Memory-Statistics Tradeoff in Continual Learning with Structural Regularization

Haoran Li[1], Jingfeng Wu[2], Vladimir Braverman[3]
[1]Shenzhen University, [2]UC Berkeley, [2]Johns Hopkins University

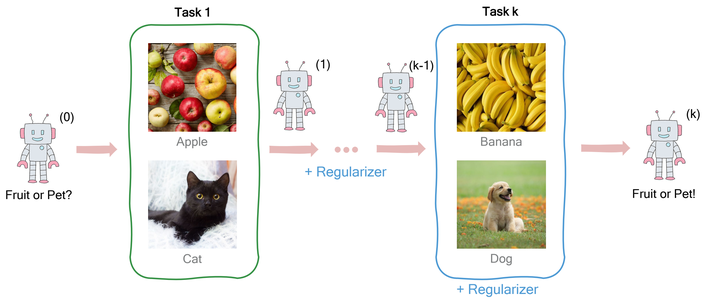
ICLR

Regularized-Based CL

Sequentially learn tasks $(\boldsymbol{x}_t^{(1)}, y_t^{(1)})_{t=1}^n \sim \mathcal{D}^{(1)}$, $(\boldsymbol{x}_t^{(2)}, y_t^{(2)})_{t=1}^n \sim \mathcal{D}^{(2)}$:

Minimize $L_1(\boldsymbol{w}) = R_1(\boldsymbol{w})$;

Minimize $L_2(\boldsymbol{w}) = R_2(\boldsymbol{w}) + L_{\text{regu}}(\boldsymbol{w}, \boldsymbol{w}^{(1)}; \boldsymbol{\Sigma})$



Problem Formulation

Two-task Linear Regression in Random Design

For tasks $k = 1,2$,

- $\boldsymbol{x}^{(k)}$ sampled from distributions $\mathcal{D}^{(k)}$ (one-hot/Gaussian)
- Shared optimal $y^{(k)} = {\boldsymbol{x}^{(k)}}^\top \boldsymbol{w}^* + \mathcal{N}(0, \sigma^2)$
- Covariance matrices $\boldsymbol{G} \coloneqq \mathbb{E}{\boldsymbol{x}^{(1)}}^\top \boldsymbol{x}^{(1)}$, $\boldsymbol{H} \coloneqq \mathbb{E}{\boldsymbol{x}^{(2)}}^\top \boldsymbol{x}^{(2)}$
- Risk for task k $R_k(\boldsymbol{w}) \coloneqq \mathbb{E}||\boldsymbol{y}^{(k)} - \boldsymbol{x}^{(k)}\boldsymbol{w}||_2^2$
- Joint population risk $R(\boldsymbol{w}) \coloneqq R_1(\boldsymbol{w}) + R_2(\boldsymbol{w})$

- One-hot setting: $\mathbb{P}(\boldsymbol{x}^{(1)} = \boldsymbol{e}_i) = \mu_i$, $\mathbb{P}(\boldsymbol{x}^{(2)} = \boldsymbol{e}_i) = \lambda_i$
- Gaussian setting: $\boldsymbol{x}^{(1)} \sim \mathcal{N}(0, \boldsymbol{G})$, $\boldsymbol{x}^{(2)} \sim \mathcal{N}(0, \boldsymbol{H})$,

Generalized ℓ_2-Regularized CL (GRCL)

$\boldsymbol{w}^{(1)} = ({\boldsymbol{X}^{(1)}}^\top \boldsymbol{X}^{(1)})^{-1} {\boldsymbol{X}^{(1)}}^\top \boldsymbol{y}^{(1)}$

GRCL: $\boldsymbol{w}^{(2)} = \arg\min_{\boldsymbol{w}} \frac{1}{n}||\boldsymbol{y}^{(2)} - \boldsymbol{X}^{(2)}\boldsymbol{w}||_2^2 + ||\boldsymbol{w} - \boldsymbol{w}^{(1)}||_{\boldsymbol{\Sigma}}^2$

- ℓ_2-RCL ($\boldsymbol{\Sigma} \to \gamma \boldsymbol{I}$): $\boldsymbol{w}^{(2)} = \arg\min_{\boldsymbol{w}} \frac{1}{n}||\boldsymbol{y}^{(2)} - \boldsymbol{X}^{(2)}\boldsymbol{w}||_2^2 + \gamma||\boldsymbol{w} - \boldsymbol{w}^{(1)}||_2^2$
- OCL ($\boldsymbol{\Sigma} \to 0$): $\boldsymbol{w}^{(2)} = \boldsymbol{w}^{(1)} + ({\boldsymbol{X}^{(2)}}^\top \boldsymbol{X}^{(2)})^{-1} {\boldsymbol{X}^{(2)}}^\top (\boldsymbol{y}^{(2)} - \boldsymbol{X}^{(2)}\boldsymbol{w}^{(1)})$,
- Joint Learning (JL): $\boldsymbol{w}_{\text{joint}} = (\boldsymbol{X}_{\text{joint}}^\top \boldsymbol{X}_{\text{joint}})^{-1} \boldsymbol{X}_{\text{joint}}^\top \boldsymbol{y}_{\text{joint}}$

Main Result

One-hot Setting

$$\mathbb{E}\big[R\big(\boldsymbol{w}^{(2)}\big)\big] - \min R(\cdot) = bias + variance$$

where

$$\texttt{bias} \eqsim \langle (\boldsymbol{G}+\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{G})^n (\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\boldsymbol{H})^{-2} + (\boldsymbol{I}-\boldsymbol{H})^n),\ \boldsymbol{w}^* {\boldsymbol{w}^*}^\top \rangle,$$

$$\texttt{variance} \eqsim \sigma^2 \big(\langle (\boldsymbol{G}+\boldsymbol{H})(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\boldsymbol{H})^{-2} + (\boldsymbol{I}-\boldsymbol{H})^n), \frac{1}{n}{\boldsymbol{G}_{\mathbb{J}}}^{-1} + n\boldsymbol{G}_{\mathbb{J}^c} \rangle$$
$$+ \langle \boldsymbol{G}+\boldsymbol{H}, \frac{1}{n}(\boldsymbol{H}_{\mathbb{K}} + \boldsymbol{\Sigma}_{\mathbb{K}})^{-2}\boldsymbol{H}_{\mathbb{K}} + n(\boldsymbol{I}_{\mathbb{K}^c} + n\boldsymbol{\Sigma}_{\mathbb{K}^c})^{-2}\boldsymbol{H}_{\mathbb{K}^c} \rangle \big).$$

[Gaussian Setting] For OCL, $bias, variance$ satisfy

$$\texttt{bias} \gtrsim \|(\frac{(\operatorname{tr}\boldsymbol{G}_{\mathbb{J}^c})^2}{n^2}\boldsymbol{G}_{\mathbb{J}}^{-2} + \boldsymbol{I}_{\mathbb{J}^c})^{\frac{1}{2}} \cdot (\frac{(\operatorname{tr}\boldsymbol{H}_{\mathbb{K}^c})^2}{n^2}\boldsymbol{H}_{\mathbb{K}}^{-2} + \boldsymbol{I}_{\mathbb{K}^c})^{\frac{1}{2}} \boldsymbol{w}^* \|_{\boldsymbol{G}}^2,$$

$$\texttt{variance} \gtrsim \frac{\sigma^2}{n} \Big\langle \boldsymbol{G},\ \boldsymbol{H}_{\mathbb{K}}^{-1} + \frac{n^2}{(\operatorname{tr}\boldsymbol{H}_{\mathbb{K}^c})^2}\boldsymbol{H}_{\mathbb{K}^c} + \Big(\frac{(\operatorname{tr}\boldsymbol{H}_{\mathbb{K}^c})^2}{n^2}\boldsymbol{H}_{\mathbb{K}}^{-2} + \boldsymbol{I}_{\mathbb{K}^c}\Big) \cdot \Big(\boldsymbol{G}_{\mathbb{J}}^{-1} + \frac{n^2}{(\operatorname{tr}\boldsymbol{G}_{\mathbb{J}^c})^2}\boldsymbol{G}_{\mathbb{J}^c}\Big) \Big\rangle.$$

[Extension to NTK regime] $bias, variance$ satisfy similar results where

$$\boldsymbol{G} := \mathbb{E}_{\mathcal{D}^{(1)}}[\phi(\boldsymbol{x}^{(1)})\phi(\boldsymbol{x}^{(1)})^\top], \boldsymbol{H} := \mathbb{E}_{\mathcal{D}^{(2)}}[\phi(\boldsymbol{x}^{(2)})\phi(\boldsymbol{x}^{(2)})^\top].$$

Catastrophic Forgetting of OCL and ℓ_2-RCL

For tasks solvable by Joint Learning,

to achieve $o(1)$ excess risk, for $\boldsymbol{G} = \operatorname{diag}(\mu_i)_{i=1}^d$, $\boldsymbol{H} = \operatorname{diag}(\lambda_i)_{i=1}^d$:

- OCL: $\sum_{i\in\mathbb{K}} \frac{\mu_i}{\lambda_i} + n^2 \sum_{i\in\mathbb{J}\cap\mathbb{K}^c} \mu_i\lambda_i = o(n)$
- ℓ_2-RCL: $\gamma = o(1), \sum_{i\in\mathbb{J}\cup\mathbb{K}} \left(\frac{\mu_i}{\lambda_i + \frac{1}{n} + \gamma} + \frac{\gamma}{\mu_i + \frac{1}{n}} \right) = o(n)$

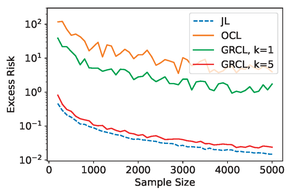
Large forgetting with dissimilar covariances $\boldsymbol{G}, \boldsymbol{H}$!

GRCL avoids Forgetting with Full Memory

For regularization $\boldsymbol{\Sigma} = \operatorname{diag}(\gamma_i)_{i=1}^d$ aligned with $\boldsymbol{G}$ ($\gamma_i = \mu_i$ for $\mu_i \geq 1/n$ and 0 otherwise),

$$\mathbb{E}\Delta(\boldsymbol{w}^{(2)}) \lesssim \mathbb{E}\Delta(\boldsymbol{w}_{\text{joint}}).$$

GRCL attain joint learning performance with sufficient memory.



[Extension to multi-task] $\mathbb{E}\Delta(\boldsymbol{w}^{(2)}) \lesssim \mathbb{E}\Delta(\boldsymbol{w}_{\text{joint}})$ holds when:

$$\gamma_i^{(t)} = \mu_i^{(t)} \text{ for } \mu_i^{(t)} \geq 1/n \text{ and 0 otherwise.}$$

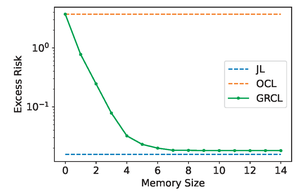
Memory-Statistics Trade-off

Catastrophic Forgetting in Low-Memory Settings

For any memory size $k << N$,
if $\mu_i = 1/(k+1)$ and $\lambda_i = 1/n$ for $1 \leq i \leq k+1$, then:
For any regularizer $\boldsymbol{\Sigma}$ of rank $\boldsymbol{k}$, the GRCL excess risk is $\Omega(1)$.

No k-dimensional regularization in GRCL
can mitigate forgetting across $k+1$ orthogonal directions.

Tempered Forgetting in Moderate Memory Settings

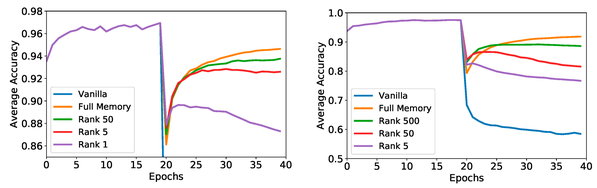


As the memory size k of GRCL increases, GRCL's excess risk decreases, approaching joint learning performance at a threshold.

GRCL with an intermediate memory constraint can partially alleviate catastrophic forgetting.

Neural Network Verification

GRCL via PCA on two-task Permuted and Rotated MNIST:



(a) Permuted MNIST (b) Rotated MNIST

Practical Algorithms

GRCL via PCA costly on NN.

Solution: GRCL via CountSketch!

← See our previous work

This work on ArXiv: