On Divergence Measures for Bayesian Pseudocoresets

Anonymous Author(s)
Affiliation
Address
email

Abstract

A Bayesian pseudocoreset is a small synthetic dataset for which the posterior over parameters approximates that of the original dataset. While promising, the scalability of Bayesian pseudocoresets is not yet validated in realistic problems such as image classification with deep neural networks. On the other hand, dataset distillation methods similarly construct a small dataset such that the optimization using the synthetic dataset converges to a solution with performance competitive with optimization using full data. Although dataset distillation has been empirically verified in large-scale settings, the framework is restricted to point estimates, and their adaptation to Bayesian inference has not been explored. This paper casts two representative dataset distillation algorithms as approximations to methods for constructing pseudocoresets by minimizing specific divergence measures: reverse KL divergence and Wasserstein distance. Furthermore, we provide a unifying view of such divergence measures in Bayesian pseudocoreset construction. Finally, we propose a novel Bayesian pseudocoreset algorithm based on minimizing forward KL divergence. Our empirical results demonstrate that the pseudocoresets constructed from these methods reflect the true posterior even in high-dimensional Bayesian inference problems.

1 Introduction

One of the main ingredients for modern statistical machine learning is large-scale data, which enable the learning of powerful and flexible models such as deep neural networks when handled correctly [2, 24]. However, in general, the larger the data are, the harder to learn with them [25]. Aside from requiring enormous computing resources for training, large-scale data often contain sensitive information that must not be shared publicly [10]. This motivates the construction of *coresets* [22, 13], a sparse subset of a large-scale dataset summarizing essential features of the original dataset. In the context of Bayesian inference, the posterior conditioned on an essential coreset should be similar to the exact posterior conditioned on the full dataset.

Coreset construction becomes more difficult as data dimensionality increases. Manousakas et al. [16] show that even the best coreset becomes suboptimal at large scales: the divergence between the posterior conditioned on the optimal coreset and the exact posterior grows with the dimension of the data. To this end, Manousakas et al. [16] propose to construct a *Bayesian pseudocoreset*, a synthetic dataset represented as a set of learnable parameters and trained to minimize the divergence between coreset posteriors and full-data posteriors. Compared to coresets, pseudocoresets scale much better with data dimension and achieve significantly lower posterior approximation error. However, the KL divergence minimization in pseudocoreset learning requires the construction of approximate posteriors on the fly during the learning process, which limits its tractability for high-dimensional models such as deep neural networks.

On the other hand, *dataset distillation* [21] aims to distill a large-scale dataset into a small synthetic dataset such that the test performance of a model trained on the distilled data is comparable to that of a

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

model trained on the original data. Even though dataset distillation methods have a similar motivation to pseudocoresets, the two methods optimize different objectives. Whereas pseudocoresets minimize the divergence between coreset and full-data posteriors, dataset distillation considers a non-Bayesian setting with heuristically set objective functions, such as matching the gradients of loss functions [28] or matching the parameters obtained from optimization trajectories [7]. Dataset distillation methods have achieved remarkable performance, showing premises on real-world problems involving high-dimensional data and models such as image classification with deep neural networks. However, since the datasets are distilled with heuristical objectives, we have no ground for using them in Bayesian inference.

In this paper, we provide a unifying view of Bayesian pseudocoresets and dataset distillation to take advantages of both approaches. We first study the choice of divergence measures being optimized in pseudocoreset learning. While the existing work [16] minimizes the reverse KL divergence between the pseudocoreset posterior distribution and the full data posterior distribution, we show that alternative divergence measures such as forward KL divergence and Wasserstein distance can also be employed effectively. Furthermore, we reveal that the approximate divergence minimization based on the specific choices of variational approximations actually recovers existing dataset distillation algorithms. This connection justifies the heuristically set objective functions being used for such dataset distillation algorithms and provides a theoretical background for using the distilled datasets for Bayesian inference. Conversely, Bayesian pseudocoreset algorithms can benefit from tricks used in the dataset distillation literature, such as reusing pre-computed optimization trajectories [7], which have already been shown to work at large scales. We empirically compare the pseudocoreset algorithms based on various divergence measures on high-dimensional image classification with deep neural networks. Our results demonstrate that we can efficiently and scalably construct Bayesian pseudocoresets with which MCMC algorithms such as Hamiltonian Monte-Carlo (HMC) [9, 18] or Stochastic Gradient Hamiltonian Monte-Carlo (SGHMC) [8] accurately approximate the full posterior.

2 Backgrounds

2.1 Bayesian pseudocoresets

Let $\mathbf{x} = \{x_1, \dots, x_n\}_{i=1}^n$ be observed data. Given a probabilistic model index by a parameter θ with some prior distribution π_0, we are interested in the posterior conditioned on the full data $\mathbf{x}$,

$$\pi_{\mathbf{x}}(\theta) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{n=1}^{N} f(x_n, \theta)\right), \ \pi_0(\theta) \coloneqq \frac{1}{Z(\mathbf{x})} \exp\left(\mathbb{1}_N^\top \mathbf{f}(\mathbf{x}, \theta)\right) \pi_0(\theta), \tag{1}$$

where $\mathbf{f}(\mathbf{x}, \theta) \coloneqq [f(x_1, \theta), \dots, f(x_N, \theta)]^\top$, $\mathbb{1}_N$ is the N-dimensional one vector, and $Z(\mathbf{x}) = \int \exp(\mathbb{1}_n^\top \mathbf{f}(\mathbf{x}, \theta)) \mathrm{d}\theta$ is a partition function. The posterior $\pi_{\mathbf{x}}$ is usually intractable due to $Z(\mathbf{x})$, so we often rely on approximate inference algorithms. Bayesian pseudocoreset methods [16] aim to construct a synthetic dataset called a *pseudocoreset* $\mathbf{u} = \{u_m\}_{m=1}^M$ with $M \ll N$ such that the posterior of θ conditioned on it approximates the original posterior $\pi_{\mathbf{x}}$. The pseudocoreset posterior is written as[1],

$$\pi_{\mathbf{u}}(\theta) = \frac{1}{Z(\mathbf{u})} \exp\left(\sum_{m=1}^{M} f(u_m, \theta)\right), \ \pi_0(\theta) = \frac{1}{Z(\mathbf{u})} \exp(\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta)) \pi_0(\theta), \tag{2}$$

where $\mathbf{f}(\mathbf{u}, \theta)$, $\mathbb{1}_M$, and $Z(\mathbf{u})$ are defined similarly. Manousakas et al. [16] suggests to learn $\mathbf{u}$ by minimizing the KL divergence between $\pi_{\mathbf{u}}$ and $\pi_{\mathbf{x}}$:

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \ D_{\mathrm{KL}}[\pi_{\mathbf{u}} \| \pi_{\mathbf{x}}], \tag{3}$$

and show that the gradient of this objective is computed as

$$\nabla_{u_m} D_{\mathrm{KL}}[\pi_{\mathbf{u}} \| \pi_{\mathbf{x}}] = -\mathrm{Cov}_{\pi_{\mathbf{u}}(\theta)} \left[\nabla_u \mathbf{f}(\mathbf{u}_m, \theta), \mathbb{1}_N^\top \mathbf{f}(\mathbf{x}, \theta) - \mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta)\right]. \tag{4}$$

The expectation over the pseudocoreset posterior $\pi_{\mathbf{u}}$ in the above equation is further approximated with the Monte-Carlo estimation with an approximate posterior $q_{\mathbf{u}} \approx \pi_{\mathbf{u}}$, for instance, from the Laplace approximation.

[1] Manousakas et al. [16] considered two sets of parameters, the coreset elements $\mathbf{u}$ and their weights $\mathbf{w} = \{w_m\}_{m=1}^M$. We empirically found that the weights $\mathbf{w}$ have negligible impact on performance, so we only learn the coreset elements.

2.2 Dataset Distillation

With a similar motivation as Bayesian pseudocoresets, dataset distillation methods construct a small synthetic dataset $\mathbf{u}$, but from the perspective of matching optimal parameters. That is, in dataset distillation, we find $\mathbf{u}$ such that

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \; d(\theta^*_{\mathbf{x}}, \theta^*_{\mathbf{u}}), \quad \theta^*_{\mathbf{x}} = \arg\min_{\theta} \; \ell(\mathbf{x}, \theta), \quad \theta^*_{\mathbf{u}} = \arg\min_{\theta} \; \ell(\mathbf{u}, \theta), \tag{5}$$

with a loss function $\ell(\cdot, \theta)$ and a distance function $d(\cdot, \cdot)$ comparing two parameters. In general, optimizing $d(\theta^*_{\mathbf{x}}, \theta^*_{\mathbf{u}})$ is intractable since it is a bilevel optimization problem requiring $\theta^\star_{\mathbf{x}}$ and $\theta^*_{\mathbf{u}}$, so existing works relax the problem in several ways, such as kernel-based approaches [19, 20] or solving only for a short horizon in inner optimization [28, 26, 7]. Below, we briefely review two representative methods — Dataset condensation (DC) [28] and Matching Traning Trajectories (MTT) [7].

Dataset Condensation Instead of fully unrolling the inner optimization process, DC instead computes the matching objective at *each* inner step, starting from some random parameter $\theta^{(0)}$.

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \; \mathbb{E}_{\theta^{(0)}} \left[\sum_{t=0}^{T} d_{\cos} \left(\nabla_\theta \ell(\mathbf{x}, \theta^{(t)}), \nabla_\theta \ell(\mathbf{u}, \theta^{(t)}) \right) \right], \tag{6}$$

$$\theta^{(t)} = \theta^{(t-1)} - \eta \nabla_\theta \ell(\mathbf{u}, \theta^{(t-1)}) \text{ for } t \geq 1, \tag{7}$$

where $d_{\cos}$ denotes the cosine similarity loss between two gradients. To implement this objective, we build a trajectory of parameters starting from a random $\theta^{(0)}$ with the current distilled dataset $\mathbf{u}$. We then compute the gradient matching objective for each layer at *each step* of the trajectory to update $\mathbf{u}$. Despite its appealing simplicity, this algorithm may suffer from short-horizon bias due to the limited trajectory length.

Matching Training Trajectories MTT extends the single step gradient matching in DC and considers longer inner optimization trajectories. Starting from a randomly initialized parameter $\theta^{(0)}$, MTT takes multiple inner optimization steps with both $\mathbf{u}$ and $\mathbf{x}$, and minimize the normalized L_2 distance between two parameters at the end of the learning trajectories.

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \; \mathbb{E}_{\theta^{(0)}} \left[d_2 \left(\theta^{(L_{\mathbf{x}})}_{\mathbf{x}}, \theta^{(L_{\mathbf{u}})}_{\mathbf{u}} \right) \right], \quad \theta^{(0)}_{\mathbf{x}} = \theta^{(0)}_{\mathbf{u}} = \theta^{(0)}, \tag{8}$$

$$\theta^{(L_{\mathbf{x}})}_{\mathbf{x}} = \theta^{(0)} - \sum_{t=1}^{L_{\mathbf{x}}} \eta_{\mathbf{x}} \nabla_\theta \ell(\mathbf{x}, \theta^{(t-1)}_{\mathbf{x}}), \quad \theta^{(L_{\mathbf{u}})}_{\mathbf{x}} = \theta^{(0)} - \sum_{t=1}^{L_{\mathbf{u}}} \eta_{\mathbf{u}} \nabla_\theta \ell(\mathbf{u}, \theta^{(t-1)}_{\mathbf{u}}), \tag{9}$$

where $d_2(\cdot, \cdot)$ is the normalized L_2 distance between parameters. Unlike DC, this algorithm starts from a common initial parameter $\theta^{(0)}$, keeps track of two separate optimization trajectories (one with $\mathbf{u}$ and another with $\mathbf{x}$), and the outer-update for $\mathbf{u}$ is done only after $L_{\mathbf{x}}$ and $L_{\mathbf{u}}$ inner steps. To reduce the heavy computational cost for the inner update with the full dataset $\mathbf{x}$, MTT employs mini-batch SGD for the inner steps, in addition to saving and re-using precomputed SGD trajectories to train $\mathbf{u}$. The use of precomputed SGD trajectories is similar to that of replay buffers for reinforcement learning [15, 17, 1], and allows MTT to handle long inner-loop sequences. While MTT significantly improves over DC, presumably due to its longer inner optimization horizon, the algorithm requires large memory for even a moderate number of inner-steps $L_{\mathbf{u}}$ because it requires backpropagation through all inner steps with $\mathbf{u}$.

3 On Divergence Measures for Bayesian Pseudocoresets

While the Bayesian pseudocoreset is originally obtained by minimizing the reverse KL divergence, in principle we can minimize an arbitrary divergence measure $D(\cdot, \cdot)$ to achieve the goal:

$$\mathbf{u}^* = \arg\min_{\mathbf{u}} \; D(\pi_{\mathbf{x}}, \pi_{\mathbf{u}}). \tag{10}$$

In this section, we study the several choices for such divergence measures.

3.1 Reverse KL divergence

Manousakas et al. [16] proposed the reverse KL divergence minimization with a Laplace approximation for the pseudocoreset posterior $\pi_{\mathbf{u}}$. Here, we show that it is possible to recast DC as a reverse KL divergence minimization with an alternative choice for the variational posterior. Let $\theta^{(t)}$ be a parameter in a learning trajectory with a loss function $\ell(\mathbf{u}, \theta) := -\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta)$. Provided that $\theta^{(t)}$ is sufficiently close to a local optimum, we can construct a naïve approximation of $\pi_{\mathbf{u}}$ as,

$$q_{\mathbf{u}}(\theta) = \mathcal{N}(\theta; \theta^{(t)}, \Sigma), \quad \theta^{(t)} = \theta^{(t-1)} - \eta \nabla_\theta \ell(\mathbf{u}, \theta^{(t-1)}). \tag{11}$$

Under this setting, we can actually show that a gradient descent for the reverse KL divergence $D_{\mathrm{KL}}[\pi_{\mathbf{u}} \| \pi_{\mathbf{x}}]$ is equivalent to a single step of gradient matching in DC.

Proposition 3.1. *Let $q_{\mathbf{u}}(\theta)$ set as Eq. 11. Under suitable conditions on η and Σ, we have*

$$\nabla_{\mathbf{u}} D_{\mathrm{KL}}[\pi_{\mathbf{u}} \| \pi_{\mathbf{x}}] \approx -\eta \nabla_{\mathbf{u}} \Big(\nabla_\theta \ell(\mathbf{x}, \theta^{(t-1)})^\top \nabla_\theta \ell(\mathbf{u}, \theta^{(t-1)}) \Big). \tag{12}$$

The proof is given in Appendix A. In other words, DC can be interpreted as a special instance of Bayesian pseudocoreset via reverse KL minimization with Gaussian approximation for the pseudocoreset posterior $\pi_{\mathbf{u}}$. Throughout the paper, we will refer to the Bayesian pseudocoreset with reverse KL minimization as BPC-rKL. To simply implement the Algorithm 1 in [16] for the image dataset, we approximated π_u as a Gaussian distribution. Please refer to the appendix for the detailed algorithm.

3.2 Wasserstein distance

We can instead choose $D(\cdot, \cdot)$ to be the Wasserstein distance between $\pi_{\mathbf{u}}$ and $\pi_{\mathbf{x}}$. The Wasserstein distance is generally intractable, but as we did earlier, if we approximate both $\pi_{\mathbf{u}}$ and $\pi_{\mathbf{x}}$ with Gaussians, it can be a closed-form expression. Specifically, let $\theta_{\mathbf{u}}$ and $\theta_{\mathbf{x}}$ be two parameters sufficiently close to local optima. Then we can construct a crude approximation for $\pi_{\mathbf{u}}$ and $\pi_{\mathbf{x}}$ as,

$$\pi_{\mathbf{u}}(\theta) \approx q_{\mathbf{u}}(\theta) := \mathcal{N}(\theta; \theta_{\mathbf{u}}, \Sigma_{\mathbf{u}}), \quad \pi_{\mathbf{x}}(\theta) \approx q_{\mathbf{x}}(\theta) := \mathcal{N}(\theta; \theta_{\mathbf{x}}, \Sigma_{\mathbf{x}}). \tag{13}$$

Then the Wasserstein distance between these two distributions are computed as

$$W_2(q_{\mathbf{u}}; q_{\mathbf{x}}) = \|\theta_{\mathbf{u}} - \theta_{\mathbf{x}}\|_2^2 + \mathrm{Tr}\Big(\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{u}} - 2(\Sigma_{\mathbf{u}}^{1/2} \Sigma_{\mathbf{x}} \Sigma_{\mathbf{u}}^{1/2})^{1/2} \Big). \tag{14}$$

Now consider a single outer-step of the MTT. Starting from a common initial θ_0, the algorithm takes the gradient steps both with $\mathbf{x}$ and $\mathbf{u}$ to get $\theta_{\mathbf{x}}^{(L_{\mathbf{x}})}$ and $\theta_{\mathbf{u}}^{(L_{\mathbf{u}})}$. If we set $q_{\mathbf{u}}(\theta) = \mathcal{N}(\theta; \theta_{\mathbf{u}}^{(L_{\mathbf{u}})}, \Sigma)$ and $q_{\mathbf{x}}(\theta) = \mathcal{N}(\theta; \theta_{\mathbf{x}}^{(L_{\mathbf{x}})}, \Sigma)$ with a common covariance Σ, then the Wasserstein distance reduces to $\|\theta_{\mathbf{x}}^{(L_{\mathbf{x}})} - \theta_{\mathbf{u}}^{(L_{\mathbf{u}})}\|_2^2$, which is proportional to the L_2 objective being optimized in the MTT. Similar to the reverse KL, we refer to the Bayesian pseudocoreset with Wasserstein distance minimization as BPC-W.

3.3 Forward KL divergence

Finally, we consider an alternative Bayesian pseudocoreset algorithm based on the forward KL divergence. The forward KL minimization is known to encourage a model to cover the entire target distribution while the reverse KL minimization favors mode capturing models, indicating that the forward KL may be a better choice for the Bayesian pseudocoreset. Under our setting, the forward KL is computed as

$$D_{\mathrm{KL}}[\pi_{\mathbf{x}} \| \pi_{\mathbf{u}}] = \log Z(\mathbf{u}) - \log Z(\mathbf{x}) + \mathbb{E}_{\pi_{\mathbf{x}}}[\mathbb{1}_N^\top \mathbf{f}(\mathbf{x}, \theta)] - \mathbb{E}_{\pi_{\mathbf{u}}}[\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta)]. \tag{15}$$

Then the gradient w.r.t. the pseudocoreset $\mathbf{u}$ is computed as,

$$\begin{aligned} \nabla_{\mathbf{u}} D_{\mathrm{KL}}[\pi_{\mathbf{x}} \| \pi_{\mathbf{u}}] &= \nabla_{\mathbf{u}} \log Z(\mathbf{u}) - \mathbb{E}_{\pi_{\mathbf{x}}}[\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta)] \\ &= \mathbb{E}_{\pi_{\mathbf{u}}}[\nabla_{\mathbf{u}}(\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta))] - \nabla_{\mathbf{u}} \mathbb{E}_{\pi_{\mathbf{x}}}[\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta)]. \end{aligned} \tag{16}$$

Again, we should approximate the expectation w.r.t. $\pi_{\mathbf{u}}$ and $\pi_{\mathbf{x}}$ to obtain this gradient. We introduce similar variational posteriors,

$$q_{\mathbf{u}}(\theta) = \mathcal{N}(\theta; \theta_{\mathbf{u}}^{(L_{\mathbf{u}})}, \Sigma_{\mathbf{u}}), \quad q_{\mathbf{x}}(\theta) = \mathcal{N}(\theta; \theta_{\mathbf{x}}^{(L_{\mathbf{x}})}, \Sigma_{\mathbf{x}}). \tag{17}$$

Algorithm 1 Bayesian Pseudocoresets with Forward KL

Require: Set of expert parameter trajectories $\{\tau\}$ trained with $\mathbf{x}$, each parameter trajectory saves parameters at the end of training epochs.
Require: Number of updates with the pseudocoreset (full data) $T_{\mathbf{u}}(T_{\mathbf{x}})$, total training steps K, maximum start epoch T^+, number of Gaussian samples S, variance σ^2, learning rate η.
Require: Differentiable augmentation function $\mathcal{A}$ (Optional).
Initialize the pseudocoreset $\mathbf{u}$ by randomly selecting a subset of size M from $\mathbf{x}$.
for $k = 1, \ldots, K$ **do**
  Sample an expert trajectory $\tau = \{\theta_*^{(r)}\}_{r=0}^{T}$.
  Randomly choose an epoch to start $r \leq T^+$ and initialize $\theta_{\mathbf{u}}^{(0)} = \theta_{\mathbf{x}}^{(0)} = \theta_*^{(r)}$.
  for $t = 1, \ldots, L_{\mathbf{u}}$ **do**
    Update the network parameter $\theta_{\mathbf{u}}^{(t)} \leftarrow \theta_{\mathbf{u}}^{(t-1)} + \eta \nabla \mathbb{1}_M^\top \mathbf{f}(\mathcal{A}(\mathbf{u}), \theta_{\mathbf{u}}^{(t-1)})$.
  end for
  Sample random Gaussian noises $\{\varepsilon_{\mathbf{x}}^{(s)}, \varepsilon_{\mathbf{u}}^{(s)}\}_{s=1}^{S} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$.
  Update the coreset with the gradient with Eq. 19.
end for

The endpoint $\theta_{\mathbf{u}}^{(L_{\mathbf{u}})}$ is a function of $\mathbf{u}$, so in principle, the expectation w.r.t. $q_{\mathbf{u}}$ involves unrolling through the parameter trajectories $\theta^{(0)} \to \theta_{\mathbf{u}}^{(1)} \to \cdots \to \theta_{\mathbf{u}}^{(L_{\mathbf{u}})}$, as for the BPC-W. To reduce memory cost due to this unrolling, we can simply block the gradient flow through $\theta^{(L_{\mathbf{u}})}$ to obtain

$$\begin{aligned}
&\nabla_{\mathbf{u}} D_{\mathrm{KL}}[\pi_{\mathbf{x}} \| \pi_{\mathbf{u}}] \approx \mathbb{E}_{q_{\mathbf{u}}}[\nabla_{\mathbf{u}}(\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta))] - \nabla_{\mathbf{u}} \mathbb{E}_{q_{\mathbf{x}}}[\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta)] \\
&= \mathbb{E}_{\varepsilon_{\mathbf{u}}}[\nabla_{\mathbf{u}}(\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \mathtt{sg}(\theta_{\mathbf{u}}^{(L_{\mathbf{u}})}) + \Sigma_{\mathbf{u}}^{1/2} \varepsilon_{\mathbf{u}}))] - \nabla_{\mathbf{u}} \mathbb{E}_{\varepsilon_{\mathbf{x}}}[\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta_{\mathbf{x}}^{(L_{\mathbf{x}})} + \Sigma_{\mathbf{x}}^{1/2} \varepsilon_{\mathbf{x}})] \\
&= \nabla_{\mathbf{u}} \Big(\mathbb{E}_{\varepsilon_{\mathbf{u}}}[\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \mathtt{sg}(\theta_{\mathbf{u}}^{(L_{\mathbf{u}})}) + \Sigma_{\mathbf{u}}^{1/2} \varepsilon_{\mathbf{u}})] - \mathbb{E}_{\varepsilon_{\mathbf{x}}}[\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta_{\mathbf{x}}^{(L_{\mathbf{x}})} + \Sigma_{\mathbf{x}}^{1/2} \varepsilon_{\mathbf{x}})] \Big),
\end{aligned} \tag{18}$$

where $\mathtt{sg}(\cdot)$ denotes the stop-grad operation. We further approximate this via Monte-Carlo estimation,

$$\approx \nabla_{\mathbf{u}} \left(\frac{1}{S} \sum_{s=1}^{S} \left(\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \mathtt{sg}(\theta_{\mathbf{u}}^{(L_{\mathbf{u}})}) + \Sigma_{\mathbf{u}}^{1/2} \varepsilon_{\mathbf{u}}^{(s)}) - \mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta_{\mathbf{x}}^{(L_{\mathbf{x}})} + \Sigma_{\mathbf{x}}^{1/2} \varepsilon_{\mathbf{x}}^{(s)}) \right) \right), \tag{19}$$

with $\varepsilon_{\mathbf{x}}^{(1)}, \ldots, \varepsilon_{\mathbf{x}}^{(S)}, \varepsilon_{\mathbf{u}}^{(1)}, \ldots, \varepsilon_{\mathbf{u}}^{(S)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$. The resulting algorithm starts from a common initial parameter $\theta^{(0)}$, keeps track of two parameter trajectories, and minimizes the difference in the log-likelihood $\mathbb{1}_M^\top \mathbf{f}(\mathbf{u}, \theta)$ evaluated at the Gaussian neighbors of the endpoints of the two trajectories. Note that even if we have blocked the gradient flow through $\theta_{\mathbf{u}}^{(L_{\mathbf{u}})}$, we still have the pseudocoreset $\mathbf{u}$ in the likelihood terms $\mathbf{f}(\mathbf{u}, \theta)$, so the gradient descent gets meaningful learning signals. On the other hand, if we apply a similar trick to the BPC-W, the matching objective would be $\|\mathtt{sg}(\theta_{\mathbf{u}}^{(L_{\mathbf{u}})}) - \theta_{\mathbf{x}}^{(L_{\mathbf{x}})}\|$, which is just a constant w.r.t. $\mathbf{u}$. Here, we can also borrow the idea from MTT, where we utilize the set of expert trajectories to efficiently evaluate the parameter trajectory with $\mathbf{x}$. The resulting algorithm which we call as BPC-fKL is summarized in Algorithm 1.

Relation to contrastive divergence BPC-fKL is closely related to contrastive divergence [12, 6], a learning algorithm to approximate maximum likelihood estimator. Suppose we want to train a model $p_\theta(x) = \frac{1}{Z(\theta)} \exp(f_\theta(x))$ by maximizing the log-likelihood function of training examples $\{x_i\}_{i=1}^n \sim p_{\text{data}}(x)$,

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i). \tag{20}$$

Then the gradient of the log-likelihood is

$$L'(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta f_\theta(x_i) - \mathbb{E}_{p_\theta(x)}[\nabla_\theta f_\theta(x)] \tag{21}$$

$$= \mathbb{E}_{p_{\text{data}}(x)}[\nabla_\theta f_\theta(x)] - \mathbb{E}_{p_\theta(x)}[\nabla_\theta f_\theta(x)]. \tag{22}$$

The contrastive divergence approximates this gradient by approximating the second term with sampling methods such as Langevin dynamics or HMC. By doing this we can obtain the MLE for

log-likelihood without knowing the exact potential function. Our proposed BPC-fKL performs the same function as the contrastive divergence if the model parameter θ and training data $\mathbf{x}$ change to the Bayesian pseudocoresets $\mathbf{u}$ and the parameters from the true posterior distribution which are approximated by the points near the expert trajectories. In this point of view, BPC-fKL is the maximum likelihood estimator that maximizes the log-likelihood of the parameters from the posterior distribution of the original dataset.

4 Related works

Bayesian coresets As running algorithms such as MCMC and VI on large datasets is challenging due to expensive computational cost, Bayesian coresets [13], methods that represent the entire dataset as a sparse and weighted sum of subsets, have been studied. Campbell and Broderick [5] interpret the coreset construction problem as sparse vector sum approximation and further extends it to the sparse regression problem in Hilbert space by considering the directionality between vectors. Campbell and Broderick [4] demonstrates that Hilbert coresets scale the coresets log-likelihood sub-optimally, and presents a modified algorithm. As Hilbert coresets include the choice of weighted L^2 inner product or finite-dimensional projection and it is difficult to choose them optimally, another method that minimizes the KL divergence between the coreset posterior and true posterior has also been studied recently [3]. Bayesian coresets have the advantages of simple and theoretically guaranteed quality of the posterior approximations. However, as the dimensions of the datasets increase, the KL divergences of the coresets consisting only of their subsets increase together, and there is also a problem of privacy, so it is difficult to apply them to the real-world conditions. Therefore, there is also a Bayesian pseudocoreset [16] consisting of synthetic data, but it has been demonstrated experimentally only in relatively easy settings such as logistic regressions, and Bayesian coresets or pseudocoresets for real data have not yet been studied.

Dataset distillation The goal of dataset distillation [21] is to create a small synthetic dataset that allows the model to have similar test performance to the original large-scale dataset even if it is trained with that small synthetic dataset. These synthetic datasets can be applicable to various fields of machine learning, such as continual learning and neural architecture search. However, dataset distillation typically involves a bilevel optimization structure in the learning process, which suffers from high computational costs or training instability. So there are existing methods such as matching one-step gradients between synthetic and real data [28, 26], solving using a closed form solution of kernel ridge regression [19, 20], or reducing the normalized distance between parameters at the end of synthetic and real training trajectories [7]. Meanwhile, these methods focus only on high test accuracy, where uncertainty is not considered and do not include how much the posterior distribution on the synthetic dataset matches the true posterior distribution. Zhao and Bilen [27] match the distributions of embedding features for many random networks of synthetic and real data, not posterior distributions. On the other hand, in this paper, we extend dataset distillation to the problem of posterior distribution matching.

5 Experiments

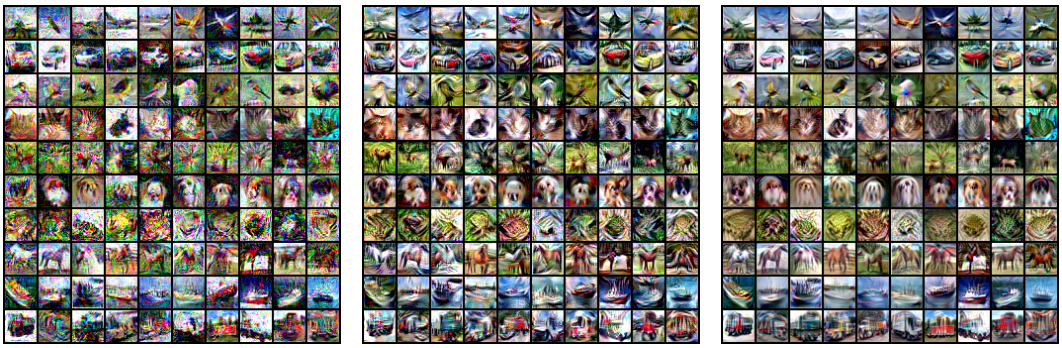
5.1 Experimental setup

Datasets and models We use the CIFAR10 dataset [14] to generate Bayesian pseudocoresets, and evaluate on the test split of CIFAR10 in addition to the CIFAR10-C dataset [11], which imposes additional uncertainty through image corruptions. Following the experimental setup of previous works [28, 26], we use a 3-layer ConvNet as the network architecture.

Evaluation We train each Bayesian coreset method using HMC [18] and SGHMC [8] and evaluate the top-1 accuracy and negative log likelihood (NLL) with respect to the ground-truth labels. Although evaluation with HMC is preferred, we use SGHMC to compare performances with and without data augmentations by considering the augmented pseudocoresets as stochastic batches. That is, even though we use the same full pseudocoresets at each step in SGHMC, the different augmentation we apply to each batch results in a different instances. We interpret this stochasticity as a batch-wise property, and we denote such evaluation Altered-SGHMC (A-SGHMC) for the remainder of this section. As our results were not sensitive to change in hyperparameters for HMC and A-SGHMC, we

Table 1: Performance of each Bayesian pseudocoreset method with $\{1, 10, 20\}$ images per class (ipc) on the CIFAR10 test dataset. We present results with HMC and A-SGHMC. Aug denotes image augmentation during training. All values are averaged over ten random seeds.

		HMC		HMC (+Aug)		A-SGHMC (+Aug)	
		Acc (↑)	NLL (↓)	Acc (↑)	NLL (↓)	Acc (↑)	NLL (↓)
1 ipc	Random	$0.1745_{\pm 0.0084}$	$2.4507_{\pm 0.0657}$	$0.1745_{\pm 0.0084}$	$2.4507_{\pm 0.0657}$	$0.1414_{\pm 0.0045}$	$2.9798_{\pm 0.0725}$
	BPC-rKL	$0.2472_{\pm 0.0120}$	$2.1635_{\pm 0.0327}$	$0.2416_{\pm 0.0133}$	$2.1688_{\pm 0.0486}$	$0.2162_{\pm 0.0083}$	$2.4617_{\pm 0.0841}$
	BPC-W	$\mathbf{0.3435}_{\pm 0.0207}$	$\mathbf{1.9311}_{\pm 0.0329}$	$0.3338_{\pm 0.0158}$	$1.9589_{\pm 0.0265}$	$\mathbf{0.2934}_{\pm 0.0121}$	$\mathbf{2.1400}_{\pm 0.0333}$
	BPC-fKL	$0.3354_{\pm 0.0066}$	$2.0253_{\pm 0.0311}$	$\mathbf{0.3468}_{\pm 0.0119}$	$\mathbf{1.9574}_{\pm 0.0269}$	$0.2823_{\pm 0.0128}$	$2.1426_{\pm 0.0343}$
10 ipc	Random	$0.2807_{\pm 0.0050}$	$2.1070_{\pm 0.0178}$	$0.2807_{\pm 0.0050}$	$2.1070_{\pm 0.0178}$	$0.3085_{\pm 0.0077}$	$2.1953_{\pm 0.0481}$
	BPC-rKL	$0.3253_{\pm 0.0075}$	$1.9627_{\pm 0.0312}$	$0.3017_{\pm 0.0074}$	$2.0251_{\pm 0.0134}$	$0.3789_{\pm 0.0154}$	$1.9492_{\pm 0.0605}$
	BPC-W	$0.3535_{\pm 0.0127}$	$1.9450_{\pm 0.0258}$	$0.3646_{\pm 0.0082}$	$1.9307_{\pm 0.0171}$	$\mathbf{0.4890}_{\pm 0.0172}$	$1.6971_{\pm 0.0392}$
	BPC-fKL	$\mathbf{0.4294}_{\pm 0.0101}$	$\mathbf{1.7292}_{\pm 0.0248}$	$\mathbf{0.4252}_{\pm 0.0091}$	$\mathbf{1.7334}_{\pm 0.0217}$	$0.4878_{\pm 0.0103}$	$\mathbf{1.6762}_{\pm 0.0390}$
20 ipc	Random	$0.3292_{\pm 0.0057}$	$2.0164_{\pm 0.0212}$	$0.3292_{\pm 0.0057}$	$2.0164_{\pm 0.0212}$	$0.3832_{\pm 0.0070}$	$1.8999_{\pm 0.0338}$
	BPC-rKL	$0.3686_{\pm 0.0115}$	$1.8848_{\pm 0.0338}$	$0.3518_{\pm 0.0102}$	$1.9540_{\pm 0.0350}$	$0.4307_{\pm 0.0073}$	$1.8555_{\pm 0.0263}$
	BPC-W	$0.4546_{\pm 0.0146}$	$1.7336_{\pm 0.0325}$	$0.4606_{\pm 0.0200}$	$1.7334_{\pm 0.0506}$	$\mathbf{0.5691}_{\pm 0.0167}$	$\mathbf{1.5439}_{\pm 0.0408}$
	BPC-fKL	$\mathbf{0.4910}_{\pm 0.0088}$	$\mathbf{1.6279}_{\pm 0.0264}$	$\mathbf{0.4833}_{\pm 0.0069}$	$\mathbf{1.6440}_{\pm 0.0231}$	$0.5153_{\pm 0.0093}$	$1.6701_{\pm 0.0256}$



(a) BPC-rKL **(b)** BPC-W **(c)** BPC-fKL

Figure 1: Examples of Bayesian pseudocoresets. Each row is the pseudocoresets for each class.

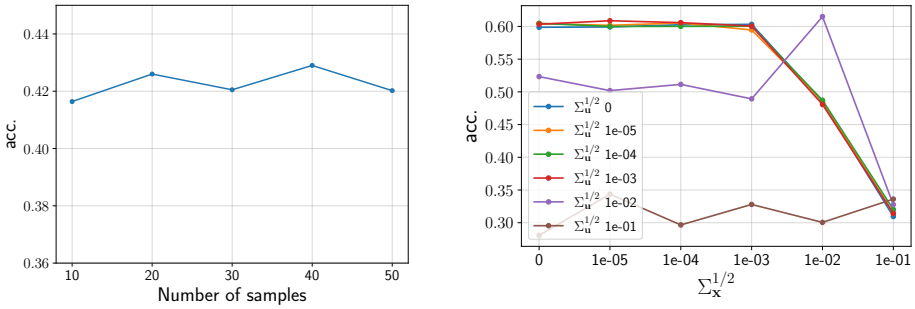
used a single set of hyperparameters which performed best in initial experiments. Please refer to the appendix for detailed evaluation settings including hyperparameters.

5.2 Main results

To validate each pseudocoreset method in various settings, we depict the HMC and A-SGHMC results on CIFAR10 dataset in Table 1. To evaluate the effectiveness of data augmentations in pseudocoresets training, we trained each pseudocoreset both with augmentations and without augmentations. Also, through the A-SGHMC, we wanted to see the effectiveness of augmentation in test time. Each row represents the results for different pseudocoreset sizes, which are 1, 10 and 20 images per class.

Table 1 shows that all Bayesian pseudocoresets have higher accuracy and lower negative log-likelihood than those of the random coresets. BPC-rKL is slightly worse than BPC-W and BPC-fKL, which we interpret as the reverse KL typically suffers from the property that covers only one major mode so that BPC-rKL posterior is somewhat sub-optimal. In 1 image per class setting, BPC-W and BPC-fKL have comparable results, and as the pseudocoreset size increases, BPC-fKL is better if there is no augmentation in test time and BPC-W is better if not. It seems that the strength of BPC-W with higher performances with test time augmentations is due to the training method that exactly matches the pseudocoreset training trajectories and expert training trajectories that also be trained with the data augmentations. However, the learning method of A-SGHMC is not an accurate Bayesian learning, and it is known that data augmentation causes cold posterior effect [23], the HMC results without augmentations are preferred.

Fig. 1 shows the examples of Bayesian pseudocoresets trained with each method. We can see that BPC-fKL looks the most smooth and BPC-rKL is the most noisy. In addition, BPC-rKL seems to have changed only in pixel space and the others seem to have more semantic changes. This may be also due to the reverse KL divergence that makes it difficult to escape the mode created by the



(a) Number of Gaussian samples S

(b) Variances of Gaussian $\Sigma_u^{1/2}$, $\Sigma_x^{1/2}$

Figure 2: Exploring the hyperparameters for Gaussian approximation in BPC-fKL. The pseudocoresets of size 10 images per class for CIFAR10.

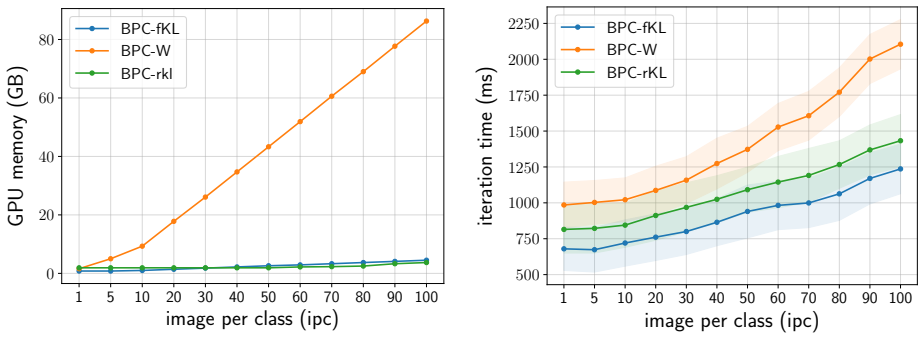
initialized images and hard to cover the distribution of the entire dataset. On the other hand, BPC-W and BPC-fKL seem to be learned as desired to include various shapes of each class.

5.3 Gaussian approximation in BPC-fKL

In this experiment, we investigate the effect of the hyperparameter of the Gaussian approximation to the performance of BPC-fKL. Firstly, we explore the number of Gaussian samples S and variances $\Sigma_u^{1/2}$, $\Sigma_x^{1/2}$ in Eq. 19. As shown in Fig. 2a, we plot the accuracy of HMC as the function of the number of samples. Even though the estimation becomes more accurate as the number of samples increases, in Fig. 2a, the number of samples do not significantly improve the performance of pseudocoresets. Thus, we use 30 samples for all the experiments. Fig. 2b, we show the accuracy with varying the variances. The value of x-axis shows $\Sigma_{\mathbf{x}}^{1/2}$'s and $\Sigma_{\mathbf{u}}^{1/2}$'s are presented as colors. As the graph shows, too small variances are not much different from using variance of 0, and when the both values are 0.01 is the best and performance drops again for the variances larger than that. So we used both $\Sigma_{\mathbf{x}}^{1/2}$ and $\Sigma_{\mathbf{u}}^{1/2}$ of 0.01.

5.4 Computational cost

In this section, we note that BPC-fKL significantly reduces the computational cost required for pseudocoreset training. We evaluate computational cost for each pseudocoreset method on CIFAR10. The computational cost is measured with fixed gradient steps with $\mathbf{u}$, $L_{\mathbf{u}}$, of 30, on 32 cores of Intel Xeon CPU Gold 5120 and 4 Tesla V100s. For BPC-W, when the number of pseudocoresets per class becomes large, it cannot be trained on single GPU due to the memory limits. Therefore, we performed parallel computations using 4 GPUs in all the methods for a fair comparison. To compare the training time, we measure a single iteration time averaged by 100 iterations, and repeat this



(a) GPU memory usage

(b) Iteration time

Figure 3: Comparison of computational costs of GPU memory and training time.

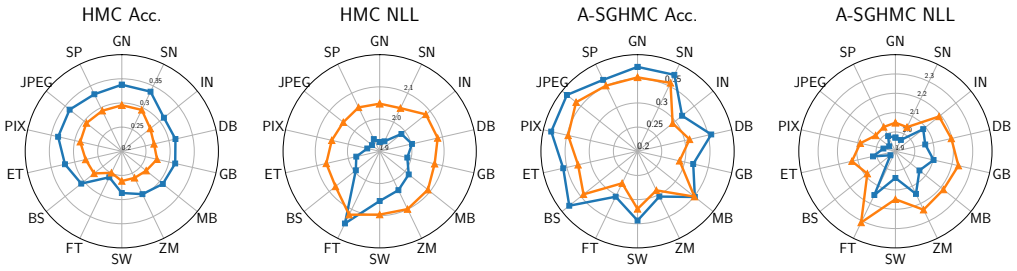


Figure 4: HMC and A-SGHMC results on CIFAR10-C. Each direction represents each type of corruptions. Blue and orange represent BPC-fKL and BPC-W, respectively. All values are averaged over 30 runs with different random seeds. ACC $\rightarrow$ higher the better, NLL $\rightarrow$ Lower the better.

process three times to select the largest value among them. Fig. 3a shows that the memory usage required for BPC-fKL compared to BPC-W is significantly reduced. This is because BPC-W must hold all gradient flows in memory for trajectory matching, although BPC-fKL does not need to. As shown in Fig. 3b, BPC-fKL also has the advantage of training time. The pseudocoreset training time increases linearly as the size of pseudocoresets increases, but the increasing rate of BPC-W is much faster than that of BPC-fKL. Combining the two results and Table 1, the pseudocoreset training using forward KL divergence is not only comparable in terms of accuracy and NLL but also remarkably efficient in both memory and training time, compared to BPC-W.

5.5 Out-of distribution dataset

Since one of the merits of Bayesian approach is Bayesian model averaging which enables robustness to distributional shift and better calibration, we evaluate the performance of each pseudocoreset method on out-of distribution dataset, CIFAR10-C [11]. Fig. 4 shows the accuracy and negative log-likelihood of each corrupted datasets for BPC-W and BPC-fKL. Note that we optimize each pseudocoreset with the clean CIFAR10 dataset and evaluate them on CIFAR10 with 14 corruptions. For A-SGHMC results, we apply differentiable augmentations to both original dataset and coreset for training, but we do not use it for running HMC. As shown in Fig. 4, for both HMC and A-SGHMC, BPC-fKL achieves higher or comparable accuracy and lower negative log-likelihood than the BPC-W for all the corruptions except for the frost corruption. It suggests that the forward KL divergence is indeed effective divergence measure for Bayesian pseudocoreset in this experimental setup.

6 Conclusion

In this paper, we explored three divergence measures for Bayesian pseudocoresets: reverse KLD, Wasserstein distance, and forward KLD. We showed that existing dataset distillation methods can be linked to the Bayesian pseudocoresets with reverse KLD and Wasserstein distance, and further proposed a novel algorithm for learning Bayesian pseudocoresets by minimizing forward KL divergence. We empirically validated all three methods in terms of their ability to approximate the posterior distribution for real-world image datasets. Bayesian pseudocoresets with both Wasserstein distance and forward KL divergence can approximate the true posterior well and BPC-fKL is more effective in terms of computational costs and robustness to the out-of distribution dataset.

Limitation Despite showing promising results for the first time on Bayesian pseudocoresets for real datasets, the performance still underperforms stochastic gradient MCMC with entire dataset. Thus, a promising future direction is to analyze whether such pseudocoresets are useful for stochastic gradient MCMC algorithms when they are used together with mini-batches of the original dataset.

Societal Impacts Our work is hardly likely to bring any negative societal impacts. However, any bias inherent in the original dataset might be transferred to coreset, which might amplify the bias on downstream task.

References

[1] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. Pieter Abbeel, and W. Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] T. Campbell and B. Beronov. Sparse variational inference: Bayesian coresets from scratch. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

[4] T. Campbell and T. Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *Proceedings of The 35th International Conference on Machine Learning (ICML 2018)*, 2018.

[5] T. Campbell and T. Broderick. Automated scalable bayesian inference via hilbert coresets. *The Journal of Machine Learning Research*, 20(1):551–588, 2019.

[6] M. A. Carreira-Perpinan and G. Hinton. On contrastive divergence learning. In *International workshop on artificial intelligence and statistics*, pages 33–40. PMLR, 2005.

[7] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[8] T. Chen, E. B. Fox, and C. Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, 2014.

[9] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216 – 222, 1987.

[10] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[11] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[12] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

[13] J. Huggins, T. Campbell, and T. Broderick. Coresets for bayesian logistic regression. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

[14] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[15] L.-J. Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3):293–321, 1992.

[16] D. Manousakas, Z. Xu, C. Mascolo, and T. Campbell. Bayesian pseudocoresets. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

[17] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[18] R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 54:113–162, 2010.

[19] T. Nguyen, Z. Chen, and J. Lee. Dataset meta-learning from kernel ridge-regression. In *International Conference on Learning Representations (ICLR)*, 2020.

[20] T. Nguyen, R. Novak, L. Xiao, and J. Lee. Dataset distillation with infinitely wide convolutional networks. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

[21] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[22] M. Welling. Herding dynamical weights to learn. In *Proceedings of the 26st International Conference on Machine Learning (ICML 2009)*, 2009.

[23] F. Wenzel, K. Roth, B. S. Veeling, J. Świątkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. How good is the bayes posterior in deep neural networks really? In *Proceedings of The 37th International Conference on Machine Learning (ICML 2020)*, 2020.

[24] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers, 2021. URL https://arxiv.org/abs/2106.04560.

[25] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[26] B. Zhao and H. Bilen. Dataset condensation with differentiable siamese augmentation. In *Proceedings of The 38th International Conference on Machine Learning (ICML 2021)*, 2021.

[27] B. Zhao and H. Bilen. Dataset condensation with distribution matching. *arXiv preprint arXiv:2110.04181*, 2021.

[28] B. Zhao, K. R. Mopuri, and H. Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Appendix A.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include our code in supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5. We reported standard deviations.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 5.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]