Continual Learning (CL) — Solution Blueprint

Classification with Concept Drift and Clear Task Boundaries

Data Scientist Postdoc Interview — Thomas Jefferson National Accelerator Facility

September 18, 2025

1. Problem Setup

We observe a sequence of T tasks with clear boundaries. Each task $\tau \in \{1, \ldots, T\}$ provides labeled data $\mathcal{D}_\tau = \{(x, y)\}_{i=1}^{n_\tau}$ from distribution $p_\tau(x, y)$. We train a *single* classifier $f_\theta : \mathcal{X} \to \Delta^C$ sequentially and cannot revisit full historical datasets. Goal: maximize performance across all tasks while minimizing forgetting.

Catastrophic forgetting. Let $R_{t,i}$ denote accuracy on task i after finishing training on task t. Forgetting occurs when $R_{T,i} \ll R_{i,i}$ for some past task $i < t$.

2. Chosen Approach

A lightweight hybrid: **Replay + Knowledge Distillation (KD) + Online EWC**.

- **Replay.** Maintain exemplar buffer $\mathcal{B}$ (fixed memory $|\mathcal{B}| \le M$); mix current and replayed samples.
- **KD (LwF).** Preserve function behavior by distilling from a frozen teacher $f_{\theta^\star}$ (previous snapshot).
- **Online EWC.** Apply Fisher-weighted quadratic penalty with a decayed importance estimate Ω.

3. Objective Function (per minibatch)

For a mixed minibatch $\mathcal{M} = \mathcal{M}_{\text{curr}} \cup \mathcal{M}_{\text{replay}}$ and a teacher snapshot $\theta^\star$:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\mathcal{M}|} \sum_{(x,y)\in\mathcal{M}} \sum_{c=1}^{C} \mathbf{1}[y = c] \log p_\theta(c \,|\, x), \tag{1}$$

$$\mathcal{L}_{\text{KD}} = \frac{T^2}{|\mathcal{M}|} \sum_{x\in\mathcal{M}} \mathrm{KL}\big(p^T_{\theta^\star}(\cdot \,|\, x) \,\|\, p^T_\theta(\cdot \,|\, x)\big)\,, \tag{2}$$

$$\mathcal{L}_{\text{EWC}} = \sum_i \frac{\lambda_{\text{EWC}}}{2}\, \Omega_i\, (\theta_i - \theta_i^\star)^2, \tag{3}$$

$$\boxed{\mathcal{L} \;=\; \mathcal{L}_{\text{CE}} + \lambda_{\text{KD}}\mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{EWC}}} \tag{4}$$

where $p^T_\theta(\cdot \,|\, x) = \mathrm{softmax}\big(z_\theta(x)/T\big)$ and $T > 0$ is the distillation temperature.

4. Training Loop (per task τ)

1. **Freeze teacher:** $\theta^\star \leftarrow \theta$.
2. **Batch mixing:** sample a minibatch with ratio α current vs. $(1-\alpha)$ replay (default $\alpha = 0.8$).
3. **Optimize** $\mathcal{L}$ with AdamW/SGD; early stopping (patience 3).
4. **Update buffer** $\mathcal{B}$ with class-balanced selection (e.g., herding/reservoir) under budget M.
5. **Update Online EWC:** estimate Fisher $\Omega^{(\tau)}$ on a few batches; merge $\Omega \leftarrow \gamma\Omega + (1-\gamma)\Omega^{(\tau)}$ with decay $\gamma \in [0, 1)$.

Default knobs. $\alpha = 0.8$ (current) / 0.2 (replay), $M = 100$ per class (adjust to memory), $T = 2$, $\lambda_{\text{KD}} = 0.5$, $\lambda_{\text{EWC}} = 50$, $\gamma = 0.9$, 10–30 epochs per task with early stopping.

5. Evaluation Protocol

Build an accuracy matrix $R \in \mathbb{R}^{T\times T}$ with $R_{t,i}$ accuracy on task i after finishing task t.

$$\text{AvgAcc} = \frac{1}{T}\sum_{i=1}^{T} R_{T,i}, \tag{5}$$

$$\text{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1} \big(R_{T,i} - R_{i,i}\big), \tag{6}$$

$$\text{FWT} = \frac{1}{T-1}\sum_{i=2}^{T} \big(R_{i-1,i} - R_{0,i}\big). \tag{7}$$

Also report footprint: buffer size (MB), model size, train/infer latency.

Baselines / Ablations. Naïve fine-tune; Replay-only; KD-only; Replay+KD; Replay+KD+Online EWC (ours-robust).

6. Minimal Toy Example (Explainable)

Split MNIST. Task 1: digits 0–4. Task 2: digits 5–9. Naïve fine-tune forgets Task 1 (BWT $\ll 0$). Replay+KD(+EWC) maintains Task 1 while learning Task 2 (BWT ≈ 0).

7. Risks & Mitigations

- **Privacy.** If raw replay is disallowed, use *feature replay* (store embeddings) or *generative replay*.
- **Imbalance.** Enforce class-balanced buffer quotas and sampling.
- **Strong conflicts.** Add small adapters per task or orthogonal gradient constraints.