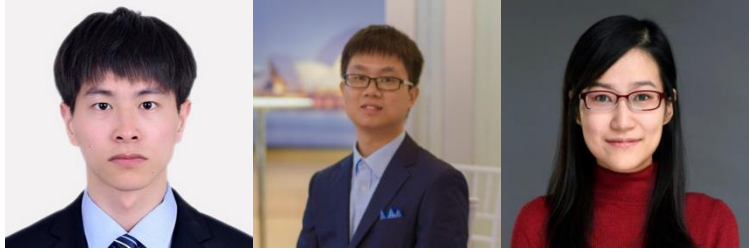


# Provable and Efficient Dataset Distillation for Kernel Ridge Regression

**Yilan Chen**[1], Wei Huang[2], Tsui-Wei Weng[1]

[1]UCSD, [2]RIEKN AIP

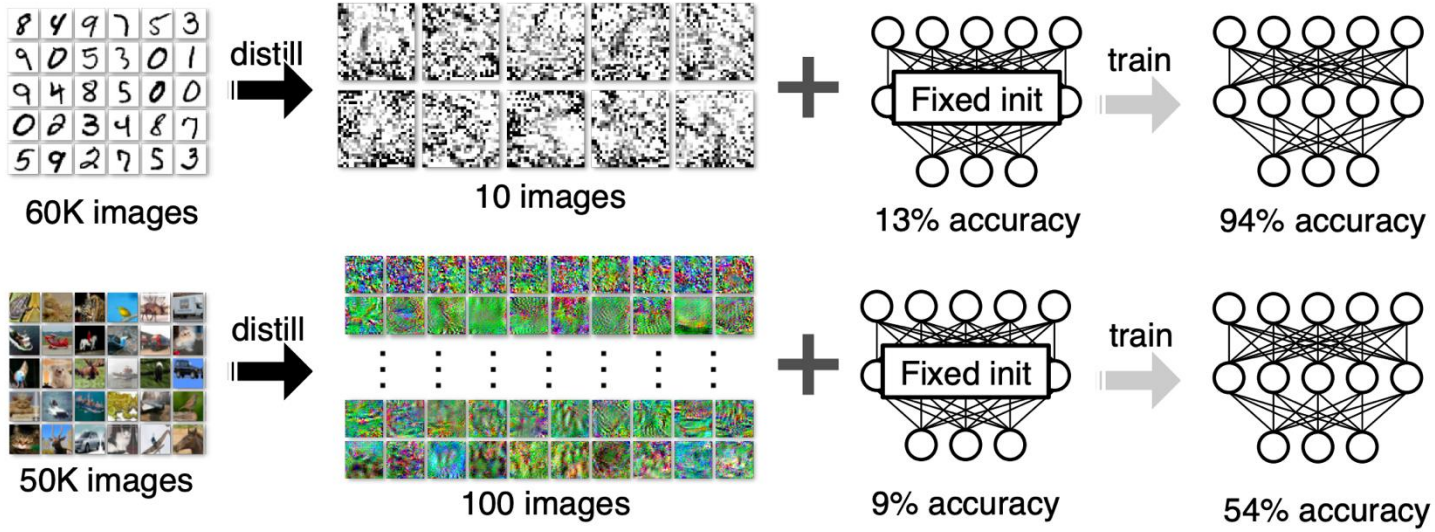
NeurIPS, December 2024

# Dataset Distillation

Motivation: Deep learning models are trained on increasingly larger datasets. It is crucial to reduce computational costs and improve data quality

- LLaMa 3 was pre-trained on over 15 trillion tokens
- Cost of training GPT-4 exceeded $100 million

Dataset distillation: distill a large dataset into a small synthesized dataset such that models trained on it can achieve similar performance to those trained on the original dataset.




(a) Dataset distillation on MNIST and CIFAR10

# Dataset Distillation

Many empirical works:

- Meta-model Matching
- Gradient Matching
- Trajectory Matching
- Distribution Matching

# Dataset Distillation

Few theoretical analysis:

- [Izzo and Zou, 2023] linear ridge regression (LRR) needs $d$ data points to recover original model's performance. Kernel ridge regression (KRR) needs $n$ data
- [Maalouf et al., 2023] use Random Fourier Features (RFF) to approximate shif-invariant kernels and construct $p$ distilled data for such RFF model. $p$: dimension of the RFF

This work:

- for KRR, one data point per class is already necessary and sufficient to recover the original model's performance in many settings
- analytical solutions for distilled dataset

Zachary Izzo and James Zou. A theoretical study of dataset distillation. NeurIPS 2023 M3L Workshop.
Alaa Maalouf, Murad Tukan, Noel Loo, Ramin Hasani, Mathias Lechner, and Daniela Rus. On the size and approximation error of distilled sets. NeurIPS 2023.

# Problem formulation

Original dataset: $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{d \times n} \times \mathbb{R}^{k \times n}$
Distilled dataset: $(\mathbf{X_S}, \mathbf{Y_S}) \in \mathbb{R}^{d \times m} \times \mathbb{R}^{k \times m}$

Kernel ridge regression: $f(\boldsymbol{x}) = \mathbf{W}\,\phi(\boldsymbol{x})$, where $\phi: \mathbb{R}^d \to \mathbb{R}^p$ and $\mathbf{W} \in \mathbb{R}^{k \times p}$

$$\min_{\mathbf{W}} \|\mathbf{W}\,\phi(\mathbf{X}) - \mathbf{Y}\|_F + \lambda \|\mathbf{W}\|_F$$

Analytical solution: $\mathbf{W} = \mathbf{Y}\phi_\lambda(\mathbf{X})^+$

$$\phi_\lambda(\mathbf{X})^+ = \begin{cases} \left(\phi(\mathbf{X})^{\mathrm{T}}\phi(\mathbf{X}) + \boldsymbol{\lambda}\,\boldsymbol{I_n}\right)^{-1}\phi(\mathbf{X}), & \boldsymbol{\lambda} > \mathbf{0} \\ \phi(\mathbf{X})^+, & \boldsymbol{\lambda} = \mathbf{0} \end{cases}$$

Similarly for distilled dataset model: $f_S(\boldsymbol{x}) = \mathbf{W_S}\,\phi(\boldsymbol{x})$

Goal of dataset distillation: find $(\mathbf{X_S}, \mathbf{Y_S})$ such that $\mathbf{W_S} = \mathbf{W}$

# Provable and Efficient Dataset Distillation for Kernel Ridge Regression

## Comparison with existing work

Table 1: Comparison with existing theoretical analysis of dataset distillation. The number of distilled data needed to recover original model's performance and models analyzed. "-" means not applicable. For linear ridge regression (LRR) and kernel ridge regression (KRR) with subjective feature mapping, our results only need one distilled data per class ($k \leq d$ in our setting), which is far less than the existing work [9, 21] that require $n$ or $p$ points. As an example, $k = 10, d = 3072, n = 50000$ for CIFAR-10. The $k, d, n$ of standard datasets are listed in Table 2. $p$ is the dimension of feature mapping $\phi : \mathbb{R}^d \mapsto \mathbb{R}^p$.

| | **LRR** | **Kernel ridge regression (KRR)** | |
|---|---|---|---|
| | | surjective $\phi$ | non-surjective $\phi$ |
| Izzo and Zou [9] | $d$ | - | $n$ (Gaussian Kernel) |
| Maalouf et al. [21] | - | - | $p$ (Shift-invariant Kernels) |
| **Our work** | $k$, $(k \leq d)$ | $k$, $(k \leq p)$ (Invertible NNs, FCNN, CNN, Random Fourier Features) | $p$ in general (Deep nonlinear NNs). $k + 1$ for deep linear NNs |

# Main result 1: Linear Ridge Regression

Linear ridge regression

**Theorem**

- When $m < k$ there is no $\mathbf{X_S}$ can guarantee $\mathbf{W_S} = \mathbf{W}$ unless the columns of $\mathbf{W}$ are in the range space of $\mathbf{Y_S}$.
- When $m \geq k$ and $\mathbf{Y_S}$ is rank $k$, let $r = \min(m, d)$ and take $\mathbf{D} = \mathbf{Y_S^+ W} + (\mathbf{I_m} - \mathbf{Y_S^+ Y_S})\mathbf{Z}$, where $\mathbf{Z} \in \mathbb{R}^{d \times m}$ is any matrix. Suppose the reduced SVD of $\mathbf{D}$ is $\mathbf{D} = \mathbf{V}\,\mathbf{diag}(\sigma_1', \ldots, \sigma_r')\mathbf{U^T}$ with $\sigma_1' \geq \cdots \geq \sigma_r' \geq 0$.
  1. $\lambda_S > 0$: $\mathbf{W_S} = \mathbf{W}$ if and only if for any $\mathbf{D}$ defined above, $\lambda_S < \frac{1}{4\sigma_1'}$ and $\mathbf{X_S} = \mathbf{U}\,\mathbf{diag}(\sigma_1, \ldots, \sigma_r)\mathbf{V^T}$ where $\sigma_i = \begin{cases} 0, & \text{if } \sigma_i' = 0, \\ \frac{1 \pm \sqrt{1 - 4\lambda_S \sigma_i'}}{2\sigma_i'}, & \text{otherwise.} \end{cases}$
  2. $\lambda_S = 0$: $\mathbf{W_S} = \mathbf{W}$ if and only if $\mathbf{X_S} = \mathbf{D^+}$ for any $\mathbf{D}$ defined above.

# Main result 1: Linear Ridge Regression

- One distilled data per class is necessary and sufficient for $\mathbf{W_S} = \mathbf{W}$.
- Intuitively, original dataset $(\mathbf{X}, \mathbf{Y})$ is compressed into $\mathbf{X_S}$ through original model's parameter $\mathbf{W}$.
  - When $m = k$, only one solution. When $\lambda_S = 0$, $\mathbf{X_S} = (\mathbf{Y_S^+ W})^+$.
  - When $m > k$, infinitely many distilled datasets since $\mathbf{Z}$ is a free variable to choose
  - When $m = n$, $(\mathbf{X}, \mathbf{Y})$ is a distilled dataset for itself.
  - When $m > n$, can generate more data than original dataset.

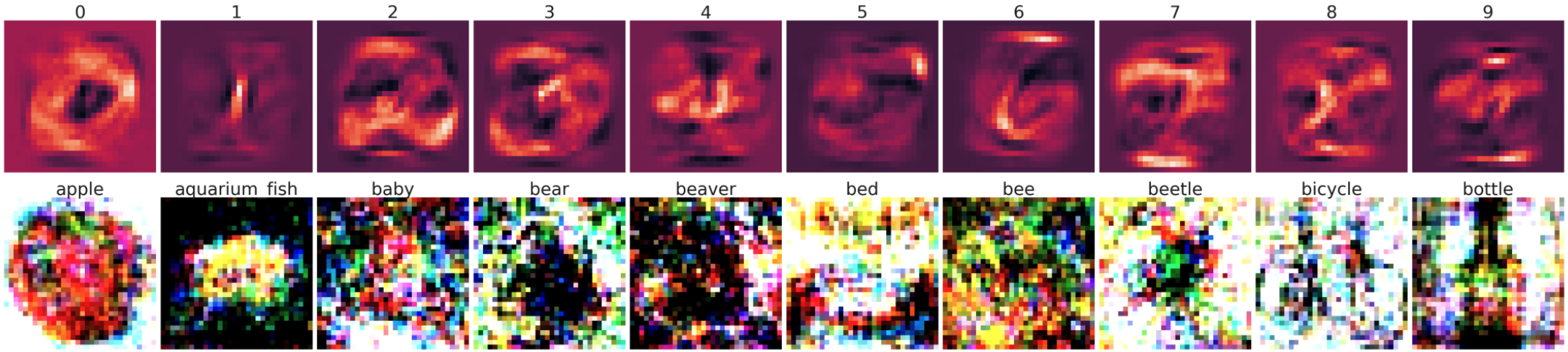
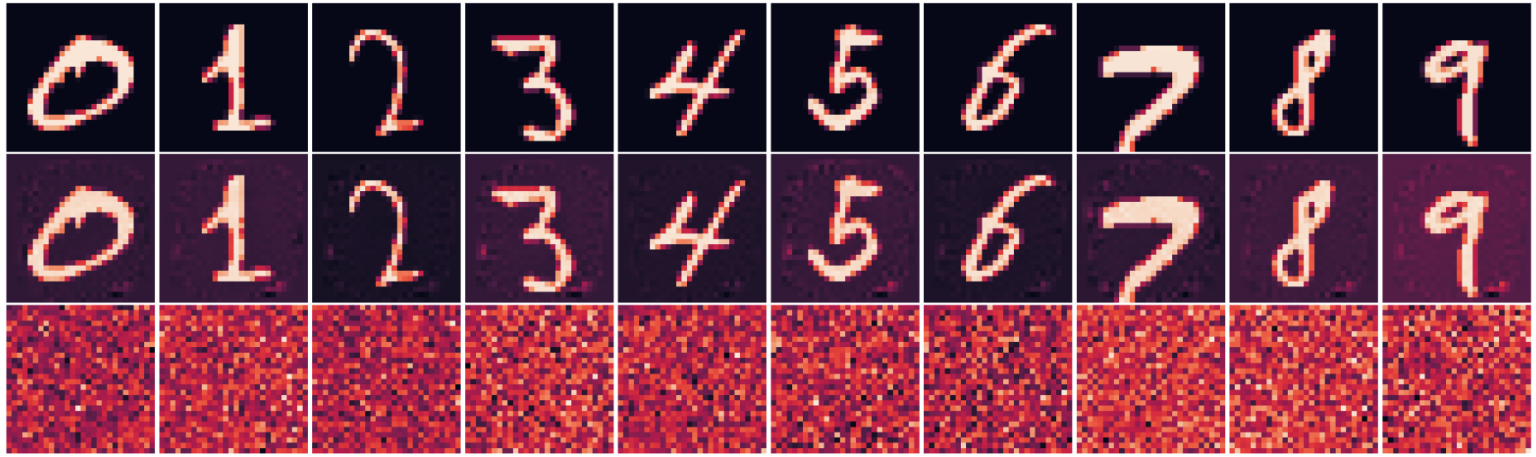



Figure 1: Distilled data of MNIST (first row) and CIFAR-100 (second row) for LRR when $m = k$.

# Main result 1: Linear Ridge Regression

1. Find realistic distilled data that is close to original data by solving closed-from solution
2. Generate distilled data from random noise



(a) MNIST with IPC=50.

(b) CIFAR-100 with IPC=5.

Figure 2: Initialized data (first row), distilled data generated from real images using techniques in Sec 4.2 (second row), and distilled data generated from random noise using techniques in Sec 4.1 (third row) for a LRR with $m = 500$ on MNIST and CIFAR-100. IPC: images per class.

# Main result 2: Kernel Ridge Regression

- The results can be extended to KRR by replacing $\mathbf{X_S}$ with $\phi(\mathbf{X_S})$
- When $\phi$ is surjective or bijective, can always find a $\mathbf{X_S}$ for a desired $\phi(\mathbf{X_S})$

Examples of surjective $\phi$: $p \leq d$

1. Invertible NNs
2. Fully-connected NN (FCNN)
3. Convolutional Neural Network (CNN)
4. Random Fourier Features (RFF)

# Main result 2: Kernel Ridge Regression

Non-surjective Feature Mapping

- For non-surjective $\phi$ such as deep nonlinear NNs, one data per class is generally not sufficient as long as $(\mathbf{Y}_S^+ \mathbf{W})^+$ is not in the range space of $\phi$
- For deep linear NNs, we show $m = k + 1$ can be sufficient under certain conditions

For surjective $\phi$, our algorithm outperforms previous work such as KIP while being significantly more efficient

Table 4: Comparison between our algorithm and KIP.

| Dataset | IPC | KIP [25] | | **Ours** | | |
|---|---|---|---|---|---|---|
| | | Accuracy ↑ | Cost ↓ (GPU Sec.) | Accuracy ↑ | Cost ↓ (GPU Sec.) | Speedup over KIP ↑ |
| MNIST | 1 | 93.44±0.17 | 159 | **93.72±0.14** | 16 | **9.9**× |
| | 10 | **93.75±0.10** | 554 | 93.69±0.17 | 16 | **34.6**× |
| | 50 | **93.72±0.11** | 3114 | 93.62±0.24 | 16 | **194.6**× |
| CIFAR-10 | 1 | 45.83±0.29 | 225 | **47.85±0.10** | 21 | **10.7**× |
| | 10 | 47.50±0.29 | 594 | **47.76±0.12** | 20 | **29.7**× |
| | 50 | 47.48±0.20 | 3510 | **47.77±0.06** | 20 | **175.5**× |
| CIFAR-100 | 1 | 20.08±0.20 | 616 | **21.58±0.15** | 20 | **30.8**× |
| | 10 | 21.56±0.16 | 9323 | **21.59±0.15** | 20 | **466.1**× |
| | 50 | - | ∼396000 | **21.58±0.13** | 25 | ∼**15840.0**× |

# Future works

- Determining minimum number of distilled data points required for non-surjective deep neural networks
- Extending our analysis to learnable feature mappings