
Surrogate Likelihoods for Variational Annealed Importance Sampling

Martin Jankowiak¹ Du Phan²

Abstract

Variational inference is a powerful paradigm for approximate Bayesian inference with a number of appealing properties, including support for model learning and data subsampling. By contrast MCMC methods like Hamiltonian Monte Carlo do not share these properties but remain attractive since, contrary to parametric methods, MCMC is asymptotically unbiased. For these reasons researchers have sought to combine the strengths of both classes of algorithms, with recent approaches coming closer to realizing this vision in practice. However, supporting data subsampling in these hybrid methods can be a challenge, a shortcoming that we address by introducing a surrogate likelihood that can be learned jointly with other variational parameters. We argue theoretically that the resulting algorithm allows an intuitive trade-off between inference fidelity and computational cost. In an extensive empirical comparison we show that our method performs well in practice and that it is well-suited for black-box inference in probabilistic programming frameworks.

1. Introduction

Bayesian modeling and inference is a powerful approach to making sense of complex datasets. This is especially the case in scientific applications, where accounting for prior knowledge and uncertainty is essential. For motivating examples we need only look to recent efforts to study COVID-19, including e.g. epidemiological models that incorporate mobility data (Miller et al., 2020; Monod et al., 2021) or link differentiable transmissibility of SARS-CoV-2 lineages to viral mutations (Obermeyer et al., 2022).

Unfortunately for many application areas the broader use of Bayesian models is hampered by the difficulty of for-

mulating scalable inference algorithms. One regime that has proven particularly challenging is the large data regime. This is essentially because powerful MCMC methods like Hamiltonian Monte Carlo (HMC) (Duane et al., 1987; Neal et al., 2011) are, at least on the face of it, incompatible with data subsampling (i.e. mini-batching) (Betancourt, 2015). While considerable effort has gone into developing MCMC methods that accommodate data subsampling,¹ developing generic MCMC methods that yield high fidelity posterior approximations while scaling to very large datasets remains challenging.

For this reason, among others, recent years have seen extensive development of approximate inference methods based on variational inference (Jordan et al., 1999; Blei et al., 2017). Besides ‘automatic’ support for data subsampling (Hoffman et al., 2013; Ranganath et al., 2014), at least for suitable model classes, variational inference has several additional favorable properties, including support for amortization (Dayan et al., 1995), log evidence estimates, and model learning, motivating researchers to combine the strengths of MCMC and variational inference (Salimans et al., 2015; Hoffman, 2017; Caterini et al., 2018; Ruiz & Titsias, 2019).

Recent work from Geffner & Domke (2021) and Zhang et al. (2021) offers a particularly elegant formulation of a variational inference method that leverages (unadjusted) HMC as well as annealed importance sampling (AIS) (Neal, 2001). Although these are fundamentally variational methods, since they make use of HMC, which does not itself readily accommodate data subsampling, these methods likewise do not automatically inherit support for data subsampling. This is unfortunate because these methods *do* automatically inherit many of the other nice features of variational inference, including support for amortization and model learning.

In this work we set out to extend the approach in Geffner & Domke (2021) and Zhang et al. (2021) to support data subsampling, thus making it applicable to the large data regime. Our basic strategy is simple and revolves around introducing a *surrogate log likelihood* that is cheap to evaluate. The surrogate log likelihood is used to guide HMC dynamics, resulting in a flexible variational distribution that is implicitly defined via the gradient of the surrogate log likelihood. The

¹See e.g. Dang et al. (2019) and Zhang et al. (2020) for recent work.

¹Broad Institute, Cambridge, MA, USA ²Google Research, Cambridge, MA, USA. Correspondence to: Martin Jankowiak <mjankowi@broadinstitute.org>.

corresponding variational objective accommodates unbiased mini-batch estimates, at least for the large class of models with appropriate conditional independence structure that we consider. As we show in experiments in Sec. 7, our method performs well in practice and is well-suited for black-box inference in probabilistic programming frameworks.

2. Problem setting

We are given a dataset \mathcal{D} and a model of the form

$$p_\theta(\mathcal{D}, \mathbf{z}) = p_\theta(\mathbf{z}) \prod_{n=1}^N p_\theta(y_n | \mathbf{z}, \mathbf{x}_n) \quad (1)$$

where the latent variable $\mathbf{z} \in \mathbb{R}^D$ is governed by a prior $p_\theta(\mathbf{z})$ and the likelihood factorizes into N terms, one for each data point $(y_n, \mathbf{x}_n) \in \mathcal{D}$, and where θ denotes any additional (non-random) parameters in the model. Eqn. 1 encompasses a large and important class of models and includes e.g. a wide variety of multi-level regression models. The log density corresponding to Eqn. 1 is given by

$$\begin{aligned} \log p_\theta(\mathcal{D}, \mathbf{z}) &= \log p_\theta(\mathbf{z}) + \sum_n \log p_\theta(y_n | \mathbf{z}, \mathbf{x}_n) \\ &\equiv \Psi_0(\mathbf{z}) + \Psi_L(\mathcal{D}, \mathbf{z}) \end{aligned} \quad (2)$$

where we define the log prior $\Psi_0(\mathbf{z})$ and log likelihood $\Psi_L(\mathcal{D}, \mathbf{z})$.²

We are interested in the regime where N is large or individual likelihood terms are costly to evaluate so that computing the full log likelihood $\Psi_L(\mathcal{D}, \mathbf{z})$ and its gradients is impractical. We aim to devise a flexible variational approximation to the posterior $p_\theta(\mathbf{z} | \mathcal{D})$ that can be fit with an algorithm that supports data subsampling. Additionally we would like our method to be generic in nature so that it can readily be incorporated into a probabilistic programming framework as a black-box inference algorithm. Finally we would like a method that supports model learning, i.e. one that allows us to learn θ in conjunction with the approximate posterior.

For simplicity in this work we primarily focus on global latent variable models with the structure in Eqn. 1. The approach we describe can also be extended to models with *local* latent variables, i.e. those local to each data point. For more discussion see Sec. 7.6 and Sec. A.1 in the appendix.

3. Background

Before describing our method in Sec. 4, we first review relevant background.

3.1. Variational inference

The simplest variants of variational inference introduce a parametric variational distribution $q_\phi(\mathbf{z})$ and proceed to

²We also use the notation $\Psi_L(\mathcal{D}_{\mathcal{I}}, \mathbf{z}) \equiv \sum_{n \in \mathcal{I}} \log p_\theta(y_n | \mathbf{z}, \mathbf{x}_n)$ for a set of indices \mathcal{I} .

choose the parameters ϕ to minimize the Kullback-Leibler (KL) divergence between the variational distribution and the posterior $p_\theta(\mathbf{z} | \mathcal{D})$, i.e. $\text{KL}(q_\phi(\mathbf{z}) | p_\theta(\mathbf{z} | \mathcal{D}))$. This can be done by maximizing the Evidence Lower Bound or ELBO

$$\text{ELBO} \equiv \mathbb{E}_{q_\phi(\mathbf{z})} [\log p_\theta(\mathcal{D}, \mathbf{z}) - \log q_\phi(\mathbf{z})] \quad (3)$$

which satisfies $\text{ELBO} \leq \log p_\theta(\mathcal{D})$. Thanks to this inequality the ELBO naturally enables joint model learning and inference, i.e. we can maximize Eqn. 3 w.r.t. both model parameters θ and variational parameters ϕ simultaneously. A potential shortcoming of the fully parametric approach described here is that it can be difficult to specify appropriate parameterizations for $q_\phi(\mathbf{z})$.

3.2. Annealed Importance Sampling

Annealed importance sampling (AIS) (Neal, 2001) is a method for estimating the evidence $p_\theta(\mathcal{D})$ that leverages a sequence of K bridging densities $\{f_k(\mathbf{z})\}_{k=1}^K$ that connect a simple base distribution $q_0(\mathbf{z})$ to the posterior. In more detail, AIS can be understood as importance sampling on an extended space. That is we can write

$$p_\theta(\mathcal{D}) = \int d\mathbf{z} p_\theta(\mathcal{D}, \mathbf{z}) = \mathbb{E}_{q_{\text{fwd}}(\mathbf{z}_{0:K})} \left[\frac{q_{\text{bwd}}(\mathbf{z}_{0:K})}{q_{\text{fwd}}(\mathbf{z}_{0:K})} \right] \quad (4)$$

where the proposal distribution q_{fwd} and the (un-normalized) target distribution q_{bwd} are given by

$$\begin{aligned} q_{\text{fwd}}(\mathbf{z}_{0:K}) &= q_0(\mathbf{z}_0) \mathcal{T}_1(\mathbf{z}_1 | \mathbf{z}_0) \cdots \mathcal{T}_K(\mathbf{z}_K | \mathbf{z}_{K-1}) \\ q_{\text{bwd}}(\mathbf{z}_{0:K}) &= p_\theta(\mathcal{D}, \mathbf{z}_K) \tilde{\mathcal{T}}_K(\mathbf{z}_{K-1} | \mathbf{z}_K) \cdots \tilde{\mathcal{T}}_1(\mathbf{z}_0 | \mathbf{z}_1) \end{aligned}$$

Here each \mathcal{T}_k is a MCMC kernel that leaves the bridging density $f_k(\mathbf{z})$ invariant. While there is considerable freedom in the choice of $q_0(\mathbf{z})$ and $\{f_k(\mathbf{z})\}$, it is natural to let $q_0(\mathbf{z}) = p_\theta(\mathbf{z})$ and $f_k(\mathbf{z}) \propto q_0(\mathbf{z})^{1-\beta_k} p_\theta(\mathcal{D}, \mathbf{z})^{\beta_k}$ where $\{\beta_k\}$ are inverse temperatures that satisfy $0 < \beta_1 < \beta_2 < \dots < \beta_K = 1$. Additionally

$$\tilde{\mathcal{T}}_k(\mathbf{z}_{k-1} | \mathbf{z}_k) = \mathcal{T}_k(\mathbf{z}_k | \mathbf{z}_{k-1}) f_k(\mathbf{z}_{k-1}) / f_k(\mathbf{z}_k) \quad (5)$$

is the reverse MCMC kernel corresponding to \mathcal{T}_k . Conceptually, the kernels $\{\mathcal{T}_k\}$ move samples from $q_0(\mathbf{z})$ towards the posterior via a sequence of moves, each of which is ‘small’ for appropriately spaced $\{\beta_k\}$ and sufficiently large K . Indeed AIS is consistent as $K \rightarrow \infty$ (Neal, 2001) and has been shown to achieve accurate estimates of $\log p_\theta(\mathcal{D})$ empirically (Grosse et al., 2015).

3.3. Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is a powerful gradient-based MCMC method (Duane et al., 1987; Neal et al., 2011). HMC proceeds by introducing an auxiliary momentum $\mathbf{v} \in \mathbb{R}^D$ and a log joint (un-normalized) density

$\log \pi(\mathbf{z}, \mathbf{v}) = -H(\mathbf{z}, \mathbf{v})$ where the Hamiltonian is defined as $H(\mathbf{z}, \mathbf{v}) = T(\mathbf{v}) + V(\mathbf{z})$, $V(\mathbf{z}) = -\log p_\theta(\mathcal{D}, \mathbf{z})$ is the potential energy, $T(\mathbf{v}) = \frac{1}{2}\mathbf{v}^\top \mathbf{M}^{-1}\mathbf{v}$ is the kinetic energy, and \mathbf{M} is the mass matrix. Hamiltonian dynamics is then defined as:

$$\frac{d\mathbf{z}}{dt} = \frac{\partial H}{\partial \mathbf{v}} \quad \frac{d\mathbf{v}}{dt} = -\frac{\partial H}{\partial \mathbf{z}} \quad (6)$$

Evolving trajectories according to Eqn. 6 for any time τ yields Markovian transitions that target the stationary distribution $\pi(\mathbf{z}, \mathbf{v})$. By alternating Hamiltonian evolution with momentum updates

$$\mathbf{v}_{t+1} \sim \mathcal{N}(\gamma \mathbf{v}_t, (1 - \gamma^2)\mathbf{M}) \quad (7)$$

and disregarding the momentum yields a MCMC chain targeting the posterior $p_\theta(\mathbf{z}|\mathcal{D})$. Here $\gamma \in [0, 1)$ controls the level of momentum refreshment; e.g. the $\gamma \approx 1$ regime suppresses random-walk behavior (Horowitz, 1991). In general we cannot simulate Hamiltonian trajectories exactly and instead use a symplectic integrator like the so-called ‘leapfrog’ integrator. Combined with momentum refreshment a single HMC step $(\mathbf{z}_{k-1}, \mathbf{v}_{k-1}) \rightarrow (\mathbf{z}_k, \mathbf{v}_k)$ is given by

$$\begin{aligned} \hat{\mathbf{z}}_k &\leftarrow \mathbf{z}_{k-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{v}_{k-1} & \hat{\mathbf{v}}_k &\leftarrow \mathbf{v}_{k-1} - \eta \nabla V(\hat{\mathbf{z}}_k) \\ \mathbf{z}_k &\leftarrow \hat{\mathbf{z}}_k + \frac{\eta}{2}\mathbf{M}^{-1}\hat{\mathbf{v}}_k & \mathbf{v}_k &\sim \mathcal{N}(\gamma \hat{\mathbf{v}}_k, (1 - \gamma^2)\mathbf{M}) \end{aligned} \quad (8)$$

where η is the step size. Since the leapfrog integrator is inexact, a Metropolis-Hastings accept/reject step is used to ensure asymptotic correctness. For a comprehensive introduction to HMC see e.g. (Betancourt, 2017).

3.4. Differentiable Annealed Importance Sampling a.k.a. Uncorrected Hamiltonian Annealing

We describe recent work, (UHA; Geffner & Domke (2021)) and (DAIS; Zhang et al. (2021)), that combines HMC and AIS in a variational inference framework. For simplicity we refer to the authors’ algorithm as DAIS and disregard any differences between the two (contemporaneous) references.³

A notable feature of HMC is that the HMC kernel $\mathcal{T}(\mathbf{z}_1, \mathbf{v}_1|\mathbf{z}_0, \mathbf{v}_0)$ can generate large moves in \mathbf{z} -space with large acceptance probabilities, which makes it an attractive kernel choice for AIS (Sohl-Dickstein & Culpepper, 2012). Moreover, as an importance sampling framework AIS naturally gives rise to a variational bound, since applying Jensen’s inequality to the log of Eqn. 4 immediately yields a lower bound to $\log p(\mathcal{D})$. Geffner & Domke (2021) and Zhang et al. (2021) note that the utility of such an AIS variational bound is severely hampered by the fact that typical MCMC kernels include a Metropolis-Hastings accept/reject step that makes the bound non-differentiable.

³A closely related approach that utilizes unadjusted overdamped Langevin steps is described in (Thin et al., 2021).

DAIS restores differentiability by removing the accept/reject step, which makes it straightforward to optimize the variational bound using gradient-based methods. While dropping the Metropolis-Hastings correction invalidates detailed balance, AIS and thus the resulting variational bound remain intact. Moreover, since DAIS employs a HMC kernel and since we expect HMC moves to have high acceptance probabilities—at least for a well-chosen step size η and mass matrix \mathbf{M} —we expect the K DAIS transitions to efficiently move samples from $q_0(\mathbf{z})$ towards the posterior.

In more detail DAIS operates on an extended space $(\mathbf{z}_0, \dots, \mathbf{z}_K, \mathbf{v}_0, \dots, \mathbf{v}_K)$, with proposal and target distributions given by

$$q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) = q_0(\mathbf{z}_0)q_0(\mathbf{v}_0) \times \quad (9)$$

$$\begin{aligned} &\prod_{k=1}^K \mathcal{T}_k(\mathbf{z}_k, \mathbf{v}_k|\mathbf{z}_{k-1}, \mathbf{v}_{k-1}) \\ q_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) &= p_\theta(\mathcal{D}, \mathbf{z}_K) \times \quad (10) \\ &\prod_{k=1}^K \tilde{\mathcal{T}}_k(\mathbf{z}_{k-1}, \mathbf{v}_{k-1}|\mathbf{z}_k, \mathbf{v}_k) \end{aligned}$$

where $q_0(\mathbf{v}_0) = \mathcal{N}(\mathbf{v}_0|\mathbf{0}, \mathbf{M})$ is the momentum distribution. Here each kernel \mathcal{T}_k performs a single leapfrog step as in Eqn. 8 using the annealed potential energy

$$\begin{aligned} V_k(\mathbf{z}) &= -(1 - \beta_k) \log q_0(\mathbf{z}) - \beta_k \log p_\theta(\mathcal{D}, \mathbf{z}) \quad (11) \\ &= -(1 - \beta_k) \log q_0(\mathbf{z}) - \beta_k (\Psi_0(\mathbf{z}) + \Psi_L(\mathcal{D}, \mathbf{z})) \end{aligned}$$

and without including a Metropolis-Hastings correction. Notably, Geffner & Domke (2021) and Zhang et al. (2021) show that the DAIS variational objective is easy to compute, as clarified by the following lemma.

Lemma 1 *The DAIS bound given by proposal and target distributions as in Eqn. 9 is differentiable and is given by*

$$\begin{aligned} \mathcal{L}_{\text{DAIS}} &\equiv \mathbb{E} \left[\log q_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) - \log q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) \right] \\ &= \mathbb{E} \left[\log p_\theta(\mathcal{D}, \mathbf{z}_K) - \log q_0(\mathbf{z}_0) + \right. \\ &\quad \left. \sum_{k=1}^K \{ \log \mathcal{N}(\hat{\mathbf{v}}_k, \mathbf{M}) - \log \mathcal{N}(\mathbf{v}_{k-1}, \mathbf{M}) \} \right] \end{aligned}$$

where the expectation is w.r.t. q_{fwd} . Moreover, gradient estimates of $\mathcal{L}_{\text{DAIS}}$ w.r.t. θ and ϕ can be computed using reparameterized gradients provided the base distribution $q_0(\mathbf{z})$ is reparameterizable. For a proof see Sec. A.3 in the supplemental materials, Geffner & Domke (2021) or Zhang et al. (2021).

Note that \mathcal{T}_k and $\tilde{\mathcal{T}}_k$ do not appear explicitly in Lemma 1; instead it suffices to compute kinetic energy differences.

4. Surrogate Likelihood DAIS

The DAIS variational objective in Lemma 1 can lead to tight bounds on the model log evidence $\log p_\theta(\mathcal{D})$, especially

for large K . However, for a model like that in Eqn. 1, optimizing $\mathcal{L}_{\text{DAIS}}$ can be prohibitively expensive, with a $\mathcal{O}(NK)$ cost per optimization step for a dataset with N data points. This is because sampling q_{fwd} requires computing the gradient of the full log likelihood, i.e. $\nabla \Psi_{\text{L}}(\mathcal{D}, \mathbf{z})$, K times, which is expensive whenever N is large or individual likelihood terms are costly to compute.

In order to speed-up training and sampling in this regime we first observe that the proof of Lemma 1 holds for *any choice* of potential energies $\{V_k(\mathbf{z})\}$.⁴ The annealed ansatz in Eqn. 11 is a natural choice, but any other choice that efficiently moves samples from $q_0(\mathbf{z})$ towards the posterior $p_{\theta}(\mathbf{z}|\mathcal{D})$ can lead to tight variational bounds. This observation naturally leads to two scalable variational bounds that are tractable in the large data regime. In the first method, a variant of which was considered theoretically in Zhang et al. (2021) and which we refer to as **NS-DAIS**, we replace $V_k(\mathbf{z})$ with a *stochastic* estimate that depends on a mini-batch of data. In the second method, which we refer to as **SL-DAIS**, we replace the full log likelihood $\Psi_{\text{L}}(\mathcal{D}, \mathbf{z})$ with a *surrogate log likelihood* $\hat{\Psi}_{\text{L}}(\mathbf{z})$ that is cheap to evaluate. Below we argue theoretically and demonstrate empirically that SL-DAIS is superior to NS-DAIS.

4.1. NS-DAIS: Naive Subsampling DAIS

For any mini-batch of indices $\mathcal{J} \subset \{1, \dots, N\}$ with $|\mathcal{J}| = B$ we can define an estimator of $\Psi_{\text{L}}(\mathcal{D}, \mathbf{z})$ as $\frac{N}{B} \Psi_{\text{L}}(\mathcal{D}_{\mathcal{J}}, \mathbf{z})$. In NS-DAIS we plug this estimator into the potential energy in Eqn. 11. With this choice of $V_k(\mathbf{z})$ sampling from q_{fwd} is $\mathcal{O}(B)$ instead of $\mathcal{O}(N)$. Moreover, by replacing the $\log p_{\theta}(\mathcal{D}, \mathbf{z}_K)$ term in $\mathcal{L}_{\text{DAIS}}$ with an estimator computed using an *independent* mini-batch of indices $\mathcal{I} \subset \{1, \dots, N\}$ with $|\mathcal{I}| = B$ we obtain an unbiased estimator that is $\mathcal{O}(B)$ instead of $\mathcal{O}(N)$. More formally, the proposal and target distributions q_{fwd} and q_{bwd} in Eqn. 9 are replaced with

$$\begin{aligned} q_{\text{fwd}} &\rightarrow q_0(\mathbf{z}_0)q_0(\mathbf{v}_0)\prod_{k=1}^K \mathcal{T}_k(\mathbf{z}_k, \mathbf{v}_k|\mathbf{z}_{k-1}, \mathbf{v}_{k-1}, \mathcal{J}) \\ q_{\text{bwd}} &\rightarrow p_{\theta}(\mathbf{z}_K|\mathcal{D}_{\mathcal{I}})^{N/B} p_{\theta}(\mathbf{z}_K)\prod_{k=1}^K \tilde{\mathcal{T}}_k(\mathbf{z}_{k-1}, \mathbf{v}_{k-1}|\cdot) \end{aligned}$$

where we omit the common factor of $q(\mathcal{I})q(\mathcal{J})$ for brevity. See Algorithm 2 and Sec. A.5 in the supplement for details.

4.2. SL-DAIS: Surrogate Likelihood DAIS

We proceed as in NS-DAIS except we plug in a *fixed non-stochastic* surrogate log likelihood $\hat{\Psi}_{\text{L}}(\mathbf{z})$ into the potential energy in Eqn. 11. More formally, the proposal and target distributions q_{fwd} and q_{bwd} in Eqn. 9 are replaced with

$$\begin{aligned} q_{\text{fwd}} &\rightarrow q(\mathcal{I})q_0(\mathbf{z}_0)q_0(\mathbf{v}_0)\prod_{k=1}^K \mathcal{T}_k(\mathbf{z}_k, \mathbf{v}_k|\mathbf{z}_{k-1}, \mathbf{v}_{k-1}, \hat{\Psi}_{\text{L}}) \\ q_{\text{bwd}} &\rightarrow q(\mathcal{I})p_{\theta}(\mathcal{D}_{\mathcal{I}}|\mathbf{z}_K)^{N/B} p_{\theta}(\mathbf{z}_K)\prod_{k=1}^K \tilde{\mathcal{T}}_k(\mathbf{z}_{k-1}, \mathbf{v}_{k-1}|\cdot) \end{aligned}$$

⁴See Sec. A.3 in the supplement for details.

Algorithm 1 SL-DAIS: Surrogate Likelihood Differentiable Annealed Importance Sampling. We highlight in blue where the algorithm differs from DAIS. To recover DAIS we substitute $\hat{\Psi}_{\text{L}}(\mathbf{z}) \rightarrow \Psi_{\text{L}}(\mathcal{D}, \mathbf{z})$ and $B \rightarrow N$. Note that $\hat{\Psi}_{\text{L}}$ is only used to guide HMC dynamics and that a stochastic estimate of $\Psi_{\text{L}}(\mathcal{D}, \mathbf{z})$ still appears on the final line.

Input: model log density $\Psi_0(\mathbf{z}) + \Psi_{\text{L}}(\mathcal{D}, \mathbf{z})$, surrogate log likelihood $\hat{\Psi}_{\text{L}}(\mathbf{z})$, base variational distribution $q_0(\mathbf{z})$, number of steps K , inverse temperatures $\{\beta_k\}$, step size η , momentum refresh parameter γ , mass matrix \mathbf{M} , dataset \mathcal{D} of size N , mini-batch size B

Initialize: $\mathbf{z}_0 \sim q_0$, $\mathbf{v}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$, $\mathcal{L} \leftarrow -\log q_0(\mathbf{z}_0)$

for $k = 1$ **to** K **do**

$$\hat{\mathbf{z}}_k \leftarrow \mathbf{z}_{k-1} + \frac{\eta}{2} \mathbf{M}^{-1} \mathbf{v}_{k-1}$$

$$\mathbf{g}_k \leftarrow \nabla_{\mathbf{z}} \left\{ \beta_k \left(\Psi_0(\mathbf{z}) + \hat{\Psi}_{\text{L}}(\mathbf{z}) \right) + (1 - \beta_k) \log q_0(\mathbf{z}) \right\} \Big|_{\mathbf{z}=\hat{\mathbf{z}}_k}$$

$$\hat{\mathbf{v}}_k \leftarrow \mathbf{v}_{k-1} + \eta \mathbf{g}_k$$

$$\mathbf{z}_k \leftarrow \hat{\mathbf{z}}_k + \frac{\eta}{2} \mathbf{M}^{-1} \hat{\mathbf{v}}_k$$

if $k < K$ **then**

$$\mathbf{v}_k \leftarrow \gamma \hat{\mathbf{v}}_k + \sqrt{1 - \gamma^2} \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$$

end if

$$\mathcal{L} \leftarrow \mathcal{L} + \log \mathcal{N}(\hat{\mathbf{v}}_k, \mathbf{M}) - \log \mathcal{N}(\mathbf{v}_{k-1}, \mathbf{M})$$

end for

Sample mini-batch indices $\mathcal{I} \subset \{1, \dots, N\}$ with $|\mathcal{I}| = B$

Return: $\mathcal{L} + \Psi_0(\mathbf{z}_K) + \frac{N}{B} \Psi_{\text{L}}(\mathcal{D}_{\mathcal{I}}, \mathbf{z}_K)$

where $q(\mathcal{I})$ encodes sampling mini-batches of B indices without replacement, and data subsampling occurs solely in the $p_{\theta}(\mathcal{D}_{\mathcal{I}}|\mathbf{z}_K)$ term. Like NS-DAIS the SL-DAIS ELBO admits a simple unbiased gradient estimator. In contrast to NS-DAIS, the HMC dynamics in SL-DAIS targets a fixed target distribution. We expect the reduced stochasticity of SL-DAIS as compared to NS-DAIS to result in superior empirical performance.⁵ See Algorithm 1 for the complete algorithm and Sec. A.6 in the supplement for a more complete formal description.

There are multiple possibilities for how to parameterize $\hat{\Psi}_{\text{L}}(\mathbf{z})$. We briefly describe the simplest possible recipe, which we refer to as **RAND**, leaving a detailed ablation study to Sec. 7.1. In **RAND** we randomly choose $N_{\text{surr}} \ll N$ surrogate data points $\{(\tilde{y}_n, \tilde{\mathbf{x}}_n)\} \subset \mathcal{D}$ and introduce a N_{surr} -dimensional vector of learnable (positive) weights $\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ can be learned jointly with other variational parameters.

⁵This expectation is closely related to a comment in Zhang et al. (2021): [DAIS] relies on all intermediate distributions, and so the error induced by stochastic gradient noise accumulates over the whole trajectory. Simply taking smaller steps fails to reduce the error. We conjecture that AIS-style algorithms are inherently fragile to gradient noise.

The surrogate log likelihood is then given by

$$\hat{\Psi}_L(\mathbf{z}) = \sum_n \omega_n \log p_\theta(\tilde{y}_n | \mathbf{z}, \tilde{\mathbf{x}}_n) \quad (12)$$

Evidently computing $\hat{\Psi}_L(\mathbf{z})$ is $\mathcal{O}(N_{\text{surr}})$. Besides its simplicity, an appealing feature of this parameterization is that it leverages the known functional form of the likelihood.

4.3. Discussion

Before taking a closer look at the theoretical properties of NS-DAIS and SL-DAIS in Sec. 5, we make two simple observations. First, as we should expect from any bona fide variational method, maximizing the variational bound does indeed lead to tighter posterior approximations, as formalized in the following proposition:

Proposition 1 *The NS-DAIS and SL-DAIS approximate posterior distributions, each of which is given by the marginal $q_{\text{fwd}}(\mathbf{z}_K)$, both satisfy the inequality*

$$\log p_\theta(\mathcal{D}) - \mathcal{L} \geq \text{KL}(q_{\text{fwd}}(\mathbf{z}_K) | p_\theta(\mathbf{z}_K | \mathcal{D})) \geq 0$$

where \mathcal{L} is $\mathcal{L}_{\text{NS-DAIS}}$ or $\mathcal{L}_{\text{SL-DAIS}}$, respectively.

See Sec. A.4 for a proof. Thus as \mathcal{L} increases for fixed θ , the KL divergence decreases and $q_{\text{fwd}}(\mathbf{z}_K)$ becomes a better approximation to the posterior. Second, sampling from q_{fwd} in the case of NS-DAIS requires the entire dataset \mathcal{D} . Conveniently in the case of SL-DAIS we only require the surrogate log likelihood $\hat{\Psi}_L(\mathbf{z})$ so that the dataset can be discarded after training.

5. Convergence analysis

As in Zhang et al. (2021) to make our analysis tractable we consider linear regression with a prior $p_\theta(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1})$ and a likelihood $\prod_n \mathcal{N}(y_n | \mathbf{z} \cdot \mathbf{x}_n, \sigma_{\text{obs}}^2)$, where each $y_n \in \mathbb{R}$ and $\mathbf{x}_n \in \mathbb{R}^D$. Furthermore we work under the following set of simplifying assumptions:

Assumption 1 *We use $\gamma = 0$, equally spaced inverse temperatures $\{\beta_k\}$, a step size that varies as $\eta \sim K^{-1/4}$, and the prior as the base distribution (i.e. $q_0(\mathbf{z}) = p_\theta(\mathbf{z})$).*

5.1. NS-DAIS

What kind of posterior approximation do we expect from NS-DAIS? As clarified by the following proposition, NS-DAIS does not directly target the posterior $p_\theta(\mathbf{z} | \mathcal{D})$.

Proposition 2 *The approximate posterior for linear regression given by running NS-DAIS with K steps under Assumption 1 converges to the ‘aggregate pseudo-posterior’*

$$p_\theta^{\text{Agg}}(\mathbf{z} | \mathcal{D}) \equiv \mathbb{E}_{q(\mathcal{J})} [p_\theta(\mathbf{z} | \mathcal{D}_{\mathcal{J}})] \quad (13)$$

as $K \rightarrow \infty$. Here $p_\theta(\mathbf{z} | \mathcal{D}_{\mathcal{J}})$ denotes the posterior corresponding to the data subset $\mathcal{D}_{\mathcal{J}}$ with appropriately scaled likelihood term, i.e. $p_\theta(\mathbf{z} | \mathcal{D}_{\mathcal{J}}) \propto p_\theta(\mathcal{D}_{\mathcal{J}} | \mathbf{z})^{N/B} p_\theta(\mathbf{z})$.

See Sec. A.5 for a proof and additional details. Thus we generally expect NS-DAIS to provide a good posterior approximation when the aggregate pseudo-posterior is a good approximation to the posterior. The poor performance of NS-DAIS in experiments in Sec. 7 suggests that this is not case for moderate mini-batch sizes in typical models.

5.2. SL-DAIS

We now analyze SL-DAIS assuming the surrogate log likelihood $\hat{\Psi}_L(\mathbf{z})$ differs from the full log likelihood. As is well known the exact posterior for linear regression is given by $\mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Lambda}_{\text{post}}^{-1})$ where $\boldsymbol{\mu}_{\text{post}} = \boldsymbol{\Lambda}_{\text{post}}^{-1}(\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \frac{1}{\sigma_{\text{obs}}^2} \mathbf{X}^\top \mathbf{y})$ and $\boldsymbol{\Lambda}_{\text{post}} = \boldsymbol{\Lambda}_0 + \frac{1}{\sigma_{\text{obs}}^2} \mathbf{X}^\top \mathbf{X}$. The gradient of the full log likelihood is given by

$$\nabla_{\mathbf{z}} \Psi_L(\mathcal{D}, \mathbf{z}) = \frac{1}{\sigma_{\text{obs}}^2} (\mathbf{y}^\top \mathbf{X} - \mathbf{X}^\top \mathbf{X} \mathbf{z}) = \mathbf{a} - \mathbf{B} \mathbf{z} \quad (14)$$

where we have defined $\mathbf{a} \equiv \frac{1}{\sigma_{\text{obs}}^2} \mathbf{y}^\top \mathbf{X}$ and $\mathbf{B} \equiv \frac{1}{\sigma_{\text{obs}}^2} \mathbf{X}^\top \mathbf{X}$. We suppose that $\hat{\Psi}_L(\mathbf{z})$ is likewise a quadratic⁶ function of \mathbf{z} but that $\hat{\Psi}_L(\mathbf{z}) \neq \Psi_L(\mathcal{D}, \mathbf{z})$ so that we can write

$$\nabla_{\mathbf{z}} \hat{\Psi}_L(\mathbf{z}) = \mathbf{a} + \delta \mathbf{a} - (\mathbf{B} + \delta \mathbf{B}) \mathbf{z} \quad (15)$$

For $\nabla_{\mathbf{z}} \hat{\Psi}_L(\mathbf{z})$ as in Eqn. 15 we can prove the following:

Proposition 3 *Running SL-DAIS for linear regression with K steps under Assumption 1 and using $\nabla_{\mathbf{z}} \hat{\Psi}_L(\mathbf{z})$ as in Eqn. 15 results in a variational gap that can be bounded as*

$$\log p(\mathcal{D}) - \mathcal{L}_{\text{SL-DAIS}} \leq \underbrace{\mathcal{O}(K^{-1/2})}_{\text{DAIS error}} + \underbrace{|\text{Tr} \boldsymbol{\Lambda}_{\text{post}}^{-1} \delta \mathbf{B}|}_{\text{surrogate likelihood error}}$$

where we have dropped higher order terms in $\delta \mathbf{a}$ and $\delta \mathbf{B}$.⁷ Furthermore the KL divergence between $q_{\text{fwd}}(\mathbf{z}_K)$ and the posterior $p_\theta(\mathbf{z}_K | \mathcal{D})$ is bounded by the same quantity.

This intuitive result can be proven by suitably adapting the results in Zhang et al. (2021); see Sec. A.6 in the supplemental materials for details.

Proposition 3 suggests that SL-DAIS offers an intuitive trade-off between inference speed and fidelity. For example, for a fixed computational cost, SL-DAIS can accommodate

⁶Note that $\hat{\Psi}_L(\mathbf{z})$ will be precisely of this form if we use the RAND parameterization. In other words Eqn. 15 follows directly from Eqn. 12, which is agnostic to the particular likelihood that appears in the model.

⁷ $\delta \mathbf{a}$ first appears at second order. Note that we drop higher order terms simply to make the bound more readily interpretable.

a value of K that is a factor $\sim N/B$ larger than DAIS. For a sufficiently good surrogate log likelihood, the benefits of larger K can more than compensate for the error introduced by an imperfect $\hat{\Psi}_L(\mathbf{z})$. See Sec. 7.3 for concrete examples.

6. Related work

Many variational objectives that leverage importance sampling (IS) have been proposed. These include the importance weighted autoencoder (IWAE) (Burda et al., 2015; Cremer et al., 2017), the thermodynamic variational objective (Masrani et al., 2019), and approaches that make use of Sequential Monte Carlo (Le et al., 2017; Maddison et al., 2017; Naesseth et al., 2018). For a general discussion of IS in variational inference see Domke & Sheldon (2018).

An early combination of MCMC methods with variational inference was proposed by Salimans et al. (2015) and Wolf et al. (2016). A disadvantage of these approaches is the need to learn reverse kernels, a shortcoming that was later addressed by Caterini et al. (2018).

Bayesian coresets enable users to run MCMC on large datasets after first distilling the data into a smaller number of weighted data points (Huggins et al., 2016; Campbell & Broderick, 2018; 2019). These methods do not support model learning. Finally a number of authors have explored MCMC methods that enable data subsampling (Maclaurin & Adams, 2015; Quiroz et al., 2018; Zhang et al., 2020), including those that leverage stochastic gradients (Welling & Teh, 2011; Chen et al., 2014; Ma et al., 2015). See Sec. A.2 for an extended discussion of related work.

7. Experiments

Next we compare the performance of SL-DAIS and NL-DAIS to various MCMC and variational baselines. Our experiments are implemented using JAX (Bradbury et al., 2020) and NumPyro (Phan et al., 2019; Bingham et al., 2019). An open source implementation of our method will be made available at <https://num.pyro.ai/en/stable/autoguide.html>.

7.1. Surrogate log likelihood comparison

We compare four ansätze for the surrogate log likelihood $\hat{\Psi}_L$ used in SL-DAIS. For concreteness we consider a logistic regression model with a bernoulli likelihood $p(y_n|\mathbf{z}, \mathbf{x}_n)$ governed by logits $\sigma(\mathbf{x}_n \cdot \mathbf{z})$, where $\sigma(\cdot)$ is the logistic function. In RAND we randomly choose N_{surr} surrogate data points $\{(\tilde{y}_n, \tilde{\mathbf{x}}_n)\} \subset \mathcal{D}$, introduce a N_{surr} -dimensional vector of learnable weights ω and let $\hat{\Psi}_L(\mathbf{z}) = \sum_n \omega_n \log p(\tilde{y}_n|\mathbf{z}, \tilde{\mathbf{x}}_n)$. In CS-INIT we proceed similarly but use a Bayesian coreset algorithm (Huggins et al., 2016) to choose the N_{surr} surrogate data points. In CS-FIX we

likewise use a coreset algorithm to choose the surrogate data points but instead of learning ω we use the weights provided by the coreset algorithm. Finally in NN we parameterize $\hat{\Psi}_L(\mathbf{z})$ as a neural network. See Table 1 for the results.

We can read off several conclusions from Table 1. First, using a neural ansatz works extremely poorly. This is not surprising, since a neural ansatz does not leverage the known likelihood function. Second, for the three methods that rely on surrogate data points $\{(\tilde{y}_n, \tilde{\mathbf{x}}_n)\}$, results improve as we increase N_{surr} , although the improvements are somewhat modest for the two inference problems we consider. Third, the simplest ansatz, namely RAND, works quite well. This is encouraging because this ansatz is simple, generic, and easily automated, which makes it an excellent candidate for probabilistic programming frameworks. Consequently we use RAND in all subsequent experiments. Finally, the poor performance of CS-FIX makes it clear that the notion of Bayesian coresets, while conceptually similar, is not congruent with our use case for surrogate likelihoods. In particular the coreset notion in (Huggins et al., 2016) aims to approximate $\Psi_L(\mathcal{D}, \mathbf{z})$ across latent space as a whole. However, the zero-avoiding behavior of variational inference implies that reproducing the precise tail behavior of $\Psi_L(\mathcal{D}, \mathbf{z})$ is less important than accurately representing the vicinity of the posterior mode. Additionally the Hamiltonian ‘integration error’ resulting from using finite η and K can be partially corrected for by learning ω jointly with $q_0(\mathbf{z})$, something that a two-stage coreset-based approach is unable to do.

Dataset	Higgs		SUSY	
N_{surr}	64	1024	64	1024
RAND	211.1	222.4	627.5	637.2
CS-INIT	209.6	218.5	625.9	633.7
CS-FIX	83.0	192.7	501.0	555.5
NN	16.8		2.6	

Table 1: We compare four ansätze for the surrogate log likelihood $\hat{\Psi}_L$ on two logistic regression datasets. For the three strategies that make use of surrogate data points, we vary $N_{\text{surr}} \in \{64, 1024\}$. We report ELBO improvements (in nats) above a mean-field Normal baseline and average results across 20 replications. See Sec. 7.1 for details and Table 6 in the supplement for expanded results, including uncertainties and additional parameterizations.

7.2. Classifying imbalanced data

To better understand the limitations of NS-DAIS and SL-DAIS we consider a binary classification problem with imbalanced data. We expect both methods to find this regime challenging, since the full log likelihood $\Psi_L(\mathcal{D}, \mathbf{z})$ exhibits elevated sensitivity to the small number of terms involving

the rare class. See Fig. 1 for the results. Test accuracies decrease substantially as the class imbalance increases; this is expected, since there is considerable overlap between the two classes in feature space. Strikingly, SL-DAIS with the RAND surrogate ansatz outperforms NS-DAIS across the board,⁸ with the gap increasing as the class imbalance increases. Indeed NS-DAIS barely outperforms a mean-field baseline (not shown for legibility). This suggests that SL-DAIS + RAND can be a viable strategy even in difficult regimes like the one explored here.

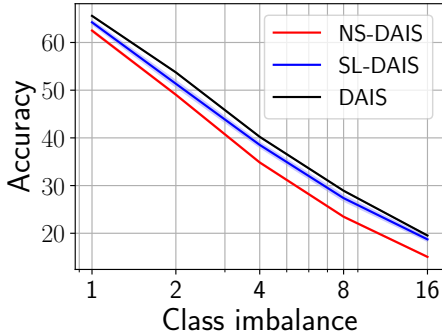


Figure 1: We depict test accuracy w.r.t. the rare class label on an imbalanced binary classification dataset. We compare DAIS to SL-DAIS with $N_{\text{surr}} = 256$ and NS-DAIS with $B = 256$. The horizontal axis encodes the ratio between the number of non-rare and rare class labels. Accuracies are averaged over 20 (3) independent replications for SL-DAIS and NS-DAIS (DAIS), respectively, and shaded bands denote 90% confidence intervals.

7.3. Logistic regression

We use five logistic regression datasets to systematically compare eight variational methods and two MCMC methods. For each dataset we consider a somewhat moderate number of training data ($N = 5 \times 10^4$) so that we can include baseline methods that are impractical for larger N . Our baselines include three variational methods that use parametric distributions: mean-field Normal (MF); multivariate Normal (MVN); and a block neural autoregressive flow (Flow) (De Cao et al., 2020). They also include DAIS-MF, NS-DAIS-MF, and SL-DAIS-MF (all of which use a mean-field base distribution) and SL-DAIS-MVN, which uses a MVN base distribution. NS-DAIS and SL-DAIS use $B = 256$ and $N_{\text{surr}} = 256$, respectively. The two MCMC methods are HMCECS (Dang et al., 2019), a variant of HMC that considers subsets of data in each iteration, and NUTS-CS, where we first form a Bayesian coreset consisting of 1024 weighted data points, and then run NUTS (Hoffman et al., 2014). We

⁸We observe similar behavior for the ELBO.

emphasize that these two MCMC methods do not offer some of the benefits of variational methods (e.g. support for model learning) but we include them so that we can better assess the quality of the approximate posterior predictive distributions returned by the variational methods.

See Fig. 2 and Table 2 for results. We find that SL-DAIS consistently outperforms NS-DAIS and in many cases it matches the performance of DAIS. Among the scalable methods Flow and SL-DAIS perform best, although SL-DAIS can be substantially faster than Flow, see Fig. 3. Fig. 3 also clarifies the computational benefits of SL-DAIS: SL-DAIS with $K = 8$ steps handily outperforms DAIS with $K = 2$ at about 2% of the computational cost. Finally we find that the best variational methods yield predictive log likelihoods that are competitive with the MCMC baselines.

7.4. Robust Gaussian process regression

We model a geospatial precipitation dataset (Lyon, 2004; Lyon & Barnston, 2005) considered in (Pleiss et al., 2020) using a Gaussian process regressor with a heavy-tailed Student’s t likelihood; see Sec. A.7.5 for details on the modeling setup. We compare SL-DAIS with $K = 4$, $N_{\text{surr}} = 256$, and a multivariate Normal (MVN) base distribution $q_0(\mathbf{z})$ to a baseline with a MVN variational distribution. Note that the dimension of the latent space D is equal to the number of inducing points. See Table 3 for the results.

There is significant variability in the data and consequently the posterior function values exhibit considerable uncertainty. Due to the non-conjugate likelihood, the posterior over the inducing points features significant non-gaussianity, especially as the number of inducing points increases and the inducing point locations become closer together. Consequently we see the largest ELBO and test log likelihood improvements for the largest number of inducing points.

latent dim.	$D = 64$	$D = 128$	$D = 256$
ΔELBO	127.3 ± 26.6	187.1 ± 26.6	529.5 ± 17.3
ΔLL	0.014 ± 0.003	0.015 ± 0.003	0.039 ± 0.006

Table 3: We report ELBO and test log likelihood (LL) improvements in nats for the regression experiment in Sec. 7.4, comparing SL-DAIS to a baseline with a multivariate Normal variational distribution. Test log likelihoods are normalized by the number of test points. Results are averaged over seven train/test splits.

7.5. Gaussian process classification

We train Gaussian process classifiers on the classification datasets considered in Sec. 7.3. We compare SL-DAIS with $K = 4$, $N_{\text{surr}} = 512$, and a MVN base distribution to a baseline with a MVN variational distribution. We use

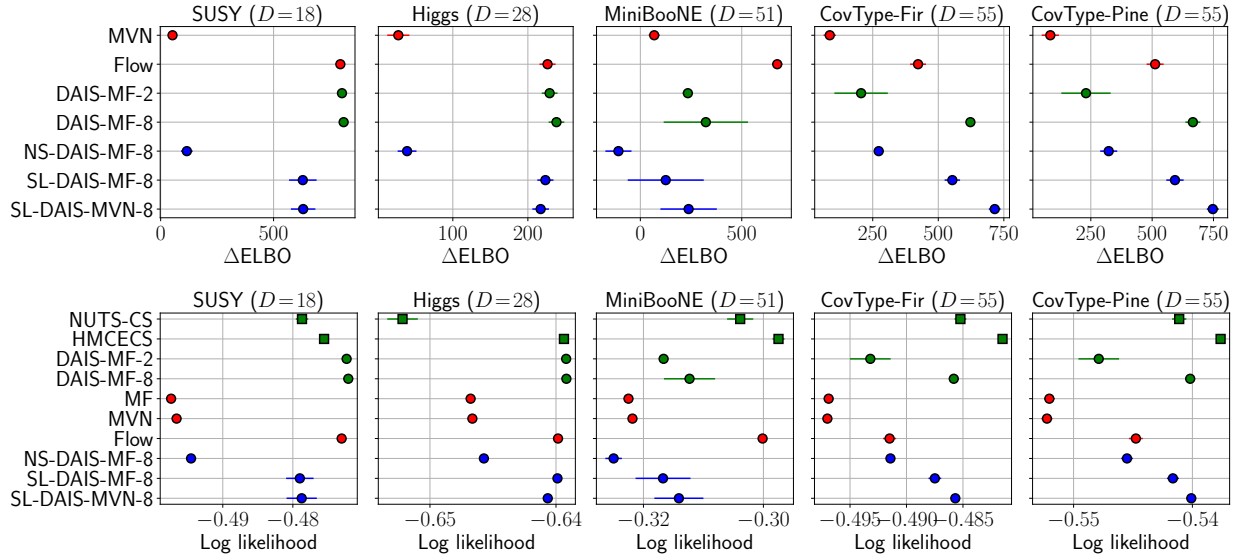


Figure 2: We report ELBO improvements and test log likelihoods for the logistic regression experiment in Sec. 7.3. ELBO improvements are with respect to the mean-field (MF) Normal baseline. Circles denote variational methods and squares denote MCMC methods. Blue methods are ours, red methods are mini-batchable variational methods, and green methods are everything else. Metrics are averaged over 7 independent replications and error bars denote standard errors (we do 3 independent replications for the most expensive methods, namely DAIS and HMCECS). See Sec. A.7.3 in the supplement for test accuracies. Note that numerals in method names indicate the number of HMC steps K used.

	MF	MVN	Flow	NS-DAIS-MF	SL-DAIS-MF	NS-DAIS-MVN	SL-DAIS-MVN
ELBO	6.60	5.36	2.08	5.00	2.60	4.16	2.20
Log likelihood	6.16	5.88	2.16	4.60	2.44	4.36	2.40
Opt. time	38.8	58.1	3420.3	113.1	83.1	152.1	131.6

Table 2: We report performance ranks w.r.t. ELBO and test log likelihood across 5 train/test splits and 5 datasets for the logistic regression experiment in Sec. 7.3. We also report time per optimization step in milliseconds as in Fig. 3. Lower is better for all metrics. The rank satisfies $1 \leq \text{rank} \leq 7$, since we compare 7 scalable variational methods.

128 inducing points so the latent dimension is $D = 128$. See Table 4 for results. We find that SL-DAIS leads to consistently higher ELBOs and that this translates to small but non-negligible gains in test accuracy for the datasets with the largest ELBO gains.

We emphasize that this and the preceding Gaussian process experiments represent difficult inference problems, since the latent dimensionality is moderately high (with D as large as 256) and since high-dimensional model parameters θ (including D inducing point locations) are learned jointly with the approximate posterior.

Dataset	Δ ELBO	Δ Accuracy
SUSY	34.0 ± 2.2	0.002 ± 0.019
Higgs	225.5 ± 7.3	0.072 ± 0.044
MiniBooNE	163.9 ± 5.4	0.121 ± 0.011
CovType-Fir	471.7 ± 42.2	0.399 ± 0.080
CovType-Pine	616.6 ± 71.8	0.425 ± 0.045

Table 4: We report ELBO and test accuracy improvements for the classification experiment in Sec. 7.5, comparing SL-DAIS to a baseline with a multivariate Normal variational distribution. Results are averaged over five train/test splits.

7.6. Local latent variable models

What about models with global and local latent variables? For simplicity we evaluate the performance of the simplest DAIS-like inference procedure that can simultaneously ac-

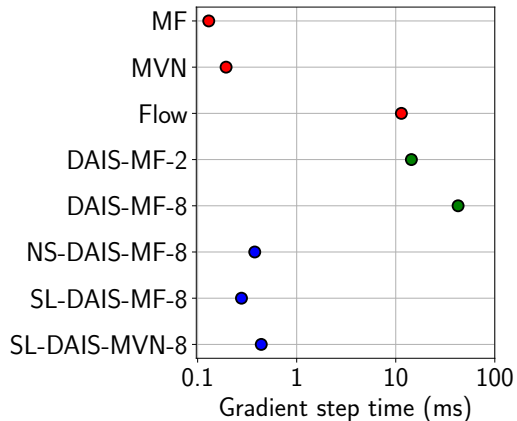


Figure 3: We report the time per optimization step for each variational method in Fig. 2 on the CovType-Fir dataset. Runtimes are for a CPU with 24 cores (Intel Xeon Gold 5220R 2.2GHz).

commodate local latent variables and data subsampling. In brief we introduce a parametric distribution for the global latent variable and use DAIS to define a distribution over the local latent variables. Assuming the local latent variables are conditionally independent once we condition on the global latent variable, this results in N *non-interacting* DAIS chains. Consequently the ELBO is amenable to data subsampling; see Sec. A.1 for an extended discussion

To evaluate this approach we consider a robust linear regression model that uses a Student’s t likelihood. We use the well-known representation of this likelihood as a continuous mixture of Normal distributions. This yields a model with local Gamma variates where the local latent variables can be integrated out exactly.

We compare two variational approaches, both of which use a mean-field Normal distribution for the global latent variable. To define an oracle baseline we integrate out the local latent variables before performing variational inference. This oracle represents an upper performance bound on the semi-parametric approach defined above, which we refer to as Semi-DAIS. See Table 5 for results. We find that for $K = 16$ Semi-DAIS ‘recovers’ about half of the gap between a fully mean-field baseline and the oracle. We expect this gap would decrease further for larger K .

Dataset	Semi-DAIS-8	Semi-DAIS-16	Oracle
Pol	91.4 ± 3.3	136.4 ± 3.6	360.2 ± 2.6
Elevators	186.3 ± 5.3	366.8 ± 5.5	758.0 ± 2.5

Table 5: We report results for the experiment in Sec. 7.6. In each case the reported ELBO improvement is above a mean field baseline. Results are averaged over 10 replications.

8. Discussion

In this work we have focused on models with global latent variables. The experiment in Sec. 7.6 only scratches the surface of what is possible for the richer and more complex case of models that contain global and local latent variables. Exploring hybrid scalable inference strategies for this class of models that combine gradient-based MCMC with variational methods is an interesting direction for future work.

Acknowledgements

We thank Tomas Geffner for answering questions about Geffner & Domke (2021). We thank Ola Rønning for contributing the open source NumPyro-based HMCECS implementation we used in our experiments.

References

Asuncion, A. and Newman, D. Uci machine learning repository, 2007.

Bachem, O., Lucic, M., and Krause, A. Practical core-set constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

Baldi, P., Sadowski, P., and Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):1–9, 2014.

Betancourt, M. The fundamental incompatibility of scalable hamiltonian monte carlo and naive data subsampling. In *International Conference on Machine Learning*, pp. 533–540. PMLR, 2015.

Betancourt, M. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

Blackard, J. A. and Dean, D. J. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151, 1999.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., and Wanderman-Milne, S. Jax: com-

- posable transformations of python+ numpy programs, 2018. URL <http://github.com/google/jax>, 4:16, 2020.
- Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.
- Campbell, T. and Broderick, T. Bayesian coresets construction via greedy iterative geodesic ascent. In *International Conference on Machine Learning*, pp. 698–706. PMLR, 2018.
- Campbell, T. and Broderick, T. Automated scalable bayesian inference via hilbert coresets. *The Journal of Machine Learning Research*, 20(1):551–588, 2019.
- Caterini, A. L., Doucet, A., and Sejdinovic, D. Hamiltonian variational auto-encoder. *arXiv preprint arXiv:1805.11328*, 2018.
- Chen, T., Fox, E., and Guestrin, C. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pp. 1683–1691. PMLR, 2014.
- Cremer, C., Morris, Q., and Duvenaud, D. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.
- Dang, K.-D., Quiroz, M., Kohn, R., Minh-Ngoc, T., and Villani, M. Hamiltonian monte carlo with energy conserving subsampling. *Journal of machine learning research*, 20, 2019.
- Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. The helmholtz machine. *Neural computation*, 7(5):889–904, 1995.
- De Cao, N., Aziz, W., and Titov, I. Block neural autoregressive flow. In *Uncertainty in Artificial Intelligence*, pp. 1263–1273. PMLR, 2020.
- Ding, X. and Freedman, D. J. Learning deep generative models with annealed importance sampling. *arXiv preprint arXiv:1906.04904*, 2019.
- Domke, J. and Sheldon, D. Importance weighting and variational inference. *arXiv preprint arXiv:1808.09034*, 2018.
- Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.
- Geffner, T. and Domke, J. Mcmc variational inference via uncorrected hamiltonian annealing. *Advances in Neural Information Processing Systems*, 34, 2021.
- Grosse, R. B., Ghahramani, Z., and Adams, R. P. Sandwiching the marginal likelihood using bidirectional monte carlo. *arXiv preprint arXiv:1511.02543*, 2015.
- Hoffman, M. D. Learning deep latent gaussian models with markov chain monte carlo. In *International conference on machine learning*, pp. 1510–1519. PMLR, 2017.
- Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.
- Hoffman, M. D., Gelman, A., et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.
- Horowitz, A. M. A generalized guided monte carlo algorithm. *Physics Letters B*, 268(2):247–252, 1991.
- Huggins, J., Campbell, T., and Broderick, T. Coresets for scalable bayesian logistic regression. *Advances in Neural Information Processing Systems*, 29:4080–4088, 2016.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Le, T. A., Igl, M., Rainforth, T., Jin, T., and Wood, F. Auto-encoding sequential monte carlo. *arXiv preprint arXiv:1705.10306*, 2017.
- Li, Y., Turner, R. E., and Liu, Q. Approximate inference with amortised mcmc. *arXiv preprint arXiv:1702.08343*, 2017.
- Lyon, B. The strength of el niño and the spatial extent of tropical drought. *Geophysical Research Letters*, 31(21), 2004.
- Lyon, B. and Barnston, A. G. Enso and the spatial extent of interannual precipitation extremes in tropical land areas. *Journal of Climate*, 18(23):5095–5109, 2005.
- Ma, Y.-A., Chen, T., and Fox, E. B. A complete recipe for stochastic gradient mcmc. *arXiv preprint arXiv:1506.04696*, 2015.
- Maclaurin, D. and Adams, R. P. Firefly monte carlo: Exact mcmc with subsets of data. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.
- Maddison, C. J., Lawson, D., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. W. Filtering variational objectives. *arXiv preprint arXiv:1705.09279*, 2017.
- Masrani, V., Le, T. A., and Wood, F. The thermodynamic variational objective. *arXiv preprint arXiv:1907.00031*, 2019.

- Miller, A. C., Foti, N. J., Lewnard, J. A., Jewell, N. P., Guestrin, C., and Fox, E. B. Mobility trends provide a leading indicator of changes in sars-cov-2 transmission. *MedRxiv*, 2020.
- Monod, M., Blenkinsop, A., Xi, X., Hebert, D., Bershan, S., Tietze, S., Baguelin, M., Bradley, V. C., Chen, Y., Coupland, H., et al. Age groups that sustain resurging covid-19 epidemics in the united states. *Science*, 371(6536):eabe8372, 2021.
- Naesseth, C., Linderman, S., Ranganath, R., and Blei, D. Variational sequential monte carlo. In *International conference on artificial intelligence and statistics*, pp. 968–977. PMLR, 2018.
- Neal, R. M. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.
- Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.
- Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S. F., Pyle, J. D., Yurkovetskiy, L., Bosso, M., Park, D. J., Babadi, M., MacInnis, B. L., Luban, J., Sabeti, P. C., and Lemieux, J. E. Analysis of 6.4 million sars-cov-2 genomes identifies mutations associated with fitness. *Science*, 2022. doi: 10.1126/science.abm1208. URL <https://www.science.org/doi/abs/10.1126/science.abm1208>.
- Phan, D., Pradhan, N., and Jankowiak, M. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.
- Pleiss, G., Jankowiak, M., Eriksson, D., Damle, A., and Gardner, J. Fast matrix square roots with applications to gaussian processes and bayesian optimization. *Advances in Neural Information Processing Systems*, 33:22268–22281, 2020.
- Quinonero-Candela, J. and Rasmussen, C. E. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6: 1939–1959, 2005.
- Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. Speeding up mcmc by efficient data subsampling. *Journal of the American Statistical Association*, 2018.
- Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Artificial intelligence and statistics*, pp. 814–822. PMLR, 2014.
- Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.
- Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I., and McGregor, G. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543(2-3):577–584, 2005.
- Ruiz, F. and Titsias, M. A contrastive divergence for combining variational inference and mcmc. In *International Conference on Machine Learning*, pp. 5537–5545. PMLR, 2019.
- Salimans, T., Kingma, D., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pp. 1218–1226. PMLR, 2015.
- Snelson, E. and Ghahramani, Z. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18:1257, 2006.
- Sohl-Dickstein, J. and Culpepper, B. J. Hamiltonian annealed importance sampling for partition function estimation. *arXiv preprint arXiv:1205.1925*, 2012.
- Thin, A., Kotelevskii, N., Doucet, A., Durmus, A., Moulines, E., and Panov, M. Monte carlo variational auto-encoders. In *International Conference on Machine Learning*, pp. 10247–10257. PMLR, 2021.
- Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. The block pseudo-marginal sampler. *arXiv preprint arXiv:1603.02485*, 2016.
- Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 681–688. Citeseer, 2011.
- Wolf, C., Karl, M., and van der Smagt, P. Variational inference with hamiltonian monte carlo. *arXiv preprint arXiv:1609.08203*, 2016.
- Zhang, G., Hsu, K., Li, J., Finn, C., and Grosse, R. B. Differentiable annealed importance sampling and the perils of gradient noise. *Advances in Neural Information Processing Systems*, 34, 2021.
- Zhang, R., Cooper, A. F., and De Sa, C. M. Asymptotically optimal exact minibatch metropolis-hastings. *Advances in Neural Information Processing Systems*, 33:19500–19510, 2020.

A. Appendix

The appendix is organized as follows. In Sec. A.1 we discuss local latent variable models. In Sec. A.2 we provide an extended discussion of related work. In Sec. A.3 we provide a proof of Lemma 1. In Sec. A.4 we discuss general properties of NS-DAIS and SL-DAIS. In Sec. A.5 we discuss NS-DAIS and in Sec. A.6 we discuss SL-DAIS. In Sec. A.7 we provide details about our experimental setup. In Sec. A.8 we present additional figures and tables that accompany those in the main text.

A.1. Local Latent Variable Models

Another important class of models includes *local* latent variables $\{\mathbf{w}_n\}$ in addition to a global latent variable \mathbf{z} , i.e. models with joint densities of the form:

$$p_\theta(\mathcal{D}, \mathbf{W}, \mathbf{z}) = p_\theta(\mathbf{z}) \prod_{n=1}^N p_\theta(\mathbf{w}_n | \mathbf{z}) p_\theta(y_n | \mathbf{w}_n, \mathbf{z}, \mathbf{x}_n) \quad (16)$$

A viable DAIS-like inference strategy for this class of models that would be scalable to large N is to adopt a *semi-parametric* approach. First, we introduce a flexible, parametric variational distribution $q_\phi(\mathbf{z})$ using, for example, a normalizing flow. After conditioning on a sample $\mathbf{z} \sim q_\phi(\mathbf{z})$ the N inference problems w.r.t. $\{\mathbf{w}_n\}$ effectively decouple. Consequently we can apply DAIS to each subproblem, resulting in an ELBO that accommodates unbiased mini-batch estimates.

In more detail we proceed as follows. Introduce a parametric variational distribution $q_\phi(\mathbf{z})$ and a mean-field distribution $q_\phi(\mathbf{W})$ that factorizes as $q_\phi(\mathbf{W}) = \prod_n q_\phi(\mathbf{w}_n)$. Then write (in this section we suppress momenta terms in the MCMC kernels for brevity)

$$\begin{aligned} q_{\text{fwd}}(\mathbf{z}, \mathbf{W}_{0:K}) &= q_\phi(\mathbf{z}) \mathcal{T}_1(\mathbf{W}_1 | \mathbf{W}_0, \mathbf{z}) \cdots \mathcal{T}_K(\mathbf{W}_K | \mathbf{W}_{K-1}, \mathbf{z}) \\ q_{\text{bwd}}(\mathbf{z}, \mathbf{w}_{0:K}) &= p_\theta(\mathcal{D}, \mathbf{W}_K, \mathbf{z}) \tilde{\mathcal{T}}_K(\mathbf{W}_{K-1} | \mathbf{W}_K, \mathbf{z}) \cdots \tilde{\mathcal{T}}_1(\mathbf{W}_0 | \mathbf{W}_1, \mathbf{z}) \\ \mathcal{L} &= \mathbb{E}_{q_{\text{fwd}}} [\log q_{\text{bwd}} - \log q_{\text{fwd}}] \end{aligned} \quad (17)$$

where the forward MCMC kernels target a log density of the form (here for $\beta = 1$)

$$\sum_{n=1}^N (\log p_\theta(\mathbf{w}_n | \mathbf{z}) + \log p_\theta(y_n | \mathbf{w}_n, \mathbf{z}, \mathbf{x}_n)) \quad (18)$$

where \mathbf{z} is kept fixed throughout HMC dynamics. Since this density splits up into a sum over N turns, we actually have N independent DAIS chains (assuming our mass matrix is appropriately factorized). Since the $\log p_\theta(\mathcal{D}, \mathbf{W}_K, \mathbf{z})$ term that enters into \mathcal{L} also splits up into such a sum, this ELBO admits mini-batch sampling. That is, we choose a mini-batch of size $B \ll N$ and run B DAIS chains forward at each optimization step so that e.g. most $\{\mathbf{w}_n\}$ are not instantiated. This of course is exactly how subsampling works with vanilla parametric mean-field variational distributions.

We can follow the same basic strategy to make this variational distribution more fully DAIS-like *while still supporting subsampling*. The basic idea is that we will have $N + 1$ DAIS chains. The N DAIS chains for each \mathbf{w}_n will be as above. But we will also introduce a DAIS chain that is responsible for producing the \mathbf{z} sample. To enable subsampling this \mathbf{z} -chain must target a surrogate likelihood.

In more detail we proceed as follows. To simplify the equations we assume $K = 1$.

$$\begin{aligned} q_{\text{fwd}}(\tilde{\mathbf{z}}_{0:1}, \tilde{\mathbf{W}}_{0:1}, \mathbf{W}_{0:1}) &= q_\phi(\tilde{\mathbf{z}}_0, \tilde{\mathbf{W}}_0) q_\phi(\mathbf{W}_0) \mathcal{T}_{\text{surr}}(\tilde{\mathbf{z}}_1, \tilde{\mathbf{W}}_1 | \tilde{\mathbf{z}}_0, \tilde{\mathbf{W}}_0) \mathcal{T}(\mathbf{W}_1 | \mathbf{W}_0, \tilde{\mathbf{z}}_1) \\ q_{\text{bwd}}(\tilde{\mathbf{z}}_{0:1}, \tilde{\mathbf{W}}_{0:1}, \mathbf{W}_{0:1}) &= p_\theta(\mathcal{D}, \mathbf{W}_1, \tilde{\mathbf{z}}_1) \tilde{\mathcal{T}}_{\text{surr}}(\tilde{\mathbf{z}}_0, \tilde{\mathbf{W}}_0 | \tilde{\mathbf{z}}_1, \tilde{\mathbf{W}}_1) \tilde{\mathcal{T}}(\mathbf{W}_0 | \mathbf{W}_1, \tilde{\mathbf{z}}_1) \\ \mathcal{L} &= \mathbb{E}_{q_{\text{fwd}}} [\log q_{\text{bwd}} - \log q_{\text{fwd}}] \end{aligned} \quad (19)$$

Here $q_\phi(\tilde{\mathbf{z}}_0, \tilde{\mathbf{W}}_0)$ and $q_\phi(\mathbf{W}_0)$ are (distinct) simple parametric variational distributions. $\mathcal{T}_{\text{surr}}$ is a DAIS chain that targets a surrogate likelihood with learnable weights that includes N_{surr} data points and also N_{surr} local latent variables $\tilde{\mathbf{W}}$. As above $\mathcal{T}(\mathbf{W}_1 | \mathbf{W}_0, \tilde{\mathbf{z}}_1)$ consists of N independent DAIS chains, all of which are conditioned on $\tilde{\mathbf{z}}_1$. Note that since these N DAIS chains factorize there is no need for a surrogate here. We note that the output of the $N + 1$ DAIS chains is $\tilde{\mathbf{z}}_1$, $\tilde{\mathbf{W}}_1$ and \mathbf{W}_1 . $\tilde{\mathbf{W}}_1$ is a N_{surr} -dimensional ‘auxiliary variable’ and so we essentially throw it out (nothing depends on it directly)—its

only purpose is to ‘integrate out’ \mathbf{W} -uncertainty in the surrogate likelihood.⁹ By contrast \mathbf{W}_1 is the ‘actual’ \mathbf{W} sample that enters into $p_\theta(\mathcal{D}, \mathbf{W}_1, \tilde{\mathbf{z}}_1)$, i.e. our approximate posterior sample is $(\mathbf{W}_1, \tilde{\mathbf{z}}_1)$. As above we can still subsample the N DAIS chains and get an ELBO that supports data subsampling. Basically we’ve replaced the parametric distribution for \mathbf{z} above with a single DAIS chain that targets a surrogate likelihood.

Yet another (somewhat simpler) variant of this approach would forego $\tilde{\mathbf{W}}$ transitions in building up the \mathbf{z} sample and instead parameterize the surrogate likelihood using a learnable point parameter $\tilde{\mathbf{W}}$. As should now be evident, the space of models with mixed local/global latent variables opens up lots of possibilities for DAIS-like inference procedures. For this reason we leave a detailed empirical exploration of the algorithms described above to future work.

A.2. Related Work (extended)

In this section we provide an extended discussion of related work. Many variational objectives that leverage importance sampling (IS) have been proposed. These include the importance weighted autoencoder (IWAE) (Burda et al., 2015; Cremer et al., 2017), the thermodynamic variational objective (Masrani et al., 2019), and approaches that make use of Sequential Monte Carlo (Le et al., 2017; Maddison et al., 2017; Naesseth et al., 2018). For a general discussion of IS in variational inference see Domke & Sheldon (2018).

As discussed in Sec. 3 in the main text our work builds on recent work, namely (UHA; Geffner & Domke (2021)) and (DAIS; Zhang et al. (2021)), which can be understood as utilizing unadjusted *underdamped* Langevin steps. A closely related approach that utilizes unadjusted *overdamped* Langevin steps is described in (Thin et al., 2021). We note that earlier attempts to incorporate AIS into a variational framework like Ding & Freedman (2019) do not benefit from a single, unified ELBO-based objective.

An early combination of MCMC methods with variational inference was proposed by Salimans et al. (2015) and Wolf et al. (2016). A disadvantage of these approaches is the need to learn reverse kernels, a shortcoming that was later addressed by Caterini et al. (2018). Additional work that explores combinations of MCMC and variational methods includes Li et al. (2017); Hoffman (2017); Ruiz & Titsias (2019)

Bayesian coresets enable users to run MCMC on large datasets after first distilling the data into a smaller number of weighted data points (Huggins et al., 2016; Bachem et al., 2017; Campbell & Broderick, 2018; 2019). These methods do not support model learning, since they rely on a two-stage approach (i.e. coreset learning and inference are done separately). Finally a number of authors have explored MCMC methods that enable data subsampling (Maclaurin & Adams, 2015; Quiroz et al., 2018; Dang et al., 2019; Zhang et al., 2020), including those that leverage stochastic gradients (Welling & Teh, 2011; Chen et al., 2014; Ma et al., 2015). Finally we note that learning flexible variational distributions parameterized by neural networks, i.e. normalizing flows, is another popular black-box approach for improving variational approximations (Rezende & Mohamed, 2015).

A.3. Computing $\mathcal{L}_{\text{DAIS}}$

We provide a proof of Lemma 1 in the main text. We reproduce the argument from Zhang et al. (2021); see Geffner & Domke (2021) for a similar derivation. The forward kernel \mathcal{T}_k in Eqn. 9 can be decomposed as

$$\mathcal{T}_k(\mathbf{z}_k, \mathbf{v}_k | \mathbf{z}_{k-1}, \mathbf{v}_{k-1}) = \mathcal{T}_k^{\text{leap}}(\mathbf{z}_k, \hat{\mathbf{v}}_k | \mathbf{z}_{k-1}, \mathbf{v}_{k-1}) \mathcal{T}_k^{\text{refresh}}(\mathbf{v}_k | \hat{\mathbf{v}}_k) \quad (20)$$

Similarly the backward kernel $\tilde{\mathcal{T}}_k$ in Eqn. 9 can be decomposed as

$$\tilde{\mathcal{T}}_k(\mathbf{z}_{k-1}, \mathbf{v}_{k-1} | \mathbf{z}_k, \mathbf{v}_k) = \tilde{\mathcal{T}}_k^{\text{refresh}}(\hat{\mathbf{v}}_k | \mathbf{v}_k) \tilde{\mathcal{T}}_k^{\text{leap}}(\mathbf{z}_{k-1}, \mathbf{v}_{k-1} | \mathbf{z}_k, -\hat{\mathbf{v}}_k) \quad (21)$$

where we flip the sign of $\hat{\mathbf{v}}_k$ to account for time reversal. Since we have

$$\mathcal{T}_k^{\text{refresh}}(\mathbf{v}_k | \hat{\mathbf{v}}_k) = \mathcal{N}(\mathbf{v}_k | \gamma \hat{\mathbf{v}}_k, (1 - \gamma^2)\mathbf{M}) \quad \tilde{\mathcal{T}}_k^{\text{refresh}}(\hat{\mathbf{v}}_k | \mathbf{v}_k) = \mathcal{N}(\hat{\mathbf{v}}_k | \gamma \mathbf{v}_k, (1 - \gamma^2)\mathbf{M}) \quad (22)$$

and

$$\|\mathbf{v}_k - \gamma \hat{\mathbf{v}}_k\|^2 - \|\hat{\mathbf{v}}_k - \gamma \mathbf{v}_k\|^2 = (1 - \gamma^2) (\|\mathbf{v}_k\|^2 - \|\hat{\mathbf{v}}_k\|^2) \quad (23)$$

⁹For this reason we will probably need to pretrain $q_\phi(\tilde{\mathbf{z}}_0, \tilde{\mathbf{W}}_0)$ —perhaps introducing an auxiliary loss function—since $\tilde{\mathbf{W}}_1$ will have a weak learning signal.

it follows that

$$\frac{\mathcal{T}_k^{\text{refresh}}(\mathbf{v}_k|\hat{\mathbf{v}}_k)}{\tilde{\mathcal{T}}_k^{\text{refresh}}(\hat{\mathbf{v}}_k|\mathbf{v}_k)} = \frac{\mathcal{N}(\mathbf{v}_k|\mathbf{0}, \mathbf{M})}{\mathcal{N}(\hat{\mathbf{v}}_k|\mathbf{0}, \mathbf{M})} \quad (24)$$

Furthermore, since the leapfrog step is volume preserving and reversible we also have that

$$\mathcal{T}_k^{\text{leap}}(\mathbf{z}_k, \hat{\mathbf{v}}_k|\mathbf{z}_{k-1}, \mathbf{v}_{k-1}) = \tilde{\mathcal{T}}_k^{\text{leap}}(\mathbf{z}_{k-1}, \mathbf{v}_{k-1}|\mathbf{z}_k, -\hat{\mathbf{v}}_k) \quad (25)$$

From these equations the only tricky part of Lemma 1, i.e. the derivation of the kinetic energy difference correction, follows. Importantly, Eqn. 25 is true regardless of the target density used to define $\mathcal{T}_k^{\text{leap}}$ and $\tilde{\mathcal{T}}_k^{\text{leap}}$; in particular using a surrogate log likelihood leaves Lemma 1 intact.

A.4. General properties of NS-DAIS and SL-DAIS

We prove Proposition 1 in the main text. In particular we show that the NS-DAIS and SL-DAIS approximate posteriors $q_{\text{fwd}}(\mathbf{z}_K)$ satisfy the inequality

$$\log p_\theta(\mathcal{D}) \geq \mathcal{L} + \text{KL}(q_{\text{fwd}}(\mathbf{z}_K)|p_\theta(\mathbf{z}_K|\mathcal{D})) \quad (26)$$

where \mathcal{L} is $\mathcal{L}_{\text{NS-DAIS}}$ or $\mathcal{L}_{\text{SL-DAIS}}$. First consider SL-DAIS and let $\tilde{q}_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K})$ be the normalized distribution corresponding to $q_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K})$ so that $\log q_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) = \log p_\theta(\mathcal{D}) + \log \tilde{q}_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K})$ and write

$$\log p_\theta(\mathcal{D}) = \mathbb{E}_{q_{\text{fwd}}} \left[\log q_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) - \log q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) \right] + \quad (27)$$

$$\begin{aligned} & \text{KL}(q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K})|\tilde{q}_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K})) \\ &= \mathcal{L}_{\text{SL-DAIS}} + \text{KL}(q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K})|\tilde{q}_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K})) \end{aligned} \quad (28)$$

$$\begin{aligned} &= \mathcal{L}_{\text{SL-DAIS}} + \text{KL}(q_{\text{fwd}}(\mathbf{z}_K)|p_\theta(\mathbf{z}_K|\mathcal{D})) + \\ & \quad \mathbb{E}_{q_{\text{fwd}}(\mathbf{z}_K)} [\text{KL}(q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}|\mathbf{z}_K)|\tilde{q}_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}|\mathbf{z}_K))] \end{aligned} \quad (29)$$

where we appealed to the chain rule of KL divergences which reads

$$\text{KL}(q(a, b)|p(a, b)) = \text{KL}(q(a)|p(a)) + \mathbb{E}_{q(b)} [\text{KL}(q(a|b)|p(a|b))] \quad (30)$$

The result then follows since the second KL divergence in Eqn. 29 is non-negative. The result for NS-DAIS follows by the same argument, with the difference that \mathcal{J} is now one of the ‘auxiliary’ variables $\{\mathbf{z}_{0:K}, \mathbf{v}_{0:K}\}$. Note that the \mathcal{I} index plays no role in this proof, since its role is to facilitate unbiased mini-batch estimates, i.e. it has no impact on the value of $\mathcal{L}_{\text{NS-DAIS}}$ or $\mathcal{L}_{\text{SL-DAIS}}$ (see the next section for details).

A.5. NS-DAIS

We first elaborate how NS-DAIS is formulated. The forward and backward kernels are defined as:

$$\begin{aligned} q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{I}, \mathcal{J}) &= q(\mathcal{I})q(\mathcal{J})q_0(\mathbf{z}_0)q_0(\mathbf{v}_0)\prod_{k=1}^K \mathcal{T}_k(\mathbf{z}_k, \mathbf{v}_k|\mathbf{z}_{k-1}, \mathbf{v}_{k-1}, \mathcal{J}) \\ q_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{I}, \mathcal{J}) &= q(\mathcal{I})q(\mathcal{J})p_\theta(\mathbf{z}_K|\mathcal{D})^{N/B} p_\theta(\mathbf{z}_K)\prod_{k=1}^K \tilde{\mathcal{T}}_k(\mathbf{z}_{k-1}, \mathbf{v}_{k-1}|\mathbf{z}_k, \mathbf{v}_k, \mathcal{J}) \end{aligned}$$

where \mathcal{J} controls the potential energy used to guide the HMC dynamics and the corresponding ELBO is given by

$$\begin{aligned} \mathcal{L}_{\text{NS-DAIS}} &\equiv \mathbb{E}_{q_{\text{fwd}}} \left[\log q_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{I}, \mathcal{J}) - \log q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{I}, \mathcal{J}) \right] \\ &= \mathbb{E}_{q'_{\text{fwd}}} \left[\log q'_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{J}) - \log q'_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{J}) \right] \end{aligned} \quad (31)$$

where

$$\begin{aligned} q'_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{J}) &= q(\mathcal{J})q_0(\mathbf{z}_0)q_0(\mathbf{v}_0)\prod_{k=1}^K \mathcal{T}_k(\mathbf{z}_k, \mathbf{v}_k|\mathbf{z}_{k-1}, \mathbf{v}_{k-1}, \mathcal{J}) \\ q'_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{J}) &= q(\mathcal{J})p_\theta(\mathbf{z}_K|\mathcal{D})p_\theta(\mathbf{z}_K)\prod_{k=1}^K \tilde{\mathcal{T}}_k(\mathbf{z}_{k-1}, \mathbf{v}_{k-1}|\mathbf{z}_k, \mathbf{v}_k, \mathcal{J}) \end{aligned} \quad (32)$$

Algorithm 2 NS-DAIS: Naive Subsampling Differentiable Annealed Importance Sampling. We highlight in blue where the algorithm differs from DAIS. To recover DAIS we substitute $\frac{N}{B}\Psi_L(\mathcal{D}_{\mathcal{J}}, \mathbf{z}) \rightarrow \Psi_L(\mathcal{D}, \mathbf{z})$, $\frac{N}{B}\Psi_L(\mathcal{D}_{\mathcal{I}}, \mathbf{z}_K) \rightarrow \Psi_L(\mathcal{D}, \mathbf{z}_K)$, and $B \rightarrow N$ (or in other words remove all mini-batch sampling).

Input: model log prior density $\Psi_0(\mathbf{z})$, model log likelihood $\Psi_L(\mathcal{D}, \mathbf{z})$, base variational distribution $q_0(\mathbf{z})$, number of steps K , inverse temperatures $\{\beta_k\}$, step size η , momentum refresh parameter γ , mass matrix \mathbf{M} , dataset \mathcal{D} of size N , mini-batch size B

Initialize: $\mathbf{z}_0 \sim q_0$, $\mathbf{v}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$, $\mathcal{L} \leftarrow -\log q_0(\mathbf{z}_0)$

Sample mini-batch indices $\mathcal{J} \subset \{1, \dots, N\}$ with $|\mathcal{J}| = B$

for $k = 1$ **to** K **do**

$$\hat{\mathbf{z}}_k \leftarrow \mathbf{z}_{k-1} + \frac{\eta}{2}\mathbf{M}^{-1}\mathbf{v}_{k-1}$$

$$\mathbf{g}_k \leftarrow \nabla_{\mathbf{z}} \left\{ \beta_k \left(\Psi_0(\mathbf{z}) + \frac{N}{B}\Psi_L(\mathcal{D}_{\mathcal{J}}, \mathbf{z}) \right) + (1 - \beta_k) \log q_0(\mathbf{z}) \right\} \Big|_{\mathbf{z}=\hat{\mathbf{z}}_k}$$

$$\hat{\mathbf{v}}_k \leftarrow \mathbf{v}_{k-1} + \eta\mathbf{g}_k$$

$$\mathbf{z}_k \leftarrow \hat{\mathbf{z}}_k + \frac{\eta}{2}\mathbf{M}^{-1}\hat{\mathbf{v}}_k$$

if $k < K$ **then**

$$\mathbf{v}_k \leftarrow \gamma\hat{\mathbf{v}}_k + \sqrt{1 - \gamma^2}\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$$

end if

$$\mathcal{L} \leftarrow \mathcal{L} + \log \mathcal{N}(\hat{\mathbf{v}}_k, \mathbf{M}) - \log \mathcal{N}(\mathbf{v}_{k-1}, \mathbf{M})$$

end for

Sample mini-batch indices $\mathcal{I} \subset \{1, \dots, N\}$ with $|\mathcal{I}| = B$

Return: $\mathcal{L} + \Psi_0(\mathbf{z}_K) + \frac{N}{B}\Psi_L(\mathcal{D}_{\mathcal{I}}, \mathbf{z}_K)$

are the forward and backward kernels without the \mathcal{I} mini-batch index.¹⁰ Note that here and elsewhere $q_0(\mathbf{v}_0) = \mathcal{N}(\mathbf{v}_0|\mathbf{0}, \mathbf{M})$ is the momentum distribution. To be more specific, for each \mathcal{J} the potential energy that guides each \mathcal{T}_k is given by

$$V_k(\mathbf{z}|\mathcal{J}) = -(1 - \beta_k) \log q_0(\mathbf{z}) - \beta_k \left(\Psi_0(\mathbf{z}) + \frac{N}{B}\Psi_L(\mathcal{D}_{\mathcal{J}}, \mathbf{z}) \right) \quad (33)$$

See Algorithm 2 for the complete procedure and see Algorithm 3 to make a direct comparison to DAIS.

Now we prove Proposition 2 in the main text. As in the main text let $q(\mathcal{J})$ denote the distribution that corresponds to sampling mini-batches of B indices $\mathcal{J} \subset \{1, \dots, N\}$ without replacement. Let $p_{\theta}(\mathbf{z}|\mathcal{D}_{\mathcal{J}})$ denote the posterior that corresponds to the data subset $\mathcal{D}_{\mathcal{J}}$ with appropriately scaled likelihood term, i.e.

$$p_{\theta}(\mathbf{z}|\mathcal{D}_{\mathcal{J}}) \propto p_{\theta}(\mathcal{D}_{\mathcal{J}}|\mathbf{z})^{N/B} p_{\theta}(\mathbf{z}) \quad (34)$$

Further denote the ‘aggregate pseudo-posterior’ by

$$p_{\theta}^{\text{Agg}}(\mathbf{z}|\mathcal{D}) \equiv \mathbb{E}_{q(\mathcal{J})} [p_{\theta}(\mathbf{z}|\mathcal{D}_{\mathcal{J}})] \quad (35)$$

We have (appealing to the convexity of the KL divergence)

$$\text{KL}(q_{\text{fwd}}(\mathbf{z}_K)|p_{\theta}^{\text{Agg}}(\mathbf{z}|\mathcal{D})) = \text{KL}(\mathbb{E}_{q(\mathcal{J})} [q_{\text{fwd}}(\mathbf{z}_K|\mathcal{J})] | \mathbb{E}_{q(\mathcal{J})} [p_{\theta}(\mathbf{z}_K|\mathcal{D}_{\mathcal{J}})]) \quad (36)$$

$$\leq \mathbb{E}_{q(\mathcal{J})} [\text{KL}(q_{\text{fwd}}(\mathbf{z}_K|\mathcal{J})|p_{\theta}(\mathbf{z}_K|\mathcal{D}_{\mathcal{J}}))] \quad (37)$$

$$= \mathcal{O}(K^{-1/2}) \quad (38)$$

where in the last line we used that $q(\mathcal{J})$ has finite support and where we have appealed to the convergence result in Zhang et al. (2021) to conclude that for each \mathcal{J} we have that $\text{KL}(q_{\text{fwd}}(\mathbf{z}_K|\mathcal{J})|p_{\theta}(\mathbf{z}_K|\mathcal{D}_{\mathcal{J}})) = \mathcal{O}(K^{-1/2})$. We note that, as is well known, convergence in KL divergence implies convergence w.r.t. the total variation distance.

¹⁰Note that these are the kernels that were used in the proof in Sec. A.4.

Algorithm 3 DAIS: Differentiable Annealed Importance Sampling. We provide a complete description of DAIS (Geffner & Domke, 2021; Zhang et al., 2021) in our notation.

Input: model log prior density $\Psi_0(\mathbf{z})$, model log likelihood $\Psi_L(\mathcal{D}, \mathbf{z})$, base variational distribution $q_0(\mathbf{z})$, number of steps K , inverse temperatures $\{\beta_k\}$, step size η , momentum refresh parameter γ , mass matrix \mathbf{M} ,

Initialize: $\mathbf{z}_0 \sim q_0$, $\mathbf{v}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$, $\mathcal{L} \leftarrow -\log q_0(\mathbf{z}_0)$

for $k = 1$ **to** K **do**

$$\hat{\mathbf{z}}_k \leftarrow \mathbf{z}_{k-1} + \frac{\eta}{2} \mathbf{M}^{-1} \mathbf{v}_{k-1}$$

$$\mathbf{g}_k \leftarrow \nabla_{\mathbf{z}} \{ \beta_k (\Psi_0(\mathbf{z}) + \Psi_L(\mathcal{D}, \mathbf{z})) + (1 - \beta_k) \log q_0(\mathbf{z}) \} \Big|_{\mathbf{z}=\hat{\mathbf{z}}_k}$$

$$\hat{\mathbf{v}}_k \leftarrow \mathbf{v}_{k-1} + \eta \mathbf{g}_k$$

$$\mathbf{z}_k \leftarrow \hat{\mathbf{z}}_k + \frac{\eta}{2} \mathbf{M}^{-1} \hat{\mathbf{v}}_k$$

if $k < K$ **then**

$$\mathbf{v}_k \leftarrow \gamma \hat{\mathbf{v}}_k + \sqrt{1 - \gamma^2} \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{M})$$

end if

$$\mathcal{L} \leftarrow \mathcal{L} + \log \mathcal{N}(\hat{\mathbf{v}}_k, \mathbf{M}) - \log \mathcal{N}(\mathbf{v}_{k-1}, \mathbf{M})$$

end for

Return: $\mathcal{L} + \Psi_0(\mathbf{z}_K) + \Psi_L(\mathcal{D}, \mathbf{z}_K)$

A.6. SL-DAIS

Before we turn to a proof of Proposition 3 in the main text, we give a more formal description of SL-DAIS. The forward and backward kernels are defined as:

$$\begin{aligned} q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{I}) &= q(\mathcal{I}) q_0(\mathbf{z}_0) q_0(\mathbf{v}_0) \prod_{k=1}^K \mathcal{T}_k(\mathbf{z}_k, \mathbf{v}_k | \mathbf{z}_{k-1}, \mathbf{v}_{k-1}, \hat{\Psi}_L) \\ q_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{I}) &= q(\mathcal{I}) p_\theta(\mathbf{z}_K | \mathcal{D})^{N/B} p_\theta(\mathbf{z}_K) \prod_{k=1}^K \tilde{\mathcal{T}}_k(\mathbf{z}_{k-1}, \mathbf{v}_{k-1} | \mathbf{z}_k, \mathbf{v}_k, \hat{\Psi}_L) \end{aligned}$$

and the corresponding ELBO is given by

$$\begin{aligned} \mathcal{L}_{\text{SL-DAIS}} &\equiv \mathbb{E}_{q_{\text{fwd}}} \left[\log q_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{I}) - \log q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}, \mathcal{I}) \right] \\ &= \mathbb{E}_{q'_{\text{fwd}}} \left[\log q'_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) - \log q_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) \right] \end{aligned} \quad (39)$$

where

$$\begin{aligned} q'_{\text{fwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) &= q_0(\mathbf{z}_0) q_0(\mathbf{v}_0) \prod_{k=1}^K \mathcal{T}_k(\mathbf{z}_k, \mathbf{v}_k | \mathbf{z}_{k-1}, \mathbf{v}_{k-1}, \hat{\Psi}_L) \\ q'_{\text{bwd}}(\mathbf{z}_{0:K}, \mathbf{v}_{0:K}) &= p_\theta(\mathbf{z}_K | \mathcal{D}) p_\theta(\mathbf{z}_K) \prod_{k=1}^K \tilde{\mathcal{T}}_k(\mathbf{z}_{k-1}, \mathbf{v}_{k-1} | \mathbf{z}_k, \mathbf{v}_k, \hat{\Psi}_L) \end{aligned}$$

are the forward and backward kernels without the \mathcal{I} mini-batch index.

We now provide a proof of Proposition 3 in the main text. First we restate some of the equations and notation from the main text. We consider Bayesian linear regression with a prior $p_\theta(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1})$ and a likelihood $\prod_n \mathcal{N}(y_n | \mathbf{z} \cdot \mathbf{x}_n, \sigma_{\text{obs}}^2)$, where each $y_n \in \mathbb{R}$ and $\mathbf{x}_n \in \mathbb{R}^D$. The exact posterior is given by $\mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{\Lambda}_{\text{post}}^{-1})$ where $\boldsymbol{\mu}_{\text{post}} = \boldsymbol{\Lambda}_{\text{post}}^{-1} (\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \frac{1}{\sigma_{\text{obs}}^2} \mathbf{X}^\top \mathbf{y})$ and $\boldsymbol{\Lambda}_{\text{post}} = \boldsymbol{\Lambda}_0 + \frac{1}{\sigma_{\text{obs}}^2} \mathbf{X}^\top \mathbf{X}$. Throughout we work under the following set of simplifying assumptions:

Assumption 1 We use full momentum refreshment ($\gamma = 0$), equally spaced inverse temperatures $\{\beta_k\}$, a step size that varies as $\eta \sim K^{-1/4}$, and the prior as the base distribution (i.e. $q_0(\mathbf{z}) \rightarrow p_\theta(\mathbf{z})$).

Next we define the parameters of the ‘surrogate posterior’ $\mathcal{N}(\tilde{\boldsymbol{\mu}}_{\text{post}}, \tilde{\boldsymbol{\Lambda}}_{\text{post}})$ that corresponds to the surrogate log likelihood $\hat{\Psi}_L(\mathbf{z})$ specified by $\delta \mathbf{a}$ and $\delta \mathbf{B}$, where $\delta \mathbf{a}$ and $\delta \mathbf{B}$ control the degree to which $\hat{\Psi}_L(\mathbf{z})$ is incorrectly specified. Given our parameterization

$$\nabla_{\mathbf{z}} \hat{\Psi}_L(\mathbf{z}) = \mathbf{a} + \delta \mathbf{a} - (\mathbf{B} + \delta \mathbf{B}) \mathbf{z} \quad (40)$$

it is easy to verify that

$$\begin{aligned} \tilde{\boldsymbol{\Lambda}}_{\text{post}} &= \boldsymbol{\Lambda}_0 + \mathbf{B} + \delta \mathbf{B} = \boldsymbol{\Lambda}_{\text{post}} + \delta \mathbf{B} \\ \tilde{\boldsymbol{\mu}}_{\text{post}} &= \tilde{\boldsymbol{\Lambda}}_{\text{post}}^{-1} (\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \mathbf{a} + \delta \mathbf{a}) = \tilde{\boldsymbol{\Lambda}}_{\text{post}}^{-1} (\boldsymbol{\Lambda}_{\text{post}} \boldsymbol{\mu}_{\text{post}} + \delta \mathbf{a}) \end{aligned} \quad (41)$$

Expanding these expressions in powers of $\delta\mathbf{a}$ and $\delta\mathbf{B}$ we have:

$$\begin{aligned}\tilde{\Lambda}_{\text{post}}^{-1} &\approx \Lambda_{\text{post}}^{-1} - \Lambda_{\text{post}}^{-1} \delta\mathbf{B} \Lambda_{\text{post}}^{-1} \\ \tilde{\boldsymbol{\mu}}_{\text{post}} &\approx \boldsymbol{\mu}_{\text{post}} + \Lambda_{\text{post}}^{-1} \delta\mathbf{a} - \Lambda_{\text{post}}^{-1} \delta\mathbf{B} \boldsymbol{\mu}_{\text{post}}\end{aligned}\quad (42)$$

These expressions also imply that

$$\det \tilde{\Lambda}_{\text{post}} = \det \Lambda_{\text{post}} \det (\mathbb{1}_D - \Lambda_{\text{post}}^{-1} \delta\mathbf{B}) \approx \det \Lambda_{\text{post}} (1 + \text{Tr} \Lambda_{\text{post}}^{-1} \delta\mathbf{B}) \quad (43)$$

which yields

$$\log |\Lambda_{\text{post}}^{-1} \tilde{\Lambda}_{\text{post}}| \approx \text{Tr} \Lambda_{\text{post}}^{-1} \delta\mathbf{B}. \quad (44)$$

Next we restate some results from Sec. 4.1 and the supplementary materials in Zhang et al. (2021), suitably adapting them to our setting and notation. Where appropriate we use subscripts/superscripts to distinguish DAIS and SL-DAIS terms.

Lemma 2 *In the case of Bayesian linear regression under Assumption 1 each of the $K + 1$ random variables $\{\mathbf{z}_k\}$ that enter DAIS are normally distributed as $\mathbf{z}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Lambda_k^{-1})$. If we run DAIS for K steps the ELBO gap is given by*

$$\begin{aligned}\log p(\mathcal{D}) - \mathcal{L}_{\text{DAIS}} = & \\ & \underbrace{\frac{1}{2} \|\boldsymbol{\mu}_K - \boldsymbol{\mu}_{\text{post}}\|_{\Lambda_{\text{post}}}^2}_{\text{(I)}} + \underbrace{\frac{1}{2} \text{Tr}(\Lambda_{\text{post}} \Lambda_K^{-1}) - \frac{D}{2}}_{\text{(II)}} + \underbrace{\frac{1}{2} \log |\Lambda_{\text{post}}^{-1} \Lambda_0| - \mathbb{E}_{q_{\text{fwd}}} \left[\sum_{k=1}^K \log \frac{\mathcal{N}(\hat{\mathbf{v}}_k, \mathbf{M})}{\mathcal{N}(\mathbf{v}_{k-1}, \mathbf{M})} \right]}_{\text{(III)}}\end{aligned}\quad (45)$$

where $\|\boldsymbol{\mu}_K - \boldsymbol{\mu}_{\text{post}}\|_{\Lambda_{\text{post}}}^2 = (\boldsymbol{\mu}_K - \boldsymbol{\mu}_{\text{post}})^{\text{T}} \Lambda_{\text{post}} (\boldsymbol{\mu}_K - \boldsymbol{\mu}_{\text{post}})$. Moreover, as is evident from the derivation in Sec. B.2 in Zhang et al. (2021), this equation also holds for the gap $\log p(\mathcal{D}) - \mathcal{L}_{\text{SL-DAIS}}$ if we substitute the SL-DAIS forward kernel q_{fwd} into the expectation in (III) and $\boldsymbol{\mu}_k$ and Λ_k are defined in terms of the $\{\mathbf{z}_k\}$ that enter SL-DAIS. Additionally for SL-DAIS¹¹ we have that

$$\begin{aligned}\|\boldsymbol{\mu}_K^{\text{SL-DAIS}} - \tilde{\boldsymbol{\mu}}_{\text{post}}\| &= \mathcal{O}(K^{-1/2}) & \|(\Lambda_K^{\text{SL-DAIS}})^{-1} - \tilde{\Lambda}_{\text{post}}^{-1}\| &= \mathcal{O}(K^{-1/2}) \\ \frac{1}{2} \log |\tilde{\Lambda}_{\text{post}}^{-1} \Lambda_0| - \mathbb{E}_{q_{\text{fwd}}} \left[\sum_{k=1}^K \log \frac{\mathcal{N}(\hat{\mathbf{v}}_k, \mathbf{M})}{\mathcal{N}(\mathbf{v}_{k-1}, \mathbf{M})} \right] &= \mathcal{O}(K^{-1/2})\end{aligned}\quad (46)$$

With these ingredients we can now proceed with the proof. From Eqn. 42 and Eqn. 46 we have

$$\|\boldsymbol{\mu}_K^{\text{SL-DAIS}} - \boldsymbol{\mu}_{\text{post}}\| \leq \|\boldsymbol{\mu}_K^{\text{SL-DAIS}} - \tilde{\boldsymbol{\mu}}_{\text{post}}\| + \|\tilde{\boldsymbol{\mu}}_{\text{post}} - \boldsymbol{\mu}_{\text{post}}\| = \mathcal{O}(K^{-1/2}) + \mathcal{O}(\|\delta\mathbf{a}\|) + \mathcal{O}(\|\delta\mathbf{B}\|). \quad (47)$$

Hence for SL-DAIS we have that

$$\text{(I)} = \mathcal{O}(K^{-1/2}) + \mathcal{O}(\|\delta\mathbf{a}\|^2 + \|\delta\mathbf{B}\|^2). \quad (48)$$

Once again appealing to Eqn. 42 and Eqn. 46 we have

$$\begin{aligned}\text{Tr}(\Lambda_{\text{post}} (\Lambda_K^{\text{SL-DAIS}})^{-1}) - D &= \text{Tr}(\Lambda_{\text{post}} ((\Lambda_K^{\text{SL-DAIS}})^{-1} - \Lambda_{\text{post}}^{-1})) \\ &= \text{Tr}(\Lambda_{\text{post}} ((\Lambda_K^{\text{SL-DAIS}})^{-1} - \tilde{\Lambda}_{\text{post}}^{-1})) + \text{Tr}(\Lambda_{\text{post}} (\tilde{\Lambda}_{\text{post}}^{-1} - \Lambda_{\text{post}}^{-1})) \\ &\approx \mathcal{O}(K^{-1/2}) + \text{Tr} \Lambda_{\text{post}}^{-1} \delta\mathbf{B}\end{aligned}\quad (49)$$

so that we get the following estimate for the second term of Eqn. 45:

$$\text{(II)} = \mathcal{O}(K^{-1/2}) + \frac{1}{2} \text{Tr} \Lambda_{\text{post}}^{-1} \delta\mathbf{B} \quad (50)$$

Finally, from Eqn. 44 and Eqn. 46 we have

$$\text{(III)} = \frac{1}{2} \log |\Lambda_{\text{post}}^{-1} \tilde{\Lambda}_{\text{post}}| + \mathcal{O}(K^{-1/2}) = \frac{1}{2} \text{Tr} \Lambda_{\text{post}}^{-1} \delta\mathbf{B} + \mathcal{O}(K^{-1/2}) \quad (51)$$

which ends the first part of the proof for Proposition 3. The fact that the KL divergence between $q_{\text{fwd}}(\mathbf{z}_K)$ and the posterior $p_{\theta}(\mathbf{z}_K | \mathcal{D})$ is bounded by the same quantity easily follows from standard arguments, see e.g. the proof of Proposition 1 in Sec. A.4.

¹¹This is evident from the proof of Lemma 1 in Zhang et al. (2021). The analogous result is also true for DAIS but we do not require this for the proof.

A.7. Experimental details

All experiments use 64-bit floating point precision. We use the Adam optimizer with default momentum hyperparameters in all experiments (Kingma & Ba, 2014). In all optimization runs we do 3×10^5 optimization steps with an initial learning rate of 10^{-3} that drops to 10^{-4} and 10^{-5} at iteration 10^5 and 2×10^5 , respectively. Similar to Geffner & Domke (2021) we parameterize the step size η_k in the k^{th} iteration of DAIS/NS-DAIS/SL-DAIS as

$$\eta_k = \text{clip}(\tilde{\eta} + \kappa\beta_k, \min = 0, \max = \eta_{\max}) \quad (52)$$

where $\tilde{\eta}$ and κ are learnable parameters and we choose $\eta_{\max} = 0.25$. The inverse temperatures $\{\beta_k\}$ are parameterized using the exponential transform to enforce positivity and a cumulative summation to enforce monotonicity. For SL-DAIS the learnable weights ω are uniformly initialized so that their total weight is equal to the number of data points, i.e. $\sum_n \omega_n = N$. In all experiments we use a diagonal mass matrix \mathbf{M} . We initialize $\tilde{\eta}$ to be small, e.g. $\tilde{\eta} \sim 10^{-4} - 10^{-2}$, and initialize κ to $\kappa = 0$. We initialize γ to $\gamma = 0.9$. Unless noted otherwise (see GP models), we learn all parameters that define the model and variational distribution jointly in a single optimization run, i.e. without any kind of pre-training (in contrast to Geffner & Domke (2021)). We found that this generally worked better, although in a few cases pre-training led to better results when the base distribution was a multivariate Normal.

A.7.1. SURROGATE LOG LIKELIHOODS

All methods use $K = 8$. We use 50k MC samples to estimate ELBOs after training. The training set has 50k data points and all replications are for the same training set. We consider eight different strategies for defining a surrogate log likelihood in the case of logistic regression:

1. **RAND**: We randomly choose N_{surr} surrogate data points $\{(\tilde{y}_n, \tilde{\mathbf{x}}_n)\} \subset \mathcal{D}$, introduce a N_{surr} -dimensional vector of learnable weights ω , and let $\hat{\Psi}_L(\mathbf{z}) = \sum_n \omega_n \log p(\tilde{y}_n | \mathbf{z}, \tilde{\mathbf{x}}_n)$.
2. **RAND \pm** : We randomly choose N_{surr} surrogate covariates $\{\tilde{\mathbf{x}}_n\}$ and introduce two N_{surr} -dimensional vectors of learnable weights, ω^- and ω^+ . We then write $\hat{\Psi}_L(\mathbf{z}) = \sum_n \omega_n^+ \log p(y_n = 1 | \mathbf{z}, \tilde{\mathbf{x}}_n) + \sum_n \omega_n^- \log p(y_n = 0 | \mathbf{z}, \tilde{\mathbf{x}}_n)$.
3. **CS-INIT**: We use a Bayesian coreset algorithm (Huggins et al., 2016) to choose N_{surr} surrogate data points, introduce a N_{surr} -dimensional vector of learnable weights ω , and let $\hat{\Psi}_L(\mathbf{z}) = \sum_n \omega_n \log p(\tilde{y}_n | \mathbf{z}, \tilde{\mathbf{x}}_n)$.
4. **CS-INIT \pm** : We proceed as in CS-INIT but ignore the $\{\tilde{y}_n\}$ returned by the coreset algorithm and instead introduce two N_{surr} -dimensional vectors of learnable weights, ω^- and ω^+ . We then write $\hat{\Psi}_L(\mathbf{z}) = \sum_n \omega_n^+ \log p(y_n = 1 | \mathbf{z}, \tilde{\mathbf{x}}_n) + \sum_n \omega_n^- \log p(y_n = 0 | \mathbf{z}, \tilde{\mathbf{x}}_n)$.
5. **CS-FIX**: As in CS-INIT we use a coreset algorithm to choose the surrogate data points but instead of learning ω we use the weights provided by the coreset algorithm.
6. **KM-INIT \pm** : We use k-means++ to extract N_{surr} cluster centroids from the set of N covariates in \mathcal{D} and introduce two N_{surr} -dimensional vectors of learnable weights, ω^- and ω^+ . We then write $\hat{\Psi}_L(\mathbf{z}) = \sum_n \omega_n^+ \log p(y_n = 1 | \mathbf{z}, \tilde{\mathbf{x}}_n) + \sum_n \omega_n^- \log p(y_n = 0 | \mathbf{z}, \tilde{\mathbf{x}}_n)$ and treat the surrogate covariates $\{\tilde{\mathbf{x}}_n\}$ as learnable parameters.
7. **KM-FIX \pm** : We proceed as in KM-INIT \pm except the surrogate covariates returned by the clustering algorithm remain fixed.
8. **NN**: We parameterize $\hat{\Psi}_L(\mathbf{z})$ as a neural network with two hidden layers, ELU non-linearities, and 100 neurons per hidden layer.

A.7.2. CLASSIFYING IMBALANCED DATA

We set $N_{\text{surr}} = B = 256$ and $K = 8$. We use the SUSY dataset with 20k data points held out for testing. The rare class corresponds to signal events. For each ratio of rare to non-rare class we keep the dataset fixed so that replications only differ with respect to the random number seed that controls optimization.

Surrogate Likelihoods for Variational Annealed Importance Sampling

Dataset	Higgs			SUSY			
	N_{surr}	64	256	1024	64	256	1024
RAND		211.1 ± 1.1	214.2 ± 1.3	222.4 ± 1.1	627.5 ± 2.1	638.9 ± 9.7	637.2 ± 2.2
RAND \pm		214.3 ± 1.0	217.0 ± 1.5	217.8 ± 1.5	635.0 ± 3.0	637.9 ± 4.0	632.2 ± 3.8
CS-INIT		209.6 ± 1.2	213.8 ± 1.1	218.5 ± 1.8	625.9 ± 2.2	645.3 ± 2.7	633.7 ± 3.8
CS-INIT \pm		213.5 ± 0.9	214.5 ± 1.4	219.0 ± 1.7	633.2 ± 3.7	634.8 ± 5.5	637.4 ± 3.6
CS-FIX		83.0 ± 12.3	136.6 ± 12.6	192.7 ± 5.5	501.0 ± 36.2	538.5 ± 14.2	555.5 ± 15.8
KM-INIT \pm		184.0 ± 2.5	187.4 ± 3.0	182.4 ± 2.8	222.3 ± 24.0	188.3 ± 11.4	169.7 ± 13.9
KM-FIX \pm		218.3 ± 1.3	217.7 ± 1.4	217.4 ± 1.8	635.5 ± 3.1	633.9 ± 2.6	632.1 ± 2.5
NN		16.8 ± 0.9			2.6 ± 1.4		

Table 6: This table is an expanded version of Table 1 in Sec 7.1. We compare eight different parameterizations for the surrogate log likelihood $\hat{\Psi}_L$. We consider two logistic regression datasets. For the seven strategies that make use of surrogate data points, we vary $N_{\text{surr}} \in \{64, 256, 1024\}$. We report ELBO improvements (in nats) above a mean-field Normal baseline. See Sec. 7.1.

A.7.3. LOGISTIC REGRESSION

When running HMCECS (Dang et al., 2019), we use the ‘perturbed’ approach with a NUTS kernel and use a control variate based on a second-order Taylor approximation around the maximum a posteriori estimate. In addition, to facilitate mixing we use block pseudo-marginal updates for data subsample indices with 25 blocks (Tran et al., 2016). We run for a total of 15×10^3 iterations, discard the first 5×10^3 samples for burn-in, thin the remaining 10×10^3 samples by a factor of 10, and use these 10^3 thinned samples to compute posterior predictive distributions on held-out test data. For NUTS we use the default settings in NumPyro, in particular a diagonal mass matrix that is adapted during warm-up and a maximum tree depth of 10.

For NUTS-CS we identify 1024 coresets using the coreset algorithm described in Huggins et al. (2016). Note that in most cases the total number of coresets returned by the algorithm is somewhat less than 1024, e.g. 1009 or 1017. We use the settings recommended in Huggins et al. (2016). In particular we use kmeans++ to construct a clustering with $k = 6$ clusters. We set the hyperparameter a to $a = 6$. We then run NUTS Hoffman et al. (2014) on a logistic regression model conditioned on the weighted coresets using the same number of NUTS iterations, mass matrix, etc., as for HMCECS.

The Higgs and SUSY datasets are described in (Baldi et al., 2014) and available from the UCI repository (Asuncion & Newman, 2007). The MiniBooNE dataset is from (Roe et al., 2005) and is likewise available from UCI. For the two CovType datasets (Blackard & Dean, 1999), which are also available from UCI, we reduce the 7-way classification problem to two binary classifications problems: i) fir versus rest; and ii) pine versus rest.

A.7.4. GAUSSIAN PROCESS CLASSIFICATION

We consider a Gaussian process model for binary classification with a logistic link function. The covariates are assumed to be d -dimensional, $\mathbf{x}_n \in \mathbb{R}^d$. To make GP inference scalable we use the FITC approximation (Snelson & Ghahramani, 2006; Quinonero-Candela & Rasmussen, 2005), which results in a model density of the following form:

$$p(\mathcal{D}, \mathbf{f}, \mathbf{u}) = p(\mathbf{u} | \mathbf{0}, \mathbf{K}_{ZZ}) \prod_n p(y_n | \sigma(f_n)) p(f_n | m_n(\mathbf{u}), v_n)$$

where $\mathbf{Z} \in \mathbb{R}^{M \times d}$ are the M inducing point locations, \mathbf{K} refers to the RBF kernel parameterized by a (scalar) kernel scale and a d -dimensional vector of length scales, $\sigma(\cdot)$ is the sigmoid function, $\mathbf{K}_{ZZ} \in \mathbb{R}^{M \times M}$ is the prior covariance, $\mathbf{u} \in \mathbb{R}^M$ are the inducing point values, f_n is the Gaussian process function value corresponding to covariate \mathbf{x}_n , and $m_n(\mathbf{u})$ and v_n are given by:

$$m_n(\mathbf{u}) = \mathbf{k}_{nZ}^T \mathbf{K}_{ZZ}^{-1} \mathbf{u} \quad v_n = k(\mathbf{x}_n, \mathbf{x}_n) - \mathbf{k}_{nZ}^T \mathbf{K}_{ZZ}^{-1} \mathbf{k}_{nZ} \quad (53)$$

Here $\mathbf{k}_{nZ} \in \mathbb{R}^M$ with $(\mathbf{k}_{nZ})_m = k(\mathbf{x}_n, \mathbf{z}_m)$ for $m = 1, \dots, M$, where $k(\cdot, \cdot)$ is the kernel function. Since by construction the $\{f_n\}$ are uncorrelated with one another, we can approximately integrate out $\{f_n\}$ using Gauss-Hermite quadrature using Q quadrature points. In all our experiments we use $Q = 16$, $M = 128$, and $N_{\text{surr}} = 512$. This results in a model density of

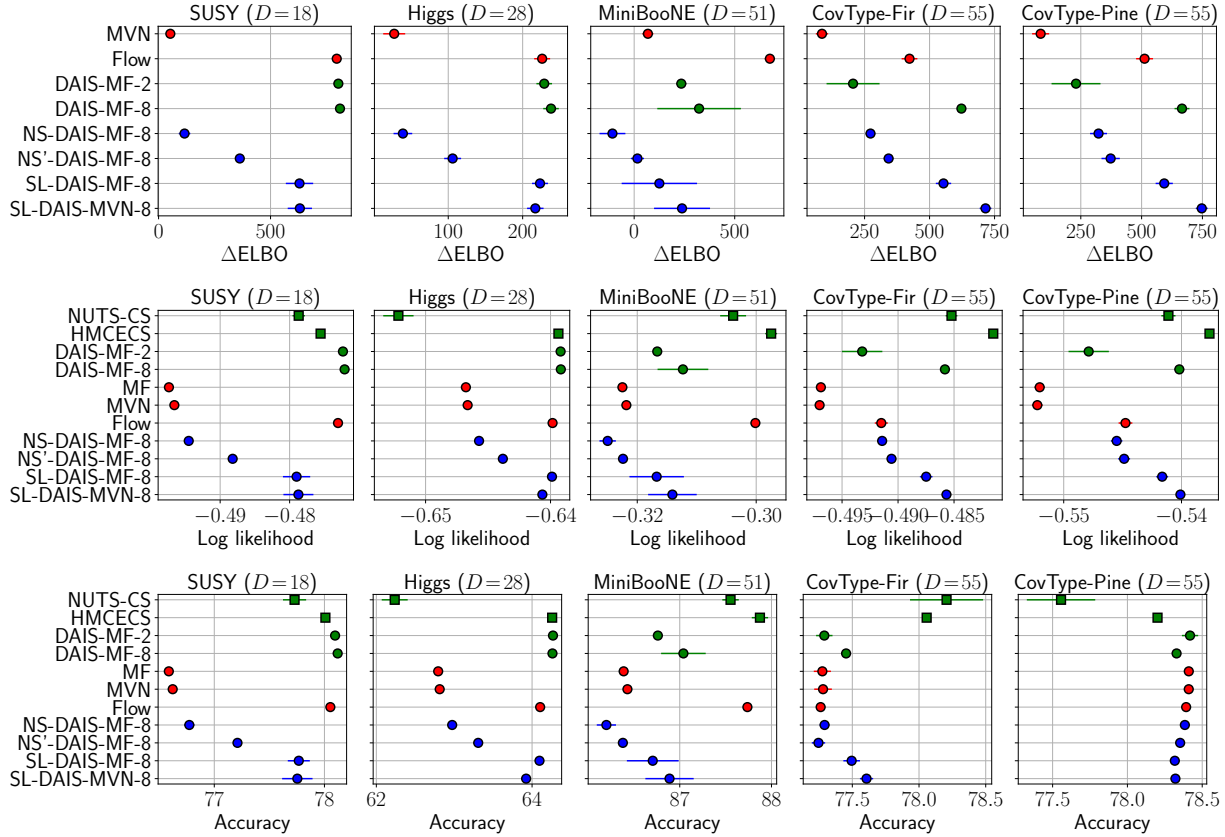


Figure 4: In this companion figure to Fig. 2 we report full results for the logistic regression experiment in Sec. 7.3. Circles denote variational methods and squares denote MCMC methods. Metrics are averaged over 7 independent replications (except for DAIS and HMCECS where we do 3 independent replications), and error bars denote standard errors. Note that these figures include an additional method, namely NS'-DAIS. This method is identical to NS-DAIS except that it utilizes independently drawn mini-batches of data for each evaluation of the annealed potential energy V_k in Eqn. 33. We note that NS'-DAIS outperforms NS-DAIS but is inferior to SL-DAIS.

the form

$$p(\mathcal{D}, \mathbf{u}) = p(\mathbf{u}|\mathbf{0}, \mathbf{K}_{ZZ}) \prod_n p(y_n|\mathbf{u}, \mathbf{x}_n) \quad (54)$$

where the cost of computing N likelihoods $p(y_n|\mathbf{u}, \mathbf{x}_n)$ is $\mathcal{O}(NM^2 + M^3 + NQ)$. We use variational inference to infer an approximate posterior over \mathbf{u} for the model density in Eqn. 54. In practice $N \gg M$ and the cost of computing the full likelihood is dominated by the $\mathcal{O}(NM^2)$ term. In all our Gaussian process experiments we initialize the inducing point locations \mathbf{Z} with k-means and treat them as learnable parameters. We use a mini-batch size $B = 128$. After training ELBOs are estimated with 20k MC samples. The MVN base distribution of SL-DAIS is initialized using the result from the MVN baseline.

A.7.5. ROBUST GAUSSIAN PROCESS REGRESSION

We proceed exactly as in Sec. A.7.4 except we use a Student's t likelihood $p(y_n|\nu, f_n, \sigma_{\text{obs}}) = \text{StudentT}(y_n|\nu, f_n, \sigma_{\text{obs}})$. Here $\nu > 0$ is the overdispersion parameter, with small ν corresponding to large overdispersion. We constrain ν to satisfy $\nu > 2$, which is equivalent to requiring that the observation noise have finite variance.

The Precipitation dataset is available from the IRI/LDEO Climate Data Library.¹² This spatio-temporal dataset represents the 'WASP' index (Weighted Anomaly Standardized Precipitation) at various latitudes and longitudes. Each data point

¹²http://iridl.ldeo.columbia.edu/maproom/Global/Precipitation/WASP_Indices.html

Surrogate Likelihoods for Variational Annealed Importance Sampling

	MF	MVN	Flow	NS-DAIS-MF	SL-DAIS-MF	NS-DAIS-MVN	SL-DAIS-MVN
ELBO	6.60	5.36	2.08	5.00	2.60	4.16	2.20
Log likelihood	6.16	5.88	2.16	4.60	2.44	4.36	2.40
Accuracy	5.48	5.24	1.96	4.52	3.40	4.12	3.28

Table 7: We report performance ranks w.r.t. three metrics across 5 train/test splits and 5 datasets for the logistic regression experiment in Sec. 7.3. Lower is better for all metrics. The rank satisfies $1 \leq \text{rank} \leq 7$, since we compare 7 scalable variational methods.

corresponds to the WASP index for a given year (which itself is the average of monthly WASP indices). We use the data from 2010, resulting in a dataset with 10127 data points. In each train/test split we randomly choose $N = 8000$ data points for training and use the remaining 2127 data points for testing.

We use a mini-batch size $B = 128$. We estimate the final ELBO using 5000 samples and test log likelihoods and accuracies using 1000 posterior samples. We use $N_{\text{surr}} = 256$ surrogate data points with the RAND surrogate parameterization (see Sec. A.7.1). We use an initial step size $\eta_{\text{init}} = 10^{-4}$ and initialize SL-DAIS using the model and variational parameters obtained with a MVN variational distribution.

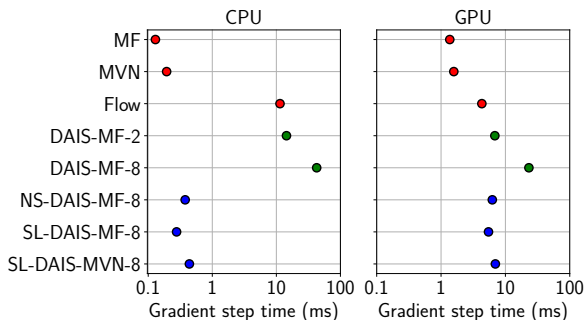


Figure 5: We report the time per optimization step for each variational method in Fig. 2 on the CovType-Fir dataset. We compare CPU runtime (24 cores; Intel Xeon Gold 5220R 2.2GHz) to GPU runtime (NVIDIA Tesla K80). We note that the runtime comparison against DAIS would be increasingly favorable to NS-DAIS and SL-DAIS as the size of the dataset increases.

A.7.6. LOCAL LATENT VARIABLE MODELS

We use the Pol UCI dataset with 15k training and 15k test data points. We use 10k MC samples to estimate ELBOs after training. We set $B = 256$.

A.8. Additional figures and tables

For additional results pertaining to the logistic regression experiment in Sec. 7.3 see Fig. 4, Fig. 5, and Table 7. For an expanded version of the table in Sec. 7.1 see Table 6.