Parabolic Continual Learning

Haoming Yang [1] Ali Hasan [1,2] Vahid Tarokh [1]

[1]Duke University [2]Morgan Stanley

Motivation

A continual learner should have the following properties:

- knowledge retention – the learner should maintain knowledge of previous tasks;
- generalization – the learner should calibrate to new data observations and distributions.

This motivates us to study an evolution of the loss function such that the learner can control:

- the rate at which information from previous data is retained;
- the rate at which the loss accumulates for any new data within a predefined region of the data space for future time points.

The Parabolic PDE

We let $u(x,t) = \mathbb{E}[\ell_{f_\theta}(x_t)]$, and assume we have access to a memory buffer $\mathcal{M}$. For each training iteration t we solve the following optimization

$$\min_{\theta_t} \mathbb{E}_{x\sim P(\mathcal{M}_t)}\left[\ell_{f_{\theta_t}}(x)\right]$$

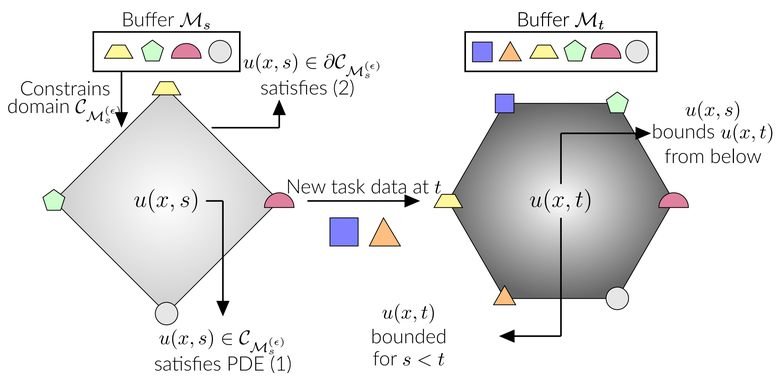
such that $u(x,t)$ will satisfy both (1) and (2)

$$\frac{\partial u}{\partial t} = \sigma\nabla^2 u(x,t) + \ell_{f_\theta}(x) \quad on \quad x,t \in \mathcal{C}_{\mathcal{M}_t^{(\epsilon)}}/\mathcal{M}_t^{(\epsilon)}, \tag{1}$$

$$u(x,t) = \ell_{f_\theta}(x) \quad on \quad x,t \in \mathcal{M}_t^{(\epsilon)} \tag{2}$$

In the above $\mathcal{M}^{(\epsilon)}$ denotes the ϵ-ball expansion of the countable set in $\mathcal{M}$ given by $\mathcal{M}^{(\epsilon)} = \cup_{x^{(i)}\in\mathcal{M}}\{x \mid \|x - x^{(i)}\|_2^2 \le \epsilon\}$ in order to provide a meaningful boundary, and $\mathcal{C}_\mathcal{M} \subset \mathcal{X}$ denotes the convex hull of $\mathcal{M}$ and provides a compact space to solve the learning problem over.

Main Workflow



Bounds On Expected Error

Using the PDE constrained optimization, we can derive the following bounds on expected forgetting error and generalization error.

Upper bound on expected forgetting error Consider $u(x_{t-\tau}, t-\tau)$ for $x_{t-\tau} \in \mathcal{C}_{\mathcal{M}_t}$ and $\tau \ge 0$. Additionally suppose that the learning landscape satisfies the following parabolic PDE

$$\frac{\partial u}{\partial t} = -\nabla^2 u(x,t) + \ell_{f_\theta}(x) \quad on \quad (x,t) \in \mathcal{C}_{\mathcal{M}_t^{(\epsilon)}}$$

Then the following inequality holds:

$$u(x,t-\tau) \le \max_{x,s\in\mathcal{M}_t^{(\epsilon)}\cup\partial\mathcal{C}_{\mathcal{M}_t^{(\epsilon)}}} u(x,s).$$

Lower bound on expected generalization error Consider $u(x_{t+\tau}, t+\tau)$ for $x_{t+\tau} \in \mathcal{C}_{\mathcal{M}_t}$ and $\tau \ge 0$. Suppose that the expected continual learning loss u satisfies the following parabolic PDE

$$\frac{\partial u}{\partial t} = \nabla^2 u(x,t) + \ell_{f_\theta}(x) \quad on \quad (x,t) \in \mathcal{C}_{\mathcal{M}^{(\epsilon)}} \times [T,\infty)$$

and that the function $\ell_{f_\theta}(x)$ is $C-$Lipshitz in x. Then the following inequality holds:

$$\min_{x,s\in\mathcal{M}_t\cup\partial\mathcal{C}_{\mathcal{M}_T^{(\epsilon)}}} u(x,s) \le u(x,t+\tau) \le C\tau + \max_{x\in\mathcal{C}_{\mathcal{M}_T^{(\epsilon)}}} u(x,s).$$

Using Feynman-Kac and Brownian Bridges

PDE in (1) can be solved by the following expectation with a simple application of Feynman-Kac where τ is the hitting time to hit any points in $\mathcal{M}^{(\epsilon)}$ [Pardoux and Raşcanu, 2014]:

$$u(x,t) = \mathbb{E}\left[\ell_{f_\theta}(X_{t\wedge\tau_{\mathcal{M}^{(\epsilon)}}}) + \int_0^{t\wedge\tau_{\mathcal{M}^{(\epsilon)}}} \ell_{f_\theta}(X_s)\mathrm{d}s \mid X_0 = x\right] \tag{3}$$

The Feynman-kac formula theoretically solves $u(x,t)$ with a Brownian motion, but to sample Brownian motions with tractable hitting times on $\epsilon-$balls is difficult.

To evaluate the expectations in (3), we use an approximation to prevent long integration times. Specifically, we compute Brownian bridges (BBs) between points in the memory buffer and the incoming data points instead of a diffusion.

Algorithm

Input: Data $X_{\text{all}}, y_{\text{all}}$, related Brownian bridge (BB) hyperparameters.
Initialize: neural network f_θ, buffer $\mathcal{M}$.
for X, y in mini-batched $X_{\text{all}}, y_{\text{all}}$ **do**
  if $\mathcal{M}$ is not empty **then**
    Sample from $\mathcal{M}$, obtain $X_\mathcal{M}, y_\mathcal{M}$.
    $X = \text{Concatenate}(X, X_\mathcal{M})$, $y = \text{Concatenate}(y, y_\mathcal{M})$.
  end if
  Obtained X', y' as BB terminal condition by shuffling X, y.
  Sample BB $X_s, y_s \sim \text{BB}_{X,y}^{X',y'}$ for arbitrary timestep s.
  Compute ℓ by integrating $\ell(f(X_s), y_s)$ using the Euler's method.
  Optimize ℓ using gradient-based optimizer
  Update Buffer $\mathcal{M}$ with reservoir sampling.
end for

Benchmarking Under Different Scenarios

We benchmark against VRMCL, the SOTA online class-incremental method based on meta-learning [Wu et al., 2024]. PCL achieved similar performance to VRMCL on regular settings but is 5× faster.

PCL is also highly robust in the following scenarios.

Label Corruption	$\lvert\mathcal{M}=200\rvert$		$\lvert\mathcal{M}=600\rvert$		$\lvert\mathcal{M}=1000\rvert$	
	AAA	Acc	AAA	Acc	AAA	Acc
VRMCL	11.32 ±0.4	5.48 ±0.9	13.39 ±0.4	7.95 ±0.5	15.16 ±0.3	8.88 ±0.5
PCL	**12.83** ±0.5	**7.05** ±0.5	13.14 ±1.1	**8.42** ±0.5	14.82 ±0.6	**9.53** ±0.7

Imbalanced CL	$\gamma=2$ Normal		$\gamma=2$ Reversed		$\gamma=2$ Random	
	AAA	Acc	AAA	Acc	AAA	Acc
VRMCL	62.13 ±2.9	50.42 ±2.4	61.41 ±3.8	49.5 ±3.0	62.31 ±2.6	50.17 ±2.3
PCL	**63.25** ±0.8	**52.73** ±1.2	**62.96** ±2.1	**51.67** ±2.9	**63.48** ±1.0	**52.09** ±1.5

Small Buffer Size	$\lvert\mathcal{M}=50\rvert$		$\lvert\mathcal{M}=100\rvert$		$\lvert\mathcal{M}=150\rvert$	
	AAA	Acc	AAA	Acc	AAA	Acc
VRMCL	9.11 ±1.2	3.36 ±0.8	9.8 ±0.2	3.95 ±0.1	11.66 ±0.6	5.18 ±0.4
PCL	**18.71** ±1.4	**9.36** ±1.7	**19.14** ±0.5	**9.71** ±0.8	**20.0** ±0.6	**10.47** ±1.2

References

E. Pardoux and A. Raşcanu. *Stochastic differential equations, backward SDEs, partial differential equations*. Stochastic Modelling and Applied Probability. Springer International Publishing, 2014. ISBN 9783319057132.

Y. Wu, L.-K. Huang, R. Wang, D. Meng, and Y. Wei. Meta continual learning revisited: Implicitly enhancing online hessian approximation via variance reduction. In *The Twelfth International Conference on Learning Representations*, 2024.

AISTATS 2025

This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-22-1-0315.

haoming.yang@duke.edu