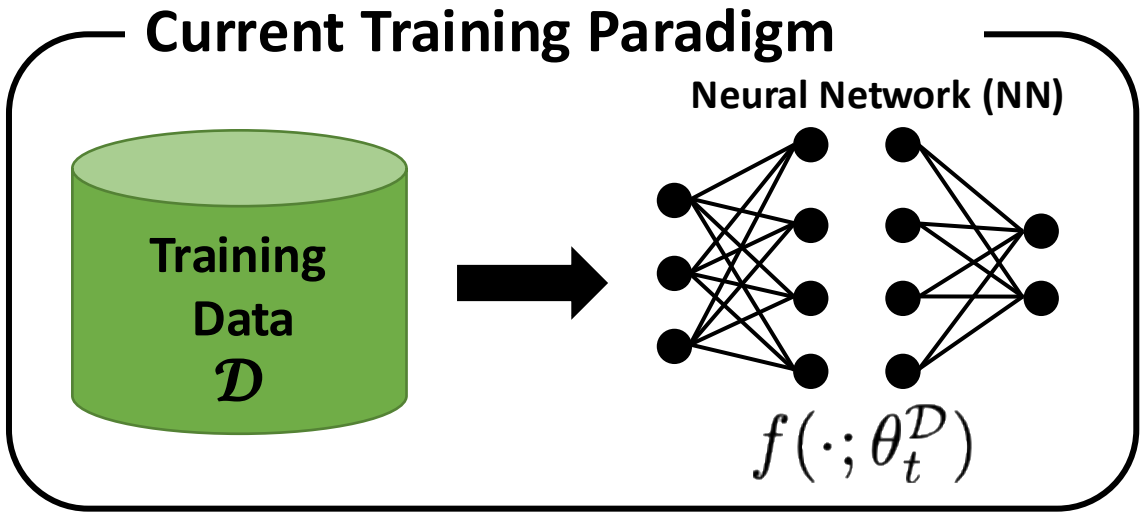
Dataset Distillation Efficiently Encodes Low-Dimensional Representations from Gradient-Based Learning of Non-Linear Tasks

Yuri Kinoshita[1,2], Naoki Nishikawa[1,3], Taro Toyoizumi[1,2]

[1]Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan.
[2]Laboratory for Neural Computation and Adaptation, RIKEN Center for Brain Science, Wako, Japan.
[3]RIKEN Center for Advanced Intelligence Project, Nihonbashi, Japan.

1. Introduction

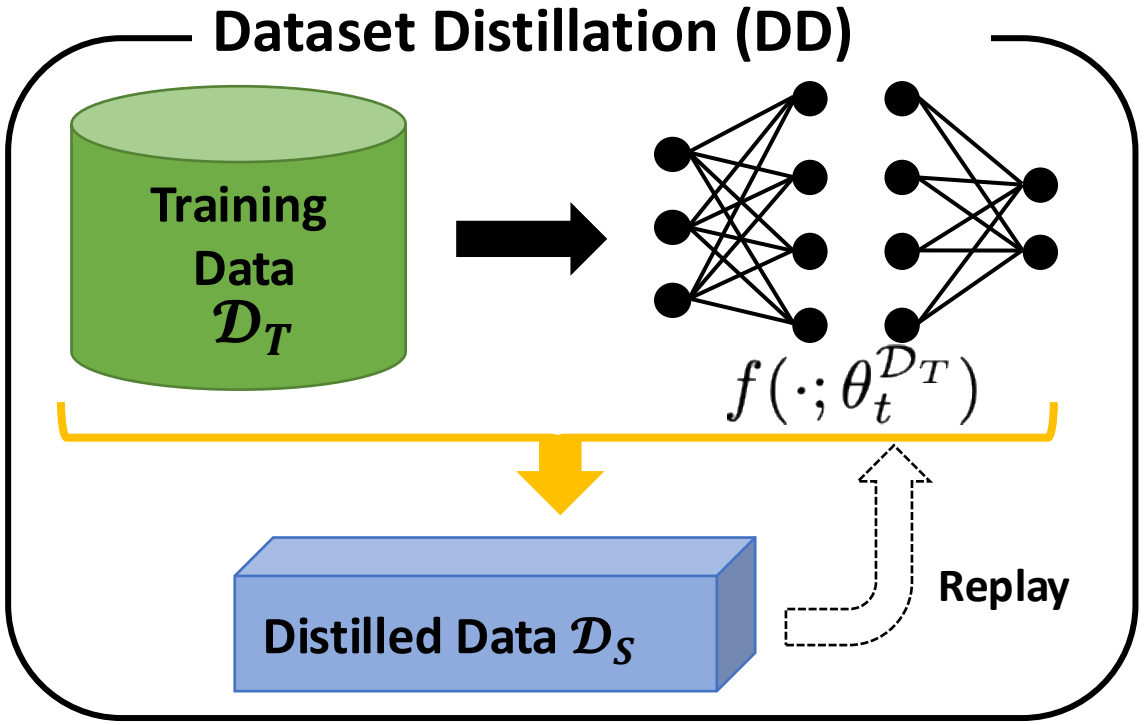
1.1 Background



Huge Dataset size → **High** Training, Storage, Transmission Cost

1. Introduction

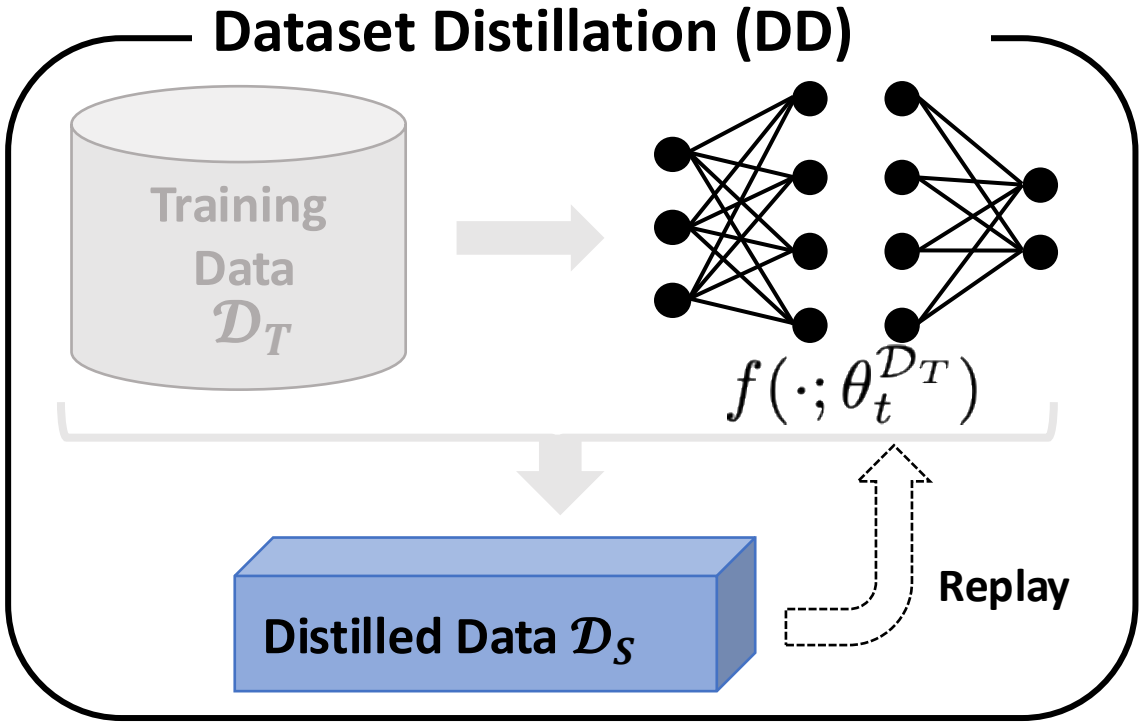
1.2 Dataset Distillation



Encode training information into synthetic data

1. Introduction

1.2 Dataset Distillation



Retraining with $\mathcal{D}_S$ mimics training with $\mathcal{D}_T$

Small Dataset size → Low Training, Storage, Transmission Cost

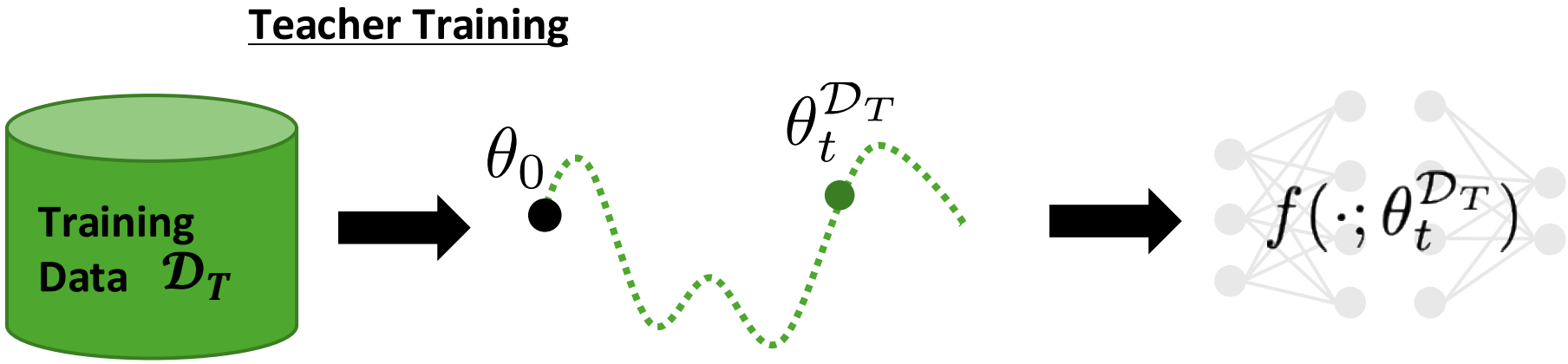
1. Introduction

1.2 Dataset Distillation

- **Reduce training, storage and transmission cost**
- **Powerful summaries for transferring task knowledge**
  - **Transfer learning** [Lee et al., 2024]
  - **Neural architecture search** [Zhao et al., 2024]
  - **Continual learning** [Liu et al., 2020]
  - **Federated learning** [Zhou et al., 2020]
  - **Data privacy** [Dong et al., 2022]
  - **Model interpretability** [Cazenavette et al., 2025]

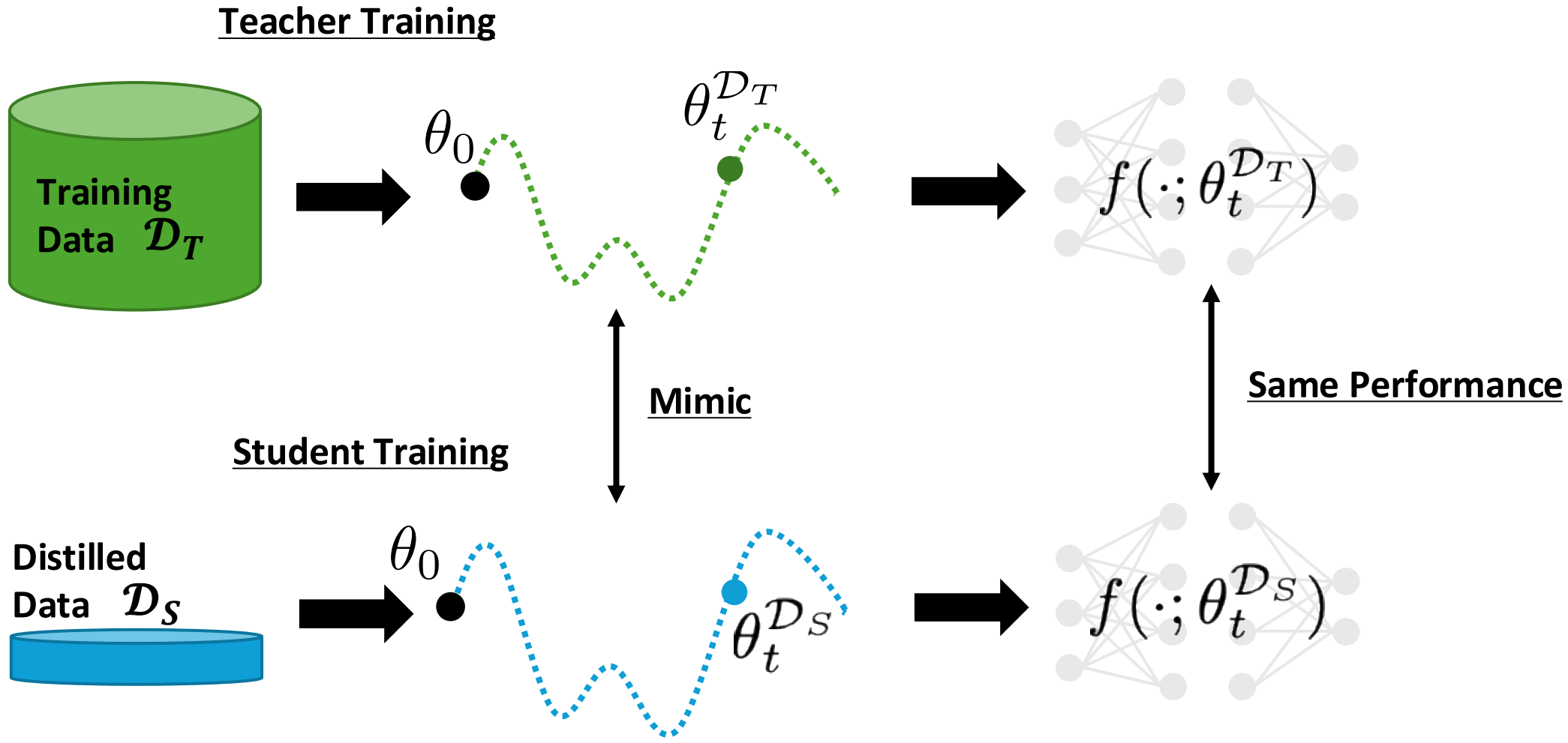
2. Dataset Distillation

2.1 Idea



2. Dataset Distillation

2.1 Idea



2. Dataset Distillation

2.2 Methods

Which aspects of teacher training should we prioritize?

- **Performance → Performance matching**
- **Gradient information → Gradient matching**
- **Data distribution → Distribution matching**
- **Optimization trajectory → Trajectory matching**

2. Dataset Distillation

2.2 Methods

Which aspects of teacher training should we prioritize?

- **Performance → Performance matching**
- **Gradient information → Gradient matching**
- **Data distribution → Distribution matching**
- **Optimization trajectory → Trajectory matching**

2. Dataset Distillation

2.3 Performance Matching (PM)

$$\min_{\mathcal{D}_S} \mathbb{E}_{\theta_0} \left[\mathcal{L}(\mathcal{D}_T; \theta_t^{\mathcal{D}_S}) \right]$$

Minimize the training loss of the model trained on the distilled data

- First DD method [Wang et al., 2018]
- Bilevel optimization
- Expectation over parameter initializations

2. Dataset Distillation

2.4 Gradient Matching (GM)

$$\min_{\mathcal{D}_S} \mathbb{E}_{\theta_0} \left[\sum_{t=0}^{T-1} d(\nabla_\theta \mathcal{L}(\mathcal{D}_T; \theta_t^{\mathcal{D}_S}), \nabla_\theta \mathcal{L}(\mathcal{D}_S; \theta_t^{\mathcal{D}_S})) \right]$$

Minimize the distance between the gradients of the teacher training and student training

- Only first order information [Zhao et al., 2021]
- More tractable
- Basis of trajectory matching

2. Dataset Distillation

2.5 Research Questions

What information is actually extracted from DD?

How many distilled data points are needed to reproduce performance?

2. Dataset Distillation

2.6 Prior Work

- Linear models
- Exact ridge regression
- Extracted information remains unclear
- No task structure
  - ➢ Distillation difficulty correlates with task difficulty

3. Our Contributions

3.1 Problem Setting

- **Nonlinear** models
- Gradient-based optimization, **closer to practice**
- Introduce **task structure**
  - ➢ Multi-index models

$$f^*(x) = \sigma^*(B^\top x), \quad B^\top \in \mathbb{R}^{r \times d},\ r \ll d$$

p-degree polynomial

Projection onto a latent dimension

Essential information is constrained on a subspace

3. Our Contributions

3.2 Theoretical Results

What information is actually extracted from DD?

How many distilled data points are needed to reproduce performance?

3. Our Contributions

3.2 Theoretical Results

What information is actually extracted from DD?

✓ **Latent structure is encoded into distilled data**

How many distilled data points are needed to reproduce performance?

3. Our Contributions

3.2 Theoretical Results

What information is actually extracted from DD?

✓ **Latent structure is encoded into distilled data**

How many distilled data points are needed to reproduce performance?

✓ **Memory Complexity of** $\sim r^2 d + L$

d: input dimension, r: latent dimension, L: width of neural network

3. Our Contributions

3.2 Theoretical Results

What information is actually extracted from DD?

✓ **Latent structure is encoded into distilled data**

How many distilled data points are needed to reproduce performance?

✓ **Memory Complexity of** $\sim r^2 d + L$

d: input dimension, r: latent dimension, L: width of neural network

To achieve error below ϵ, raw samples would need memory complexity of

$$\sim \max\{d^3, d^2/\epsilon, d/\epsilon^4\}$$

[Damian et al., 2022]