A Theoretical Study of Dataset Distillation

Zachary Izzo
Machine Learning Department
NEC Labs America
zach@nec-labs.com

James Zou
Department of Biomedical Data Science
Stanford University
jamesz@stanford.edu

Abstract

Modern machine learning models are often trained using massive amounts of data. Such large datasets come at a high cost in terms of both storage and computation, especially when the data will need to be used repeatedly (e.g., for neural architecture search or continual learning). *Dataset distillation* (DD) describes the process of constructing a smaller "distilled" dataset (usually consisting of synthetic examples), such that models trained on the distilled dataset will be similar to models trained on the original dataset. In this paper, we study DD from a theoretical perspective. We show that for generalized linear models, it is possible to construct a distilled dataset with only a *single point* which will exactly recover the model trained on the original dataset, regardless of the original number of points. We provide a specialized distillation for linear regression with size independent of the original number of points, but which perfectly reconstructs the model obtained from the original dataset with *any* data-independent regularizer, or by combining the original dataset with any additional data. We also provide impossibility results showing that similar constructions are impossible for logistic regression, and that DD cannot be accomplished in general for kernel regression, even if the goal is only to recover a single model.

1 Introduction

Training data is the fuel which powers machine learning. As the amount of data collected increases day by day, the computational efficiency of learning from such massive datasets becomes an ever more pressing problem. Training a single model on a vast amount of training data may itself be feasible, but this is often only a fraction of the computational load. Practitioners will frequently need to train many different models with thousands of possible hyperparameter settings in order to train a performant model. If the data are not collected all at once, old data must be stored and periodically reused in order to prevent the model from forgetting the knowledge it contains. Some methods, such as kernel-based approaches, also have inference costs which scale with the size of the training data. In these instances, massive training datasets can impose an undesirable, or even infeasible, computational burden.

To address these problems, [25] introduced dataset distillation. The goal is to "distill" a large training dataset into a smaller synthetic set, such that training ML models on the smaller distilled set approximates model training on the much larger original dataset. Besides the purely scientific question of how much the original data can be compressed, effective dataset distillation methods can be applied to hyperparameter tuning, neural architecture search, and continual learning, reducing the computational cost of these tasks. There are other benefits such as reduced memory footprint and even improved data privacy [4].

While there has been a great deal of empirical progress on DD, many of its theoretical properties still have not been studied in the literature. In this work, we seek to begin to fill this theoretical gap. Aside from answering questions of purely academic interest, a more complete theory can eventually

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

lead to methods which are more principled and computationally efficient than the current SOTA. Our contributions are as follows:

- For GLMs, we show that we can achieve perfect single-model dataset distillation (i.e., construct a distilled dataset which exactly recovers the model trained on the full training set) with only a *single* synthetic point.
- For linear regression, we show that a synthetic dataset of size d (where d is the dimension of the features) is necessary and sufficient to recover the *full regularization path* of model training on the original data, with *any* data-independent regularizer.
- We show that it is not possible to compress the original dataset while preserving the regularization path for logistic regression. We also show that compression is impossible for kernel regression, even if we only desire to recover a single model.

2 Related Work

Dataset distillation was originally introduced by [25]. Since their original paper, there has been a growing body of work on the topic, seeking to make improvements both on downstream performance and computational efficiency. Many works perform distillation by simultaneously training a "student" model and optimizing the distilled dataset, so that the training trajectory of the student model matches that of the original model on the original data [1, 5]. Another approach which is simimlar in spirit seeks to match loss gradients between the synthetic and real data, which implicitly results in similar training trajectories; this approach is commonly referred to as dataset *condensation* [32, 30, 8, 7]. Other approaches seek to directly capture properties of the data distribution, rather than enforcing similar training trajectories [24, 31]. There are also works which specifically address the computational or memory burdens of previous methods, allowing them to scale to larger datasets such as CIFAR-100 or ImageNet [14, 3]. Dataset distillation has found applications in a variety of areas. Two commonly cited use cases are continual learning [12, 16, 27, 18] and neural architecture search [21], but other uses include privacy [22, 2, 4, 11], robust learning [28, 23] and federated learning [6, 10, 20]. We refer the reader to the works of [9, 17, 29] for a more comprehensive review, and to the blog post of [26] for an up-to-date list of works in the area.

While most works in the area are empirical in nature, there are some exceptions. [13] provide theoretical results for the case of single-model distillation in ridge-regularized linear regression. We extend their results to the case of logistic regression with cross-entropy loss, and we improve on their results for linear regression to account for not only single-model distillation, but also recover regularization paths and dataset combination.

3 Notation & Formal Problem Statement

In this section, we introduce notation and formally state the types of DD we will study. We will use $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R}$ to refer to a generic dataset. Generally, $\mathcal{D}_{\text{orig}}$ will refer to the original training set and $\mathcal{D}_{\text{dist}}$ will referred to the distilled dataset. We use θ to refer to model parameters. We overload notation and use $\theta(\mathcal{D})$ to refer to the model resulting from training on the dataset $\mathcal{D}$. The loss function used for training will be supplied in context. We also use the notation $\theta^R(\mathcal{D})$ to refer to the model resulting from training of $\mathcal{D}$ with (data-independent) regularizer $R(\theta)$. We use the notation θ^λ to refer to training with ridge regularization with strength λ, i.e. $R(\theta) = \lambda\|\theta\|^2$, where $\|\cdot\|$ denotes the Euclidean norm. We will sometimes abbreviate the notation and write $\theta^\lambda_{\text{orig}} = \theta^\lambda(\mathcal{D}_{\text{orig}})$ and $\theta^\lambda_{\text{dist}} = \theta^\lambda(\mathcal{D}_{\text{dist}})$.

Throughout the paper, we will consider three possible tasks for DD. In *single-model distillation*, we seek to just recover the single model obtained from a fixed training procedure on the original dataset. While this is not one of the practical use cases for DD, studying single-model distillation can help inform our intuition for more realistic DD scenarios.

Definition 1. *A dataset* $\mathcal{D}_{\text{dist}}$ *is a* single-model distillation (SD) *of* $\mathcal{D}_{\text{orig}}$ *if* $\theta(\mathcal{D}_{\text{dist}}) = \theta(\mathcal{D}_{\text{orig}})$.

In *hyperparameter distillation*, we require the ability to recover the models obtained from the original dataset with different choices of hyperparameters. In this work, the hyperparameters will be different regularization strengths. This setting is a simplified version of using DD to speed up more general hyperparameter tuning tasks during model training.

Definition 2. *A dataset $\mathcal{D}_{\text{dist}}$ is a* hyperparameter distillation (HD) *of $\mathcal{D}_{\text{orig}}$ if $\theta^\lambda(\mathcal{D}_{\text{dist}}) = \theta^\lambda(\mathcal{D}_{\text{orig}})$ for all $\lambda \geq 0$.*

Finally, in *combination distillation*, we require that combining the distilled dataset with any other dataset $\mathcal{D}$ should produce the same result as combining the original dataset with $\mathcal{D}$. This is desirable if dataset distillation is used to prevent catastrophic forgetting in continual learning.

Definition 3. *A dataset $\mathcal{D}_{\text{dist}}$ is a* combination distillation (CD) *of $\mathcal{D}_{\text{orig}}$ if $\theta(\mathcal{D}_{\text{dist}} \cup \mathcal{D}) = \theta(\mathcal{D}_{\text{orig}} \cup \mathcal{D})$ for any other dataset $\mathcal{D}$. Here, the union is considered as a union of multisets.*

It is tautologically clear that $\mathcal{D}_{\text{orig}}$ is always a distillation of itself, for any of the three DD tasks. Thus we will be interested in distillation procedures for which $|\mathcal{D}_{\text{dist}}| \ll |\mathcal{D}_{\text{orig}}|$.

4 Theoretical Results

4.1 Single-Model Distillation

We begin by considering single-model distillation, the simplest task for DD. We show that for a broad class of models (GLMs), the model trained on the original dataset can be encoded using only a distilled dataset containing only a single-point (Theorem 1). On the other hand, any amount of compression is impossible in general for kernel regression models (Theorem 2). We begin with the positive result.

We refer the reader to [15] for an introduction to GLMs. We consider GLMs of the following form: we use the canonical link function g so that $\mathbb{E}[y|x] = g^{-1}(x^\top \theta)$ for some θ. Conditional on x, y belongs to a generalized exponential family with density given by

$$f(y; x, \theta) = h(y) \exp(yx^\top \theta - b(x^\top \theta))$$

for some functions h and b. Given a dataset $\{(x_i, y_i)\}_{i=1}^n$, the model parameters θ_{orig} can be trained by minimizing the negative log-likelihood, which takes the following form:

$$\mathcal{L}(\theta) = -\sum_{i=1}^n (y_i x_i^\top \theta - b(x_i^\top \theta)) + (\text{terms independent of } \theta).$$

Theorem 1. *For any GLM initialized at 0 and trained via gradient descent on the negative log-likelihood, the single-point dataset $\mathcal{D}_{\text{dist}} = \{(\theta_{\text{orig}}, g^{-1}(\|\theta_{\text{orig}}\|^2))\}$ accomplishes single-model dataset distillation.*

The setting of Theorem 1 includes unregularized linear and logistic regression. Single-model distillation is also possible for these special cases when ridge regularization is applied. Due to space constraints, we defer these results to the appendix.

We have shown that it is possible to encode the weights for any GLM in a single datapoint, the best compression result which could possibly hold. However, for kernel regression models, we find the opposite extreme: in general, it is impossible to produce a condensed dataset which is smaller than the original dataset.

We consider a standard kernel regression setting [19]. Let $\mathcal{H}$ be a RKHS with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, norm $\|w\|_{\mathcal{H}}^2 = \langle w, w \rangle_{\mathcal{H}}$, associated kernel K, and feature map $\phi : \mathbb{R}^d \to \mathcal{H}$, so that $\langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = K(x, x')$ for all $x, x' \in \mathbb{R}^d$. In kernel regression, given a dataset $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R}$, the goal is to find

$$\min \|w\|_{\mathcal{H}}^2 \quad \text{s.t.} \quad w \in \operatorname{argmin} \sum_{i=1}^n (y_i - \langle w, x_i \rangle_{\mathcal{H}})^2.$$

By the representer theorem, the solution w^* has the form $w^* = \sum_{i=1}^n \alpha_i K(\cdot, x_i)$ for some coefficients α_i. Kernel regression allows us to model nonlinear relationships between the covariates and target variable. However, this increase in expressive power also means that the information contained in the original dataset cannot be perfectly distilled into a smaller dataset.

Theorem 2. *Suppose that the original dataset $\mathcal{D}_{\text{orig}} = \{(x_i, y_i)\}_{i=1}^n$ is such that (x_i, y_i) are independent for each i, and x_i and $y_i|x_i$ have distributions which are absolutely continuous with respect to the Lebesgue measure. Then with probability 1, any exact single-model distillation for Gaussian kernel regression must contain at least n points.*

We remark that the proof does not rely heavily on the specific choice of the Gaussian kernel. Any kernel K where the functions $K(\cdot, x_i)$ are linearly independent and the coefficients α_i are nonzero with probability 1 will suffice. The same result also holds for kernel ridge regression.

4.2 Hyperparameter and Combination Distillation

Next, we examine the more challenging scenarios of hyperparameter and combination distillation. The special structure of the loss for linear regression means that both of these can be accomplished simultaneously with the same distillation (Theorem 3), the size of which is independent of the original number of datapoints. However, this does not generalize to all GLMs, and distillation with fewer than the original number of datapoints is provably impossible for logistic regression (Theorem 4).

For linear regression, we will consider the unnormalized square loss $\mathcal{L}(\theta) = \sum_{i=1}^{n}(x_i^\top \theta - y_i)^2$. (The lack of a normalizing n^{-1} does not change any of our results, as this can simply be absorbed into the (x_i, y_i); we use this convention for notational convenience in the proofs.)

Theorem 3. *Let $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \mathbb{R}$ be a linear regression dataset, and let $X \in \mathbb{R}^{n \times d}$ be the associated design matrix. We assume that $n \geq d$ and the x_i are in general position. Let $v_1, \ldots, v_d \in \mathbb{R}^d$ be the right singular vectors of X with associated singular values $\sigma_1, \ldots, \sigma_d$. Finally, define $\tilde{x}_i = \sigma_i v_i$ and $\tilde{y}_i = \theta_{\text{orig}}^\top \tilde{x}_i$. Then the distilled dataset $\mathcal{D}_{\text{dist}} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^d$ accomplishes hyperparameter and combination distillation.*

We can actually show a stronger result, namely that $\mathcal{D}_{\text{dist}}$ satisfies $\theta^R(\mathcal{D}_{\text{dist}}) = \theta^R(\theta_{\text{orig}})$ for *any* data-independent regularizer $R(\theta)$. Unforunately, this strong positive result for linear regression does not extend to the slightly more complicated case of logistic regression.

We consider a logistic regression model trained with the (normalized) cross-entropy loss: $\mathcal{L}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log h_\theta(x_i) + (1 - y_i)\log(1 - h_\theta(x_i)))$, where $h_\theta(x) = 1/(1 + e^{-\theta^\top x})$. (Again, our decision to include the normalizing constant in this case does not change the result.) Unlike linear regression, no compression via DD is possible in this case.

Theorem 4. *Let $\{(x_i, y_i)\}_{i=1}^n \subseteq \mathbb{R}^d \times \{0, 1\}$ be a binary classification dataset such that all of the x_i are distinct. Then any distilled dataset which accomplishes exact hyperparameter or combination distillation must contain at least n points.*

Linear and logistic regression both belong to the class of generalized linear models. The conflicting results in these two cases regarding hyperparameter and combination distillation mean that we cannot generally conclude whether HD and CD are possible (or impossible) for GLMs.

5 Discussion

In this work, we studied the problem of dataset distillation from a theoretical perspective. We showed that it is possible to distill the training set for a GLM down to a single point, regardless of the original number of training points, when the goal is to simply recover the original model by training on the distilled data. We further showed exact hyperparameter and combination distillation (two commonly cited practical use cases for dataset distillation) are possible for linear regression, but compression is provably impossible for even the slightly more complicated case of logistic regression.

Limitations & Future Work For the case of logistic regression, Theorem 4 shows that compression is not possible for exact hyperparameter distillation with ridge regularization, provided that the *same regularization parameter is used* on both the original and distilled datasets. We conjecture that a stronger result should hold, namely that the regularization *paths* of the original and distilled datasets cannot coincide for any correspondence between the original and distilled regularization parameters.

Deriving results for *approximate* distillation would also be of practical interest. For instance, suppose we want to enforce a tolerance $\|\theta_{\text{orig}}^\lambda - \theta_{\text{dist}}^\lambda\| \leq \varepsilon$ for all $\lambda \geq 0$; how much compression is possible in this case, for either kernel or logistic regression (and similarly for combination distillation)?

Lastly, a common use case for DD is neural architecture search. In this setting, it is important that the same distilled dataset will perform well across different network architectures. This case is not covered by any of our theory. Extending our analyses to this setting (perhaps by revisiting the kernel regression setting and deriving approximate results for kernels which are sufficiently similar, corresponding to sufficiently similar network architectures) is also a potentially fruitful line of future inquiry.

References

[1] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4750–4759, 2022.

[2] Dingfan Chen, Raouf Kerkouche, and Mario Fritz. Private set generation with discriminative information. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[3] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. *arXiv preprint arXiv:2211.10586*, 2022.

[4] Tian Dong, Bo Zhao, and Lingjuan Liu. Privacy for free: How does dataset condensation help privacy? In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5378–5396, 2022.

[5] Jiawei Du, Yidi Jiang, Vincent T. F. Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. *arXiv preprint arXiv:2211.11004*, 2022.

[6] Jack Goetz and Ambuj Tewari. Federated learning via synthetic data. *arXiv preprint arXiv:2008.04489*, 2020.

[7] Zixuan Jiang, Jiaqi Gu, Mingjie Liu, and David Z. Pan. Delving into effective gradient matching for dataset condensation. *arXiv preprint arXiv:2208.00311*, 2022.

[8] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12352–12364, 2022.

[9] Shiye Lei and Dacheng Tao. A comprehensive survey to dataset distillation. *arXiv preprint arXiv:2301.05603*, 2023.

[10] Ping Liu, Xin Yu, and Joey Tianyi Zhou. Meta knowledge condensation for federated learning. *arXiv preprint arXiv:2209.14851*, 2022.

[11] Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Backdoor attacks against dataset distillation. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2023.

[12] Wojciech Masarczyk and Ivona Tautkute. Reducing catastrophic forgetting with learning on synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Workshop*, 2020.

[13] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[14] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 5186–5198, 2021.

[15] Philippe Rigollet. 18.650 (statistics for applications) lecture 10: Generalized linear models, 2016. URL `https://ocw.mit.edu/courses/18-650-statistics-for-applications-fall-2016/resources/mit18_650f16_glm/`.

[16] Andrea Rosasco, Antonio Carta, Andrea Cossu, Vincenzo Lomonaco, and Davide Bacciu. Distilled replay: Overcoming forgetting through synthetic samples. *arXiv preprint arXiv:2103.15851*, 2021.

[17] Noveen Sachdeva and Julian McAuley. Data distillation: A survey. *arXiv preprint arXiv:2301.04272*, 2023.

[18] Mattia Sangermano, Antonio Carta, Andrea Cossu, and Davide Bacciu. Sample condensation in online continual learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022.

[19] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2002.

[20] Rui Song, Dai Liu, Dave Zhenyu Chen, Andreas Festag, Carsten Trinitis, Martin Schulz, and Alois Knoll. Federated learning via decentralized dataset distillation in resource-constrained edge environments. *arXiv preprint arXiv:2208.11311*, 2022.

[21] Felipe Petroski Such, Aditya Rawal, Joel Lehman, Kenneth Stanley, and Jeffrey Clune. Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 9206–9216, 2020.

[22] Ilia Sucholutsky and Matthias Schonlau. SecDD: Efficient and secure method for remotely training neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 15897–15898, 2021.

[23] Nikolaos Tsilivis, Jingtong Su, and Julia Kempe. Can we achieve robustness from data alone? *arXiv preprint arXiv:2207.11727*, 2022.

[24] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. CAFE: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12196–12205, 2022.

[25] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[26] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation project blog. `https://www.tongzhouwang.info/dataset_distillation/`, 2022. Accessed: 2022-12-12.

[27] Felix Wiewel and Bin Yang. Condensed composite memory continual learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[28] Yihan Wu, Xinda Li, Florian Kerschbaum, Heng Huang, and Hongyang Zhang. Towards robust dataset learning. *arXiv preprint arXiv:2211.10752*, 2022.

[29] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *arXiv preprint arXiv:2301.07014*, 2023.

[30] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021.

[31] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[32] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *ICLR*, 1(2):3, 2021.

A Proofs Deferred from Section 4.1

Proof of Theorem 1. The negative log-likelihood for a GLM on a dataset $\{(x_i, y_i)\}_{i=1}^n$ is given by $\ell(\theta) = -\sum_{i=1}^n (y_i x_i^\top \theta - b(x_i^\top \theta))$. (We ignore the term which does not depend on θ.) Its gradient is therefore

$$\nabla \ell(\theta) = -\sum_{i=1}^n (y_i - b'(x_i^\top \theta)) x_i = \sum_{i=1}^n (y_i - g^{-1}(x_i^\top \theta)) x_i.$$

Here we have used the fact that $b'(x^\top \theta) = g^{-1}(x^\top \theta)$ (see [15]). A simple inductive argument shows that the parameters of a model initialized at 0 and trained via GD on ℓ remains in the span of the x_i. In particular, for the distillation $\mathcal{D}_{\text{dist}}$, GD will arrive at parameters $\theta_{\text{dist}} = c\theta_{\text{orig}}$ for some scalar c. We claim that in fact $\theta_{\text{dist}} = \theta_{\text{orig}}$. If $\theta_{\text{orig}} = 0$, we are already done, so assume that $\theta_{\text{orig}} \neq 0$. Then the scalar c satisfies

$$0 = \nabla \ell(c\theta_{\text{orig}}) = (g^{-1}(\|\theta_{\text{orig}}\|^2) - g^{-1}(c\|\theta_{\text{orig}}\|^2))\theta_{\text{orig}}.$$

Since $\theta_{\text{orig}} \neq 0$, the result will follow immediately if g^{-1} is injective. This follows immediately from the fact that g^{-1} exists (i.e., g is invertible). This is because

$$g^{-1}(x) = g^{-1}(y) \implies x = g(g^{-1}(x)) = g(g^{-1}(y)) = y.$$

This completes the proof. □

Proof of Theorem 2. Let $K = (e^{-\|x_i - x_j\|^2})_{i,j=1}^n$ denote the kernel matrix and $y = (y_1, \ldots, y_n)^\top$ denote the response vector. The predictions for the kernel regression model trained on this dataset are given by

$$f(x) = \sum_{i=1}^n \alpha_i e^{-\|x - x_i\|^2},$$

where $\alpha = K^\dagger y$. (In fact, $\alpha = K^{-1} y$ since K is invertible with probability 1.) Note that the independence and absolute continuity conditions on x_i and $y_i | x_i$ imply that the x_i are distinct and $\alpha_i \neq 0$ for all i with probability 1.

For a Gaussian kernel regression model trained on any distilled dataset $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^k$, by the representer theorem, the resulting predictor takes the form

$$g(x) = \sum_{i=1}^k \beta_i e^{-\|x - \tilde{x}_i\|^2}$$

for some coefficients β_i. Setting $f(x) = g(x)$, the result holds since any collection $\{e^{-\|x - x_i\|^2}\}_{i=1}^N$ is linearly independent provided the x_i are unique. □

B Proofs Deferred from Section 4.2

Proof of Theorem 3. Let $X \in \mathbb{R}^{n \times d}$ denote the original data matrix and $Y \in \mathbb{R}^n$ denote the response vector. The training loss function is given by

$$\sum_{i=1}^n (\theta^\top x_i - y_i)^2 + R(\theta) = \theta^\top (X^\top X)\theta - 2\theta^\top (X^\top Y) + R(\theta) + Y^\top Y.$$

From this expression, we see that if $\widetilde{X}^\top \widetilde{X} = X^\top X$ and $\widetilde{X}^\top \widetilde{Y} = X^\top Y$ (where $\widetilde{X}$ and $\widetilde{Y}$ are the data matrix and response vector for the distilled dataset), then the losses on the original and distilled datasets are the same as a function of θ (modulo a constant offset) and thus have the same minima.

Let $X = U\Sigma V^\top$ be the SVD of X, where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_d) \in \mathbb{R}^{d \times d}$ and $V = [v_1 \cdots v_d] \in \mathbb{R}^{d \times d}$ with $v_i \in \mathbb{R}^d$. Since V is orthogonal, it is easy to see that if $\widetilde{X} = \Sigma V^\top$ and $\widetilde{Y} = \Sigma^{-1} V^\top X^\top Y$, we have the desired equalities. This can be accomplished with the distilled dataset

$$\{(\tilde{x}_i, \tilde{y}_i) \,:\, i = 1, \ldots, d\}, \qquad \tilde{x}_i = \sigma_i v_i, \quad \tilde{y}_i = \sigma_i^{-1} v_i^\top X^\top Y.$$

We remark that $\widetilde{Y}$ has a simpler expression, namely $\widetilde{Y} = \widetilde{X}\theta^0_{\text{orig}}$. That is, $\widetilde{Y}$ is just the predictions of the unregularized model on the synthetic dataset $\widetilde{X}$. This is because

$$\widetilde{Y} = \Sigma^{-1}V^\top X^\top Y = (X^\top X)^{-1/2}X^\top Y = \underbrace{(X^\top X)^{1/2}}_{\widetilde{X}} \cdot \underbrace{(X^\top X)^{-1}X^\top Y}_{\theta^0_{\text{orig}}}.$$

Since in this case we have $\widetilde{X}^\top\widetilde{X} = X^\top X$ and $\widetilde{X}^\top\widetilde{Y} = X^\top Y$, it follows that $\theta^R_{\text{orig}} = \theta^R_{\text{dist}}$ for any data-independent regularizer $R(\theta)$.

Note that this analysis also implies that the same distilled dataset also works for combination distillation. In particular, let X', Y' be the data matrix and response vector for any other dataset $\mathcal{D}$. The combined data matrix for the original datasets is $\mathbf{X} = [X^\top\ X'^\top]^\top$, while the combined response vector is $\mathbf{Y} = [Y^\top\ Y'^\top]^\top$; for the dataset resulting from combination with the distilled dataset, these quantities are $\widetilde{\mathbf{X}} = [\widetilde{X}^\top\ X'^\top]^\top$ and $\widetilde{\mathbf{Y}} = [\widetilde{Y}^\top\ Y'^\top]^\top$, respectively. Then we have

$$\mathbf{X}^\top\mathbf{X} = X^\top X + X'^\top X' = \widetilde{X}^\top\widetilde{X} + X'^\top X' = \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}},$$
$$\mathbf{X}^\top\mathbf{Y} = X^\top Y + X'^\top Y' = \widetilde{X}^\top\widetilde{Y} + X'^\top Y' = \widetilde{\mathbf{X}}^\top\widetilde{\mathbf{Y}}.$$

It directly follows that $\theta(\mathcal{D}_{\text{orig}} \cup \mathcal{D}) = \theta(\mathcal{D}_{\text{dist}} \cup \mathcal{D})$. □

Proof of Theorem 4. We begin with hyperparameter distillation. Suppose that $\theta^\lambda_{\text{orig}} = \theta^\lambda_{\text{dist}} = \theta^\lambda$. Since $\theta^\lambda_{\text{orig}}, \theta^\lambda_{\text{dist}}$ are stationary points of their respective losses, we have

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{1}{1+e^{-\theta^\lambda \cdot x_i}} - y_i\right) x_i + \lambda\theta^\lambda = \frac{1}{k}\sum_{i=1}^k \left(\frac{1}{1+e^{-\theta^\lambda \cdot \tilde{x}_i}} - \tilde{y}_i\right) \tilde{x}_i + \lambda\theta^\lambda = 0.$$

In particular, we see that

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{1}{1+e^{-\theta^\lambda \cdot x_i}} - y_i\right) x_i = \frac{1}{k}\sum_{i=1}^k \left(\frac{1}{1+e^{-\theta^\lambda \cdot \tilde{x}_i}} - \tilde{y}_i\right) \tilde{x}_i \tag{1}$$

for all θ^λ. Since $\{\theta^\lambda : \lambda \geq 0\}$ has an accumulation point at 0 and both functions in equation (1) are analytic functions of θ, by the identity theorem, the equality (1) must hold for *all* values of θ (not just those in the regularization path θ^λ). The desired result then follows since any collection $\{\frac{1}{1+e^{-\theta\cdot\bar{x}_i}}\}_{i=1}^N$ are linearly independent as functions of θ, provided that the $\bar{x}_i$ are unique.

The proof for combination distillation is similar. Let $\{(x_i, y_i)\}_{i=1}^n$ be the original dataset and $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^k$ be the distilled dataset, and fix a regularization strenght $\lambda > 0$. We will use $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^m$ to denote an arbitrary second dataset that we wish to combine with the first.

By the definition of combination distillation, the model obtained by training on the union of the original dataset with the second dataset must be the same as the model obtained by training on the union of the distilled dataset and the second dataset. In particular, this implies that

$$\frac{1}{n+m}\left[\sum_{i=1}^n \left(\frac{1}{1+e^{-x_i^\top\theta}} - y_i\right) x_i + \sum_{i=1}^m \left(\frac{1}{1+e^{-\bar{x}_i^\top\theta}} - \bar{y}_i\right)\bar{x}_i\right] + \lambda\theta = 0$$
$$\implies \quad \frac{1}{k+m}\left[\sum_{i=1}^k \left(\frac{1}{1+e^{-\tilde{x}_i^\top\theta}} - \tilde{y}_i\right) \tilde{x}_i + \sum_{i=1}^m \left(\frac{1}{1+e^{-\bar{x}_i^\top\theta}} - \bar{y}_i\right)\bar{x}_i\right] + \lambda\theta = 0. \tag{2}$$

For each $m \geq 0$, define $\theta^{(m)}$ to be the (shared) solution to (2) when the second dataset is $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^m$ has $\bar{x}_i = 0$, $\bar{y}_i = 0$ for all i. In particular, we see that

$$\sum_{i=1}^n \left(\frac{1}{1+e^{-x_i^\top\theta^{(m)}}} - y_i\right) x_i - \sum_{i=1}^k \left(\frac{1}{1+e^{-\tilde{x}_i^\top\theta^{(m)}}} - \tilde{y}_i\right) \tilde{x}_i + (n-k)\lambda\theta^{(m)} = 0. \tag{3}$$

Since $\{\theta^{(m)}\}_{m=0}^\infty$ has an accumulation point at 0 and the LHS of (3) is analytic as a function of θ, it follows that (3) holds for all θ (not just $\theta^{(m)}$) by the indentity theorem. Note that the norm of the two summation terms are bounded by a constant, while $\|(n-k)\lambda\theta\|$ is unbounded provided that $n - k \neq 0$. In particular, this means that we must have $k = n$, as desired. □