Information-theoretic Online Memory Selection for Continual Learning

Shengyang Sun, Daniele Calandriello, Clara Huiyi Hu, Ang Li, Michalis Titsias

Overview

- Motivation

- Online memory selection
  - Information-theoretic criteria
  - An efficient Bayesian model

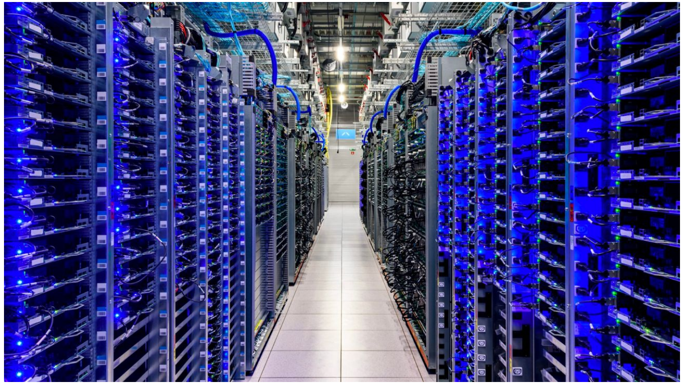
- Continual learning
  - The timing of memory updates
  - Information-theoretic Reservoir Sampling (InfoRS)

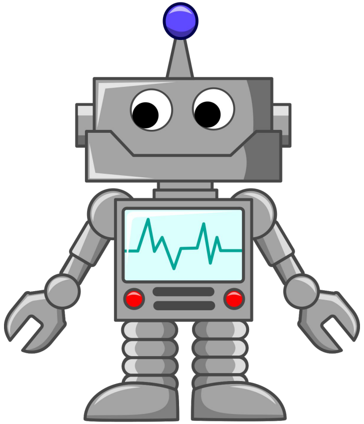
- Open questions and future works

Motivation

Learning Efficiency

Memory Computation Data



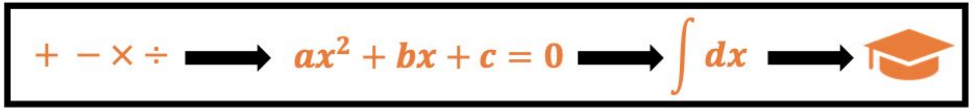


https://blog.google/inside-google/infrastructure/how-data-center-security-works/ https://commons.wikimedia.org/wiki/File:Cartoon_Robot.svg

Motivation

Learning Efficiency

Memory **Computation** Data

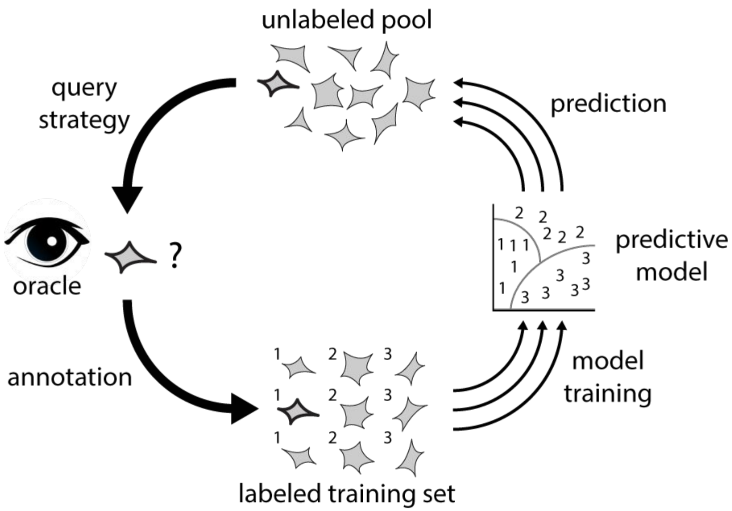


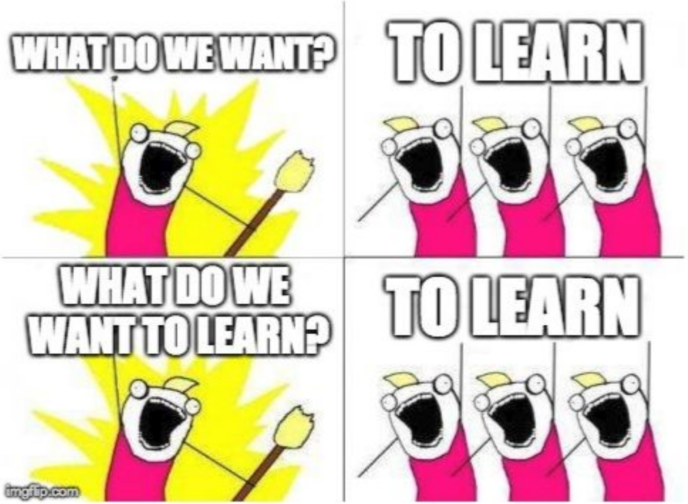
https://www.hpcwire.com/2017/04/10/nvidia-responds-google-tpu-benchmarking/ https://pixabay.com/vectors/metal-blocks-steel-commodity-iron-36867/

Motivation

Learning Efficiency

Memory Computation **Data**





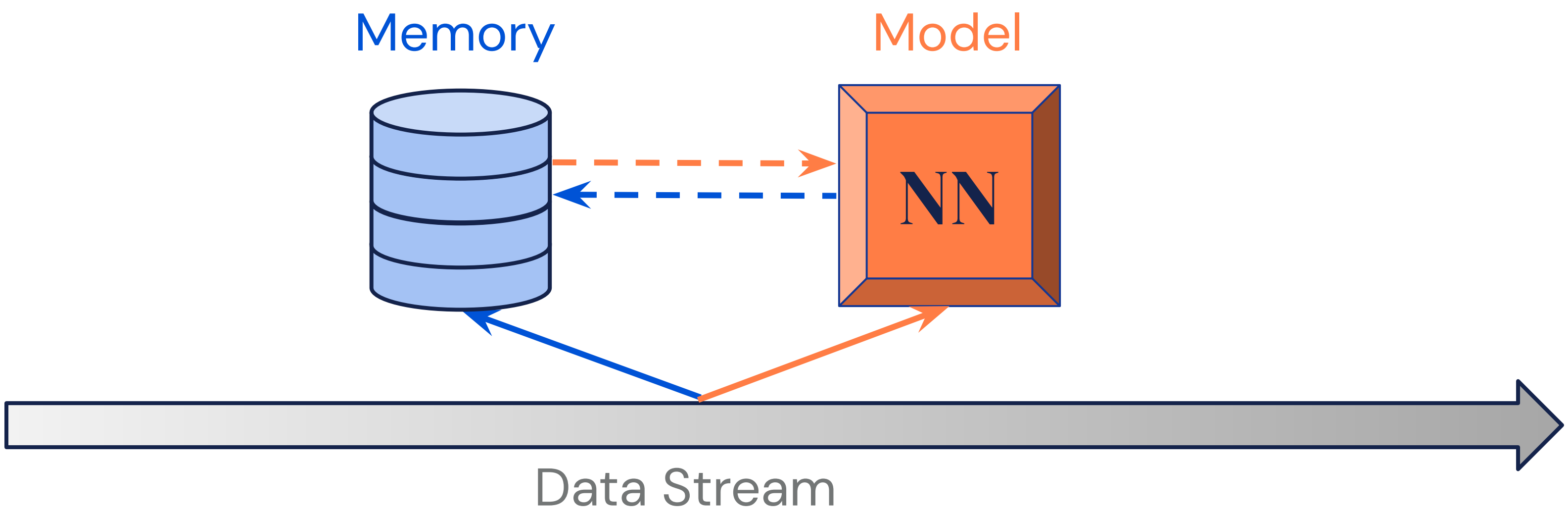
https://www.biodataanalysis.de/showcase-actlearn

Pesah and Wehenkel, Improve your Scientific Models with Meta-Learning and Likelihood-Free Inference

Online Memory Selection

- Online Memory Selection is a key ingredient for learning efficiency.
  - **Continual Learning**, Reinforcement Learning, "Sample-Efficient" Learning, ...
- The agent updates both the memory and the model based on the instant observation,

$$(f_\theta, \mathcal{M}) \longleftarrow (f_\theta, \mathcal{M}, (\mathbf{x}_\star, y_\star))$$



Online Memory Selection

- Challenges
  - The **purely online** constraint calls for both effectiveness and efficiency.

  - To select a representative memory needs to deal with **data imbalance**.

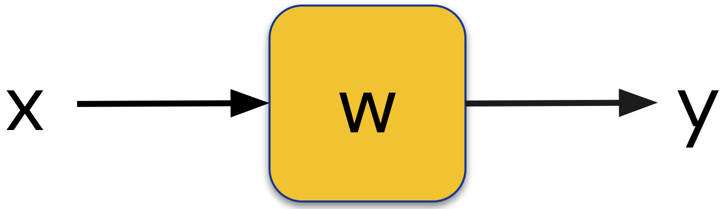
- Existing Approaches
  - Reservoir sampling (RS) [1] draws uniform samples in a single pass.
  - GSS [2] encourages diversity by minimizing gradient similarities.

[1] Vitter, 1985; [2] Aljundi et al., 2019

Memorable Information Criterion

- We approach the problem from an information-theoretic perspective.

- Consider a Bayesian model $p(y|\mathbf{w};\mathbf{x})$ for the target function,

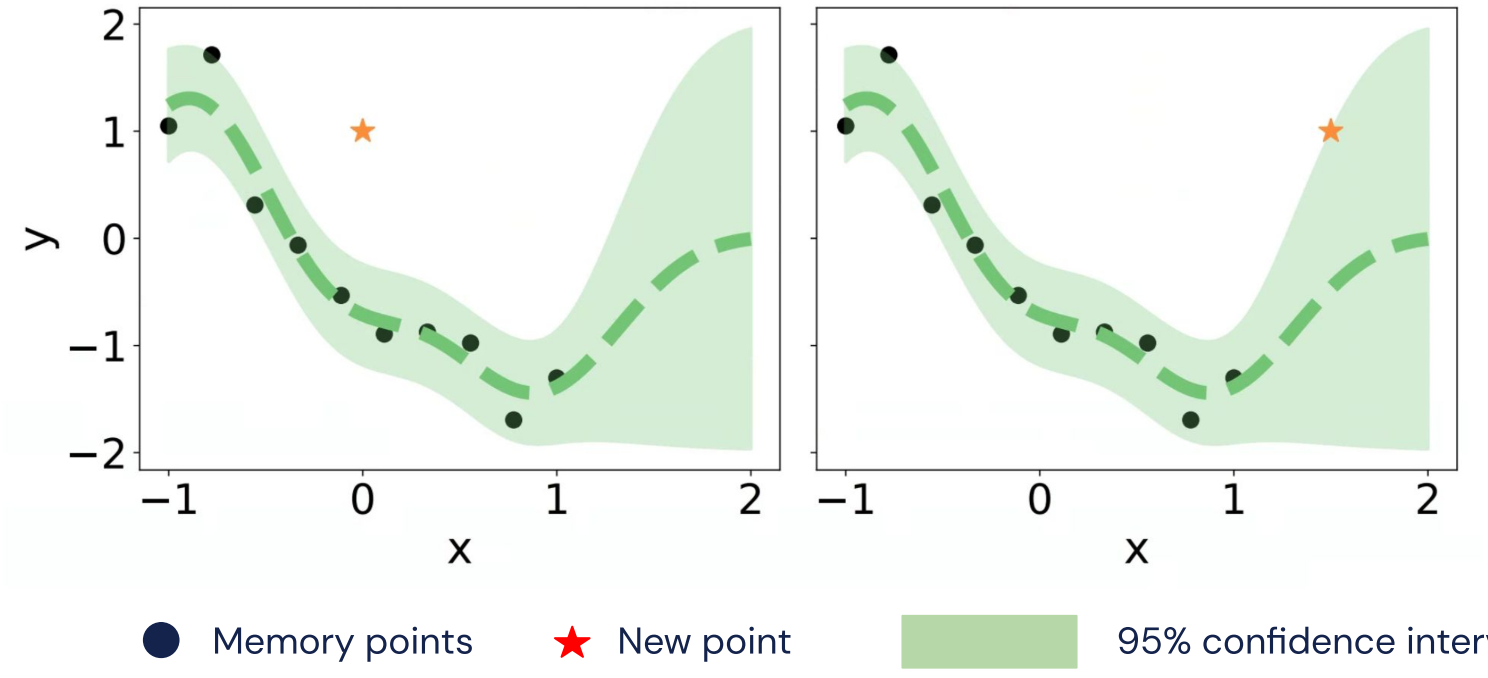


- Intuition: incorporating "**surprising**" data points brings new information **to the memory.**

$$s_{\text{surp}}((\mathbf{x}_\star, y_\star); \mathcal{M}) = \log p(y_\star|\mathbf{y}_\mathcal{M})$$

Memorable Information Criterion

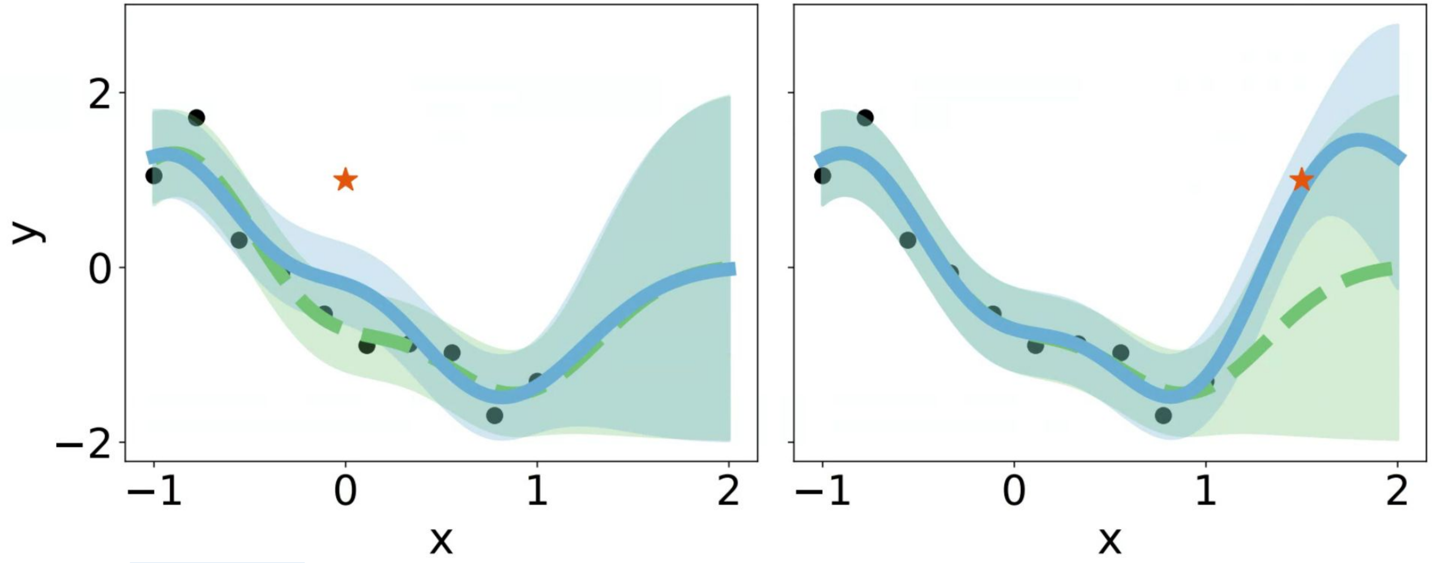
- Surprising points: “harmful” outliers & “Helpful” unfamiliar points



Memorable Information Criterion

- We propose the ***learnability***

$$s_{\text{learn}}((\mathbf{x}_\star, y_\star); \mathcal{M}) = \log p(y_\star | y_\star, \mathbf{y}_\mathcal{M})$$



95% confidence interval given the memories and the new point

Memorable Information Criterion

- *Surprise*

$$s_{\text{surp}}((\mathbf{x}_\star, y_\star); \mathcal{M}) = \log p(y_\star | \mathbf{y}_\mathcal{M})$$

- *Learnability*

$$s_{\text{learn}}((\mathbf{x}_\star, y_\star); \mathcal{M}) = \log p(y_\star | y_\star, \mathbf{y}_\mathcal{M})$$

- *Memorable Information Criterion (MIC)*

$$\text{MIC}_\eta((\mathbf{x}_\star, y_\star); \mathcal{M}) = \eta s_{\text{learn}}((\mathbf{x}_\star, y_\star); \mathcal{M}) + s_{\text{surp}}((\mathbf{x}_\star, y_\star); \mathcal{M})$$

An Efficient Bayesian Linear Model

- A Bayesian linear model,

$$y = \mathbf{w}^\top \mathbf{x} + \epsilon, \mathbf{w} \sim \mathcal{N}(0, \sigma_w^2 \mathbf{I}), \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Analytic weight posterior,

$$p(\mathbf{w}|\mathbf{y}_\mathcal{M}; \mathbf{X}_\mathcal{M}) = \mathcal{N}(\mathbf{A}_\mathcal{M}^{-1}\mathbf{b}_\mathcal{M}, \sigma^2 \mathbf{A}_\mathcal{M}^{-1}),$$
$$\mathbf{A}_\mathcal{M}^{-1} = \left(\mathbf{X}_\mathcal{M}^\top \mathbf{X}_\mathcal{M} + c\mathbf{I}_d\right)^{-1}, \mathbf{b}_\mathcal{M} = \mathbf{X}_\mathcal{M}^\top \mathbf{y}_\mathcal{M}$$

- The memory buffer can be summarized by the matrix $\mathbf{A}_\mathcal{M}^{-1}$ and $\mathbf{b}_\mathcal{M}$.

An Efficient Bayesian Linear Model

- The MIC can be computed explicitly,

$$\mathrm{MIC}_\eta((\mathbf{x}_\star, y_\star); \mathcal{M}) = \eta \log \mathcal{N}(y_\star | \mathbf{x}_\star^\top \mathbf{A}_{\mathcal{M}^+}^{-1} \mathbf{b}_{\mathcal{M}^+}, \sigma^2 \mathbf{x}_\star^\top \mathbf{A}_{\mathcal{M}^+}^{-1} \mathbf{x}_\star + \sigma^2) - \log \mathcal{N}(y_\star | \mathbf{x}_\star^\top \mathbf{A}_{\mathcal{M}}^{-1} \mathbf{b}_{\mathcal{M}}, \sigma^2 \mathbf{x}_\star^\top \mathbf{A}_{\mathcal{M}}^{-1} \mathbf{x}_\star + \sigma^2)$$

- The updated statistics matrices,

$$\mathbf{A}_{\mathcal{M}^+} = \mathbf{A}_{\mathcal{M}} + \mathbf{x}_\star \mathbf{x}_\star^\top, \mathbf{b}_{\mathcal{M}^+} = \mathbf{b}_{\mathcal{M}} + \mathbf{x}_\star y_\star$$

- The rank-one difference allows to use the Sherman-Morrison formula [1]

$$\mathbf{A}_{\mathcal{M}^+}^{-1} = \mathbf{A}_{\mathcal{M}}^{-1} - \frac{\mathbf{A}_{\mathcal{M}}^{-1} \mathbf{x}_\star \mathbf{x}_\star^\top \mathbf{A}_{\mathcal{M}}^{-1}}{1 + \mathbf{x}_\star^\top \mathbf{A}_{\mathcal{M}}^{-1} \mathbf{x}_\star}$$

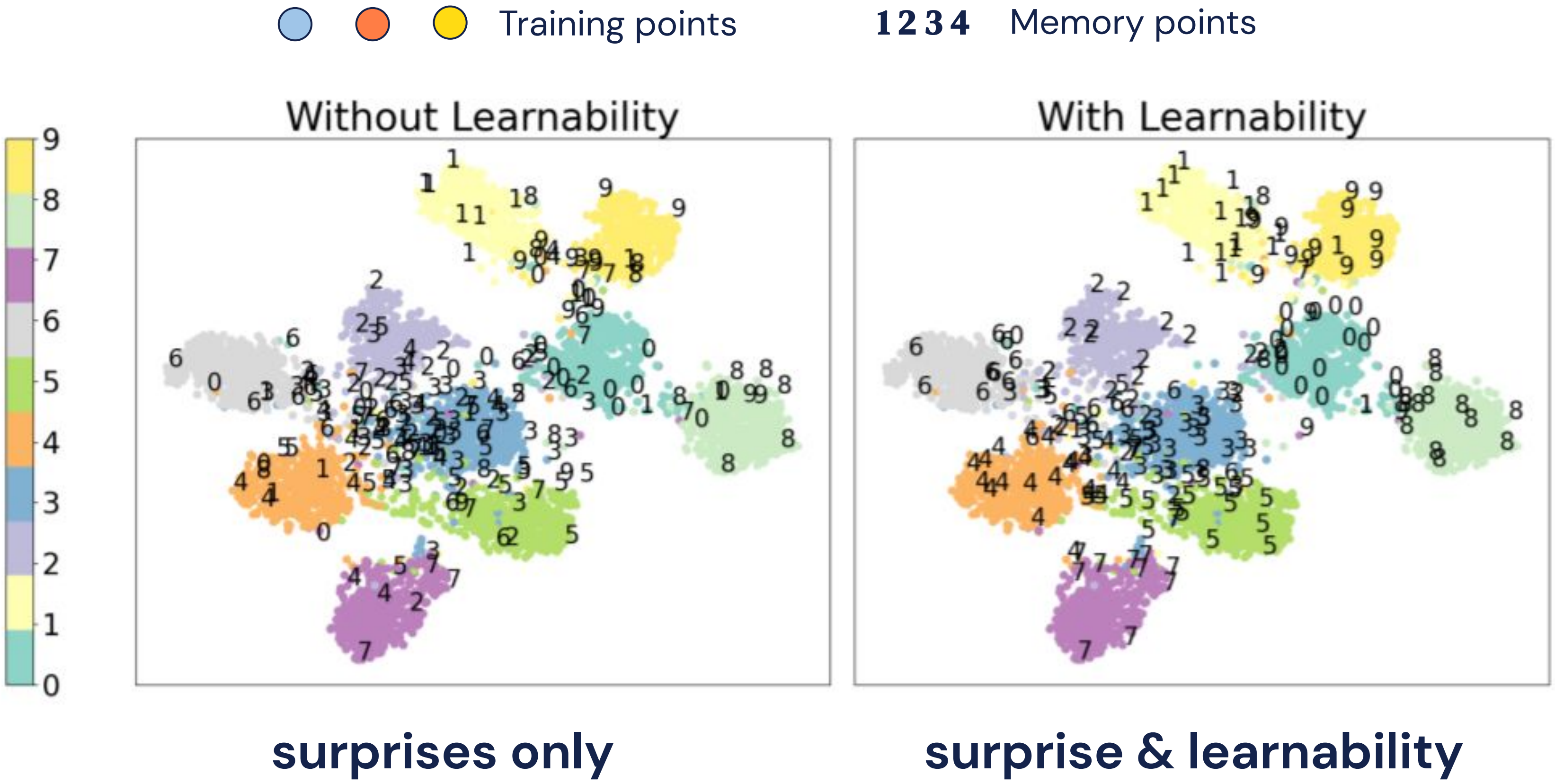
Demonstrating the Proposed Criteria

- Algorithm: A greedy algorithm (InfoGS) to replace the informative new point with the least informative memory point.

- Dataset: pretrained ResNet features for CIFAR-10 classification.

- Problem: Split-CIFAR10 with 5 tasks.

Demonstrating the Proposed Criteria



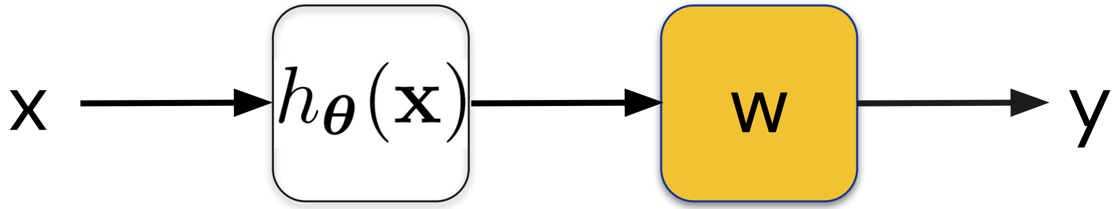
Bayesian Linear Model in Neural Networks

- If the model is in the following form,

$$f_{\boldsymbol{\theta}}(\cdot) = g_{\boldsymbol{\theta}}(h_{\boldsymbol{\theta}}(\cdot))$$

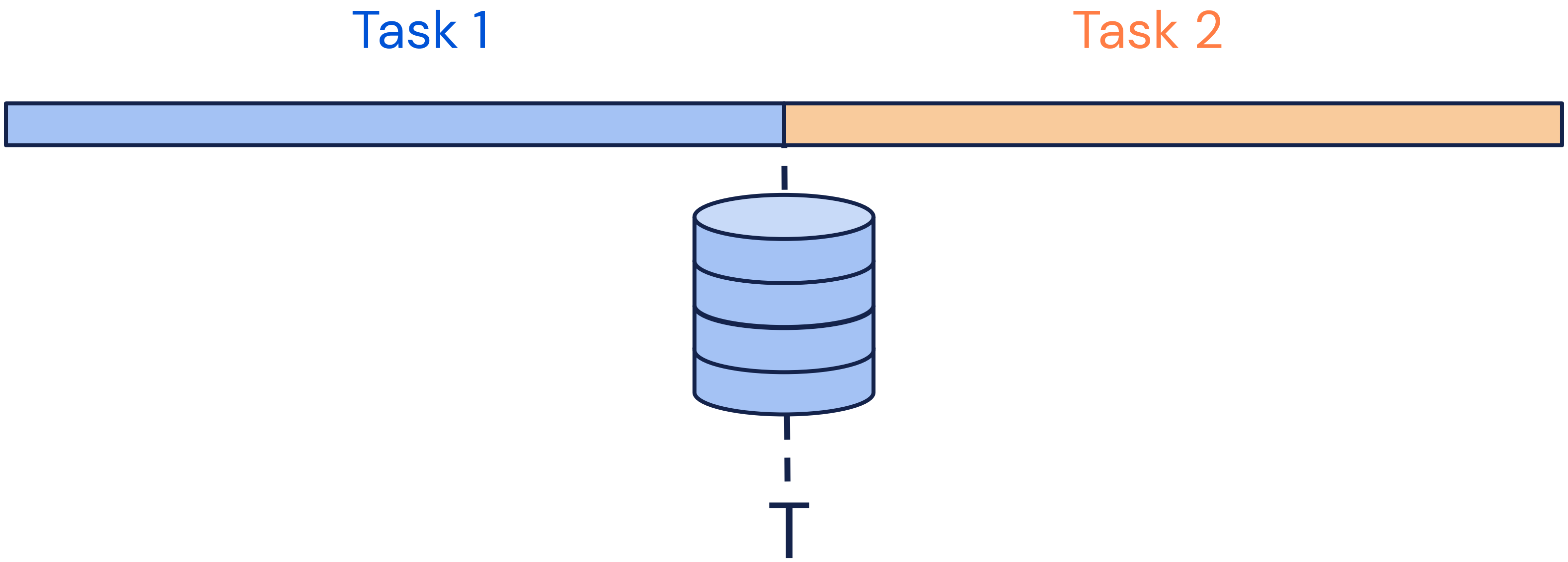
- Neural networks are well-known for learning meaningful representations $h_{\boldsymbol{\theta}}$

- We apply the Bayesian linear model from the network feature to the targets.



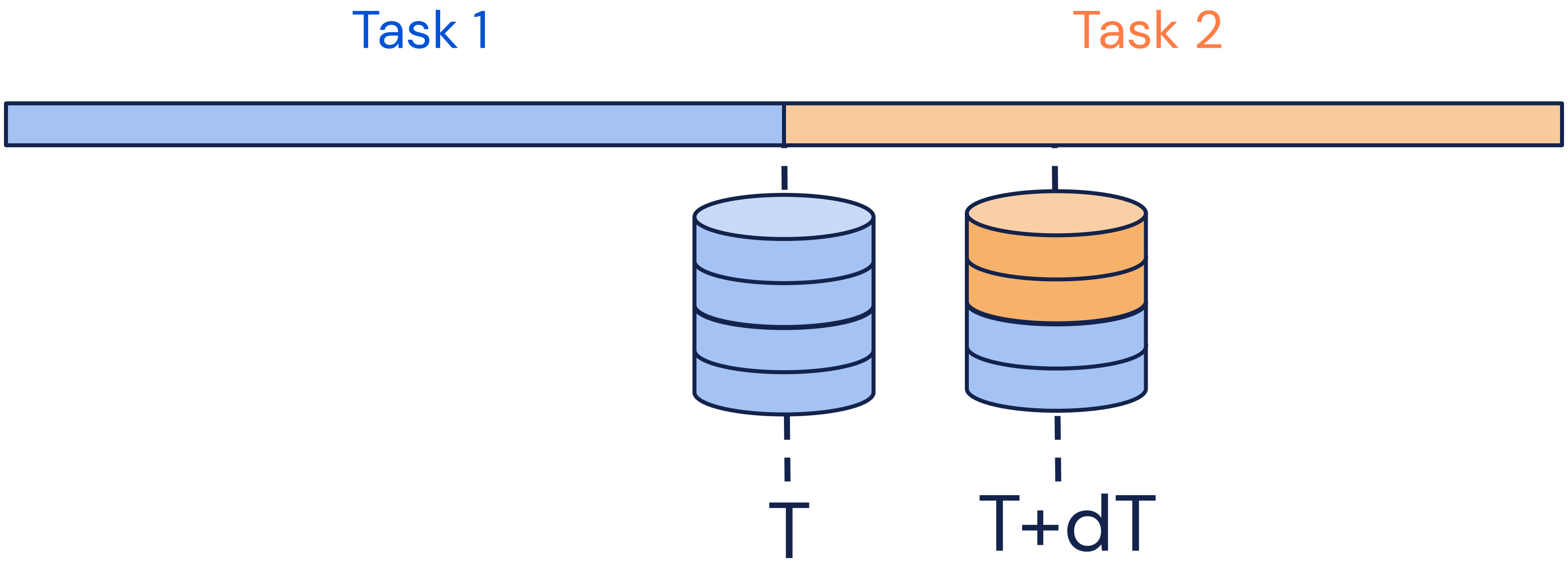
Continual Learning: the timing to update memory

- Besides *how to update the memory*, *when to update the memory* is also important,



Continual Learning: the timing to update memory

- Besides *how to update the memory*, *when to update the memory* is also important,



- Better to choose a large **dT** !

Continual Learning: InfoRS

☹ The greedy algorithm updates the memory urgently.
☺ Reservoir sampling updates the memory stochastically.

- We propose the Information-theoretic Reservoir Sampling (InfoRS), which combines the merits of both the information-theoretic criteria and the reservoir sampling.

- InfoRS conducts reservoir sampling over informative points.

Continual Learning: InfoRS

Algorithm 1 Information-theoretic Reservoir Sampling (InfoRS)

Input: Memory $\mathcal{M}$ and matrices $\mathbf{A}_{\mathcal{M}}^{-1}, \mathbf{b}_{\mathcal{M}}$, the batch $\mathcal{B}$, the predictor $f_{\boldsymbol{\theta}}$.
Input: The reservoir count n and the budget M.
Input: Running mean and stddev for the MIC: $\hat{\mu}_i, \hat{\sigma}_i$. The thresholding ratio γ_i.
Update $f_{\boldsymbol{\theta}}$ based on $\mathcal{M}$ and $\mathcal{B}$ [3]. // Predictor Update
Update the features for the memory points used in replay, and update $\mathbf{A}_{\mathcal{M}}^{-1}, \mathbf{b}_{\mathcal{M}}$ accordingly.
for $(\mathbf{x}_\star, y_\star)$ in $\mathcal{B}$ **do**
  if $|\mathcal{M}| < M$ **or** $\text{MIC}_\eta((\mathbf{x}_\star, y_\star); \mathcal{M}) \geq \hat{\mu}_i + \hat{\sigma}_i * \gamma_i$ // Information Thresholding
    Update $\mathcal{M}, n \longleftarrow$ **ReservoirSampling**$(\mathcal{M}, M, n, (\mathbf{x}_\star, y_\star))$. // Memory Update
    Update $\mathbf{A}_{\mathcal{M}}^{-1}, \mathbf{b}_{\mathcal{M}}$ based on the Sherman-Morrison formula if $\mathcal{M}$ is updated.
  Update $\hat{\mu}_i, \hat{\sigma}_i$ using the criterion $\text{MIC}_\eta((\mathbf{x}_\star, y_\star); \mathcal{M})$. // Running Moments Update
return Buffer $\mathcal{M}$ and $\mathbf{A}_{\mathcal{M}}^{-1}, \mathbf{b}_{\mathcal{M}}$. The reservoir count n and statistics $\hat{\mu}_i, \hat{\sigma}_i$. The updated $f_{\boldsymbol{\theta}}$.

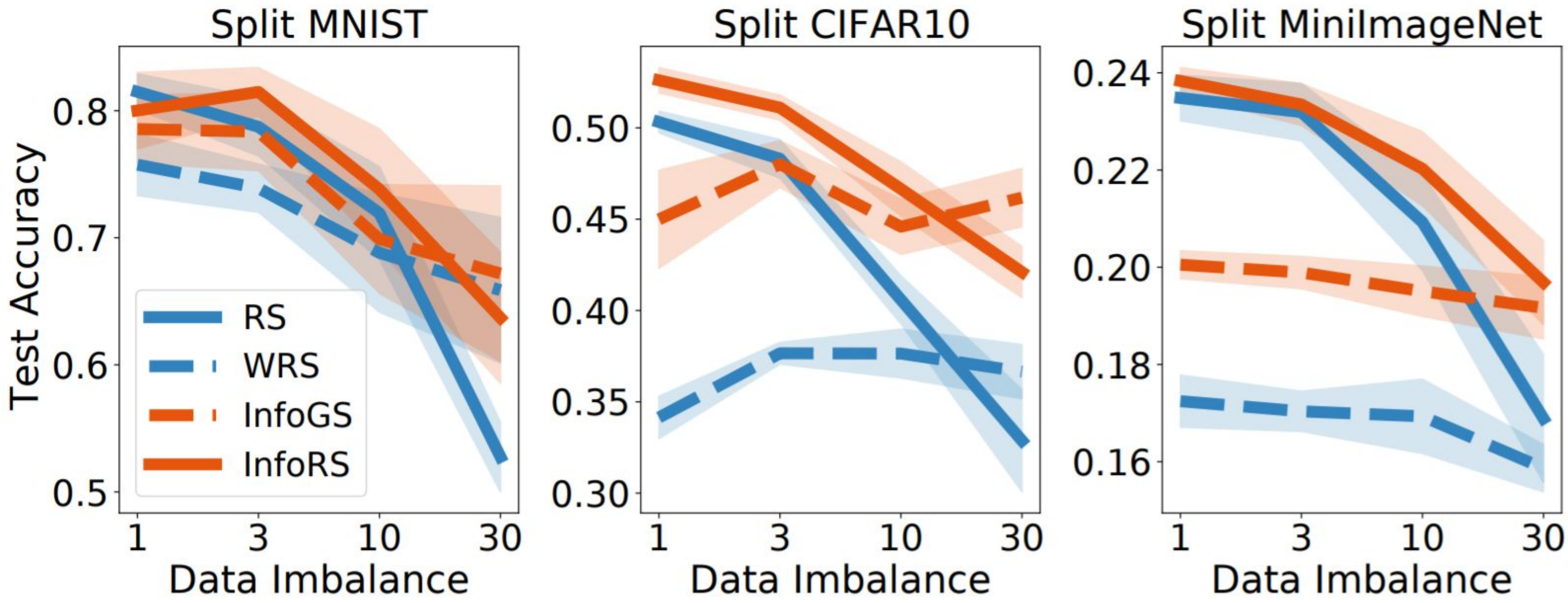
Continual Learning: Experiments

- Dataset: Split-MNIST, Split-CIFAR10, Split-MiniImageNet.
- Data ImBalance: one specific task are trained **R** times epochs than other tasks.
- Learning the model: dark experience replay [1]
- Online memory selection: RS, WRS, InfoGS, InfoRS

[1] *Buzzega et al., 2020*

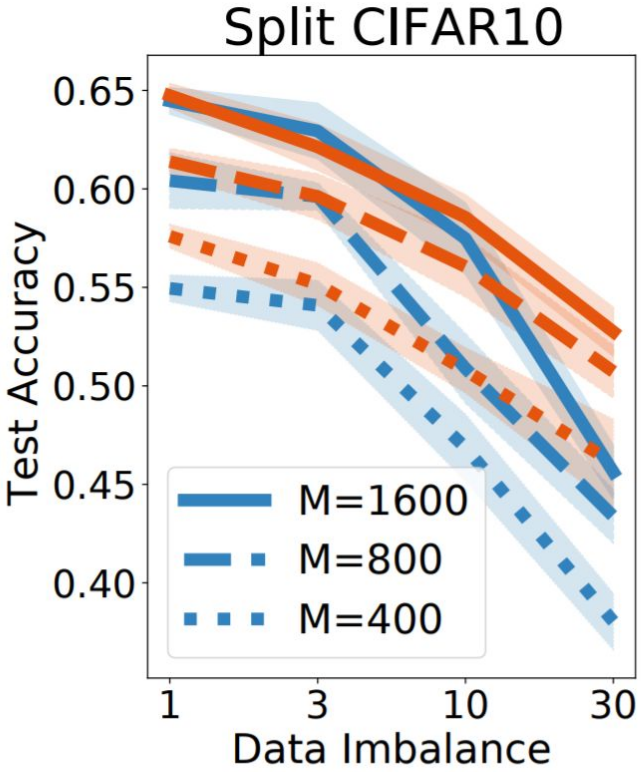
Continual Learning: Experiments

- InfoRS improves the robustness over imbalanced data from RS.



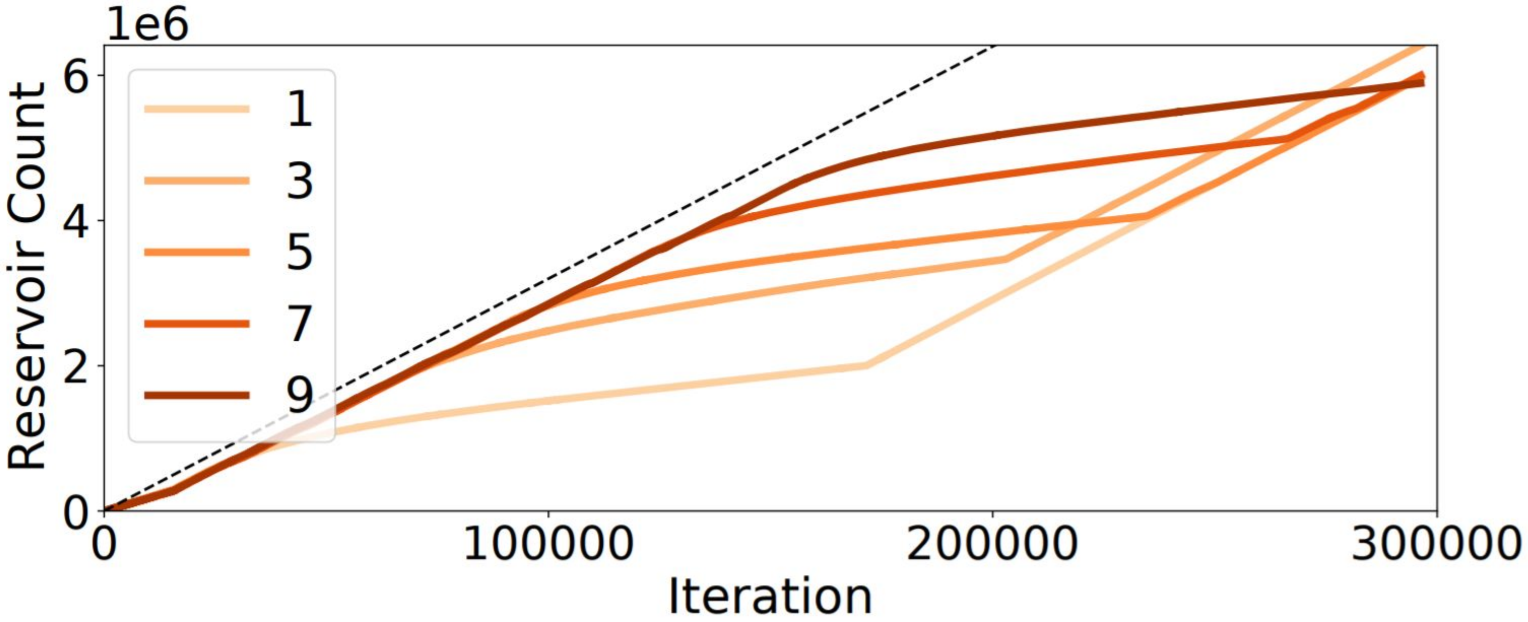
Continual Learning: Experiments

- InfoRS improves the robustness over imbalanced data from RS.



Continual Learning: Experiments

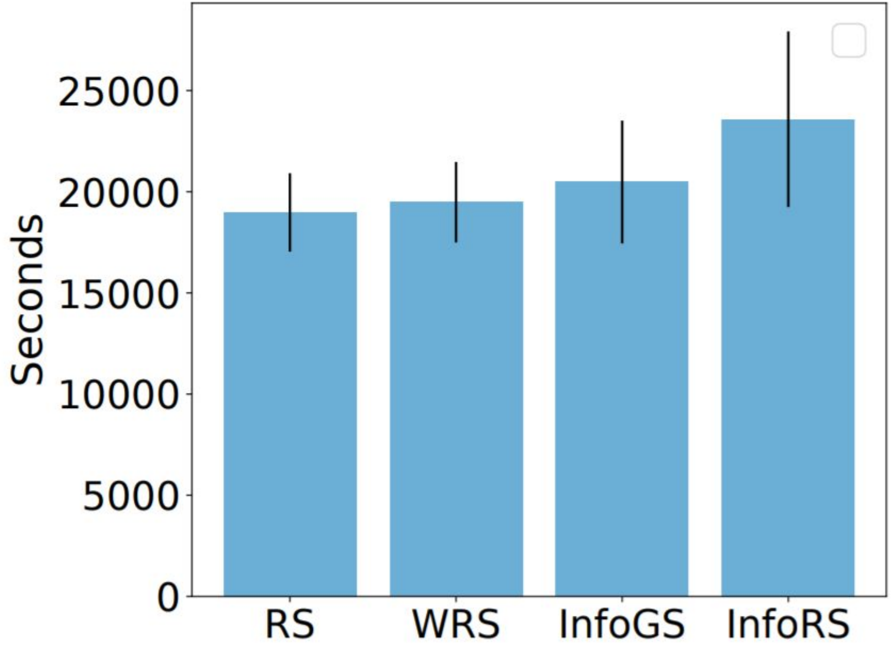
- InfoRS adapts the speed to incorporate new points based on the informativeness.



Each number represents the index of the task which has 10 times more data than the other tasks.

Continual Learning: Experiments

- The computational efficiency,



Open Questions and Future Directions

- How to improve over uniform sampling for balanced data streams ?
- How to deal with representation shifting along the process ?
- To combine learnability and surprise, is the weighted summation the best, particularly when the model is misspecified ?
- How does InfoRS perform over other problems: RL, “sample-efficient” learning, … ?

Thanks

Information-theoretic algorithms for online memory selection in task-free continual learning, with improved robustness against data imbalance.

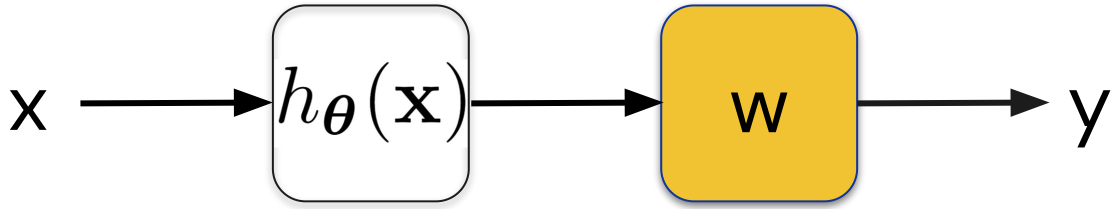
Bayesian Linear Model in Neural Networks

- We assume that the model is in the following form,

$$f_{\boldsymbol{\theta}}(\cdot) = g_{\boldsymbol{\theta}}(h_{\boldsymbol{\theta}}(\cdot))$$

- Neural networks are well-known for learning meaningful representations $h_{\boldsymbol{\theta}}$

- We apply the Bayesian linear model from the network feature to the targets.



- How to obtain the features of the memories ?
  - ☹ Using the stored features suffer from the representation shifting.
  - ☹ Computing the features in each iteration is computationally intensive.
  - ☺ Store and update the memory features in experience replay.

MIC for Memory Points

- We can evaluate the memorable information criterion for points in the memory,
- Let $(\mathbf{x}_m, y_m)$ be a data point within the memory, define the *pseudo-memory*,

$$\mathcal{M}_{*,-m} := \mathcal{M} \cup (\mathbf{x}_\star, y_\star) \backslash (\mathbf{x}_m, y_m)$$

- Its MIC is computed with respect to the pseudo-memory,

$$\mathrm{MIC}_\eta((\mathbf{x}_m, y_m); \mathcal{M}_{\star,-m})$$

- The MIC for memories is comparable to the MIC for new points,

$$\mathrm{MIC}_\eta((\mathbf{x}_\star, y_\star); \mathcal{M}) = \mathrm{MIC}_\eta((\mathbf{x}_\star, y_\star); \mathcal{M}_{\star,-\star})$$

Learnability + Surprise

- The Information Gain (IG),

$$\mathrm{KL}\left(p(\mathbf{w}|y_\star, \mathbf{y}_\mathcal{M})||p(\mathbf{w}|\mathbf{y}_\mathcal{M})\right)$$

- IG can be rewritten as the combination of "learnability" and "surprise" as well.

$$\mathbb{E}_{p(\mathbf{w}|y_\star, \mathbf{y}_\mathcal{M})}\left[\log p(y_\star|\mathbf{w})\right] - \log p(y_\star|\mathbf{y}_\mathcal{M})$$

- The prediction gain (PG),

$$\mathcal{L}((\mathbf{x}_\star, y_\star); \boldsymbol{\theta}) - \mathcal{L}((\mathbf{x}_\star, y_\star); \boldsymbol{\theta}')$$

Cover, 1999; Bellemare et al., 2016; Graves et al., 2017

Information-theoretic Criteria

- *Entropy Reduction (ER)*

$$\mathrm{ER}((\mathbf{x}_\star, y_\star); \mathcal{M}) := \mathbb{H}[p(\mathbf{w}|\mathcal{M})] - \mathbb{H}[p(\mathbf{w}|\mathcal{M}, (\mathbf{x}, y))]$$

- *Weighted Information Gain (IG)*

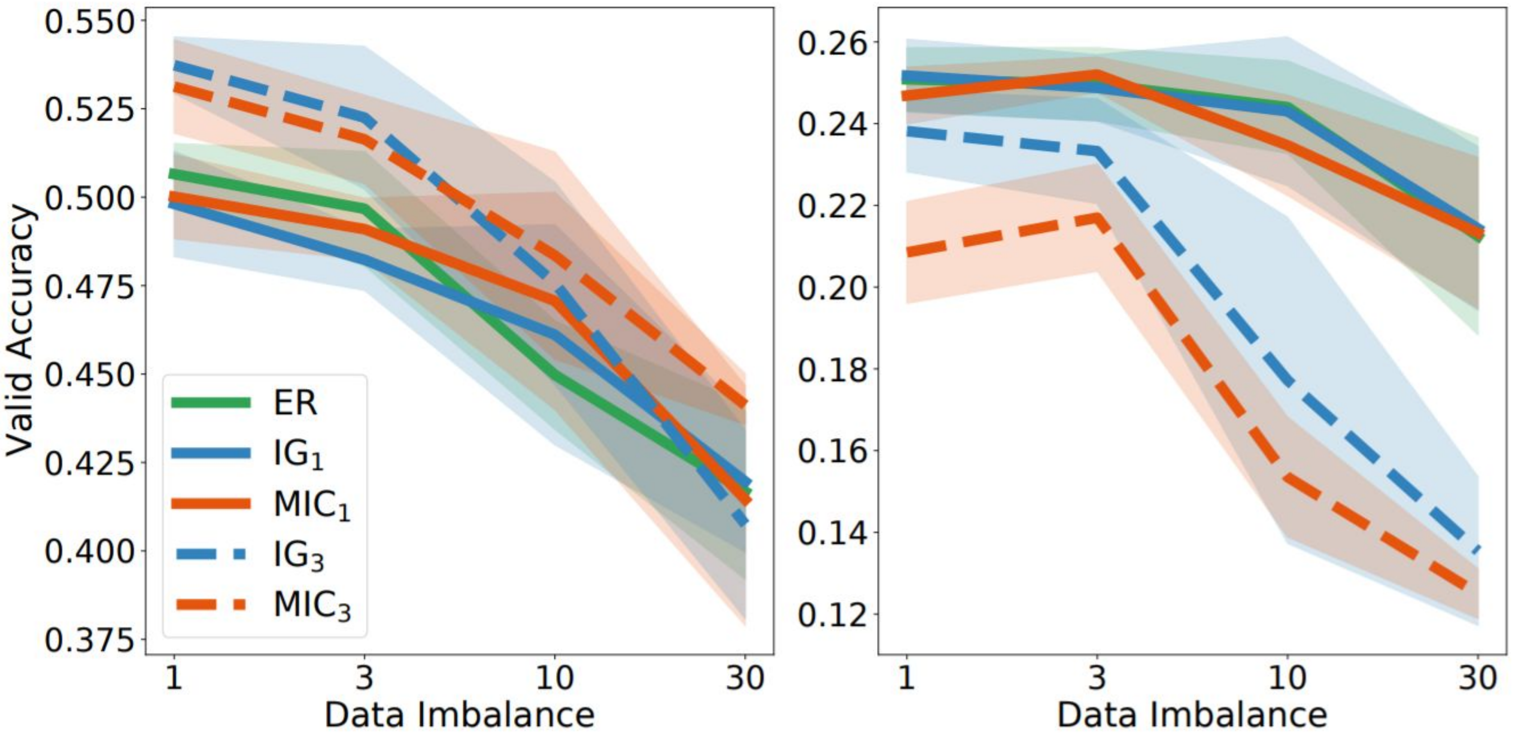
$$\mathrm{IG}_\eta((\mathbf{x}_\star, y_\star); \mathcal{M}) = \eta \mathbb{E}_{p(\mathbf{w}|y_\star, \mathbf{y}_\mathcal{M})} [\log p(y_\star|\mathbf{w})] - \log p(y_\star|\mathbf{y}_\mathcal{M})$$

- *Memorable Information Criterion (MIC)*

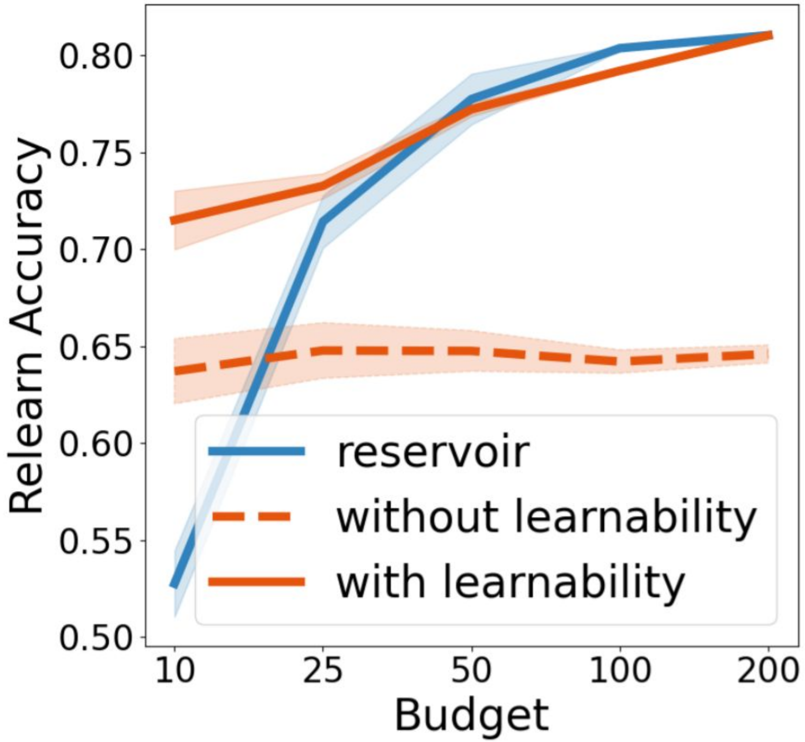
$$\mathrm{MIC}_\eta \; ((\mathbf{x}_\star, y_\star); \mathcal{M}) = \eta \log p(y_\star|y_\star, \mathbf{y}_\mathcal{M}) - \log p(y_\star|\mathbf{y}_\mathcal{M})$$

Continual Learning: Experiments

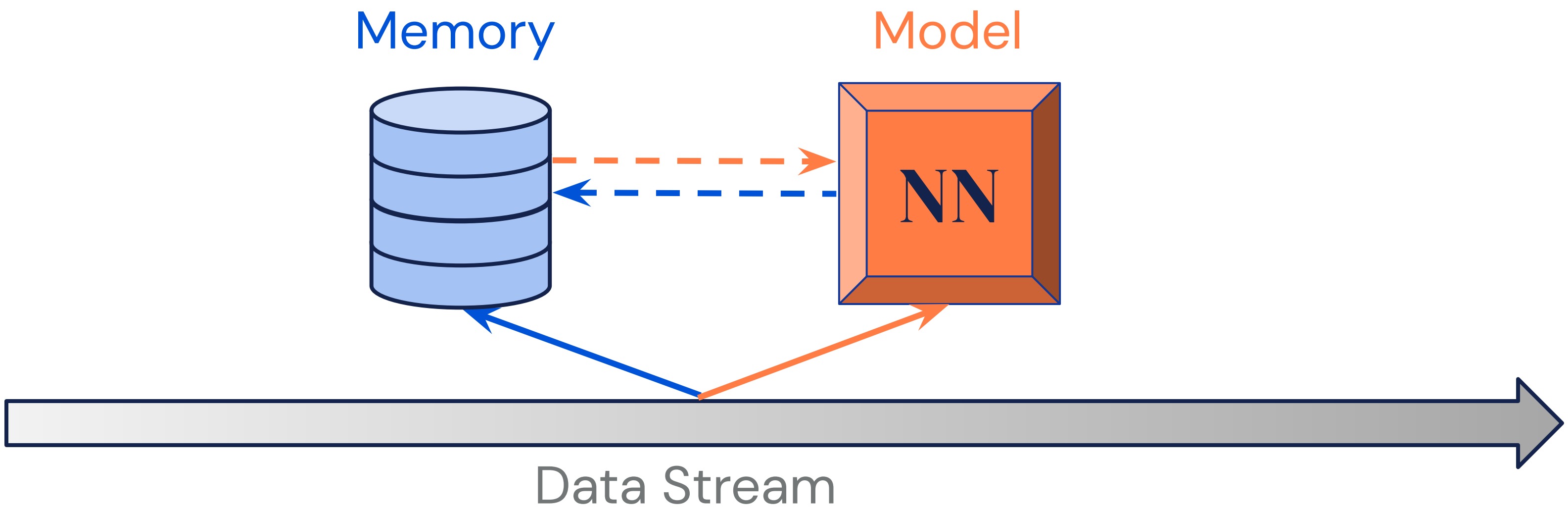
- Comparing information-theoretic criteria,



Demonstrating the Proposed Criteria



Continual Learning



- We use dark experience replay [1] for learning the model, which optimizes the objective,

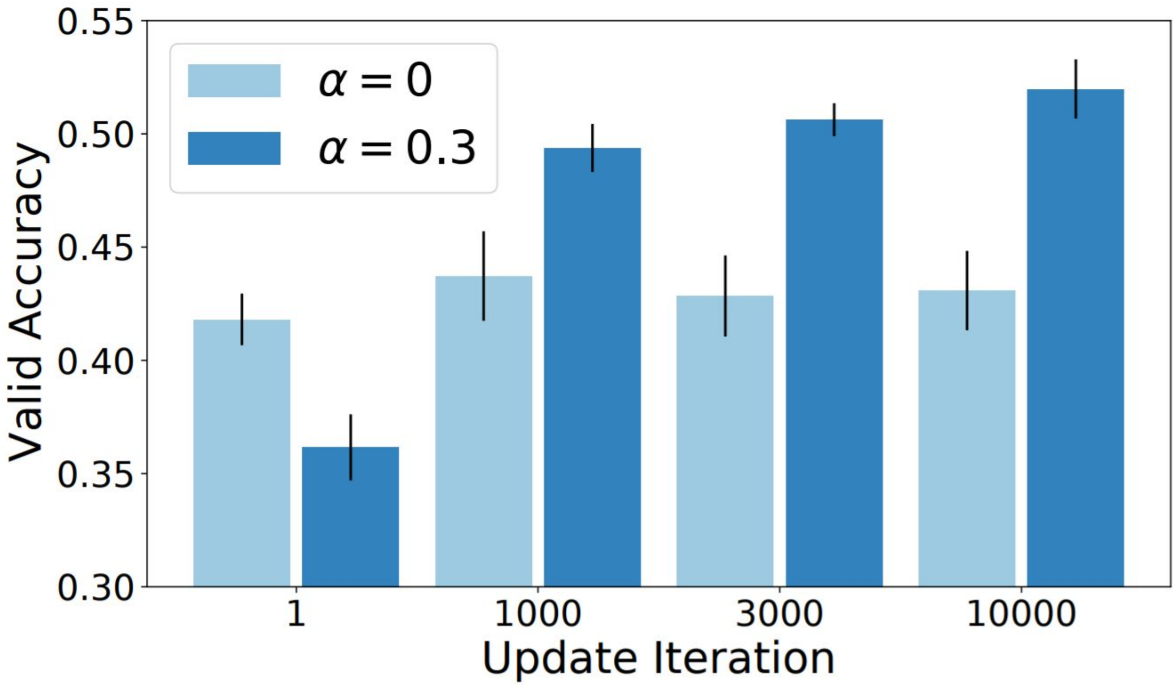
$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{M}) = l(\mathcal{B}; \boldsymbol{\theta}) + \alpha \sum_{m=1}^{M} \| f_{\boldsymbol{\theta}}(\mathbf{x}_m) - \mathbf{g}_m \|_2^2 + \beta \sum_{m=1}^{M} l((\mathbf{x}_m, y_m); \boldsymbol{\theta})$$

fitting loss — logit regularization — label regularization

[1] *Buzzega et al., 2020*

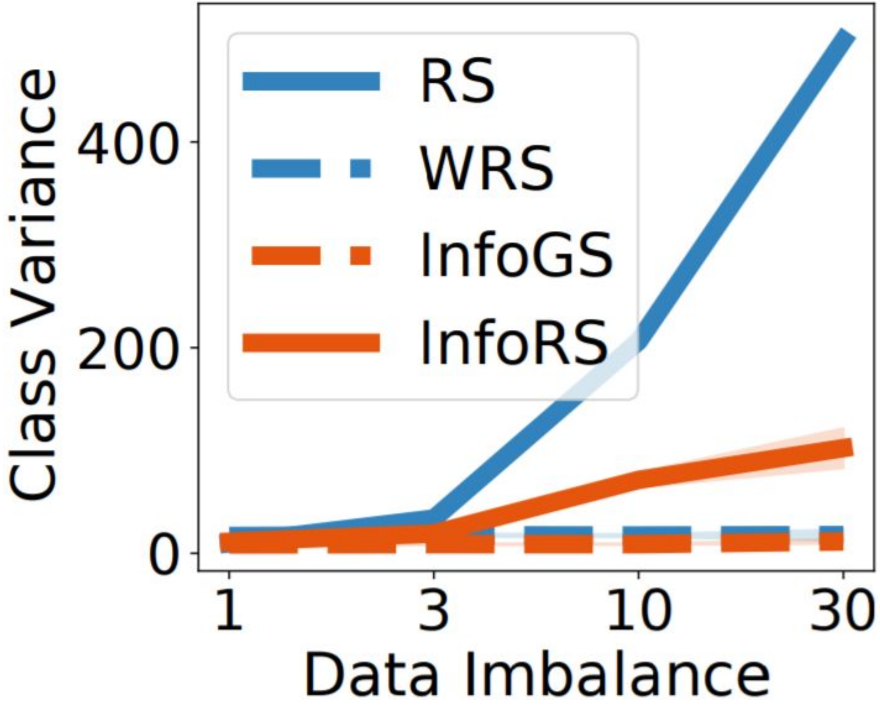
Continual Learning: the timing to update memory

- Besides *how to update the memory*, *when to update the memory* is also important,



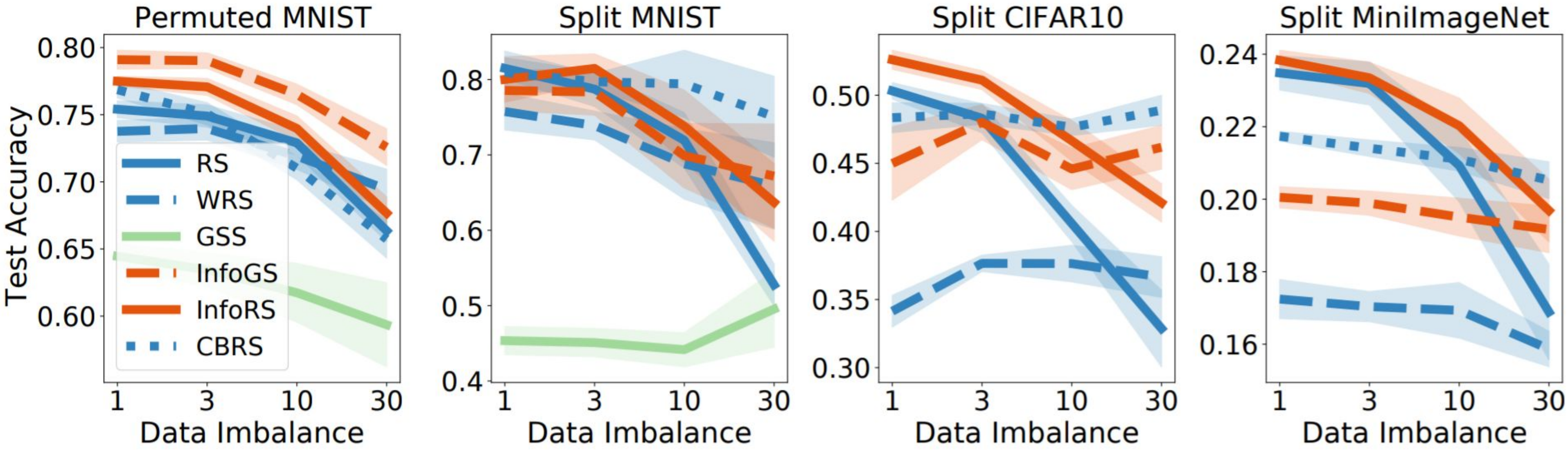
Continual Learning: Experiments

- InfoRS achieves a more balanced buffer.



Continual Learning: Experiments

- More baselines,



Continual Learning: Experiments

- An ablation study for InfoGS and InfoRS,