Convergence and Implicit Bias of Gradient Descents on Continual Linear Classification

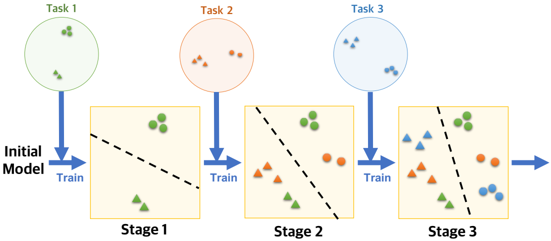
ICLR 2025 Poster

Hyunji Jung* Hanseul Cho* Chulhee Yun

2025.03.31

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY
KAIST AI
Kim Jaechul Graduate School
OptiML
Optimization & Machine Learning

Theoretical Understanding of Continual Learning



- **Practice:** Many applications based on **classification** tasks.

Theory-Practice Gap!

- **Theory:** Mostly focusing on continual (linear) **regression**.
- Q. How about usual gradient-based optimization algorithm, without projection nor regularization?

Problem Setup

- Continually learn M binary classification tasks
  - Data point $\boldsymbol{x} \in \mathbb{R}^d$, label $y \in \{-1, 1\}$
  - Linear model $f(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{x}^\top \boldsymbol{w}$
  - Prediction: $+1$ iff $y_i f(\boldsymbol{x}_i; \boldsymbol{w}) \geq 0$ (and -1 otherwise)
  - We consider partitioned index set: $I = I_1 \cup I_2 \cup \cdots \cup I_M$
  - Each task m has their own dataset $\mathcal{D}_m = (x_i, y_i)_{i \in I_m}$

- Task Ordering: which task do we solve at stage t?
  - **Cyclic task ordering.** Tasks are presented in a fixed cyclic order. That is, $m_t = t \bmod M$.
  - **Random task ordering.** The next task is independently sampled uniformly at random. That is, $m_t \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0 : M-1])$.

Problem Setup

- GOAL
  - We aim to minimize the joint training loss
    $$\mathcal{L}(\boldsymbol{w}) = \sum_{i \in I} \ell(y_i \boldsymbol{x}_i^\top \boldsymbol{w})$$
  - But we only have access to current training loss at stage t
    $$\mathcal{L}^{(t)}(\boldsymbol{w}) = \sum_{i \in I_{m_t}} \ell(y_i \boldsymbol{x}_i^\top \boldsymbol{w})$$

- Sequential Gradient Descent (or, Sequential GD)
  - For each stage $t = 0, 1, \dots,$ we update our linear classifier K times using the corresponding task m_t.

$$\boldsymbol{w}_{k+1}^{(t)} = \boldsymbol{w}_k^{(t)} - \eta \nabla \mathcal{L}^{(t)}(\boldsymbol{w}_k^{(t)}) \quad \text{for } k \in [0 : K-1].$$

  - At a new stage $t+1$, we initialize the model with the last iterate of the previous stage.

$$\boldsymbol{w}_0^{(t+1)} = \boldsymbol{w}_K^{(t)}.$$

Cyclic Ordering: Asymptotic Result

Theorem (Loss / Directional Convergence, Informal)

If the learning rate $\eta < \frac{\phi^2}{4K\beta\sigma_{\max}^3(M\phi+\sigma_{\max})}$, *then*

1. *Loss converges to zero:* $\lim_{t\to\infty}\mathcal{L}(\boldsymbol{w}_k^{(t)}) = 0$.
2. *Every data point is eventually classified correctly:* $\lim_{t\to\infty}\boldsymbol{x}_i^\top\boldsymbol{w}_k^{(t)} = \infty$.
3. *Converge in the direction of joint max-margin solution:*

$$\boldsymbol{w}_k^{(t)} = \ln\left(\frac{K}{M}t\right)\cdot\hat{\boldsymbol{w}} + \boldsymbol{\rho}_k^{(t)},$$

where $\boldsymbol{\rho}_k^{(t)}$ *is bounded.*

Non-asymptotic Convergence of Loss & Forgetting

We define Cycle-averaged Forgetting on cycle J as:

$$\mathcal{F}_{\text{cyc}}(J) := \frac{1}{M} \sum_{m=0}^{M-1} \mathcal{L}_m(\boldsymbol{w}_0^{(MJ+M)}) - \mathcal{L}_m(\boldsymbol{w}_K^{(MJ+m)})$$

Theorem (Loss & Forgetting Convergence, Informal)

Under the same conditions, we have

1. $\mathcal{L}(\boldsymbol{w}_k^{(MJ+m)}) = O\left(\frac{\ln^2(J)}{J}\right)$
2. $-\eta K \cdot A^+ \cdot O\left(\frac{\ln^4 J}{J^2}\right) \leq \mathcal{F}_{\text{cyc}}(J) \leq \eta K \cdot A^- \cdot O\left(\frac{\ln^4 J}{J^2}\right),$

where A_{Pos} & A_{Neg}: positive & negative task alignments.

Random ordering: Almost-sure Asymptotic Result

Theorem (Loss / Directional Convergence, Informal)

If the learning rate $\eta < \frac{2\phi^2}{\beta\sigma_{\max}^4}$, *then with probability 1,*

1. *Loss converges to zero:* $\lim_{t\to\infty} \mathcal{L}(\boldsymbol{w}_k^{(t)}) = 0$.
2. *Every data point is eventually classified correctly:* $\lim_{t\to\infty} \boldsymbol{x}_i^\top \boldsymbol{w}_k^{(t)} = \infty$.
3. *Converge in the direction of joint max-margin solution:*

$$\boldsymbol{w}_k^{(t)} = \ln\left(\frac{K}{M}t\right) \cdot \hat{\boldsymbol{w}} + \text{(bounded)}$$

Beyond Jointly Separable Tasks

Theorem (Iterate Convergence, Informal)

Suppose we learn M tasks cyclically for $J > 1$ cycles. If we choose a step size

$$\eta = \min\left\{\frac{1}{2\sqrt{2}KB}, \frac{1+2\sqrt{2}}{2\sqrt{2}KJ}\ln\left(J^2 \cdot \max\left\{1, \frac{\|\boldsymbol{w}_0^{(0)} - \boldsymbol{w}_\star\|^2 \mu^3}{B^2 V_\star}\right\}\right)\right\},$$

then the final iterate of sequential GD satisfies

$$\left\|\boldsymbol{w}_0^{(MJ)} - \boldsymbol{w}_\star\right\|^2 \leq \mathcal{O}\left(\frac{\ln^2 J}{J^2}\right), \tag{1}$$

Conclusion

- CL on multiple linear classification tasks by Sequential GD.
- Main Result:
  1. Jointly Separable, Cyclic:
     Convergence Analysis / Implicit Bias / Forgetting Analysis
  2. Jointly Separable, Random:
     Asymptotic Convergence Analysis / Implicit Bias
  3. Beyond Separability:
     Non-asymptotic Convergence Analysis
- Fr, Apr 25, 10:00 SGT – Poster Session 3