{1}{N_A} \sum_{j=1}^{N_A} h(a_j, \theta)},$$

In P-CLOC, this is simply absorbed into the penalty function. We recover the IEF from [23] with

$$\bar{h}(a_i, \theta) = 1,$$

The empirical Fisher corresponds instead to the unnormalized weighting profile

$$h(a_i, \theta) = w_i^G = r_i^\top G_i^{-1} r_i.$$

With mean-normalized $\bar{h}$, this only differs by a global scale that can be absorbed into λ. This extension yields the final form of the importance matrix that we use:

$$M_{\mathrm{IEF},G} = \frac{1}{N} \sum_{n=1}^{N} \bar{h}(a_n, \theta) \frac{g_n g_n^\top}{r_n^\top G_n^{-1} r_n + \tau}, \qquad g_n = J_n^\top r_n.$$

The free parameters of an implementation can be described more precisely as the four-tuple

$$(\lambda, \tau, \Gamma, h).$$

Here λ is the strength of the preservation constraint from CLOC, τ controls damping for small-gradient samples, h maps samples to their importance coefficient and Γ specifies the output geometry. Concretely, we treat G_n as the samplewise metric induced by the chosen geometry model Γ at the old output. Γ maps an output to the positive-definite matrix that constitutes the local quadratic approximation of the distance:

$$\Gamma : \mathcal{Z} \to \mathbb{S}_{++}^m, \qquad G_n = \Gamma\big(f_{\theta_A^\star}(a_n)\big).$$

As motivated earlier, the canonical choice is to take $\Gamma(\cdot) = I$, $\tau \approx 0$ and $h(\cdot, \cdot) = 1$. Empirically, $\tau = 10^{-2}$ is a practical default for the diagonal CIFAR-100 runs, and Appendix C shows that different values measurably affect forgetting. Conversely, the remaining parameter is then λ that intuitively controls the strength of the solution preservation constraint. For low-rank approximations of M_{IEF}, one also has to determine the rank k, turning this into a five-tuple. **Together, these derivations and implementation choices define an implementable mechanism derived from PL-CLOC: EWC arises from scalar loss preservation, and IEWC extends this mechanism through normalization, damping, sample weighting, and output-geometry selection. We call this mechanism Improved Elastic Weight Consolidation (IEWC)**.

14 Empirical Evaluation

Unless stated otherwise, we instantiate a canonical IEWC configuration: $(\lambda, \tau, \Gamma, h) = (\lambda, 10^{-2}, I, 1)$. Here λ is task-dependent, $\tau = 10^{-2}$ is selected from a CIFAR-100 sensitivity sweep, $\Gamma = I$ gives a Euclidean output geometry, and $h = 1$ uses unweighted empirical aggregation over the IEF summands. The task-specific λ values, comparator grids, and selection procedure are given in Appendix C.

Continual Learning Across Domains The classification benchmark uses FACIL on CIFAR-100 with the iCaRL class order [32], ten class-incremental distributions, ResNet-32, and task-aware evaluation. Task-aware accuracy means that the distribution identity is known at evaluation time, so predictions are compared only among the classes belonging to the current distribution. The remaining experiments use phase-shift regression, class-conditional denoising diffusion probabilistic model (DDPM) denoising on MNIST digits [33], and VOC2012 binary semantic segmentation with

Task type	Metric	Sequential	EF-EWC	EWC-DR	IEWC	IEWC-SW
Classification	Final avg. TAw accuracy ↑	0.3974 ± 0.0294	0.5617 ± 0.0062	0.5932 ± 0.0296	**0.5978 ± 0.0260**	–
Classification	Avg. TAw forgetting ↓	0.4783 ± 0.0325	0.1997 ± 0.0107	0.1039 ± 0.0085	**0.0624 ± 0.0025**	–
Regression	Final avg. MSE across distributions ↓	0.2243 ± 0.0021	0.2199 ± 0.0043	–	**0.2013 ± 0.0047**	–
Regression	Old-distribution MSE after new distribution ↓	0.4484 ± 0.0041	0.4374 ± 0.0075	–	**0.3951 ± 0.0107**	–
Regression	Forgetting (MSE increase) ↓	0.4479 ± 0.0039	0.4369 ± 0.0076	–	**0.3946 ± 0.0109**	–
Regression	New-distribution MSE ↓	**1.71e-04 ± 1.43e-04**	0.0024 ± 0.0020	–	0.0075 ± 0.0016	–
Diffusion	Final avg. denoising MSE across distributions ↓	0.0622 ± 0.0059	0.0314 ± 0.0028	–	**0.0310 ± 0.0027**	0.0313 ± 0.0012
Diffusion	Old-distribution denoising MSE after new distribution ↓	0.0705 ± 0.0086	0.0344 ± 0.0049	–	0.0332 ± 0.0045	**0.0279 ± 4.60e-04**
Diffusion	Forgetting (MSE increase) ↓	0.0420 ± 0.0095	0.0058 ± 0.0066	–	0.0048 ± 0.0061	**-6.52e-04 ± 0.0020**
Diffusion	New-distribution denoising MSE ↓	0.0540 ± 0.0041	**0.0284 ± 0.0019**	–	0.0288 ± 0.0021	0.0348 ± 0.0023
Segmentation	Final avg. foreground IoU across class sets ↑	0.3708 ± 0.0331	**0.3754 ± 0.0211**	–	0.3735 ± 0.0176	–
Segmentation	Old class-set foreground IoU after new class set ↑	0.2623 ± 0.0360	0.2798 ± 0.0531	–	**0.2882 ± 0.0508**	–
Segmentation	Forgetting (IoU decrease) ↓	0.1633 ± 0.0273	0.1517 ± 0.0480	–	**0.1371 ± 0.0463**	–
Segmentation	New class-set foreground IoU ↑	**0.4792 ± 0.0302**	0.4709 ± 0.0158	–	0.4587 ± 0.0157	–

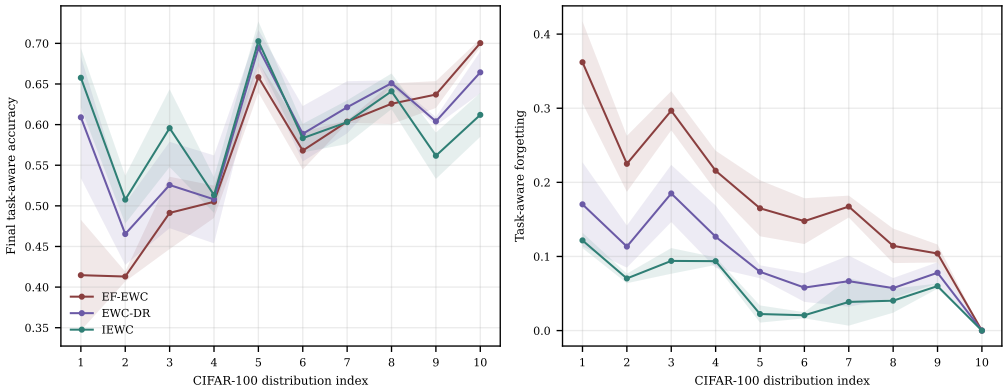

Figure 1: Classification distribution retention. IEWC keeps higher final accuracy than EF-EWC on most distributions and has much lower forgetting, while EWC-DR also improves over EF-EWC on average forgetting.

animal and vehicle foreground class sets. Classification and segmentation use three seeds; the controlled regression and diffusion experiments use five seeds. Unless stated otherwise, the EWC matrices are diagonal. EF-EWC is tuned over the tested regularization grid for old-distribution retention; EWC-DR is shown only for classification because logits reversal is classification-specific.

Here IEWC-SW denotes IEWC with the sliced-Wasserstein output metric. The main comparison is the retention row within each task type. IEWC improves old-distribution retention over EF-EWC in classification, regression, diffusion, and segmentation, and improves final task-aware accuracy on CIFAR-100. After tuning, EWC-DR is competitive on CIFAR-100, while IEWC gives lower average forgetting in these runs. In diffusion, tuned EF-EWC and IEWC are close on denoising MSE, while IEWC-SW gives the strongest old-distribution retention at a visible plasticity cost.

Additionally, Appendix C contains the exact training protocol, the CIFAR-100 tau-sensitivity curve, and the EWC-DR parameter check. Appendix D uses controlled label noise as an empirical analogue of the contamination model in Appendix B, comparing the resulting accuracy and diagonal-matrix stability. Appendix E reports qualitative DDPM samples, and Appendix F records the rejected sample-weighting variants.

Code Release

We release an accompanying repository with a readily usable IEWC implementation. The `IEWCConfig` class encapsulates the four-tuple parametrization introduced above. The same configuration interface is used by the diagonal and low-rank IEWC paths and by the empirical scripts, with a FACIL-based entry point for the CIFAR-100 continual classification experiments. The repository is available at `https://github.com/Axym-Labs/iewc`.

15 Conclusion

We formulated continual learning as constrained optimization through CLOC and its population-setting, low-order variants. Within this framework, the empirical Fisher matrix used in EF-EWC is recovered exactly if we choose preserving the old loss value as constraint. This places the empirical Fisher matrix on solid theoretical grounding without relying on the true Fisher matrix or KL-type geometry. The same framework also instantiates preservation

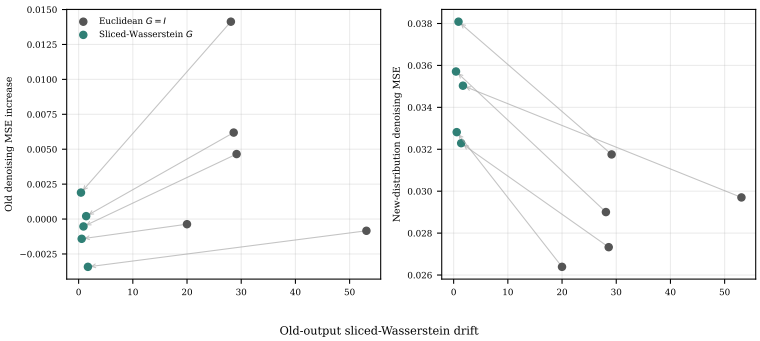

Figure 2: Diffusion output-geometry tradeoff. The diffusion geometry experiment replaces the Euclidean output metric in the IEWC denominator with a sliced-Wasserstein surrogate over denoised MNIST images, changing which local output displacements the quadratic penalty treats as large. Across five seeds, sliced-Wasserstein IEWC reduces old-output sliced-Wasserstein drift from 31.7878 ± 12.4909 to 0.9981 ± 0.5442. The paired-seed arrows show the resulting geometry-selective stability tradeoff: the old denoising map moves much less in sliced-Wasserstein geometry, while new-distribution denoising MSE increases.

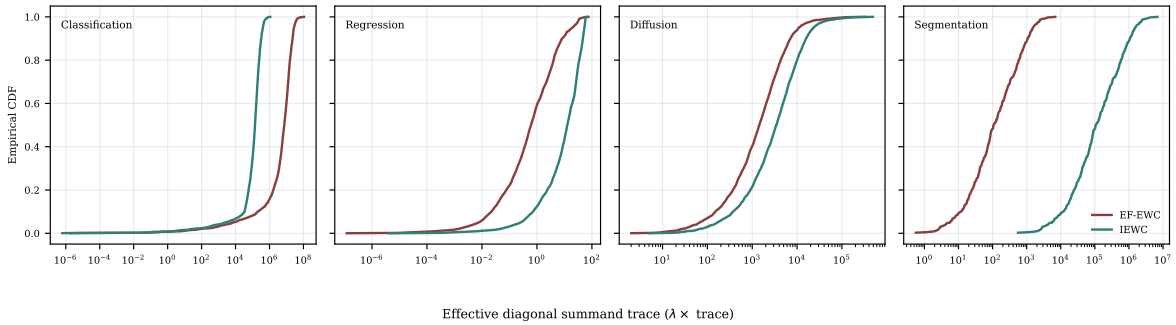

Figure 3: Effective diagonal summand norms. Each panel reports the empirical CDF of λ times the per-sample diagonal summand trace before averaging over old-distribution samples, with EF-EWC and IEWC shown on the same task type.

as distance to the old-task loss level set in output space; after first-order approximation, the resulting importance matrix is the improved empirical Fisher. A central result of our analysis is that the classification-specific failure mode addressed by EWC-DR is not an isolated pathology. Instead, it is one manifestation of a more general effect: EF-EWC scales old sample importance by the magnitude of the output loss gradient, suppressing samples that are already well fit. We show this explicitly in Θ-Notation. The improved empirical Fisher removes this dependence and thereby yields a general solution that extends beyond classification to well-fitted old-task settings and arbitrary local output geometries. These design choices define a concrete continual-learning mechanism derived from first principles. In particular, this yields IEWC as a practical successor to diagonal Fisher regularization in EF-EWC. Across classification, regression, diffusion, and segmentation experiments, IEWC improves old-distribution retention over EF-EWC; the sliced-Wasserstein diffusion experiment further illustrates that the output-geometry parameter can target a chosen notion of stability.

References

[1] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24. Academic Press, 1989.

[2] Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4), 1999.

[3] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 2019.

[4] Gido M. van de Ven and Andreas S. Tolias. Three scenarios for continual learning, 2019.

[5] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3), 2014.

[6] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *NeurIPS*, 2017.

[7] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. In *ICLR*, 2019.

[8] James Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 2017.

[9] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *ICML*, 2017.

[10] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, 2018.

[11] Ferenc Huszar. Note on the quadratic penalties in elastic weight consolidation. *Proceedings of the National Academy of Sciences*, 115(11), 2018.

[12] Jonathan Schwarz, Jelena Luketina, Wojciech M. Czarnecki, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *ICML*, 2018.

[13] Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *ICLR*, 2018.

[14] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 1998.

[15] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21, 2020.

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

[17] Z. Wang, A. Gupta, Z. Dong, and C. J. MacLellan. Avoid catastrophic forgetting with rank-1 fisher from diffusion models, 2025.

[18] Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent. In *NeurIPS*, 2019.

[19] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *ECCV*, 2016.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[21] Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *ICLR*, 2020.

[22] Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard E. Turner, and Mohammad Emtiyaz Khan. Continual deep learning by functional regularisation of memorable past. In *NeurIPS*, 2020.

[23] X. Wu, W. Yu, C. Zhang, and P. Woodland. An improved empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, 2024.

[24] Nicol N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7), 2002.

[25] James Martens. Deep learning via hessian-free optimization. In *ICML*, 2010.

[26] X. Liu and X. Chang. Elastic weight consolidation done right for continual learning. In *CVPR*, 2026.

[27] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 2019.

[28] D. Haviv, A.-A. Pooladian, D. Pe'er, and B. Amos. Wasserstein flow matching: Generative modeling over families of distributions. In *ICML*, 2025.

[29] Gabriel Peyre and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6), 2019.

[30] Nicolas Bonneel, Julien Rabin, Gabriel Peyre, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1), 2015.

[31] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *NeurIPS*, 2019.

[32] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017.

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[34] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *ICML*, 2017.

[35] Yuan Cao, Zhiying Fang, Yue Wu, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning, 2020.

A Invariance to Loss Reparameterization

This appendix treats the undamped case $\tau = 0$.

The output-level feasible set

$$\mathcal{C}_n = \{z : \ell_n(z) = \ell_n(z_n^A)\}$$

depends on the level sets of ℓ_n, not on the numerical labels assigned to those level sets. If we apply a local monotone scalar transformation

$$\tilde{\ell} = \phi(\ell), \qquad \phi'(\ell) > 0,$$

then the output gradient rescales as

$$\tilde{r}_n = \phi'(\ell_n(z_n^A))\, r_n.$$

For EF, the sample contribution changes by

$$J_n^\top \tilde{r}_n \tilde{r}_n^\top J_n = \phi'(\ell_n(z_n^A))^2 J_n^\top r_n r_n^\top J_n.$$

For IEF, the same scalar factor appears in numerator and denominator:

$$\frac{\tilde{r}_n \tilde{r}_n^\top}{\tilde{r}_n^\top G_n^{-1} \tilde{r}_n} = \frac{r_n r_n^\top}{r_n^\top G_n^{-1} r_n}.$$

Thus IEF is invariant to local scalar relabelings of the constraint function. This includes loss rescaling, unit changes in regression losses, reduction conventions, and smooth monotone transformations that preserve the local level-set geometry. This invariance can also improve hyperparameter interpretability, as EF's penalty scale depends on the loss magnitudes. IEF removes this dependence, while regularization strength and damping still require tuning.

B Spectral Stability Under Contamination

The normalization difference between EF and IEF has consequences for their spectrum. This matters in practice because any diagonal or low-rank approximation is built from the leading part of the matrix. Consider for example [17]. To see how this difference can affect the leading eigenspace, consider the following simple contamination model: Contaminated samples, such as mislabeled or atypical data, can have large output-loss gradients because their signal aligns with directions that gradient descent learns only very slowly; in the spectral-bias / kernel-eigenmode framework, low-alignment or low-eigenvalue components can remain effectively unlearned over the training horizon [34, 35]. With probability $1 - \varepsilon$, a sample is clean and its normalized sensitivity direction contributes to a stable old-task preservation matrix $\Sigma_{\text{clean}} = \mathbb{E}[aa^\top \mid a \in D_{A,\text{clean}}]$. Assume also that clean weights are independent of these normalized directions, or are sufficiently concentrated around their clean mean. With probability ε, a sample is contaminated. Assume its direction is concentrated near a unit vector u outside that clean leading subspace, and that its output gradient scale is large,

$$w \approx W, \qquad W \gg 1.$$

Under this model, the population version of IEF is

$$\mathbb{E}[M_{\text{IEF}}] = (1 - \varepsilon)\Sigma_{\text{clean}} + \varepsilon u u^\top$$

while the corresponding EF matrix becomes

$$\mathbb{E}[F_{\mathrm{EF}}] = (1-\varepsilon)\mathbb{E}[w \mid \text{clean}]\,\Sigma_{\text{clean}} + \varepsilon W u u^{\top}.$$

In terms of the spectrum, let λ_k^{clean} denote the k-th eigenvalue of the clean component. If

$$\varepsilon W \gtrsim (1-\varepsilon)\lambda_k^{\text{clean}}\mathbb{E}[w \mid \text{clean}],$$

then the contaminated direction u will enter the leading eigenspace of F_{EF}. The insight is that the normalization can realistically affect which directions dominate the spectrum. In settings with noisy labels, hard examples, underfit outliers, or other rare but high-loss samples, we should expect the leading eigenspace of F_{EF} to be less stable than that of M_{IEF}, relevant to low-rank approximations thereof.

C Experimental Details and Damping Sensitivity

Empirical Setup Details The ten-distribution CIFAR-100 classification runs use FACIL, the iCaRL class order, ResNet-32, SGD with momentum 0.9, learning rate 0.05, weight decay 0.0002, batch size 128, 60 epochs per distribution, $\lambda = 10000$, and 512 old-distribution importance samples. Diagonal IEWC uses $\tau = 10^{-2}$ with EF trace matching. EF-EWC uses the best setting from the tested regularization grid for each task type. The controlled regression and diffusion experiments use five seeds and diagonal importances. IEWC uses $\lambda = 1$ for regression and $\lambda = 25$ for MNIST DDPM diffusion; the selected EF-EWC comparator values are $\lambda = 30$ and $\lambda = 10000$, respectively. The segmentation runs use VOC2012, binary animal/vehicle class-set foreground masks, a small U-Net, three seeds, and the tuned values $\lambda = 1$ for EF-EWC and $\lambda = 10$ for IEWC. The diffusion output-geometry analysis uses the same MNIST DDPM denoising setup and sliced-Wasserstein IEWC with $\lambda = 10^7$. The qualitative diffusion samples use the same MNIST architecture with a longer EMA training budget for sample quality. The summand-norm analysis uses stored diagonal per-sample traces from the matched CIFAR-100 prefix run and from the accepted regression, diffusion, and segmentation runs, multiplied by the selected λ for each method and task type.

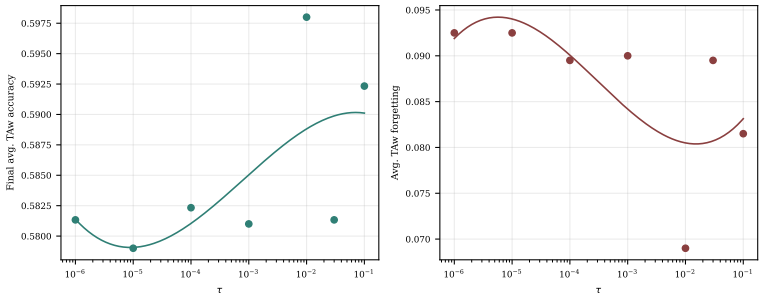

Figure 4: Tau sensitivity on CIFAR-100. We use FACIL with CIFAR-100, the iCaRL class order, ResNet-32, and the first three distributions of the ten-distribution task-aware protocol. Each distribution is trained for 60 epochs. IEWC uses a diagonal matrix surrogate, 512 old-distribution samples for the importance estimate, and importances scale-matched to the corresponding EF estimate so that the EF regularization strength is reused. The tested values are $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 3 \cdot 10^{-2}, 10^{-1}\}$. The damping parameter has a measurable effect on this prefix protocol: $\tau = 10^{-2}$ gives the highest final task-aware accuracy and lowest forgetting among the tested values, so we use it as the practical default in the diagonal CIFAR-100 runs.

For EWC-DR, we also checked $\lambda \in \{100, 300, 1000, 3000, 10000\}$ on the same three-distribution CIFAR-100 prefix. The best tested value on this prefix was $\lambda = 100$; we use this check only as a local parameter-sensitivity check.

D Label-Noise Stress Test

The contamination model in Appendix B predicts that rare high-loss samples can dominate EF-EWC because their loss-gradient scales enter the empirical Fisher directly. Controlled label noise instantiates this mechanism: corrupted labels create atypical high-loss training samples, while clean task-aware test sets measure whether the stored preservation matrix still protects the original classification structure. We therefore add uniform label noise to the CIFAR-100 training data in FACIL and report final average task-aware accuracy. IEWC degrades much more slowly than EF-EWC as the label-noise rate increases.

Method	$p = 0$	$p = 0.1$	$p = 0.25$
EF-EWC	0.5617 ± 0.0062	0.5132 ± 0.0112	0.2773 ± 0.0320
EWC-DR	0.5932 ± 0.0296	0.5310 ± 0.0226	0.4598 ± 0.0095
IEWC	0.5978 ± 0.0260	0.5642 ± 0.0215	0.5322 ± 0.0239

Method	Noise	Entries	Stored-tail mass ratio	Stored-tail profile L1	Stored-tail log-scale RMSE
EF-EWC	0.1	512	0.0288	0.2610	1.4660
EF-EWC	0.25	512	0.0244	0.2737	1.5337
IEWC	0.1	512	1.0833	0.0350	0.0350
IEWC	0.25	512	1.1003	0.0154	0.0415

To connect this performance effect to the matrix-level prediction of the contamination model, we also record the largest stored diagonal entries of the EF and IEWC matrices on the same CIFAR-100 prefix protocol. These are not low-rank approximations. For diagonal matrices, the relevant stability object is the high-importance coordinate tail, so the table uses all 512 exported entries per matrix and reports the stored-tail mass ratio, the normalized stored-tail profile change, and the log-scale RMSE of the sorted entries. EF-EWC changes much more under contamination than IEWC on the log-scale metric.

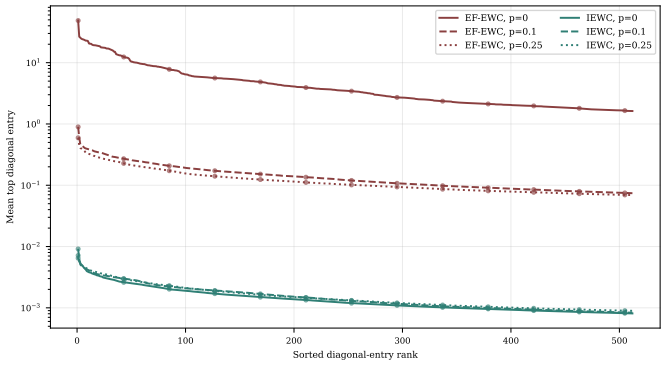

Figure 5: Label-noise diagonal spectra. The plot visualizes the high-importance coordinate tail underlying the matrix-level comparison above; EF-EWC changes much more under contamination than IEWC on the log scale.

E Qualitative DDPM Samples

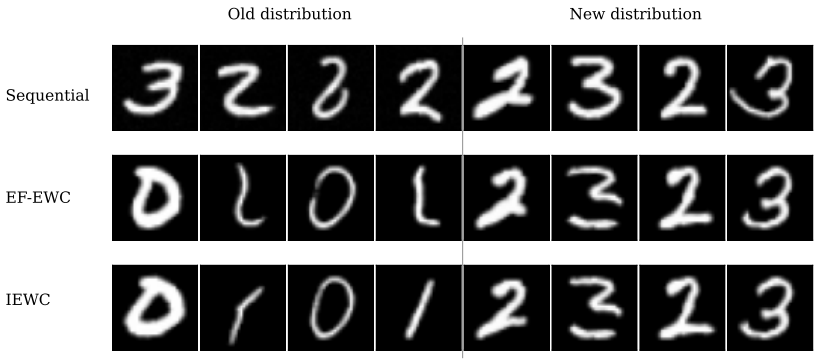

Figure 6: Diffusion generated samples. The grid shows class-conditional MNIST generations from the DDPM architecture after a longer EMA training budget and continual training on both distributions. Rows compare sequential training, EF-EWC, and IEWC; columns show four old-distribution samples and four new-distribution samples. The old column alternates labels 0 and 1, and the new column alternates labels 2 and 3. Sequential training produces recognizable new-distribution digits but old-label samples drift toward new-distribution digits. EF-EWC and IEWC preserve old-label generations while learning recognizable new-label samples. This MNIST setting does not support the pattern reported in Rank1Diffusion, where EWC alone struggles in diffusion continual learning and diagonal EWC is ineffective relative to rank-1 Fisher regularization with generative distillation [17].

F Additional Sample-Weighting Experiments

The formulation in Section 13 allows an optional sample-weighting function h, while the accepted experiments use $h_n = 1$. We tested two faithful IEWC weighting adaptations and one diagonal surrogate. The GSS adaptation used the residual-novelty score $\|(I - P_S)a_n^G\|_2^2$ on normalized IEWC directions. The FROMP adaptation used the predictive Hessian trace of the classifier output as the sample weight. The row-Frobenius diagonal surrogate set diagonal scale from each normalized direction's outer-product mass. In initial experiments, none improved downstream performance after scale matching. We therefore keep $h_n = 1$ in the main experiments, and note that h_n retains its use of capturing the implicit weights of the empirical Fisher in our framework.