Differentiable Annealed Importance Sampling Minimizes The Jensen-Shannon Divergence Between Initial and Target Distribution

Johannes Zenn [1 2 3]   **Robert Bamler** [1 2]

Abstract

Differentiable annealed importance sampling (DAIS), proposed by Geffner & Domke (2021) and Zhang et al. (2021), allows optimizing, among others, over the initial distribution of AIS. In this paper, we show that, in the limit of many transitions, DAIS minimizes the symmetrized KL divergence (Jensen-Shannon divergence) between the initial and target distribution. Thus, DAIS can be seen as a form of variational inference (VI) in that its initial distribution is a parametric fit to an intractable target distribution. We empirically evaluate the usefulness of the initial distribution as a variational distribution on synthetic and real-world data, observing that it often provides more accurate uncertainty estimates than standard VI (optimizing the reverse KL divergence), importance weighted VI, and Markovian score climbing (optimizing the forward KL divergence).

1. Introduction

Annealed importance sampling (AIS) (Neal, 2001) allows estimating the normalization constant $Z := \int f(z)\,\mathrm{d}z$ of an unnormalized distribution $f(z)$. In probabilistic machine learning, AIS can be used for Bayesian model selection (Knuth et al., 2015), where the goal is to maximize the marginal likelihood $p(x) = \int p(z, x)\,\mathrm{d}z$ over a family of candidate probabilistic models p. Here, $p(z, x)$ is the model's joint distribution over latent variables z and (fixed) observed data x. In this paper, we investigate the initial distribution of AIS or, more specifically, *differentiable* AIS (DAIS) (Geffner & Domke, 2021; Zhang et al., 2021). DAIS combines aspects of Markov Chain Monte Carlo (MCMC) and Variational Inference (VI): it draws samples from a (variational) initial distribution q_0 and then follows MCMC

dynamics to move the samples towards a target distribution.

We show theoretically that, in the limit of many transitions, DAIS minimizes the symmetrized Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), also known as Jensen-Shannon divergence, between its initial and target distribution. The Jensen-Shannon divergence is the sum of the reverse and the forward KL divergence. While the reverse KL divergence, used in VI, is known to be *mode-seeking*, the forward KL divergence, used in Markovian Score Climbing (MSC) (Naesseth et al., 2020), is typically associated with a *mass-covering* behavior (Murphy, 2023). The Jensen-Shannon divergence averages between both divergences.

We empirically analyze implications of this theoretical result by asking: "How useful is the initial distribution q_0 of DAIS as a parametric approximation of the target distribution?" This question corresponds to an inference scheme (which we denote by DAIS$_0$) that is identical to DAIS at training time (i.e., it moves particles sampled from q_0 along an MCMC chain). At inference time, however, DAIS$_0$ only uses q_0, omitting expensive AIS steps. Here, we refer to fitting the model parameters as "training" and running the model on test data as "inference". Generally, we do not expect DAIS$_0$ to match the performance of DAIS. However, DAIS$_0$ has the advantage of providing an *explicit* and *compact* representation of the approximate target distribution.

We call a distribution *explicit* if it has an analytic expression (as opposed to, e.g., an algorithmic prescription for sampling from it). It is often stated that an explicit expression for the exact Bayesian posterior $p(z \mid x) = p(z, x)/p(x)$ is intractable in large models, since calculating $p(x)$ is prohibitively expensive. However, this adage misses an important point. Even if we had an oracle that told us the value of $p(x)$, the explicit expression for the exact posterior would typically be far too complicated to be of any practical use in downstream tasks because it would, in general, have one term per data point. We argue that we are actually interested in a *compact* approximation of the posterior distribution, i.e., a tractable parameterized distribution that one can efficiently evaluate and sample from. Having a compact approximate posterior enables various downstream applications, such as continual learning (Nguyen et al., 2018), pruning in BNNs (Xiao et al., 2023), and compression of neural networks

---

[1]Tübingen AI Center [2]University of Tübingen [3]IMPRS-IS. Correspondence to: Johannes Zenn <johannes.zenn@uni-tuebingen.de>.

(Tan & Bamler, 2022) or other data (Yang et al., 2020).

Starting from our theoretical contribution—showing that DAIS minimizes the Jensen-Shannon divergence between q_0 and f / Z—we provide an empirical analysis of DAIS_0 for approximate Bayesian inference. We compare DAIS_0, DAIS, VI, importance weighted VI (IWVI), and MSC. We find that DAIS_0 often provides more accurate uncertainty estimates than VI, IWVI, and MSC on Gaussian process regression and real-world datasets for logistic regression.

2. Related Work

In the following, we discuss related work on variational distributions that are augmented by MCMC transitions, the forward and reverse KL divergence, and DAIS. We cover related work on methods to estimate normalization constants in Section 3. Related work on bounds with respect to the Jensen-Shannon divergence is given in Section 4.

MCMC-Augmented Variational Distributions. Sequential Monte Carlo samplers (SMCS) (Del Moral et al., 2006) are methods derived from prarticle filters (Doucet et al., 2001) to estimate normalization constants. While SMCS and annealed importance sampling (AIS) both describe the same mathematical framework, SMCS typically integrate a resampling step that lets the model focus on "important" particles. AIS can be seen as a finite-difference approximation to thermodynamic integration (TI) (Gelman & Meng, 1998) which computes ratios of normalization constants by solving a one-dimensional integration problem. Masrani et al. (2019) connect TI and variational inference (VI) resulting in tighter variational bounds by using importance sampling to compute the integral. Thin et al. (2021) propose a variant of the algorithm based on sequential importance sampling. Hamiltonian variational inference combines variational inference and MCMC iterations differentiably by augmenting the variational distribution with MCMC steps (Salimans et al., 2015; Wolf et al., 2016). The Hamiltonian VAE (Caterini et al., 2018) builds on Hamiltonian Importance Sampling (Neal, 2005) and improves on HVI by using optimally chosen reverse MCMC kernels. DAIS_0, investigated here, uses an MCMC-augmented variational distribution during training but not during inference.

VI With Forward and Reverse KL Divergence. Black box VI (Ranganath et al., 2014) is typically associated with the reverse KL divergence (between the variational distribution and the real posterior distribution). However, also various other divergences have been investigated to obtain an approximation of the posterior distribution (Hernandez-Lobato et al., 2016; Li & Turner, 2016; Dieng et al., 2017; Wan et al., 2020). Most related to DAIS_0, Ruiz & Titsias (2019) refine the variational distribution by running MCMC
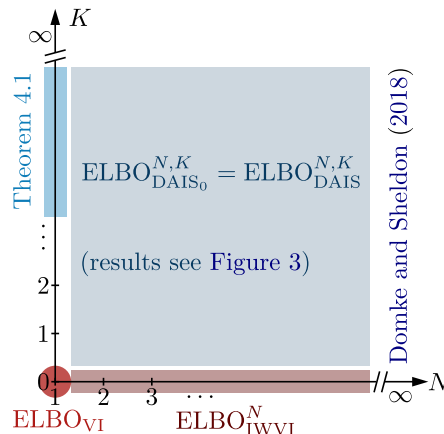

Figure 1. The landscape spanned by various lower bounds to the normalization constant $Z = \int f(z) \, dz$ where N denotes number of particles and K the number of importance sampling transitions. A discussion and further details can be found in Section 3.

transitions. They minimize a contrastive divergence which, in the limit, converges to the Jensen-Shannon divergence. Markovian score climbing (MSC) (Naesseth et al., 2020), that we also compare to in Section 5, optimizes the forward KL divergence using unbiased stochastic gradients. MSC samples a Markov chain and uses the samples to follow the score of the variational distribution. The Markov kernel for the MCMC dynamics is based on importance sampling.

Differentiable Annealed Importance Sampling (DAIS). Geffner & Domke (2021) and Zhang et al. (2021) propose DAIS concurrently focusing on different aspects of the model (empirical results and a convergence analysis for linear regression, respectively). Doucet et al. (2022) identify that the backward transitions of AIS are conveniently chosen but suboptimal. They propose Monte Carlo diffusion (MCD) which parameterizes the time reversal of the forward dynamics and which can be learned by maximizing a lower bound on the marginal likelihood, or equivalently, a denoising score matching loss. Geffner & Domke (2023) provide a more general framework for MCD and investigate various dynamics and numerical simulation schemes. Zenn & Bamler (2023) extend DAIS to a sequential Monte Carlo sampler and provide a theoretical argument for leaving out the gradients corresponding to the resampling operation.

3. Estimating Normalization Constants

In this section, we discuss how variational inference (VI), importance weighted VI (IWVI), and (differentiable) AIS ((D)AIS) can be understood as approaches to reducing the variance of importance sampling (IS). Figure 1 summarizes the space spanned by these methods, highlights a limiting behavior (Domke & Sheldon, 2018), and shows how our

2

theoretical result (Theorem 4.1) fits into this unifying framework. Section 3.1 and Section 3.2 give an overview over (IW)VI. Section 3.3 then introduces our notation for (D)AIS.

Importance Sampling is a principal technique for estimating the integral over a (nonnegative) function f by sampling from a normalized proposal distribution q,

$$Z := \int f(\boldsymbol{z}) \, d\boldsymbol{z} = \mathbb{E}_{\boldsymbol{z} \sim q} \left[\frac{f(\boldsymbol{z})}{q(\boldsymbol{z})} \right], \qquad (1)$$

where we assume that the support of q contains the support of f. We need to be able to efficiently draw samples (or *particles*) $\boldsymbol{z}$ from q, and to evaluate $f(\boldsymbol{z})$ and $q(\boldsymbol{z})$ for these samples. How many samples are necessary to obtain an accurate estimate of Z depends on how well q approximates $f \, / \, Z$. In high dimensions, the mismatch between $f \, / \, Z$ and any simple proposal distribution q grows, which leads to high variance of the importance weight $w(\boldsymbol{z}) := f(\boldsymbol{z}) \, / \, q(\boldsymbol{z})$, and the number of samples has to grow exponentially in the dimension (Agapiou et al., 2017). We now discuss how VI, IWVI, and AIS all aim to reduce this variance.

3.1. Variational Inference

Variational inference (VI) applies the logarithm to the importance weights $w(\boldsymbol{z}) := f(\boldsymbol{z}) \, / \, q(\boldsymbol{z})$ in Equation (1). This reduces exponentially growing variances to linearly growing ones, but it introduces bias, resulting in a lower bound on $\log Z$ by Jensen's inequality (Blei et al., 2017),

$$\log Z \geq \mathbb{E}_{\boldsymbol{z} \sim q} \left[\log \left(\frac{f(\boldsymbol{z})}{q(\boldsymbol{z})} \right) \right] =: \text{ELBO}_{\text{VI}}(f, q). \quad (2)$$

Equation (2) is called the evidence lower bound (ELBO) since VI is often used to estimate the evidence $\log p(\boldsymbol{x}) = \log \int p(\boldsymbol{z}, \boldsymbol{x}) \, d\boldsymbol{z}$ of a probabilistic model $p(\boldsymbol{z}, \boldsymbol{x})$ (latent variables $\boldsymbol{z}$ and observed data $\boldsymbol{x}$). VI maximizes the ELBO over parameters of q, leading to the best approximation of $\log p(\boldsymbol{x})$, which is useful for approximate Bayesian model selection (Beal & Ghahramani, 2003; Kingma & Welling, 2014). In addition, VI is a popular method for approximate Bayesian inference as the distribution q that maximizes the ELBO also approximates the (intractable) posterior $p(\boldsymbol{z} \, | \, \boldsymbol{x}) = p(\boldsymbol{z}, \boldsymbol{x}) \, / \, Z$. This is because maximizing the ELBO over parameters of q is equivalent to minimizing the gap $\Delta_{\text{VI}} := \log Z - \text{ELBO}_{\text{VI}}(f, q) \geq 0$, which turns out to be the KL divergence from the true posterior to q, i.e., $\Delta_{\text{VI}} = D_{\text{KL}}(q(\boldsymbol{z}) \, \| \, p(\boldsymbol{z} \, | \, \boldsymbol{x}))$ (Blei et al., 2017).

As we discuss next, IWVI and AIS can both be seen as methods that reduce the gap of VI by reducing the sampling variance. AIS is typically considered a sampling method, prioritizing the quality of samples or an accurate estimate of Z over a good approximate posterior distribution q, which would be the perspective of VI. In Section 4, we analyze (differentiable) AIS from the perspective of VI.

3.2. Importance Weighted Variational Inference

Importance weighted variational inference (IWVI) (Burda et al., 2016; Domke & Sheldon, 2018) reduces the variance of the importance weight $w(\boldsymbol{z})$ by averaging N independent samples from q, where $N \geq 2$, i.e., it uses the weights

$$w_{\text{IWVI}}^N(\boldsymbol{z}^{(1:N)}) := \frac{1}{N} \sum_{i=1}^N \frac{f(\boldsymbol{z}^{(i)})}{q(\boldsymbol{z}^{(i)})} \qquad (3)$$

where $\boldsymbol{z}^{(i)} \sim q$ for all i. The resulting bound,

$$\log Z \geq \mathbb{E}_{\boldsymbol{z}^{(1:N)} \sim q} \left[\log \left(w_{\text{IWVI}}^N(\boldsymbol{z}^{(1:N)}) \right) \right]$$
$$=: \text{ELBO}_{\text{IWVI}}^N(f, q), \qquad (4)$$

recovers $\text{ELBO}_{\text{VI}}(f, q)$ for $N = 1$ (see Equation (2) and red highlights on the x axis of Figure 1). Sampling from q at inference time requires a sampling-importance-resampling (SIR) procedure (Cremer et al., 2017) which we denote by IWVI_{SIR}. The corresponding Markov kernel is known as i-SIR and studied in detail by Andrieu et al. (2018). As N grows, the bound becomes tighter. In the limit of $N \to \infty$, Domke & Sheldon (2018) find (based on results due to Maddison et al. (2017, Proposition 1)) the following.

Theorem 3.1 (Theorem 3 in Domke & Sheldon (2018)). *For large N, the gap $\Delta_{\text{IWVI}}^N := \log Z - \text{ELBO}_{\text{IWVI}}^N$ of importance weighted VI is proportional to the variance of w_{IWVI}^N, defined in Equation (3). Formally, if $\limsup_{N \to \infty} \mathbb{E}_q[1 \, / \, w_{\text{IWVI}}^N] < \infty$ and there exists some $\alpha > 0$ such that $\mathbb{E}_q\big[|w_{\text{IWVI}}^N - \log Z|^{2+\alpha}\big] < \infty$, then*

$$\lim_{N \to \infty} N \Delta_{\text{IWVI}}^N = \frac{\text{Var}_q\big[w_{\text{IWVI}}^N(\boldsymbol{z}^{(1:N)})\big]}{2Z^2}. \qquad (5)$$

Proof. See Domke & Sheldon (2018, Theorem 3). $\qquad \square$

Thus, for large N, a higher variance of w_{IWVI}^N corresponds to a larger gap (see Section 3.1). Annealed importance sampling, discussed next, provides a way to further reduce the sampling variance for a fixed number of particles N.

3.3. (Differentiable) Annealed Importance Sampling

Annealed importance Sampling (AIS) reduces the variance of $w_{\text{IWVI}}^N(\boldsymbol{z}^{(1:N)})$ further by recalling that the variance of the importance weights $f(\boldsymbol{z}^{(i)}) \, / \, q(\boldsymbol{z}^{(i)})$ in Equation (3) is a consequence of a distribution mismatch between q and f/Z. To reduce the distribution mismatch, AIS interpolates between q and $f \, / \, Z$ with distributions $\{\pi_{\beta_k}(\boldsymbol{z}_k)\}_{k=0}^K$ over auxiliary variables $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_K$ for K interpolation steps.

In detail, AIS estimates the normalization constant of an unnormalized distribution f_{AIS} over all $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_K$,

$$f_{\text{AIS}}(\boldsymbol{z}_{0:K}) := f(\boldsymbol{z}_K) \prod_{k=1}^K B_k(\boldsymbol{z}_{k-1} \, | \, \boldsymbol{z}_k). \qquad (6)$$

Here, the so-called backward transition kernels B_k have to be normalized in their first argument so that f_{AIS} and f have the same normalization constant, $\int f_{AIS}(z_{0:K}) \, dz_{0:K} = \int f(z_K) \, dz_K = Z$. Originally, Neal (2001) estimates Z with IS, using a proposal distribution of the form

$$q_{AIS}(z_{0:K}) := q_0(z_0) \prod_{k=1}^{K} F_k(z_k \mid z_{k-1}), \qquad (7)$$

where we call q_0 the initial distribution and F_k a forward transition kernel. Instead of using q_{AIS} with IS to estimate Z, we can use q_{AIS} also with IWVI to obtain a bound on Z,

$$\log Z \geq \mathbb{E}_{z_{0:K}^{(1:N)} \sim q_{AIS}} \left[\log \left(w_{AIS}\left(z_{0:K}^{(1:N)}\right) \right) \right]$$
$$=: \mathrm{ELBO}_{AIS}^{N,K}(f_{AIS}, q_{AIS}) \qquad (8)$$

with the annealed importance weights

$$w_{AIS}^{N,K}\left(z_{0:K}^{(1:N)}\right) = \frac{1}{N} \sum_{i=1}^{N} \frac{f(z_K^{(i)})}{q_0(z_0^{(i)})} \prod_{k=1}^{K} \frac{B_k(z_{k-1}^{(i)} \mid z_k^{(i)})}{F_k(z_k^{(i)} \mid z_{k-1}^{(i)})}. \qquad (9)$$

Equation (8) and Equation (9) hold for arbitrary normalized forward and backward kernels F_k and B_k (as long as $\mathrm{support}(q_{AIS}) \supseteq \mathrm{support}(f_{AIS})$). In practice, however, we want to draw samples $z_k^{(i)}$ from distributions that interpolate smoothly between $q_0(z_0)$ and $f(z_K)$ so that none of the factors on the right-hand side of Equation (9) has a large variance. One typically realizes each $F_k(z_k \mid z_{k-1})$ as a Markov Chain Monte Carlo process (typically Hamiltonian Monte Carlo (HMC)), which leaves a distribution $\pi_{\beta_k}(z_k)$ invariant, where the distributions $\{\pi_{\beta_k}\}_{k=0}^{K}$ interpolate between $\pi_{\beta_0} := q_0$ and $\pi_{\beta_K} := f / Z$. The most common choice of π_{β_k} uses the geometric mean (aka annealing), i.e.,

$$\pi_{\beta_k}(z) := \frac{\gamma_{\beta_k}(z)}{Z_{\beta_k}} \quad \text{with} \quad \gamma_{\beta_k}(z) := q_0(z) \left(\frac{f(z)}{q_0(z)} \right)^{\beta_k} \qquad (10)$$

where $Z_{\beta_k} = \int \gamma_{\beta_k}(z) \, dz$ and $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$. To ensure that Equation (9) can be evaluated for samples $z_{1:K} \sim q_{AIS}$, one typically sets B_k to the reversal of F_k, i.e., $B_k(z_{k-1} \mid z_k) := F(z_k \mid z_{k-1}) \gamma_{\beta_k}(z_{k-1}) / \gamma_{\beta_k}(z_k)$, which is properly normalized. With this choice, Equation (9) simplifies to

$$w_{AIS}^{N,K}\left(z_{0:K}^{(1:N)}\right) = \frac{1}{N} \sum_{i=1}^{N} \prod_{k=1}^{K} \frac{\gamma_{\beta_k}\left(z_{k-1}^{(i)}\right)}{\gamma_{\beta_{k-1}}\left(z_{k-1}^{(i)}\right)}. \qquad (11)$$

Differentiable Annealed Importance Sampling. To ensure that $F_k(z_k \mid z_{k-1})$ leaves π_{β_k} invariant, the HMC implementation of F_k involves a Metropolis-Hastings (MH) acceptance or rejection step, which makes the method non-differentiable. Differentiable annealed importance sampling

(DAIS) drops this MH step. The resulting "uncorrected" transition kernels F_k do not leave π_{β_k} invariant. Instead, the backward kernels B_k are defined by starting from an exact sample and reversing the forward chain. With these modifications, the resulting lower bound $\mathrm{ELBO}_{DAIS}^{N,K}(f_{DAIS}, q_{DAIS})$ has the same functional form as $\mathrm{ELBO}_{AIS}^{N,K}(f_{AIS}, q_{AIS})$, where only the kernels F_k and B_k differ. Furthermore, it can be optimized with reparameterization gradients.

Theorem 3.1 (Domke & Sheldon, 2018, Theorem 3) also applies to this bound. Therefore, for large N, the gap of AIS is proportional to the variance of $w_{AIS}^{N,K}$, which, for good choices of the forward and backward kernels F_k and B_k, is smaller than the variance of the estimator of IWVI.

4. Analyzing the Initial Distribution of DAIS

In Theorem 4.1 of this section we present our main theoretical result that DAIS minimizes the Jensen-Shannon divergence between its initial distribution and its target distribution. Then, we discuss compact representations in terms of their sampling complexity. Finally, we frame finding the initial distribution of DAIS as a form of VI.

4.1. DAIS Minimizes the Jensen-Shannon Divergence

Theorem 4.1 below presents our main theoretical contribution showing that DAIS minimizes the Jensen-Shannon divergence between its initial and target distribution. We see Theorem 4.1 as starting point to motivate DAIS$_0$ and the analysis of the shape of the fitted initial distribution q_0. The result also holds true for AIS, but it is most relevant for DAIS, which learns parameters of its initial distribution. One can also show Theorem 4.1 by combining results of Grosse et al. (2013) and Crooks (2007) from the physics literature. Brekelmans et al. (2022) state a related bound on the difference between an upper and a lower bound on the evidence (whereas Theorem 4.1 makes an asymptotic statement about the gap between lower bound and evidence).

Theorem 4.1. *Let* $\mathrm{support}(q_0) \supset \mathrm{support}(f)$. *We assume that the transitions between consecutive annealing distributions are perfect and that* β_k *are equally spaced, i.e.,* $\beta_k = k / K$. *Then, for large K and $N = 1$, the gap* $\Delta_{AIS}^{1,K} := \log Z - \mathrm{ELBO}_{AIS}^{1,K}(f_{AIS}, q_{AIS})$ *is a divergence between the initial distribution* q_0 *and the target distribution* f / Z. *Up to corrections of* $\mathcal{O}(1 / K^3)$, *the divergence is proportional to the Jensen-Shannon divergence,*

$$\Delta_{AIS}^{1,K} = \frac{1}{K} \left(\frac{1}{2} D_{KL}(f(z) / Z \,\|\, q_0(z)) \right.$$
$$\left. + \frac{1}{2} D_{KL}(q_0(z) \,\|\, f(z) / Z) \right) \qquad (12)$$
$$+ \mathcal{O}(1 / K^3).$$

Proof. We prove the theorem in Appendix A. □

Statement for $N > 1$. For a general $N > 1$ we find that $\Delta_{\text{AIS}}^{N,K} \leq \Delta_{\text{AIS}}^{1,K}$ and therefore

$$\lim_{K \to \infty} \Delta_{\text{AIS}}^{N,K} \leq \frac{1}{K} D_{\text{JS}}(q_0, f/Z) + \mathcal{O}(1/K^3), \quad (13)$$

where D_{JS} denotes the Jensen-Shannon divergence. This shows that the bound for $N > 1$ can be upper-bounded by the Jensen-Shannon divergence but does not give additional insights on, e.g., symmetry. For $N = 1$ the inequality is an equality. See Appendix A for a discussion in greater detail.

Reverse KL Divergence. VI is known to underestimate uncertainties (Blei et al., 2017) which is due to an asymmetry in the ELBO. More specifically, ELBO_{VI} (Equation (2)) minimizes the reverse KL divergence $D_{\text{KL}}(q \,\|\, f/Z)$, which takes the expectation over q. Therefore, penalties for regions in z-space where $q(z) > f(z)/Z$ are weighted higher than penalties for regions where $q(z) < f(z)/Z$. As a result, ELBO_{VI} is mode-seeking and tends to underestimate the true posterior variance. For a small N, IWVI also shows a mode-seeking behavior, which can be overcome by increasing N. Theorem 4.1 states that, at least for $K \to \infty$ and $N = 1$, $\text{ELBO}_{\text{DAIS}}^{N,K}$ does not suffer from such an asymmetry. We investigate the situation for $K < \infty$ and $N \geq 1$ empirically in various experiments in Section 5.

Forward KL Divergence. In contrast, the forward KL divergence $D_{\text{KL}}(f/Z \,\|\, q)$ is mass-covering since the expectation is taken over f/Z and regions in z-space where $q(z) < f(z)/Z$ are weighted higher than the regions where $q(z) > f(z)/Z$. While the forward KL divergence is less likely to underestimate posterior variances, it often performs poorly in moderate to high dimensions when posterior correlations increase (Dhaka et al., 2021). Section 5 provides empirical evidence that DAIS_0 (implicitly minimizing the Jensen-Shannon divergence) indeed outperforms MSC (minimizing the forward KL divergence).
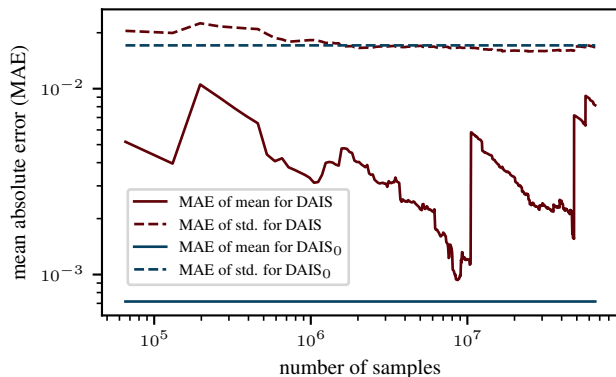
Figure 2. Mean absolute error of estimated mean and standard deviation as a function of the number of DAIS samples used for the estimate. The estimator converges poorly (see Section 4.2).

Limitations. We show Theorem 4.1 for large K which might be prohibitive in real world experiments. Additionally, we generally use more than $N = 1$ particles in DAIS. Therefore, we expect that Theorem 4.1 holds only approximately in practice. For large N, Theorem 3.1 (Domke & Sheldon, 2018, Theorem 3) holds and the gap closes. As a result, we see a combination of effects in our experiments with practical (moderate) values for K and N (see Section 5).

4.2. Compact Representation of the Initial Distribution

The initial distribution of DAIS, q_0, provides both an analytical expression (i.e., it is *explicit*) and allows for tractable sampling and evaluation (i.e., it is *compact*). In the following, we discuss positive implications of this representation.

Importance sampling is likely to suffer from inefficiencies due to the sampling complexity in high dimensions. Namely, if f becomes increasingly complicated, and thus the mismatch between f/Z and the proposal distribution q grows, the variance of the importance weight grows exponentially in its dimension (Bamler et al., 2017). Agapiou et al. (2017) show that also the number of samples grows exponentially in the dimension of the problem. Furthermore, Chatterjee & Diaconis (2018) show that the sample size is exponential in the KL divergence between the two measures (here: $D_{\text{KL}}(f/Z \,\|\, q)$) if they are nearly singular with respect to each other (which is often the case in practice).

As we discuss in Section 3 above, DAIS can be seen as importance sampling on an augmented space. Therefore, DAIS can also suffer from high sampling complexity in high dimensions. While we should expect samples from DAIS to technically follow the target distribution more faithfully than samples from the initial distribution q_0, obtaining accurate estimates of summary statistics (e.g., mean and variance) of the target distribution from DAIS samples can be prohibitively expensive in high dimensions. By contrast, q_0 is typically parameterized by interpretable summary statistics that can be read off without requiring empirical estimates over exponentially many expensive MCMC samples.

In Section 5 below, we observe how difficult it can be in practice to estimate summary statistics from DAIS samples. Figure 2 shows the mean absolute error (MAE) of the estimated mean and standard deviation for the Gaussian process experiment (RBF$_1$, $d = 25$ in Section 5.2) as a function of the number of DAIS samples used for the estimate (red curves). Note the sudden jumps of the estimator. We compare to the mean and standard deviations read off directly from the initial distribution q_0 (horizontal blue lines).

4.3. The Initial Distribution of DAIS for Inference

In the following, we investigate the initial distribution q_0 of DAIS as a candidate for an approximate posterior dis-
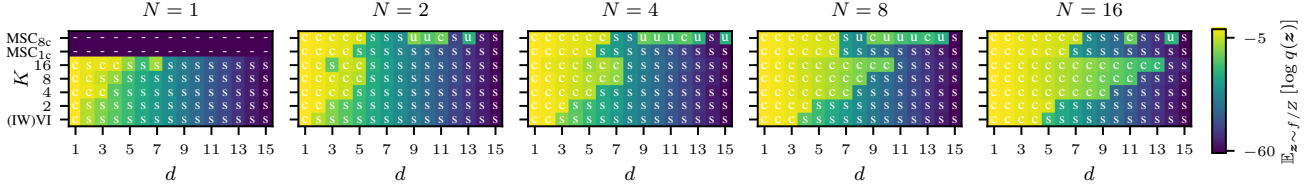
Figure 3. Density of variational distributions of VI, IWVI, MSC, and DAIS$_0$ (K) evaluated on samples from a d-dimensional bimodal Gaussian target distribution. "-": unable to find an optimum, "c": mass-covering distribution, "s": mode-seeking distribution, "u": undecidable whether "c" or "s". DAIS$_0$ achieves higher densities in higher dimension d for increasing K across all considered N. MSC does not converge for $N = 1$. MSC$_{1c}$ learns variational distributions that are less mass-covering for larger d than DAIS$_0$ (16). MSC$_{8c}$ achieves sometimes higher densities in higher dimension d but performs inconsistent across N. Results are discussed in Section 5.1.

tribution. We denote this scheme as DAIS$_0$: At training time, DAIS$_0$ mirrors DAIS and maximizes ELBO$_{DAIS}^{N,K}$. At inference time, DAIS$_0$ drops the AIS steps and solely uses q_0 as a variational approximation to the target distribution.

Following the approach by Zhang et al. (2021) we use DAIS with HMC dynamics. Thus, Equation (11) simplifies to

$$\frac{1}{N} \sum_{i=1}^{N} \frac{f(\boldsymbol{z}_K)}{q_0(\boldsymbol{z}_0)} \prod_{k=1}^{K} \frac{\mathcal{N}(\mathbf{v}_k^{(i)}; 0, \boldsymbol{M})}{\mathcal{N}(\mathbf{v}_{k-1}^{(i)}; 0, \boldsymbol{M})}, \qquad (14)$$

where $\mathcal{N}(\mathbf{v}_k^{(i)}; 0, \boldsymbol{M})$ denotes the density of the normal distribution at point $\mathbf{v}_k^{(i)}$ with covariance matrix $\boldsymbol{M}$ and mean 0. $\mathbf{v}_k^{(i)}$ and $\boldsymbol{M}$ are the momenta and the (diagonal) mass matrix of HMC, respectively. The initial distribution q_0 is a fully factorized normal distribution. We use gradient updates to learn the means and variances of q_0, the annealing schedule ($\beta_1, \ldots, \beta_{K-1}$), and the step width of the sampler.

As highlighted in Section 4.2, we do not expect that the initial distribution of DAIS (DAIS$_0$) generally outperforms the DAIS method. However, we highlight that DAIS$_0$ provides various desirable properties due to its compact representation. Additionally, DAIS$_0$ is much more computationally efficient at inference time compared to DAIS. In comparison to VI, DAIS$_0$ is expected be less mode-seeking (Section 4.1), and we pose that the Jensen-Shannon divergence is less susceptible to problems optimizing the forward KL divergence. Section 5 provides experimental evidence.

5. Experiments

In Section 5.1, we investigate DAIS$_0$ qualitatively for various dimensions, N, and K on toy data by analyzing its mode-seeking or mass-covering behavior. In Section 5.2 and Section 5.3, we show quantitative results on both generated and real world data. Throughout this section, we compare DAIS$_0$, IWVI, MSC, IWVI$_{SIR}$, and DAIS, where the first three find a compact representation of the approximate posterior while the latter two require sampling at inference time. As MSC is known to have convergence issues due to high variance (Kim et al., 2022) we report results using 8

parallel chains as MSC$_{8c}$ and results with 1 chain as MSC$_{1c}$.

5.1. Bimodal Target Distribution

This experiment is designed to investigate the behavior of DAIS$_0$ for $N \geq 1$ and $K \geq 1$. More specifically, we investigate the relationship between the parameters N and K alongside the dimension d of the variational distribution of DAIS$_0$, and compare it to the compact methods (IW(VI), MSC). Thus, the experiment covers the entire space spanned in Figure 1. For further details, see Appendix B.

We consider a bimodal target distribution $f(\boldsymbol{z}) / Z$ that is an equally-weighted mixture of two d-dimensional Gaussians with means $(0, \ldots, 0)$ and $(1, \ldots, 1)$, respectively, and covariance matrices $0.25^2 \boldsymbol{I}$. We learn the variational distributions of VI, IWVI, DAIS$_0$, and MSC (both, with 1 chain and 8 chains). To evaluate the learned variational distributions, we draw $1,000$ samples from the target distribution and compute the log density under q, i.e., $\mathbb{E}_{\boldsymbol{z} \sim f/Z}[\log(q(\boldsymbol{z}))]$. In Figure 3 we plot results for $N \in \{1, 2, 4, 8, 16\}$ particles and $K \in \{2, 4, 8, 16\}$ transitions in $d \in \{1, \ldots, 15\}$ dimensions. Figure 4 shows the density along the line from $(0, \ldots, 0)$ to $(1, \ldots, 1)$ for $N = 1$ and $d = 3$.
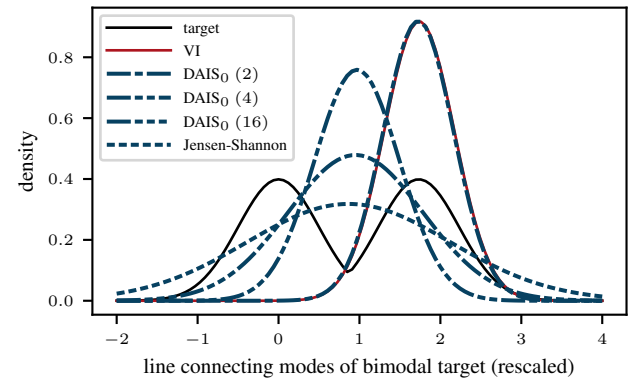
Figure 4. Densities on the diagonal between the two modes of the bimodal Gaussian distribution (same experiment as Figure 3 with $N = 1, d = 3$). VI covers a single mode while DAIS$_0$ covers both modes with increasing K (details in Section 5.1).
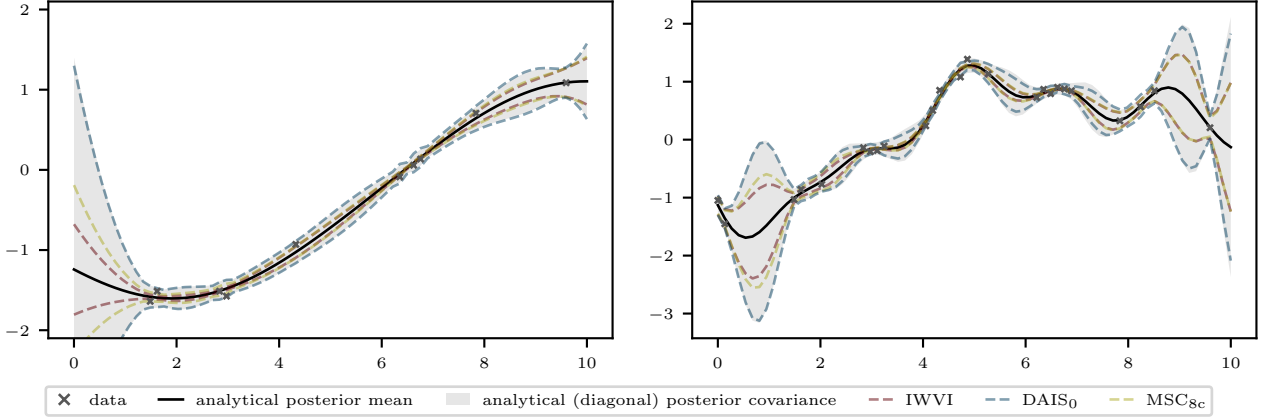
Figure 5. Gaussian process regression on generated data, using a prior with RBF kernel with two different sets of parameters (see $\star$ in Table 1). We show 97.5% quantiles of the posterior covariance for analytic (shaded gray; covariance matrix is diagonalized on data points, see Appendix C), IWVI (red), DAIS_0 (blue), and MSC_{8c} (yellow). Learned means are indistinguishable from the analytic mean (black) at this line width. DAIS_0 often provides more accurate uncertainty estimates compared to the other methods (details in Section 5.2).

Mode-Seeking Versus Mass-Covering. In Figure 4 we can clearly see that VI and DAIS_0 with $K = 2$ transitions lead to a mode-seeking distribution. For an increasing K (i.e., $K \in \{4, 16\}$), we find that the variational distribution of DAIS_0 becomes more mass-covering and increasingly similar to the solution that (numerically) minimizes the Jensen-Shannon divergence (regularly dashed line). This is consistent with Theorem 4.1. We find that we can unambiguously classify almost all of the learned variational distributions into being mode-seeking ("s") or mass-covering ("c") by calculating the distance of the mean of the variational distribution to both modes of the target distribution, and to their mid point (Appendix B provides a histogram of these distances and more details on the classification). Figure 3 shows corresponding classifications as labels "s" and "c" for each combination (N, K, d). We label the few cases where the classification is not perfectly unambiguous with "u" which is short for for "undecidable".

Comparing (IW)VI and DAIS_0 (16) in Figure 3, we find that DAIS_0 (16) typically covers both modes of the target distribution for larger dimensions (across all N). We attribute this finding to the Jensen-Shannon divergence that is implicitly minimized by DAIS_0. MSC does not converge to reasonable values ("-") for $N = 1$. With a single chain MSC_{1c} outperforms (IW)VI over all N but seems to be less mass-covering than DAIS_0. With 8 parallel chains MSC_{8c} sometimes outperforms the other methods in finding a mass-covering distribution. However, we find that the variational distributions of MSC_{8c} are often not centered on the mid-point between the two modes which we attribute to the high variance of the method (see Appendix B). This is in accordance with Kim et al. (2022) who report high variance for MSC and Dhaka et al. (2021) who note that the forward KL divergence is difficult to optimize for large dimensions.

Number of Transitions K. For a fixed N, we find that with increasing K, DAIS_0 (q_0) achieves higher log densities compared to VI, IWVI, and MSC, especially in higher dimensions. For example, for $N = 4$, we find that q_0 covers both modes with mass for a dimension up to $d = 7$ while VI and IWVI collapse to a single mode after $d = 1$ and $d = 2$, respectively. These results align with our theoretical findings. MSC_{1c} collapses after $d = 5$ which we attribute to the optimization of the forward KL divergence. MSC_{8c} performs better but inconsistently for higher dimensions.

Number of Particles N. With an increasing number of particles N, we see that IWVI and DAIS_0 improve. This follows the theoretical result on the number of particles (presented in Theorem 3.1 (Domke & Sheldon, 2018, Theorem 3)) which suggests to choose N as large as possible. Also, MSC shows higher densities with an increasing N; for MSC_{8c} the results are often inconsistent or undecidable.

5.2. Gaussian Processes Regression

We investigate a Gaussian process (GP) regression task on synthetic data. We generate the data by (i) sampling a ground truth process from a GP prior $\boldsymbol{f} \sim \mathcal{GP}(\boldsymbol{0}, k)$, (ii) randomly sampling positions $\mathcal{I}$ on the domain $[0, 10]$, and (iii) generating data from a Gaussian likelihood model $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{f}_{\mathcal{I}}, 0.1 \cdot \boldsymbol{I})$. We set k to be a Radial Basis Function (RBF) kernel (Williams & Rasmussen, 2006) and investigate two sets of hyperparameters (RBF_1 and RBF_2, see Appendix C). We learn a variational distribution for IWVI, MSC, and DAIS_0 on the positions $\mathcal{I}$. While this setup might sound artificial as the analytic posterior can be computed in closed form (due to the Gaussian likelihood), one can easily think of problems with non-Gaussian likelihoods or large latent spaces that can be phrased in the same framework.
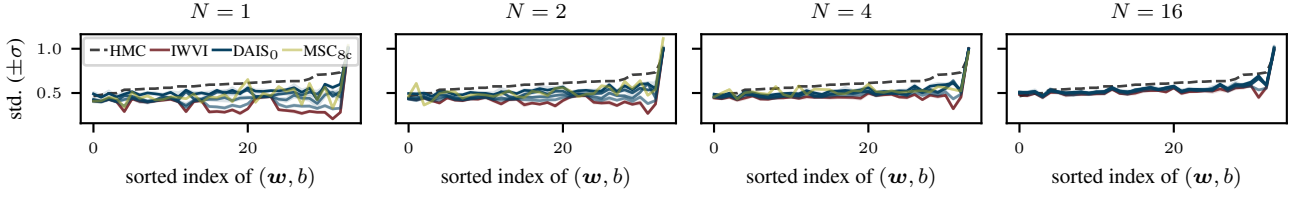
Figure 6. Bayesian logistic regression: standard deviations of learned posterior weight vector (indices are plotted on the x-axis, we report $\pm\sigma$) for IWVI (red), MSC_{8c} (yellow) and $DAIS_0$ (blue), compared to HMC samples (dashed) on *ionosphere*. $DAIS_0$ provides uncertainty estimates that improve with increasing $K \in \{2, 4, 8, 16\}$ (shown by increasing opacity). Details in Section 5.3.

Figure 5 shows two GPs that are marked with "$\star$" in Table 1. We train each GP with $N = 16$ particles. $DAIS_0$ uses $K = 16$ transitions. The black line and gray area show the theoretically best posterior mean and the corresponding posterior covariance that any Gaussian mean field approach can possibly achieve. We calculate the posterior approximation by finding the analytic posterior on $\mathcal{I}$, diagonalizing its covariance matrix, and then calculating the predictive posterior on the full range $[0, 10]$. By visual inspection we find that the uncertainties of $DAIS_0$ are more accurate compared to IWVI and MSC_{8c}. MSC_{8c} visually outperforms IWVI.

Table 1 shows mean absolute errors (MAEs) for a larger set of GP models. We compare models with compact distributions (IWVI, $DAIS_0$, MSC_{8c}) and sampling based approaches (DAIS using 10^5 and $IWVI_{SIR}$ using $10^3 \times 10^3$ samples). Best results are written bold. The best method with compact representation is written in italic font. Within the methods that have compact representations, $DAIS_0$ achieves the best MAE with respect to the standard deviation in most cases (except for RBF_2, 10 where relative errors are a magnitude smaller in general). In general, $DAIS_0$ performs competitively across all methods. Our theoretical result Theorem 4.1 might explain this, as the Jensen-Shannon divergence (minimized by $DAIS_0$) does not suffer from a mode-seeking behavior but learns a q_0 that covers more mass of the target distribution. For the mean estimation, IWVI outperforms $DAIS_0$ while $IWVI_{SIR}$ performs best. Comparing $DAIS_0$ and DAIS, we find that a compact

representation indeed helps in some cases (e.g., mean in RBF_1, $d = 10, 25$). MSC_{8c}, minimizing the forward KL divergence, performs competitively but slightly worse in comparison to the other methods. We attribute this finding to the instability of the forward KL divergence during optimization. We provide a larger study in Appendix C.

5.3. Bayesian Logistic Regression

We investigate a Bayesian logistic regression task with five real-world data sets (Gorman & Sejnowski, 1988; Sigillito et al., 1989; Kelly et al.). We model the Bayesian logistic regression with a factorizing standard normal prior on weights $\boldsymbol{w}$ and bias b. The likelihood is chosen to be Bernoulli distributed $\mathrm{Bern}(r)$ with parameter $r = \mathrm{sigmoid}(\boldsymbol{w}^\top \boldsymbol{x} + b)$. Here, $\boldsymbol{x}$ denotes a data point. We learn a factorizing Normal variational distribution for $\boldsymbol{z} = (\boldsymbol{w}, b)$.

Table 2 shows MAEs between learned means and learned standard deviations and means and standard deviations of HMC samples for various methods. $DAIS_0$ reaches the best errors for three out of five datasets with comparable mean MAE and lower MAE for the standard deviation. $DAIS_0$ outperforms DAIS in several cases by a significant margin (e.g., *sonar* and *ionosphere*) in terms of standard deviation which we mainly attribute to the larger sampling complexity of DAIS (see also Section 4.2). Similarly, when comparing IWVI and the sample-based $IWVI_{SIR}$, we find that IWVI (compact) outperforms the sampling based method for larger dimensional problems. Although MSC_{8c} shows errors com-

Table 1. Mean absolute error (compared to analytic solution) of mean and standard deviation of GP regression ($N = 16$). $DAIS_0$ (K) gives accurate uncertainty estimates for most cases (within all compact methods) while degrading the predicted mean only negligibly (details in Section 5.2). $RBF_{\{1,2\}}$ denote different prior parameters. Figure 5 shows "$\star$" visually. Results are averaged over 3 runs.

	d	MAE	IWVI	$IWVI_{SIR}$	$DAIS_0$ (16)	DAIS (16)	MSC_{8c}
RBF_1	$\star$ 10	mean	$5.60_{\pm 1.10} \cdot 10^{-4}$	$\mathbf{2.83_{\pm 1.70} \cdot 10^{-4}}$	$1.77_{\pm .31} \cdot 10^{-3}$	$7.72_{\pm .23} \cdot 10^{-3}$	$1.74_{\pm .24} \cdot 10^{-2}$
		std.	$4.34_{\pm .0020} \cdot 10^{-2}$	$3.50_{\pm .027} \cdot 10^{-2}$	$\mathit{4.54_{\pm .071} \cdot 10^{-3}}$	$\mathbf{3.028_{\pm .25} \cdot 10^{-3}}$	$3.66_{\pm .14} \cdot 10^{-2}$
	25	mean	$3.57_{\pm 1.10} \cdot 10^{-4}$	$\mathbf{4.17_{\pm 4.50} \cdot 10^{-5}}$	$9.58_{\pm 1.40} \cdot 10^{-4}$	$1.39_{.036} \cdot 10^{-2}$	$1.14_{\pm .16} \cdot 10^{-2}$
		std.	$3.83_{\pm .00030} \cdot 10^{-2}$	$3.62_{\pm .042} \cdot 10^{-2}$	$\mathbf{\mathit{1.03_{\pm .0010} \cdot 10^{-2}}}$	$2.16_{\pm .027} \cdot 10^{-2}$	$3.77_{\pm .015} \cdot 10^{-2}$
RBF_2	10	mean	$1.77_{\pm .14} \cdot 10^{-3}$	$\mathbf{4.00_{\pm 1.00} \cdot 10^{-5}}$	$2.44_{\pm 1.10} \cdot 10^{-3}$	$9.64_{\pm 1.20} \cdot 10^{-4}$	$1.57_{\pm .27} \cdot 10^{-3}$
		std.	$2.54_{\pm .10} \cdot 10^{-3}$	$9.71_{\pm 1.40} \cdot 10^{-3}$	$6.72_{\pm .91} \cdot 10^{-3}$	$\mathbf{4.50_{\pm .30} \cdot 10^{-4}}$	$7.85_{\pm 2.40} \cdot 10^{-4}$
	$\star$ 25	mean	$5.21_{\pm .84} \cdot 10^{-4}$	$\mathbf{1.11_{\pm .39} \cdot 10^{-4}}$	$1.15_{\pm .046} \cdot 10^{-3}$	$1.99_{\pm .066} \cdot 10^{-2}$	$2.16_{\pm .34} \cdot 10^{-2}$
		std.	$4.14_{\pm .002} \cdot 10^{-2}$	$3.81_{\pm .049} \cdot 10^{-2}$	$\mathit{1.31_{\pm .0030} \cdot 10^{-2}}$	$\mathbf{1.34_{\pm .013} \cdot 10^{-2}}$	$4.07_{\pm .057} \cdot 10^{-2}$

Table 2. Mean absolute error calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression ($N = 16$). DAIS$_0$ (16) outperforms other methods on 3/5 datasets (details in Section 5.3). Results are averaged over 3 runs.

	d	MAE	IWVI	IWVI$_{\text{SIR}}$	DAIS$_0$ (16)	DAIS (16)	MSC$_{8c}$
sonar	61	mean	$\boldsymbol{8.66}_{\pm.075} \cdot \boldsymbol{10^{-2}}$	$8.86_{\pm.17} \cdot 10^{-2}$	$\boldsymbol{8.58}_{\pm.19} \cdot \boldsymbol{10^{-2}}$	$1.10_{\pm.089} \cdot 10^{-1}$	$1.33_{\pm.081} \cdot 10^{-1}$
		std.	$7.95_{\pm.089} \cdot 10^{-2}$	$8.26_{\pm.071} \cdot 10^{-2}$	$\boldsymbol{4.27}_{\pm.13} \cdot \boldsymbol{10^{-2}}$	$6.23_{\pm.13} \cdot 10^{-2}$	$7.26_{\pm.16} \cdot 10^{-2}$
ionosphere	35	mean	$\boldsymbol{4.39}_{\pm.098} \cdot \boldsymbol{10^{-2}}$	$7.18_{\pm.14} \cdot 10^{-2}$	$\boldsymbol{4.34}_{\pm.023} \cdot \boldsymbol{10^{-2}}$	$1.06_{\pm.10} \cdot 10^{-1}$	$2.14_{\pm.050} \cdot 10^{-1}$
		std.	$4.73_{\pm.048} \cdot 10^{-2}$	$8.34_{\pm.070} \cdot 10^{-2}$	$\boldsymbol{3.25}_{\pm.36} \cdot \boldsymbol{10^{-2}}$	$7.64_{\pm.49} \cdot 10^{-2}$	$8.55_{\pm.024} \cdot 10^{-2}$
heart disease	16	mean	$\boldsymbol{2.16}_{\pm.23} \cdot \boldsymbol{10^{-2}}$	$\boldsymbol{2.30}_{\pm.15} \cdot \boldsymbol{10^{-2}}$	$\boldsymbol{2.26}_{\pm.26} \cdot \boldsymbol{10^{-2}}$	$\boldsymbol{2.22}_{\pm.23} \cdot \boldsymbol{10^{-2}}$	$3.37_{\pm.032} \cdot 10^{-1}$
		std.	$2.59_{\pm.072} \cdot 10^{-2}$	$2.67_{\pm.060} \cdot 10^{-2}$	$\boldsymbol{1.08}_{\pm.28} \cdot \boldsymbol{10^{-2}}$	$\boldsymbol{1.18}_{\pm.29} \cdot \boldsymbol{10^{-2}}$	$2.81_{\pm.10} \cdot 10^{-2}$
heart attack	14	mean	$5.05_{\pm.75} \cdot 10^{-2}$	$4.81_{\pm.57} \cdot 10^{-2}$	$\boldsymbol{4.68}_{\pm.16} \cdot \boldsymbol{10^{-2}}$	$\boldsymbol{4.67}_{\pm.067} \cdot \boldsymbol{10^{-2}}$	$1.75_{\pm.011} \cdot 10^{-1}$
		std.	$\boldsymbol{2.80}_{\pm.15} \cdot \boldsymbol{10^{-2}}$	$3.11_{\pm.14} \cdot 10^{-2}$	$3.35_{\pm.11} \cdot 10^{-2}$	$3.28_{\pm.094} \cdot 10^{-2}$	$7.41_{\pm.35} \cdot 10^{-2}$
loan	12	mean	$\boldsymbol{1.67}_{\pm.76} \cdot \boldsymbol{10^{-2}}$	$\boldsymbol{2.38}_{\pm.47} \cdot \boldsymbol{10^{-2}}$	$\boldsymbol{2.02}_{\pm0.079} \cdot \boldsymbol{10^{-2}}$	$\boldsymbol{2.06}_{\pm.26} \cdot \boldsymbol{10^{-2}}$	$1.76_{\pm.0089} \cdot 10^{-1}$
		std.	$\boldsymbol{9.23}_{\pm.68} \cdot \boldsymbol{10^{-3}}$	$1.14_{\pm.077} \cdot 10^{-2}$	$1.28_{\pm.060} \cdot 10^{-2}$	$1.32_{\pm.13} \cdot 10^{-2}$	$5.08_{\pm.16} \cdot 10^{-2}$

parable to the other methods, it falls behind on most datasets. The results are consistent with the previous experiments and Theorem 4.1. More results can be found in Appendix D.

Figure 6 depicts standard deviations of HMC samples (black) and the learned standard deviations of the compact methods on the *ionosphere* dataset (IWVI: red, MSC$_{8c}$: yellow, DAIS$_0$: blue, increasing opacity corresponds to increasing $K \in \{2, 4, 8, 16\}$ for training). We use $N = 16$ particles for training. We find that DAIS$_0$ outperforms IWVI and MSC$_{8c}$ for all N (see also Table 2). For a larger N, the differences between IWVI and DAIS$_0$ are visually indistinct. MSC$_{8c}$ seems to overestimate some variances.

6. Conclusion

In this work, we investigate the initial distribution of DAIS. We show theoretically that, for many transition steps, it minimizes the Jensen-Shannon divergence (symmetrized KL divergence) to the target distribution. Therefore, by using q_0 as an approximate posterior (a method that we call DAIS$_0$), we expect it to be less prone to underestimating variances (compared to VI) and easier to optimize in higher-dimensional settings (compared to MSC). Additionally, we argue that q_0 is an explicit and compact representation of the target distribution, which provides advantages over sampling based methods for various downstream tasks. In experiments on both synthetic and real-world data, we verify that DAIS$_0$ indeed often fulfills our expectations by finding distributions with more accurate variances in higher dimensions (compared to other compact and sampling based methods). While DAIS$_0$ is more expensive than VI at training time, inference with DAIS$_0$ is just as expensive as VI but it is significantly cheaper than DAIS.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Software and Data

The datasets that we run experiments on are all either publicly available or can be generated by code. We release the code to run all experiments at `https://github.com/jzenn/DAIS0`.

References

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 2017.

Andrieu, C., Lee, A., and Vihola, M. Uniform ergodicity of the iterated conditional smc and geometric ergodicity of

particle gibbs samplers. *Bernoulli*, pp. 842–872, 2018.

Bamler, R., Zhang, C., Opper, M., and Mandt, S. Perturbative black box variational inference. In *Advances in Neural Information Processing Systems*, 2017.

Beal, M. J. and Ghahramani, Z. The variational bayesian em algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, 7, 2003.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112, 2017.

Brekelmans, R., Huang, S., Ghassemi, M., Steeg, G. V., Grosse, R. B., and Makhzani, A. Improving mutual information estimation with annealed and energy-based bounds. In *International Conference on Learning Representations*, 2022.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

Caterini, A. L., Doucet, A., and Sejdinovic, D. Hamiltonian variational auto-encoder. In *Advances in Neural Information Processing Systems*, 2018.

Chatterjee, S. and Diaconis, P. The sample size required in importance sampling. *The Annals of Applied Probability*, 2018.

Cremer, C., Morris, Q., and Duvenaud, D. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

Crooks, G. E. Measuring thermodynamic length. *Physical Review Letters*, 99(10):100602, 2007.

Del Moral, P., Doucet, A., and Jasra, A. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68, 2006.

Dhaka, A. K., Catalina, A., Welandawe, M., Andersen, M. R., Huggins, J., and Vehtari, A. Challenges and opportunities in high dimensional variational inference. In *Advances in Neural Information Processing Systems*, 2021.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, 2017.

Domke, J. and Sheldon, D. R. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*, 2018.

Doucet, A., De Freitas, N., and Gordon, N. An introduction to sequential monte carlo methods. *Sequential Monte Carlo methods in practice*, 2001.

Doucet, A., Grathwohl, W. S., Matthews, A. G. D. G., and Strathmann, H. Score-based diffusion meets annealed importance sampling. In *Advances in Neural Information Processing Systems*, 2022.

Geffner, T. and Domke, J. Mcmc variational inference via uncorrected hamiltonian annealing. In *Advances in Neural Information Processing Systems*, 2021.

Geffner, T. and Domke, J. Langevin diffusion variational inference. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Gelman, A. and Meng, X.-L. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 1998.

Gorman, R. P. and Sejnowski, T. J. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural networks*, 1, 1988.

Grosse, R. B., Maddison, C. J., and Salakhutdinov, R. R. Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems*, 2013.

Hernandez-Lobato, J., Li, Y., Rowland, M., Bui, T., Hernández-Lobato, D., and Turner, R. Black-box alpha divergence minimization. In *International Conference on Machine Learning*, 2016.

Kelly, M., Longjohn, R., and Nottingham, K. The uci machine learning repository. URL https://archive.ics.uci.edu.

Kim, K., Oh, J., Gardner, J., Dieng, A. B., and Kim, H. Markov chain score ascent: A unifying framework of variational inference with markovian gradients. In *Advances in Neural Information Processing Systems*, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Knuth, K. H., Habeck, M., Malakar, N. K., Mubeen, A. M., and Placek, B. Bayesian evidence and model selection. *Digital Signal Processing*, 47, 2015.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 1951.

Li, Y. and Turner, R. E. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, 2016.

Maddison, C. J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, 2017.

Masrani, V., Le, T. A., and Wood, F. The thermodynamic variational objective. In *Advances in Neural Information Processing Systems*, 2019.

Murphy, K. P. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023.

Naesseth, C., Lindsten, F., and Blei, D. Markovian score climbing: Variational inference with kl (p—— q). In *Advances in Neural Information Processing Systems*, 2020.

Neal, R. M. Annealed importance sampling. *Statistics and Computing*, 11, 2001.

Neal, R. M. Hamiltonian importance sampling. In *Banff International Research Station (BIRS) Workshop on Mathematical Issues in Molecular Dynamics*, 2005.

Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. Variational continual learning. In *International Conference on Learning Representations*, 2018.

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, 2014.

Ruiz, F. and Titsias, M. A contrastive divergence for combining variational inference and mcmc. In *International Conference on Machine Learning*, 2019.

Salimans, T., Kingma, D., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, 2015.

Sigillito, V. G., Wing, S. P., Hutton, L. V., and Baker, K. B. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10, 1989.

Tan, Z. and Bamler, R. Post-Training Neural Network Compression With Variational Bayesian Quantization. In *Advances in Neural Information Processing Systems*, Workshop on Challenges in Deploying and Monitoring Machine Learning Systems, 2022.

Thin, A., Kotelevskii, N., Doucet, A., Durmus, A., Moulines, E., and Panov, M. Monte carlo variational auto-encoders. In *International Conference on Machine Learning*, 2021.

Wan, N., Li, D., and Hovakimyan, N. F-divergence variational inference. In *Advances in Neural Information Processing Systems*, 2020.

Williams, C. K. and Rasmussen, C. E. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Wolf, C., Karl, M., and van der Smagt, P. Variational inference with hamiltonian monte carlo. *arXiv preprint arXiv:1609.08203*, 2016.

Xiao, T. Z., Liu, W., and Bamler, R. A compact representation for bayesian neural networks by removing permutation symmetry. In *UniReps: the First Workshop on Unifying Representations in Neural Models*, 2023.

Yang, Y., Bamler, R., and Mandt, S. Variational bayesian quantization. In *International Conference on Machine Learning*, 2020.

Zenn, J. and Bamler, R. Resampling gradients vanish in differentiable sequential monte carlo samplers. In *The First Tiny Papers Track at International Conference on Learning Representations*, 2023.

Zhang, G., Hsu, K., Li, J., Finn, C., and Grosse, R. B. Differentiable annealed importance sampling and the perils of gradient noise. In *Advances in Neural Information Processing Systems*, 2021.

A. Proof of the Main Theoretical Result

We prove Theorem 4.1 of the main paper. For the reader's convenience, we restate the theorem below.

Theorem 4.1. *We assume that* $\mathrm{support}(q_0) \supset \mathrm{support}(f)$, *perfect transitions between two annealing distributions and equally spaced* β_k, *i.e.,* $\beta_k = k/K$. *Then, for large K and $N = 1$, the gap* $\Delta_{\mathrm{AIS}}^{1,K} := \log Z - \mathrm{ELBO}_{\mathrm{AIS}}^{1,K}(f_{\mathrm{AIS}}, q_{\mathrm{AIS}})$ *is a divergence between the initial distribution q_0 and the target distribution f/Z. Up to corrections of order $\mathcal{O}(1/K^3)$, this divergence is proportional to the Jensen-Shannon divergence (symmetrized KL-divergence),*

$$\Delta_{\mathrm{AIS}}^{1,K} = \frac{1}{K}\left(\frac{1}{2}D_{\mathrm{KL}}\big(f(\boldsymbol{z})/Z \,\|\, q_0(\boldsymbol{z})\big) + \frac{1}{2}D_{\mathrm{KL}}\big(q_0(\boldsymbol{z}) \,\|\, f(\boldsymbol{z})/Z\big)\right) + \mathcal{O}\big(1/K^3\big). \tag{15}$$

Proof. We start from Equation (8) and Equation (11) of the main text, which we reproduce here for convenience,

$$\log Z \geq \mathbb{E}_{\boldsymbol{z}_{0:K}^{(1:N)} \sim q_{\mathrm{AIS}}}\left[\log\left(w_{\mathrm{AIS}}\big(\boldsymbol{z}_{0:K}^{(1:N)}\big)\right)\right] =: \mathrm{ELBO}_{\mathrm{AIS}}^{N,K}(f_{\mathrm{AIS}}, q_{\mathrm{AIS}}) \tag{16}$$

and

$$w_{\mathrm{AIS}}^{N,K}\big(\boldsymbol{z}_{0:K}^{(1:N)}\big) = \frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\frac{\gamma_{\beta_k}\big(\boldsymbol{z}_{k-1}^{(i)}\big)}{\gamma_{\beta_{k-1}}\big(\boldsymbol{z}_{k-1}^{(i)}\big)} \tag{17}$$

Equation (17) holds for DAIS if we assume perfect transitions (Grosse et al., 2013). Inserting $\gamma_{\beta_k}(\boldsymbol{z}) = Z_{\beta_k}\pi_{\beta_k}(\boldsymbol{z})$, $Z_{\beta_K} = Z_1 = Z$, and $Z_{\beta_0} = Z_0 = 1$, we obtain for $N = 1$ (Grosse et al., 2013),

$$\Delta_{\mathrm{AIS}}^{1,K} = \log Z - \mathbb{E}_{\boldsymbol{z}_{0:K} \sim q_{\mathrm{AIS}}}\left[\log Z_{\beta_K} - \log Z_{\beta_0} + \sum_{k=1}^{K}\log\frac{\pi_{\beta_k}(\boldsymbol{z}_{k-1})}{\pi_{\beta_{k-1}}(\boldsymbol{z}_{k-1})}\right] = \sum_{k=1}^{K}D_{\mathrm{KL}}(\pi_{\beta_{k-1}} \,\|\, \pi_{\beta_k}). \tag{18}$$

As stated in the main text, $\Delta_{\mathrm{AIS}}^{1,K}$ is a divergence since it is a sum of divergences and since $\Delta_{\mathrm{AIS}}^{1,K} = 0$ for $q_0 = f/Z$ as, in this case, $\pi_{\beta_k} = q_0 \; \forall k \in \{0, \dots, K\}$ since, by definition,

$$\pi_{\beta_k}(\boldsymbol{z}) := \frac{\gamma_{\beta_k}(\boldsymbol{z})}{Z_{\beta_k}} \qquad \text{with} \qquad \gamma_{\beta_k}(\boldsymbol{z}) := q_0(\boldsymbol{z})\left(\frac{f(\boldsymbol{z})}{q_0(\boldsymbol{z})}\right)^{\beta_k} \qquad \text{and} \qquad Z_{\beta_k} = \int \gamma_{\beta_k}(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}. \tag{19}$$

We now show Equation (15). For each term on the right-hand side of Equation (18), we have

$$D_{\mathrm{KL}}(\pi_{\beta_{k-1}} \,\|\, \pi_{\beta_k}) = h_{k-1}(\beta_k) - h_{k-1}(\beta_{k-1}), \qquad \text{where} \qquad h_{k-1}(\eta) := -\mathbb{E}_{\pi_{\beta_{k-1}}(\boldsymbol{z})}[\log \pi_\eta(\boldsymbol{z})]. \tag{20}$$

Using Taylor's theorem, we can express

$$h_{k-1}(\beta_k) - h_{k-1}(\beta_{k-1}) = (\beta_k - \beta_{k-1})h_{k-1}'(\beta_k) + R_1(\beta_k), \tag{21}$$

where primes denote derivatives. We use the Lagrange from for the remainder

$$R_1(\beta_k) = \frac{1}{2}(\beta_k - \beta_{k-1})^2 h_{k-1}''(\eta_{k-1}), \tag{22}$$

for some $\eta_{k-1} \in [\beta_{k-1}, \beta_k]$.

We can show that the first term of the right-hand side of Equation (21) is zero by

$$\begin{aligned}
h_{k-1}'(\beta_{k-1}) &= -\nabla_\eta\left(\mathbb{E}_{\pi_{\beta_{k-1}}(\boldsymbol{z})}[\log \pi_\eta(\boldsymbol{z})]\right)_{\eta=\beta_{k-1}} \\
&= -\int \pi_{\beta_{k-1}}(\boldsymbol{z})\frac{\nabla_\eta\left(\pi_\eta(\boldsymbol{z})\right)_{\eta=\beta_{k-1}}}{\pi_{\beta_{k-1}}(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z} = -\nabla_\eta\left(\int \pi_\eta(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}\right)_{\eta=\beta_{k-1}} = -\nabla_\eta(1) = 0.
\end{aligned} \tag{23}$$

Therefore, only the second term on the right-hand side of Equation (21) contributes.

12

From Equation (19), we combine the leftmost and the middle equation and arrive at

$$\log \pi_\eta(\boldsymbol{z}) = \log q_0(\boldsymbol{z}) - \eta\big(\log f(\boldsymbol{z}) - \log q_0(\boldsymbol{z})\big) - \log Z_\eta. \tag{24}$$

Inserting Equation (24) into Equation (20) and taking the second derivative we get

$$h''_{k-1}(\eta_{k-1}) = \partial^2 \log Z_{\eta_{k-1}} / \partial\eta_{k-1}^2. \tag{25}$$

Inserting Equation (23) and Equation (25) into Equation (21) and the result into Equation (18), we find

$$\Delta_{\mathrm{AIS}}^{1,K} = \sum_{k=1}^{K} D_{\mathrm{KL}}(\pi_{\beta_{k-1}} \,\|\, \pi_{\beta_k}) = \frac{1}{2K}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\partial^2 \log Z_{\eta_{k-1}}}{\partial\eta_{k-1}^2}\right). \tag{26}$$

Here, the term in the parentheses approximates an integral (since $\eta_{k-1} \in [\beta_{k-1}, \beta_k] = \left[\frac{k-1}{K}, \frac{k}{K}\right] \forall k$). Using that the error for such numerical integration scales as $\mathcal{O}(1/K^2)$, we thus find

$$\Delta_{\mathrm{AIS}}^{1,K} = \frac{1}{2K}\left(\int_0^1 \frac{\partial^2 \log Z_\eta}{\partial^2 \eta}\,\mathrm{d}\eta + \mathcal{O}(1/K^2)\right) \tag{27}$$

$$= \frac{1}{2K}\left(\left(\frac{\partial \log Z_\eta}{\partial\eta}\right)_{\eta=1} - \left(\frac{\partial \log Z_\eta}{\partial\eta}\right)_{\eta=0}\right) + \mathcal{O}(1/K^3). \tag{28}$$

Thus, for large K, only the boundary terms remain. We calculate them by inserting the definitions of Z_η from Equation (10) and using the relation $\partial x^\eta / \partial\eta = x^\eta \log x$. We find

$$\frac{\partial \log Z_\eta}{\partial\eta} = \frac{1}{Z_\eta}\int q_0(\boldsymbol{z})\left(\frac{f(\boldsymbol{z})}{q_0(\boldsymbol{z})}\right)^\eta \log\frac{f(\boldsymbol{z})}{q_0(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}. \tag{29}$$

Thus, recalling that $Z_0 = 1$ and $Z_1 = Z$ are the normalization constants of q_0 and f, respectively,

$$\left(\frac{\partial \log Z_\eta}{\partial\eta}\right)_{\eta=1} = \mathbb{E}_{\boldsymbol{z}\sim f(\boldsymbol{z})/Z}\left[\log\frac{f(\boldsymbol{z})}{q_0(\boldsymbol{z})}\right] = D_{\mathrm{KL}}\big(f(\boldsymbol{z})/Z \,\|\, q_0(\boldsymbol{z})\big); \tag{30}$$

$$-\left(\frac{\partial \log Z_\eta}{\partial\eta}\right)_{\eta=0} = \mathbb{E}_{\boldsymbol{z}\sim q_0(\boldsymbol{z})}\left[\log\frac{q_0(\boldsymbol{z})}{f(\boldsymbol{z})}\right] = D_{\mathrm{KL}}\big(q_0(\boldsymbol{z}) \,\|\, f(\boldsymbol{z})/Z\big). \tag{31}$$

Inserting Equations (30) and (31) into Equation (28) leads to the claim in Equation (15). $\qquad\square$

Statement for $N > 1$. Starting from Equation (17) we obtain for the gap with general N,

$$\Delta_{\mathrm{AIS}}^{N,K} = -\mathbb{E}_{\boldsymbol{z}_{0:K}^{(1:N)}\sim q_{\mathrm{AIS}}}\left[\log\left(\frac{1}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\frac{\gamma_{\beta_k}(\boldsymbol{z}_{k-1}^{(i)})}{\gamma_{\beta_{k-1}}(\boldsymbol{z}_{k-1}^{(i)})}\right)\right]. \tag{32}$$

Due to the sum over N, the right-hand side can now no longer be separated into a sum of K terms. However, we can still bound the right-hand side by pulling the sum out of the logarithm using Jensen's inequality, resulting in

$$\Delta_{\mathrm{AIS}}^{N,K} \leq \Delta_{\mathrm{AIS}}^{1,K}, \tag{33}$$

which implies

$$\lim_{K\to\infty}\Delta_{\mathrm{AIS}}^{N,K} \leq \frac{1}{K}D_{\mathrm{JS}}(q_0, f/Z) + \mathcal{O}(1/K^3), \tag{34}$$

where D_{JS} denotes the Jensen-Shannon divergence. This shows that the bound for $N > 1$ can be upper-bounded by the Jensen-Shannon divergence but does not give additional insights on, e.g., symmetry. For $N = 1$ the inequality is an equality.

B. Multidimensional Bimodal Target Distributions

We provide further details for the experiment on multidimensional bimodal Gaussian target distributions. The experiment is discussed in Section 5.1 of the main text. We use an Intel XEON CPU E5-2650 v4 for running the small-scale experiments and a single NVIDIA GeForce GTX 1080 Ti for the large-scale experiments.

The model p is defined as follows

$$p(\boldsymbol{z}) = \frac{1}{2}\mathcal{N}(\boldsymbol{z}; \mathbf{0}, 0.25^2 \boldsymbol{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{z}; \mathbf{1}, 0.25^2 \boldsymbol{I}). \tag{35}$$

We initialize the variational distributions of VI, IWVI and DAIS_0 with

$$q_\star(\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{z}; \frac{1}{2}\mathbf{1}, \boldsymbol{I}\right), \tag{36}$$

where the mean and the diagonal of the covariance matrix are learnable parameters. We train the models for $7,500$ iterations with the Adam optimizer (Kingma & Ba, 2015) and a learning rate of 10^{-2}.

DAIS utilizes the parameterization of Zhang et al. (2021). In addition to the parameters of q_0, we learn the scale c of the mass matrix $\boldsymbol{M} = c\boldsymbol{I}$, the annealing schedule $(\beta_1, \ldots, \beta_{K-1})$, and the step widths of the sampler.

MSC is trained for a comparable number of iterations with a learning rate of 10^{-4}.

Classification Into Mode-Seeking "s" and Mass-Covering "c" and Undecidable "u".   To get a first classification on "c" or "s", we measure the distance of the mean of the variational distribution to both modes of the target distribution, and to their mid point $(0.5, \ldots, 0.5)$. It turns out that this criterion clearly identifies in almost all cases whether a distribution is mode-seeking (see Figure 7). For cases where we do not clearly identify whether a distribution is mass-covering or mode-seeking, we use the letter "u", short for "undecidable". Afterwards, we manually verified the classifications by making a plot similar to Figure 4 for each "square" of Figure 3.

projection of mean onto the the $\mathbf{1}$-direction, scaled such that the two modes of the target distribution are at $\mathbf{0}$ and $\mathbf{1}$
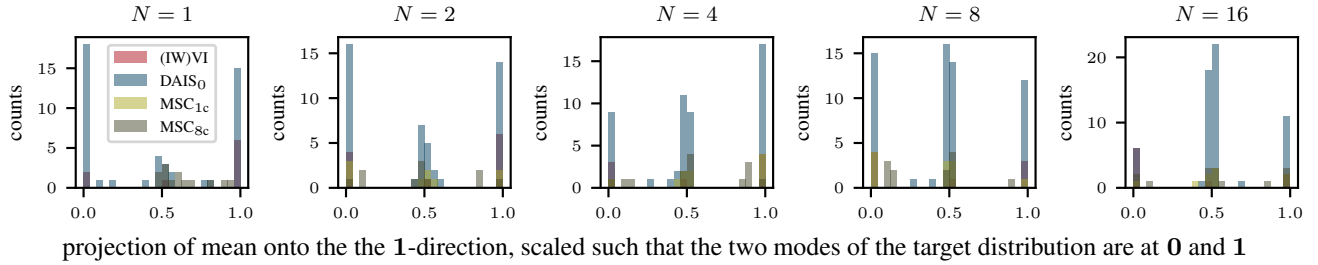
Figure 7. Distance of the mean of the variational distribution to both modes of p and its mid point. Distances are mostly 0.0, 1.0, or 0.5 for (IW)VI, DAIS_0, and MSC_{1c} indicating that the mean of the variational distribution falls on one of the two modes or the mid point. For MSC_{8c} there are more outliers.

C. Gaussian Process Regression

We provide further details on model, the joint distribution we plot in the main text, and quantitative results of the GP regression experiment on synthetic data discussed in Section 5.2 of the main text. We use an Intel XEON CPU E5-2650 v4 for running the small-scale experiments and a single NVIDIA GeForce GTX 1080 Ti for the large-scale experiments.

C.1. Model

The kernels, called RBF_1 and RBF_2 in the main text, are instances of the following RBF kernel with different lengthscales $\rho_1 = 0.8$ and $\rho_2 = 3.0$ and

$$k(t, s) = \exp\left(-\frac{(t - s)^2}{2\rho^2}\right). \tag{37}$$

We discretize the domain $[0, 10]$ on 75 points. We initialize the variational distributions of VI, IWVI and DAIS_0 with standard normal distributions.

All models are trained for $50,000$ iterations with the Adam optimizer and a learning rate of 10^{-3}. DAIS utilizes the parameterization of Zhang et al. (2021). In addition to the parameters of q_0, we learn the diagonal $\boldsymbol{d}$ of the mass matrix $\boldsymbol{M} = \text{diag}(\boldsymbol{d})$, the annealing schedule $(\beta_1, \ldots, \beta_{K-1})$, and the step widths of the sampler.

C.2. Joint Distribution

This section provides more background on how we produce the plots of the GP regression experiment in the main text.

Let $\boldsymbol{f}$ denote the latent process of interest. Let $\boldsymbol{f}_o$ denote the part of $\boldsymbol{f}$ that we have data on and let $\boldsymbol{f}_u$ denote the remaining part. The prior on $\boldsymbol{f}$ can then be written as follows.

$$p(\boldsymbol{f}) = p\left(\begin{pmatrix}\boldsymbol{f}_o \\ \boldsymbol{f}_u\end{pmatrix}\right) = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{f}_o \\ \boldsymbol{f}_u\end{pmatrix}; \begin{pmatrix}\boldsymbol{m}_o \\ \boldsymbol{m}_u\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{o,o} & \boldsymbol{\Sigma}_{o,u} \\ \boldsymbol{\Sigma}_{u,o} & \boldsymbol{\Sigma}_{u,u}\end{pmatrix}\right) \tag{38}$$

$$p(\boldsymbol{f}_o)p(\boldsymbol{f}_u \mid \boldsymbol{f}_o) = \mathcal{N}(\boldsymbol{f}_o; \boldsymbol{m}_o, \boldsymbol{\Sigma}_{o,o})\mathcal{N}(\boldsymbol{f}_u; \boldsymbol{m}_u + \boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1}(\boldsymbol{f}_o - \boldsymbol{m}_o), \boldsymbol{\Sigma}_{u,u} - \boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1}\boldsymbol{\Sigma}_{o,u}) \tag{39}$$

We model the data $\boldsymbol{y}$ with a Gaussian likelihood model with fixed variance $\sigma^2 = 0.1$.

$$p(\boldsymbol{f}_o \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}; f_o, \sigma^2 \boldsymbol{I}) \tag{40}$$

If we condition on data $\boldsymbol{y}$ we compute the joint as follows.

$$p(\boldsymbol{f} \mid \boldsymbol{y}) = p(\boldsymbol{f}_o \mid \boldsymbol{y})p(\boldsymbol{f}_u \mid \boldsymbol{f}_o, \boldsymbol{y}) = p(\boldsymbol{f}_o \mid \boldsymbol{y})p(\boldsymbol{f}_u \mid \boldsymbol{f}_o) \tag{41}$$

$$= \mathcal{N}\left(\begin{pmatrix}\boldsymbol{f}_o \\ \boldsymbol{f}_u\end{pmatrix}; \begin{pmatrix}\boldsymbol{m}^+ \\ \boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1}\boldsymbol{m}^+ + \boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1}\boldsymbol{m}_o + \boldsymbol{m}_u\end{pmatrix}, \right.$$
$$\left. \begin{pmatrix}\boldsymbol{\Sigma}^+ & \boldsymbol{\Sigma}^+(\boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1})^\top \\ \boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1}\boldsymbol{\Sigma}^+ & (\boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1}\boldsymbol{\Sigma}^+(\boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1})^\top + \boldsymbol{\Sigma}_{u,u} - \boldsymbol{\Sigma}_{u,o}\boldsymbol{\Sigma}_{o,o}^{-1}\boldsymbol{\Sigma}_{o,u}\end{pmatrix}\right) \tag{42}$$

where $\boldsymbol{m}^+$ and $\boldsymbol{\Sigma}^+$ are either (a) the analytically inferred mean and variance of the posterior Gaussian given the data or (b) the learned variational approximation given data.

For the first case, we get

$$\boldsymbol{m}^+ = \boldsymbol{m}_o + \boldsymbol{\Sigma}_{o,o}(\boldsymbol{\Sigma}_{o,o} + \sigma^2 \boldsymbol{I})^{-1}(\boldsymbol{x} - \boldsymbol{m}_o) \quad \text{and} \tag{43}$$

$$\boldsymbol{\Sigma}^+ = \boldsymbol{\Sigma}_{o,o} - \boldsymbol{\Sigma}_{o,o}(\boldsymbol{\Sigma}_{o,o} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_{o,o}. \tag{44}$$

In the second case, we model the Gaussian distribution by a mean vector and a variance vector

$$\boldsymbol{m}^+ = \check{\boldsymbol{m}} \quad \text{and} \quad \boldsymbol{\Sigma}^+ = \text{diag}(\check{\boldsymbol{s}}). \tag{45}$$

We plot the shaded region by calculating Equation (42) with

$$\boldsymbol{\Sigma}^+ = \text{diag}(\text{diag}(\boldsymbol{\Sigma}_{o,o} - \boldsymbol{\Sigma}_{o,o}(\boldsymbol{\Sigma}_{o,o} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_{o,o})). \tag{46}$$

C.3. Complementary Quantitative Results

provide additional results on our the experiment.

Table 3. MAE (compared to analytic solution) of mean and standard deviation of GP regression. More details are provided in Section 5.2. Results are averaged over 3 runs. $\mathrm{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	IWVI	DAIS$_0$ (2)	DAIS$_0$ (4)	DAIS$_0$ (8)	DAIS$_0$ (16)
					$N=1$		
RBF_1	10	mean	$2.29 \cdot 10^{-4}{}_{\pm 3.9 \cdot 10^{-5}}$	$2.30 \cdot 10^{-4}{}_{\pm 7.9 \cdot 10^{-5}}$	$6.65 \cdot 10^{-4}{}_{\pm 9.5 \cdot 10^{-5}}$	$1.30 \cdot 10^{-3}{}_{\pm 6.9 \cdot 10^{-5}}$	$6.11 \cdot 10^{-4}{}_{\pm 2.2 \cdot 10^{-4}}$
		std.	$5.16 \cdot 10^{-2}{}_{\pm 1.7 \cdot 10^{-5}}$	$4.84 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-5}}$	$3.97 \cdot 10^{-2}{}_{\pm 2.1 \cdot 10^{-5}}$	$3.18 \cdot 10^{-2}{}_{\pm 3.0 \cdot 10^{-5}}$	$2.51 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-5}}$
	25	mean	$1.73 \cdot 10^{-4}{}_{\pm 1.3 \cdot 10^{-5}}$	$1.58 \cdot 10^{-4}{}_{\pm 2.2 \cdot 10^{-5}}$	$2.46 \cdot 10^{-4}{}_{\pm 4.8 \cdot 10^{-5}}$	$4.95 \cdot 10^{-4}{}_{\pm 9.0 \cdot 10^{-5}}$	$5.53 \cdot 10^{-4}{}_{\pm 6.5 \cdot 10^{-5}}$
		std.	$4.03 \cdot 10^{-2}{}_{\pm 2.8 \cdot 10^{-6}}$	$3.97 \cdot 10^{-2}{}_{\pm 5.1 \cdot 10^{-6}}$	$3.79 \cdot 10^{-2}{}_{\pm 4.0 \cdot 10^{-6}}$	$3.51 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-4}}$	$2.61 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-5}}$
RBF_2	10	mean	$1.08 \cdot 10^{-3}{}_{\pm 2.6 \cdot 10^{-4}}$	$9.84 \cdot 10^{-4}{}_{\pm 1.5 \cdot 10^{-4}}$	$1.39 \cdot 10^{-3}{}_{\pm 9.5 \cdot 10^{-5}}$	$1.26 \cdot 10^{-3}{}_{\pm 2.9 \cdot 10^{-4}}$	$1.57 \cdot 10^{-3}{}_{\pm 5.8 \cdot 10^{-4}}$
		std.	$1.40 \cdot 10^{-2}{}_{\pm 8.7 \cdot 10^{-5}}$	$1.00 \cdot 10^{-2}{}_{\pm 7.2 \cdot 10^{-5}}$	$5.71 \cdot 10^{-3}{}_{\pm 4.1 \cdot 10^{-5}}$	$5.53 \cdot 10^{-3}{}_{\pm 9.5 \cdot 10^{-5}}$	$5.45 \cdot 10^{-3}{}_{\pm 9.5 \cdot 10^{-5}}$
	25	mean	$3.00 \cdot 10^{-4}{}_{\pm 8.3 \cdot 10^{-5}}$	$4.76 \cdot 10^{-4}{}_{\pm 7.0 \cdot 10^{-5}}$	$6.65 \cdot 10^{-4}{}_{\pm 1.7 \cdot 10^{-4}}$	$8.92 \cdot 10^{-4}{}_{\pm 6.5 \cdot 10^{-5}}$	$9.36 \cdot 10^{-4}{}_{\pm 1.7 \cdot 10^{-4}}$
		std.	$4.87 \cdot 10^{-2}{}_{\pm 2.6 \cdot 10^{-5}}$	$4.33 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-5}}$	$3.55 \cdot 10^{-2}{}_{\pm 9.5 \cdot 10^{-6}}$	$2.88 \cdot 10^{-2}{}_{\pm 4.4 \cdot 10^{-5}}$	$2.27 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-5}}$

Table 4. MAE (compared to analytic solution) of mean and standard deviation of GP regression. More details are provided in Section 5.2. Results are averaged over 3 runs. $\mathrm{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	IWVI$_{\mathrm{SIR}}$	DAIS (2)	DAIS (4)	DAIS (8)	DAIS (16)
					$N=1$		
RBF_1	10	mean	$1.13 \cdot 10^{-3}{}_{\pm 5.0 \cdot 10^{-5}}$	$4.64 \cdot 10^{-3}{}_{\pm 1.9 \cdot 10^{-4}}$	$2.19 \cdot 10^{-2}{}_{\pm 3.1 \cdot 10^{-4}}$	$7.55 \cdot 10^{-3}{}_{\pm 4.5 \cdot 10^{-4}}$	$1.16 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-4}}$
		std.	$1.48 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-4}}$	$4.72 \cdot 10^{-2}{}_{\pm 9.2 \cdot 10^{-5}}$	$2.40 \cdot 10^{-2}{}_{\pm 4.1 \cdot 10^{-5}}$	$6.95 \cdot 10^{-3}{}_{\pm 2.3 \cdot 10^{-4}}$	$2.25 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-4}}$
	25	mean	$2.44 \cdot 10^{-4}{}_{\pm 1.4 \cdot 10^{-5}}$	$3.20 \cdot 10^{-3}{}_{\pm 7.3 \cdot 10^{-5}}$	$2.29 \cdot 10^{-3}{}_{\pm 2.9 \cdot 10^{-4}}$	$5.78 \cdot 10^{-3}{}_{\pm 2.7 \cdot 10^{-3}}$	$4.41 \cdot 10^{-3}{}_{\pm 3.8 \cdot 10^{-4}}$
		std.	$2.75 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-4}}$	$3.97 \cdot 10^{-2}{}_{\pm 5.1 \cdot 10^{-5}}$	$3.16 \cdot 10^{-2}{}_{\pm 9.0 \cdot 10^{-5}}$	$2.17 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-3}}$	$2.71 \cdot 10^{-2}{}_{\pm 3.6 \cdot 10^{-5}}$
RBF_2	10	mean	$1.54 \cdot 10^{-5}{}_{\pm 5.2 \cdot 10^{-6}}$	$1.15 \cdot 10^{-2}{}_{\pm 4.7 \cdot 10^{-4}}$	$1.90 \cdot 10^{-3}{}_{\pm 3.0 \cdot 10^{-5}}$	$1.87 \cdot 10^{-3}{}_{\pm 4.0 \cdot 10^{-5}}$	$2.78 \cdot 10^{-3}{}_{\pm 9.3 \cdot 10^{-5}}$
		std.	$1.69 \cdot 10^{-3}{}_{\pm 9.4 \cdot 10^{-5}}$	$6.40 \cdot 10^{-3}{}_{\pm 1.1 \cdot 10^{-4}}$	$2.25 \cdot 10^{-3}{}_{\pm 1.6 \cdot 10^{-5}}$	$1.64 \cdot 10^{-3}{}_{\pm 1.4 \cdot 10^{-5}}$	$1.99 \cdot 10^{-3}{}_{\pm 2.0 \cdot 10^{-4}}$
	25	mean	$3.33 \cdot 10^{-4}{}_{\pm 2.2 \cdot 10^{-5}}$	$2.07 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-3}}$	$4.83 \cdot 10^{-3}{}_{\pm 2.4 \cdot 10^{-4}}$	$2.12 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$	$3.14 \cdot 10^{-2}{}_{\pm 3.9 \cdot 10^{-4}}$
		std.	$1.20 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-4}}$	$5.19 \cdot 10^{-2}{}_{\pm 2.0 \cdot 10^{-3}}$	$2.46 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-4}}$	$1.82 \cdot 10^{-2}{}_{\pm 6.2 \cdot 10^{-4}}$	$3.21 \cdot 10^{-2}{}_{\pm 5.5 \cdot 10^{-4}}$

Table 5. MAE (compared to analytic solution) of mean and standard deviation of GP regression. More details are provided in Section 5.2. Results are averaged over 3 runs. $\mathrm{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	IWVI	DAIS$_0$ (2)	DAIS$_0$ (4)	DAIS$_0$ (8)	DAIS$_0$ (16)
					$N=2$		
RBF_1	10	mean	$4.21 \cdot 10^{-4}{}_{\pm 9.6 \cdot 10^{-5}}$	$3.19 \cdot 10^{-4}{}_{\pm 6.8 \cdot 10^{-5}}$	$7.78 \cdot 10^{-4}{}_{\pm 3.4 \cdot 10^{-4}}$	$6.10 \cdot 10^{-4}{}_{\pm 3.6 \cdot 10^{-4}}$	$8.30 \cdot 10^{-4}{}_{\pm 2.2 \cdot 10^{-4}}$
		std.	$4.95 \cdot 10^{-2}{}_{\pm 2.1 \cdot 10^{-5}}$	$4.53 \cdot 10^{-2}{}_{\pm 5.8 \cdot 10^{-6}}$	$3.59 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-5}}$	$2.76 \cdot 10^{-2}{}_{\pm 2.5 \cdot 10^{-5}}$	$1.84 \cdot 10^{-2}{}_{\pm 6.8 \cdot 10^{-5}}$
	25	mean	$1.77 \cdot 10^{-4}{}_{\pm 1.7 \cdot 10^{-5}}$	$2.53 \cdot 10^{-4}{}_{\pm 2.9 \cdot 10^{-5}}$	$4.10 \cdot 10^{-4}{}_{\pm 5.8 \cdot 10^{-5}}$	$3.60 \cdot 10^{-4}{}_{\pm 9.6 \cdot 10^{-5}}$	$6.93 \cdot 10^{-4}{}_{\pm 5.0 \cdot 10^{-5}}$
		std.	$3.98 \cdot 10^{-2}{}_{\pm 4.9 \cdot 10^{-6}}$	$3.89 \cdot 10^{-2}{}_{\pm 4.1 \cdot 10^{-6}}$	$3.75 \cdot 10^{-2}{}_{\pm 3.4 \cdot 10^{-5}}$	$3.39 \cdot 10^{-2}{}_{\pm 7.5 \cdot 10^{-5}}$	$2.12 \cdot 10^{-2}{}_{\pm 4.7 \cdot 10^{-5}}$
RBF_2	10	mean	$1.36 \cdot 10^{-3}{}_{\pm 2.5 \cdot 10^{-4}}$	$1.02 \cdot 10^{-3}{}_{\pm 1.5 \cdot 10^{-4}}$	$2.01 \cdot 10^{-3}{}_{\pm 6.8 \cdot 10^{-4}}$	$1.45 \cdot 10^{-3}{}_{\pm 1.2 \cdot 10^{-4}}$	$1.47 \cdot 10^{-3}{}_{\pm 5.3 \cdot 10^{-5}}$
		std.	$8.80 \cdot 10^{-3}{}_{\pm 6.4 \cdot 10^{-5}}$	$4.07 \cdot 10^{-3}{}_{\pm 7.1 \cdot 10^{-5}}$	$5.10 \cdot 10^{-3}{}_{\pm 4.4 \cdot 10^{-5}}$	$5.23 \cdot 10^{-3}{}_{\pm 5.3 \cdot 10^{-5}}$	$5.47 \cdot 10^{-3}{}_{\pm 9.0 \cdot 10^{-5}}$
	25	mean	$3.20 \cdot 10^{-4}{}_{\pm 4.7 \cdot 10^{-5}}$	$4.68 \cdot 10^{-4}{}_{\pm 6.0 \cdot 10^{-5}}$	$6.18 \cdot 10^{-4}{}_{\pm 8.9 \cdot 10^{-5}}$	$9.13 \cdot 10^{-4}{}_{\pm 7.7 \cdot 10^{-5}}$	$1.05 \cdot 10^{-3}{}_{\pm 1.6 \cdot 10^{-4}}$
		std.	$4.66 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-5}}$	$4.07 \cdot 10^{-2}{}_{\pm 2.1 \cdot 10^{-5}}$	$3.45 \cdot 10^{-2}{}_{\pm 3.0 \cdot 10^{-5}}$	$2.66 \cdot 10^{-2}{}_{\pm 2.3 \cdot 10^{-5}}$	$1.95 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-5}}$

Table 6. MAE (compared to analytic solution) of mean and standard deviation of GP regression. More details are provided in Section 5.2. Results are averaged over 3 runs. $\text{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	IWVI_{SIR}	DAIS (2)	DAIS (4)	DAIS (8)	DAIS (16)
				$N = 2$			
RBF_1	10	mean	$8.62 \cdot 10^{-4}{}_{\pm 4.5 \cdot 10^{-5}}$	$5.69 \cdot 10^{-3}{}_{\pm 3.3 \cdot 10^{-4}}$	$1.67 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$	$1.15 \cdot 10^{-2}{}_{\pm 3.2 \cdot 10^{-4}}$	$1.64 \cdot 10^{-2}{}_{\pm 8.9 \cdot 10^{-4}}$
		std.	$1.12 \cdot 10^{-2}{}_{\pm 4.6 \cdot 10^{-4}}$	$4.44 \cdot 10^{-2}{}_{\pm 5.3 \cdot 10^{-5}}$	$1.96 \cdot 10^{-2}{}_{\pm 3.5 \cdot 10^{-4}}$	$2.58 \cdot 10^{-2}{}_{\pm 5.5 \cdot 10^{-5}}$	$2.12 \cdot 10^{-2}{}_{\pm 2.0 \cdot 10^{-4}}$
	25	mean	$7.90 \cdot 10^{-4}{}_{\pm 4.5 \cdot 10^{-6}}$	$4.04 \cdot 10^{-3}{}_{\pm 7.8 \cdot 10^{-4}}$	$5.74 \cdot 10^{-3}{}_{\pm 4.9 \cdot 10^{-4}}$	$8.46 \cdot 10^{-3}{}_{\pm 1.0 \cdot 10^{-4}}$	$3.44 \cdot 10^{-3}{}_{\pm 4.7 \cdot 10^{-5}}$
		std.	$3.54 \cdot 10^{-2}{}_{\pm 4.6 \cdot 10^{-4}}$	$3.94 \cdot 10^{-2}{}_{\pm 6.6 \cdot 10^{-4}}$	$3.27 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-4}}$	$3.64 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-5}}$	$2.69 \cdot 10^{-2}{}_{\pm 1.7 \cdot 10^{-4}}$
RBF_2	10	mean	$1.00 \cdot 10^{-5}{}_{\pm 2.0 \cdot 10^{-6}}$	$7.30 \cdot 10^{-3}{}_{\pm 1.9 \cdot 10^{-4}}$	$1.97 \cdot 10^{-3}{}_{\pm 7.6 \cdot 10^{-6}}$	$1.85 \cdot 10^{-3}{}_{\pm 5.7 \cdot 10^{-5}}$	$2.52 \cdot 10^{-3}{}_{\pm 3.7 \cdot 10^{-5}}$
		std.	$1.64 \cdot 10^{-3}{}_{\pm 1.3 \cdot 10^{-4}}$	$3.41 \cdot 10^{-3}{}_{\pm 1.5 \cdot 10^{-4}}$	$2.23 \cdot 10^{-3}{}_{\pm 3.0 \cdot 10^{-5}}$	$1.71 \cdot 10^{-3}{}_{\pm 1.2 \cdot 10^{-5}}$	$1.95 \cdot 10^{-3}{}_{\pm 7.7 \cdot 10^{-5}}$
	25	mean	$1.65 \cdot 10^{-4}{}_{\pm 1.1 \cdot 10^{-5}}$	$1.34 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-3}}$	$1.12 \cdot 10^{-2}{}_{\pm 1.7 \cdot 10^{-3}}$	$1.25 \cdot 10^{-2}{}_{\pm 5.2 \cdot 10^{-4}}$	$1.98 \cdot 10^{-2}{}_{\pm 3.4 \cdot 10^{-4}}$
		std.	$6.84 \cdot 10^{-3}{}_{\pm 2.3 \cdot 10^{-4}}$	$4.29 \cdot 10^{-2}{}_{\pm 1.9 \cdot 10^{-3}}$	$2.33 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-3}}$	$1.35 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-4}}$	$1.87 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-4}}$

Table 7. MAE (compared to analytic solution) of mean and standard deviation of GP regression. More details are provided in Section 5.2. Results are averaged over 3 runs. $\text{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	IWVI	DAIS_0 (2)	DAIS_0 (4)	DAIS_0 (8)	DAIS_0 (16)
				$N = 4$			
RBF_1	10	mean	$3.09 \cdot 10^{-4}{}_{\pm 9.1 \cdot 10^{-5}}$	$4.36 \cdot 10^{-4}{}_{\pm 6.9 \cdot 10^{-5}}$	$7.07 \cdot 10^{-4}{}_{\pm 1.6 \cdot 10^{-4}}$	$7.60 \cdot 10^{-4}{}_{\pm 2.2 \cdot 10^{-4}}$	$8.57 \cdot 10^{-4}{}_{\pm 1.8 \cdot 10^{-4}}$
		std.	$4.75 \cdot 10^{-2}{}_{\pm 2.5 \cdot 10^{-5}}$	$4.24 \cdot 10^{-2}{}_{\pm 2.8 \cdot 10^{-5}}$	$3.21 \cdot 10^{-2}{}_{\pm 4.6 \cdot 10^{-5}}$	$2.21 \cdot 10^{-2}{}_{\pm 1.7 \cdot 10^{-5}}$	$1.20 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-5}}$
	25	mean	$2.68 \cdot 10^{-4}{}_{\pm 3.0 \cdot 10^{-5}}$	$2.67 \cdot 10^{-4}{}_{\pm 4.5 \cdot 10^{-5}}$	$4.48 \cdot 10^{-4}{}_{\pm 4.0 \cdot 10^{-5}}$	$4.77 \cdot 10^{-4}{}_{\pm 7.9 \cdot 10^{-5}}$	$7.16 \cdot 10^{-4}{}_{\pm 1.5 \cdot 10^{-4}}$
		std.	$3.93 \cdot 10^{-2}{}_{\pm 6.0 \cdot 10^{-6}}$	$3.81 \cdot 10^{-2}{}_{\pm 7.0 \cdot 10^{-6}}$	$3.68 \cdot 10^{-2}{}_{\pm 9.7 \cdot 10^{-6}}$	$2.94 \cdot 10^{-2}{}_{\pm 6.9 \cdot 10^{-6}}$	$1.71 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-5}}$
RBF_2	10	mean	$9.76 \cdot 10^{-4}{}_{\pm 1.6 \cdot 10^{-4}}$	$1.23 \cdot 10^{-3}{}_{\pm 2.1 \cdot 10^{-4}}$	$2.04 \cdot 10^{-3}{}_{\pm 4.5 \cdot 10^{-4}}$	$1.88 \cdot 10^{-3}{}_{\pm 6.4 \cdot 10^{-4}}$	$2.25 \cdot 10^{-3}{}_{\pm 3.4 \cdot 10^{-4}}$
		std.	$3.84 \cdot 10^{-3}{}_{\pm 8.0 \cdot 10^{-5}}$	$7.37 \cdot 10^{-4}{}_{\pm 8.6 \cdot 10^{-5}}$	$4.73 \cdot 10^{-3}{}_{\pm 5.7 \cdot 10^{-5}}$	$5.18 \cdot 10^{-3}{}_{\pm 9.3 \cdot 10^{-5}}$	$5.46 \cdot 10^{-3}{}_{\pm 6.3 \cdot 10^{-5}}$
	25	mean	$4.41 \cdot 10^{-4}{}_{\pm 7.8 \cdot 10^{-5}}$	$4.90 \cdot 10^{-4}{}_{\pm 7.6 \cdot 10^{-5}}$	$6.79 \cdot 10^{-4}{}_{\pm 6.3 \cdot 10^{-5}}$	$1.31 \cdot 10^{-3}{}_{\pm 1.1 \cdot 10^{-4}}$	$9.90 \cdot 10^{-4}{}_{\pm 1.3 \cdot 10^{-4}}$
		std.	$4.47 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-5}}$	$3.85 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-5}}$	$3.34 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-5}}$	$2.49 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-5}}$	$1.71 \cdot 10^{-2}{}_{\pm 3.6 \cdot 10^{-5}}$

Table 8. MAE (compared to analytic solution) of mean and standard deviation of GP regression. More details are provided in Section 5.2. Results are averaged over 3 runs. $\text{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	IWVI_{SIR}	DAIS (2)	DAIS (4)	DAIS (8)	DAIS (16)
				$N = 4$			
RBF_1	10	mean	$2.44 \cdot 10^{-4}{}_{\pm 2.1 \cdot 10^{-5}}$	$6.90 \cdot 10^{-3}{}_{\pm 3.4 \cdot 10^{-4}}$	$6.45 \cdot 10^{-3}{}_{\pm 5.0 \cdot 10^{-5}}$	$1.50 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-4}}$	$1.65 \cdot 10^{-2}{}_{\pm 5.2 \cdot 10^{-4}}$
		std.	$9.93 \cdot 10^{-3}{}_{\pm 1.2 \cdot 10^{-4}}$	$4.21 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-4}}$	$1.33 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-4}}$	$2.28 \cdot 10^{-2}{}_{\pm 1.9 \cdot 10^{-4}}$	$1.44 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-4}}$
	25	mean	$8.54 \cdot 10^{-4}{}_{\pm 3.6 \cdot 10^{-5}}$	$5.02 \cdot 10^{-3}{}_{\pm 7.3 \cdot 10^{-4}}$	$6.88 \cdot 10^{-3}{}_{\pm 3.7 \cdot 10^{-4}}$	$8.42 \cdot 10^{-3}{}_{\pm 4.0 \cdot 10^{-4}}$	$6.77 \cdot 10^{-3}{}_{\pm 1.2 \cdot 10^{-4}}$
		std.	$3.24 \cdot 10^{-2}{}_{\pm 8.6 \cdot 10^{-4}}$	$3.91 \cdot 10^{-2}{}_{\pm 6.1 \cdot 10^{-4}}$	$3.12 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-4}}$	$3.43 \cdot 10^{-2}{}_{\pm 3.5 \cdot 10^{-4}}$	$2.79 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-4}}$
RBF_2	10	mean	$1.34 \cdot 10^{-5}{}_{\pm 4.3 \cdot 10^{-6}}$	$4.98 \cdot 10^{-3}{}_{\pm 2.1 \cdot 10^{-5}}$	$2.00 \cdot 10^{-3}{}_{\pm 5.1 \cdot 10^{-5}}$	$1.84 \cdot 10^{-3}{}_{\pm 1.1 \cdot 10^{-4}}$	$2.66 \cdot 10^{-3}{}_{\pm 7.9 \cdot 10^{-5}}$
		std.	$1.50 \cdot 10^{-3}{}_{\pm 1.5 \cdot 10^{-4}}$	$3.14 \cdot 10^{-3}{}_{\pm 1.2 \cdot 10^{-4}}$	$1.97 \cdot 10^{-3}{}_{\pm 1.4 \cdot 10^{-4}}$	$1.71 \cdot 10^{-3}{}_{\pm 4.1 \cdot 10^{-5}}$	$1.80 \cdot 10^{-3}{}_{\pm 5.7 \cdot 10^{-5}}$
	25	mean	$1.30 \cdot 10^{-4}{}_{\pm 8.1 \cdot 10^{-6}}$	$7.57 \cdot 10^{-3}{}_{\pm 1.6 \cdot 10^{-3}}$	$3.16 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-3}}$	$1.02 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$	$1.88 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-4}}$
		std.	$5.58 \cdot 10^{-3}{}_{\pm 3.3 \cdot 10^{-4}}$	$3.92 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-4}}$	$3.75 \cdot 10^{-2}{}_{\pm 1.9 \cdot 10^{-3}}$	$1.08 \cdot 10^{-2}{}_{\pm 4.5 \cdot 10^{-4}}$	$1.56 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-4}}$

Table 9. MAE (compared to analytic solution) of mean and standard deviation of GP regression. More details are provided in Section 5.2. Results are averaged over 3 runs. $\text{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	DAIS_0 (2)	DAIS_0 (4)	DAIS_0 (8)
				$N = 16$	
RBF_1	10	mean	$7.50 \cdot 10^{-4}{}_{\pm 2.1 \cdot 10^{-4}}$	$1.35 \cdot 10^{-3}{}_{\pm 3.4 \cdot 10^{-4}}$	$1.02 \cdot 10^{-3}{}_{\pm 1.5 \cdot 10^{-4}}$
		std.	$3.64 \cdot 10^{-2}{}_{\pm 7.9 \cdot 10^{-5}}$	$2.56 \cdot 10^{-2}{}_{\pm 6.9 \cdot 10^{-6}}$	$1.17 \cdot 10^{-2}{}_{\pm 5.3 \cdot 10^{-5}}$
	25	mean	$3.86 \cdot 10^{-4}{}_{\pm 5.8 \cdot 10^{-5}}$	$6.58 \cdot 10^{-4}{}_{\pm 5.4 \cdot 10^{-5}}$	$6.81 \cdot 10^{-4}{}_{\pm 9.9 \cdot 10^{-5}}$
		std.	$3.66 \cdot 10^{-2}{}_{\pm 4.5 \cdot 10^{-6}}$	$3.55 \cdot 10^{-2}{}_{\pm 5.7 \cdot 10^{-5}}$	$2.64 \cdot 10^{-2}{}_{\pm 9.0 \cdot 10^{-6}}$
RBF_2	10	mean	$1.94 \cdot 10^{-3}{}_{\pm 2.5 \cdot 10^{-4}}$	$1.80 \cdot 10^{-3}{}_{\pm 2.0 \cdot 10^{-4}}$	$2.88 \cdot 10^{-3}{}_{\pm 7.4 \cdot 10^{-4}}$
		std.	$4.15 \cdot 10^{-3}{}_{\pm 1.3 \cdot 10^{-4}}$	$6.04 \cdot 10^{-3}{}_{\pm 5.3 \cdot 10^{-5}}$	$4.95 \cdot 10^{-3}{}_{\pm 6.3 \cdot 10^{-5}}$
	25	mean	$6.14 \cdot 10^{-4}{}_{\pm 6.0 \cdot 10^{-5}}$	$8.55 \cdot 10^{-4}{}_{\pm 9.8 \cdot 10^{-5}}$	$1.31 \cdot 10^{-3}{}_{\pm 2.5 \cdot 10^{-4}}$
		std.	$3.48 \cdot 10^{-2}{}_{\pm 4.1 \cdot 10^{-5}}$	$2.90 \cdot 10^{-2}{}_{\pm 4.1 \cdot 10^{-5}}$	$2.20 \cdot 10^{-2}{}_{\pm 5.3 \cdot 10^{-6}}$

Table 10. MAE (compared to analytic solution) of mean and standard deviation of GP regression. More details are provided in Section 5.2. Results are averaged over 3 runs. $\text{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	$N = 16$		
			DAIS (2)	DAIS (4)	DAIS (8)
RBF_1	10	mean	$1.15 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-3}}$	$7.61 \cdot 10^{-3}{}_{\pm 7.1 \cdot 10^{-4}}$	$1.60 \cdot 10^{-2}{}_{\pm 4.0 \cdot 10^{-4}}$
		std.	$4.06 \cdot 10^{-2}{}_{\pm 3.0 \cdot 10^{-4}}$	$9.66 \cdot 10^{-3}{}_{\pm 2.3 \cdot 10^{-4}}$	$1.82 \cdot 10^{-2}{}_{\pm 4.0 \cdot 10^{-4}}$
	25	mean	$4.92 \cdot 10^{-3}{}_{\pm 7.2 \cdot 10^{-4}}$	$6.26 \cdot 10^{-3}{}_{\pm 9.3 \cdot 10^{-4}}$	$7.93 \cdot 10^{-3}{}_{\pm 4.8 \cdot 10^{-4}}$
		std.	$3.77 \cdot 10^{-2}{}_{\pm 3.7 \cdot 10^{-4}}$	$3.15 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-3}}$	$3.08 \cdot 10^{-2}{}_{\pm 3.4 \cdot 10^{-4}}$
RBF_2	10	mean	$4.55 \cdot 10^{-3}{}_{\pm 2.6 \cdot 10^{-4}}$	$2.46 \cdot 10^{-3}{}_{\pm 1.3 \cdot 10^{-4}}$	$2.38 \cdot 10^{-3}{}_{\pm 4.0 \cdot 10^{-4}}$
		std.	$3.29 \cdot 10^{-3}{}_{\pm 1.4 \cdot 10^{-4}}$	$2.33 \cdot 10^{-3}{}_{\pm 2.6 \cdot 10^{-5}}$	$2.09 \cdot 10^{-3}{}_{\pm 7.3 \cdot 10^{-4}}$
	25	mean	$6.01 \cdot 10^{-3}{}_{\pm 4.5 \cdot 10^{-4}}$	$1.90 \cdot 10^{-2}{}_{\pm 1.7 \cdot 10^{-3}}$	$7.22 \cdot 10^{-3}{}_{\pm 7.7 \cdot 10^{-4}}$
		std.	$3.88 \cdot 10^{-2}{}_{\pm 8.5 \cdot 10^{-4}}$	$2.73 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$	$5.17 \cdot 10^{-3}{}_{\pm 1.7 \cdot 10^{-4}}$

Table 11. MAE (compared to analytic solution) of mean and standard deviation of GP regression. More details are provided in Section 5.2. Results are averaged over 3 runs. $\text{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	$\text{MSC}_{1c}\ (N = 1)$	$\text{MSC}_{1c}\ (N = 2)$	$\text{MSC}_{1c}\ (N = 4)$	$\text{MSC}_{1c}\ (N = 16)$
RBF_1	10	mean	$9.22 \cdot 10^{-1}{}_{\pm 0.0}$	$4.10 \cdot 10^{-2}{}_{\pm 7.3 \cdot 10^{-3}}$	$3.11 \cdot 10^{-2}{}_{\pm 7.6 \cdot 10^{-3}}$	$1.87 \cdot 10^{-2}{}_{\pm 4.5 \cdot 10^{-3}}$
		std.	$6.66 \cdot 10^{-2}{}_{\pm 0.0}$	$5.02 \cdot 10^{-2}{}_{\pm 5.2 \cdot 10^{-4}}$	$4.68 \cdot 10^{-2}{}_{\pm 3.0 \cdot 10^{-4}}$	$4.25 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-4}}$
	25	mean	$9.20 \cdot 10^{-1}{}_{\pm 0.0}$	$2.72 \cdot 10^{-1}{}_{\pm 4.6 \cdot 10^{-2}}$	$5.02 \cdot 10^{-2}{}_{\pm 6.4 \cdot 10^{-3}}$	$2.72 \cdot 10^{-2}{}_{\pm 0.77 \cdot 10^{-3}}$
		std.	$4.54 \cdot 10^{-2}{}_{\pm 0.0}$	$4.20 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-4}}$	$4.02 \cdot 10^{-2}{}_{\pm 2.7 \cdot 10^{-5}}$	$3.89 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-4}}$
RBF_2	10	mean	$6.23 \cdot 10^{-1}{}_{\pm 0.0}$	$1.72 \cdot 10^{-2}{}_{\pm 2.6 \cdot 10^{-3}}$	$7.30 \cdot 10^{-3}{}_{\pm 1.6 \cdot 10^{-3}}$	$4.16 \cdot 10^{-3}{}_{\pm 2.9 \cdot 10^{-4}}$
		std.	$8.80 \cdot 10^{-2}{}_{\pm 0.0}$	$7.70 \cdot 10^{-3}{}_{\pm 5.2 \cdot 10^{-4}}$	$3.57 \cdot 10^{-3}{}_{\pm 3.3 \cdot 10^{-4}}$	$1.67 \cdot 10^{-3}{}_{\pm 5.7 \cdot 10^{-5}}$
	25	mean	$7.17 \cdot 10^{-1}{}_{\pm 0.0}$	$1.11 \cdot 10^{-1}{}_{\pm 7.1 \cdot 10^{-3}}$	$4.43 \cdot 10^{-2}{}_{\pm 2.5 \cdot 10^{-3}}$	$3.61 \cdot 10^{-2}{}_{\pm .73 \cdot 10^{-3}}$
		std.	$7.11 \cdot 10^{-2}{}_{\pm 0.0}$	$4.81 \cdot 10^{-2}{}_{\pm 5.7 \cdot 10^{-4}}$	$4.54 \cdot 10^{-2}{}_{\pm 3.8 \cdot 10^{-4}}$	$4.18 \cdot 10^{-2}{}_{\pm 3.1 \cdot 10^{-4}}$

Table 12. MAE (compared to analytic solution) of mean and standard deviation of GP regression for MSC with 8 parallel chains (MSC_{8c}). Results are averaged over 3 runs. $\text{RBF}_{\{1,2\}}$: different prior parameters.

	d	MAE	$\text{MSC}_{8c}\ (N = 1)$	$\text{MSC}_{8c}\ (N = 2)$	$\text{MSC}_{8c}\ (N = 4)$	$\text{MSC}_{8c}\ (N = 16)$
RBF_1	10	mean	$9.22 \cdot 10^{-1}{}_{\pm 0.0}$	$1.66 \cdot 10^{-2}{}_{\pm 3.6 \cdot 10^{-3}}$	$1.66 \cdot 10^{-2}{}_{\pm 5.2 \cdot 10^{-3}}$	$1.74 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-3}}$
		std.	$6.66 \cdot 10^{-2}{}_{\pm 0.0}$	$4.57 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$	$4.29 \cdot 10^{-2}{}_{\pm 9.6 \cdot 10^{-4}}$	$3.66 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-3}}$
	25	mean	$9.20 \cdot 10^{-1}{}_{\pm 0.0}$	$4.17 \cdot 10^{-2}{}_{\pm 6.3 \cdot 10^{-3}}$	$2.03 \cdot 10^{-2}{}_{\pm 3.7 \cdot 10^{-3}}$	$1.14 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-3}}$
		std.	$4.54 \cdot 10^{-2}{}_{\pm 0.0}$	$3.95 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-4}}$	$3.89 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-4}}$	$3.77 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-4}}$
RBF_2	10	mean	$6.23 \cdot 10^{-1}{}_{\pm 0.0}$	$7.54 \cdot 10^{-3}{}_{\pm 2.5 \cdot 10^{-3}}$	$4.07 \cdot 10^{-3}{}_{\pm 9.2 \cdot 10^{-4}}$	$1.57 \cdot 10^{-3}{}_{\pm 2.7 \cdot 10^{-4}}$
		std.	$8.80 \cdot 10^{-2}{}_{\pm 0.0}$	$4.04 \cdot 10^{-3}{}_{\pm 7.7 \cdot 10^{-4}}$	$2.34 \cdot 10^{-3}{}_{\pm 6.6 \cdot 10^{-4}}$	$7.85 \cdot 10^{-4}{}_{\pm 2.4 \cdot 10^{-4}}$
	25	mean	$7.18 \cdot 10^{-1}{}_{\pm 0.0}$	$2.77 \cdot 10^{-2}{}_{\pm 2.0 \cdot 10^{-3}}$	$1.95 \cdot 10^{-2}{}_{\pm 2.3 \cdot 10^{-3}}$	$2.16 \cdot 10^{-2}{}_{\pm 3.4 \cdot 10^{-3}}$
		std.	$7.11 \cdot 10^{-2}{}_{\pm 0.0}$	$4.61 \cdot 10^{-2}{}_{\pm 3.6 \cdot 10^{-4}}$	$4.28 \cdot 10^{-2}{}_{\pm 5.8 \cdot 10^{-4}}$	$4.07 \cdot 10^{-2}{}_{\pm 5.7 \cdot 10^{-4}}$

D. Bayesian Logistic Regression

We provide further details on model, the sampling procedure, and quantitative results of the logistic regression experiments on the five datasets considered. The experiment is discussed in Section 5.3 of the main text. We use an Intel XEON CPU E5-2650 v4 for running the small-scale experiments and a single NVIDIA GeForce GTX 1080 Ti for the large-scale experiments.

D.1. Model

We train our models for $100,000$ iterations on full-batch gradients. We use the Adam optimizer with a learning rate of 10^{-3}. DAIS utilizes the parameterization of Zhang et al. (2021). In addition to the parameters of q_0, we learn the diagonal $\boldsymbol{d}$ of the mass matrix $\boldsymbol{M} = \text{diag}(\boldsymbol{d})$, the annealing schedule $(\beta_1, \ldots, \beta_{K-1})$, and the step widths of the sampler.

D.2. Sampling

We sample the model described in the main text with HMC and compare the learned variational distributions of VI, IWVI, and DAIS$_0$ to the samples (that we treat as "ground truth"). We use the leapfrog integrator with an identity mass matrix for n_l steps and a step size of ϵ_{HMC}, n_b burn-in steps, take every n_e-th sample and sample n_t in total. We apply MH correction steps. Table 13 shows the corresponding hyperparameters.

Table 13. Hyperparameters of HMC sampling for the *sonar* dataset and the *ionosphere* dataset.

	ϵ_{HMC}	n_l	n_b	n_e	n_t
sonar	0.001	50	10,000	10	10,000
ionosphere	0.001	50	10,000	5	10,000

D.3. Complementary Quantitative Results

Table 14, Table 15, Table 16, Table 17, Table 18, Table 19, Table 20, Table 21, and Table 22 provide additional results on the experiment.

Table 14. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression. Details can be found in Section 5.3. Results are averaged over 3 runs.

	d	MAE	IWVI	DAIS$_0$ (2)	DAIS$_0$ (4)	DAIS$_0$ (8)	DAIS$_0$ (16)
sonar	61	mean	$9.00 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-3}}$	$8.67 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-3}}$	$8.76 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-3}}$	$8.69 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$	$8.68 \cdot 10^{-2}{}_{\pm 9.7 \cdot 10^{-4}}$
		std.	$2.86 \cdot 10^{-1}{}_{\pm 1.9 \cdot 10^{-3}}$	$1.77 \cdot 10^{-1}{}_{\pm 9.5 \cdot 10^{-4}}$	$1.23 \cdot 10^{-1}{}_{\pm 1.5 \cdot 10^{-3}}$	$1.22 \cdot 10^{-1}{}_{\pm 1.4 \cdot 10^{-3}}$	$1.20 \cdot 10^{-1}{}_{\pm 4.3 \cdot 10^{-4}}$
ionosphere	35	mean	$9.23 \cdot 10^{-2}{}_{\pm 5.8 \cdot 10^{-4}}$	$8.66 \cdot 10^{-2}{}_{\pm 5.0 \cdot 10^{-4}}$	$7.87 \cdot 10^{-2}{}_{\pm 3.5 \cdot 10^{-3}}$	$8.38 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-3}}$	$8.57 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-3}}$
		std.	$2.32 \cdot 10^{-1}{}_{\pm 8.5 \cdot 10^{-4}}$	$2.01 \cdot 10^{-1}{}_{\pm 1.1 \cdot 10^{-3}}$	$1.57 \cdot 10^{-1}{}_{\pm 6.5 \cdot 10^{-3}}$	$1.20 \cdot 10^{-1}{}_{\pm 1.3 \cdot 10^{-3}}$	$8.40 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$
heart-disease	16	mean	$1.80 \cdot 10^{-2}{}_{\pm 7.9 \cdot 10^{-5}}$	$1.65 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$	$1.93 \cdot 10^{-2}{}_{\pm 2.1 \cdot 10^{-3}}$	$1.82 \cdot 10^{-2}{}_{\pm 5.6 \cdot 10^{-4}}$	$1.69 \cdot 10^{-2}{}_{\pm 8.8 \cdot 10^{-4}}$
		std.	$1.72 \cdot 10^{-1}{}_{\pm 7.2 \cdot 10^{-4}}$	$1.15 \cdot 10^{-1}{}_{\pm 1.4 \cdot 10^{-3}}$	$4.82 \cdot 10^{-2}{}_{\pm 3.8 \cdot 10^{-4}}$	$4.66 \cdot 10^{-2}{}_{\pm 8.6 \cdot 10^{-4}}$	$3.61 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-4}}$
heart-attack	14	mean	$5.05 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-3}}$	$4.97 \cdot 10^{-2}{}_{\pm 2.5 \cdot 10^{-3}}$	$5.01 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-3}}$	$4.68 \cdot 10^{-2}{}_{\pm 3.2 \cdot 10^{-3}}$	$5.25 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-3}}$
		std.	$2.19 \cdot 10^{-1}{}_{\pm 1.1 \cdot 10^{-3}}$	$1.25 \cdot 10^{-1}{}_{\pm 8.5 \cdot 10^{-4}}$	$3.23 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-4}}$	$3.14 \cdot 10^{-2}{}_{\pm 5.9 \cdot 10^{-4}}$	$2.69 \cdot 10^{-2}{}_{\pm 7.5 \cdot 10^{-4}}$
loan	12	mean	$1.03 \cdot 10^{-2}{}_{\pm 6.5 \cdot 10^{-4}}$	$1.26 \cdot 10^{-2}{}_{\pm 8.5 \cdot 10^{-4}}$	$1.27 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-3}}$	$1.34 \cdot 10^{-2}{}_{\pm 7.4 \cdot 10^{-4}}$	$1.17 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-3}}$
		std.	$1.46 \cdot 10^{-1}{}_{\pm 1.3 \cdot 10^{-3}}$	$9.30 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-3}}$	$2.47 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-3}}$	$2.75 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-3}}$	$2.25 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$

The header $N = 1$ spans the DAIS columns.

Table 15. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression. Details can be found in Section 5.3. Results are averaged over 3 runs.

	d	MAE	IWVI$_{\text{SIR}}$	DAIS (2)	DAIS (4)	DAIS (8)	DAIS (16)
					$N = 1$		
sonar	61	mean	$9.28 \cdot 10^{-2}_{\pm 7.6 \cdot 10^{-4}}$	$3.77 \cdot 10^{-1}_{\pm 2.2 \cdot 10^{-2}}$	$2.06 \cdot 10^{-1}_{\pm 1.4 \cdot 10^{-2}}$	$1.77 \cdot 10^{-1}_{\pm 1.5 \cdot 10^{-2}}$	$1.68 \cdot 10^{-1}_{\pm 8.2 \cdot 10^{-3}}$
		std.	$2.90 \cdot 10^{-1}_{\pm 2.1 \cdot 10^{-3}}$	$3.38 \cdot 10^{-1}_{\pm 2.1 \cdot 10^{-2}}$	$1.40 \cdot 10^{-1}_{\pm 7.1 \cdot 10^{-3}}$	$1.25 \cdot 10^{-1}_{\pm 7.9 \cdot 10^{-3}}$	$1.02 \cdot 10^{-1}_{\pm 8.1 \cdot 10^{-3}}$
ionosphere	35	mean	$8.80 \cdot 10^{-2}_{\pm 5.0 \cdot 10^{-4}}$	$9.75 \cdot 10^{-2}_{\pm 3.2 \cdot 10^{-3}}$	$1.08 \cdot 10^{-1}_{\pm 1.9 \cdot 10^{-2}}$	$1.11 \cdot 10^{-1}_{\pm 2.4 \cdot 10^{-2}}$	$8.90 \cdot 10^{-2}_{\pm 2.3 \cdot 10^{-3}}$
		std.	$2.32 \cdot 10^{-1}_{\pm 8.7 \cdot 10^{-4}}$	$1.55 \cdot 10^{-1}_{\pm 8.3 \cdot 10^{-4}}$	$1.13 \cdot 10^{-1}_{\pm 1.1 \cdot 10^{-2}}$	$8.57 \cdot 10^{-2}_{\pm 9.8 \cdot 10^{-3}}$	$4.32 \cdot 10^{-2}_{\pm 3.8 \cdot 10^{-3}}$
heart-disease	16	mean	$1.82 \cdot 10^{-2}_{\pm 1.8 \cdot 10^{-4}}$	$1.68 \cdot 10^{-2}_{\pm 1.2 \cdot 10^{-3}}$	$1.84 \cdot 10^{-2}_{\pm 1.4 \cdot 10^{-3}}$	$1.85 \cdot 10^{-2}_{\pm 7.7 \cdot 10^{-4}}$	$1.62 \cdot 10^{-2}_{\pm 7.8 \cdot 10^{-4}}$
		std.	$1.71 \cdot 10^{-1}_{\pm 6.2 \cdot 10^{-4}}$	$1.17 \cdot 10^{-1}_{\pm 1.4 \cdot 10^{-3}}$	$4.31 \cdot 10^{-2}_{\pm 4.2 \cdot 10^{-4}}$	$4.64 \cdot 10^{-2}_{\pm 9.4 \cdot 10^{-4}}$	$3.63 \cdot 10^{-2}_{\pm 2.3 \cdot 10^{-4}}$
heart-attack	14	mean	$4.80 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$4.96 \cdot 10^{-2}_{\pm 2.9 \cdot 10^{-3}}$	$5.14 \cdot 10^{-2}_{\pm 2.3 \cdot 10^{-3}}$	$4.72 \cdot 10^{-2}_{\pm 3.6 \cdot 10^{-3}}$	$5.33 \cdot 10^{-2}_{\pm 1.0 \cdot 10^{-3}}$
		std.	$2.20 \cdot 10^{-1}_{\pm 1.2 \cdot 10^{-3}}$	$1.25 \cdot 10^{-1}_{\pm 9.0 \cdot 10^{-4}}$	$2.76 \cdot 10^{-2}_{\pm 2.5 \cdot 10^{-4}}$	$2.95 \cdot 10^{-2}_{\pm 5.7 \cdot 10^{-4}}$	$2.55 \cdot 10^{-2}_{\pm 7.4 \cdot 10^{-4}}$
loan	12	mean	$1.78 \cdot 10^{-2}_{\pm 6.2 \cdot 10^{-4}}$	$1.42 \cdot 10^{-2}_{\pm 1.2 \cdot 10^{-3}}$	$1.18 \cdot 10^{-2}_{\pm 8.0 \cdot 10^{-4}}$	$1.49 \cdot 10^{-2}_{\pm 8.6 \cdot 10^{-4}}$	$1.18 \cdot 10^{-2}_{\pm 2.9 \cdot 10^{-3}}$
		std.	$1.43 \cdot 10^{-1}_{\pm 1.3 \cdot 10^{-3}}$	$9.46 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$2.02 \cdot 10^{-2}_{\pm 1.2 \cdot 10^{-3}}$	$2.59 \cdot 10^{-2}_{\pm 2.5 \cdot 10^{-3}}$	$2.15 \cdot 10^{-2}_{\pm 1.4 \cdot 10^{-3}}$

Table 16. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression. Details can be found in Section 5.3. Results are averaged over 3 runs.

	d	MAE	IWVI	DAIS$_0$ (2)	DAIS$_0$ (4)	DAIS$_0$ (8)	DAIS$_0$ (16)
					$N = 2$		
sonar	61	mean	$8.81 \cdot 10^{-2}_{\pm 8.7 \cdot 10^{-4}}$	$8.53 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$8.65 \cdot 10^{-2}_{\pm 3.5 \cdot 10^{-4}}$	$8.51 \cdot 10^{-2}_{\pm 2.2 \cdot 10^{-3}}$	$8.49 \cdot 10^{-2}_{\pm 1.5 \cdot 10^{-3}}$
		std.	$2.09 \cdot 10^{-1}_{\pm 1.2 \cdot 10^{-3}}$	$1.31 \cdot 10^{-1}_{\pm 1.2 \cdot 10^{-3}}$	$9.08 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$9.08 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$8.86 \cdot 10^{-2}_{\pm 8.2 \cdot 10^{-4}}$
ionosphere	35	mean	$4.84 \cdot 10^{-2}_{\pm 3.3 \cdot 10^{-4}}$	$4.59 \cdot 10^{-2}_{\pm 3.2 \cdot 10^{-4}}$	$4.46 \cdot 10^{-2}_{\pm 5.5 \cdot 10^{-4}}$	$4.69 \cdot 10^{-2}_{\pm 4.7 \cdot 10^{-4}}$	$4.87 \cdot 10^{-2}_{\pm 5.9 \cdot 10^{-4}}$
		std.	$1.04 \cdot 10^{-1}_{\pm 3.6 \cdot 10^{-4}}$	$9.00 \cdot 10^{-2}_{\pm 2.1 \cdot 10^{-4}}$	$7.53 \cdot 10^{-2}_{\pm 1.6 \cdot 10^{-4}}$	$5.94 \cdot 10^{-2}_{\pm 8.2 \cdot 10^{-4}}$	$4.14 \cdot 10^{-2}_{\pm 1.7 \cdot 10^{-3}}$
heart-disease	16	mean	$1.56 \cdot 10^{-2}_{\pm 1.7 \cdot 10^{-3}}$	$1.83 \cdot 10^{-2}_{\pm 1.6 \cdot 10^{-3}}$	$1.92 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$1.88 \cdot 10^{-2}_{\pm 1.5 \cdot 10^{-3}}$	$2.08 \cdot 10^{-2}_{\pm 3.3 \cdot 10^{-3}}$
		std.	$1.28 \cdot 10^{-1}_{\pm 5.0 \cdot 10^{-4}}$	$7.57 \cdot 10^{-2}_{\pm 1.0 \cdot 10^{-3}}$	$3.28 \cdot 10^{-2}_{\pm 8.8 \cdot 10^{-4}}$	$3.15 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$1.98 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$
heart-attack	14	mean	$4.95 \cdot 10^{-2}_{\pm 3.9 \cdot 10^{-3}}$	$4.86 \cdot 10^{-2}_{\pm 2.5 \cdot 10^{-3}}$	$4.74 \cdot 10^{-2}_{\pm 1.8 \cdot 10^{-3}}$	$4.60 \cdot 10^{-2}_{\pm 2.9 \cdot 10^{-3}}$	$5.05 \cdot 10^{-2}_{\pm 1.5 \cdot 10^{-3}}$
		std.	$1.45 \cdot 10^{-1}_{\pm 1.8 \cdot 10^{-3}}$	$6.95 \cdot 10^{-2}_{\pm 1.5 \cdot 10^{-3}}$	$3.02 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$2.62 \cdot 10^{-2}_{\pm 1.8 \cdot 10^{-3}}$	$2.81 \cdot 10^{-2}_{\pm 9.8 \cdot 10^{-4}}$
loan	12	mean	$9.57 \cdot 10^{-3}_{\pm 1.0 \cdot 10^{-3}}$	$1.35 \cdot 10^{-2}_{\pm 2.4 \cdot 10^{-3}}$	$1.81 \cdot 10^{-2}_{\pm 3.3 \cdot 10^{-3}}$	$9.02 \cdot 10^{-3}_{\pm 7.1 \cdot 10^{-4}}$	$1.53 \cdot 10^{-2}_{\pm 1.1 \cdot 10^{-3}}$
		std.	$9.55 \cdot 10^{-2}_{\pm 7.8 \cdot 10^{-4}}$	$5.30 \cdot 10^{-2}_{\pm 1.7 \cdot 10^{-3}}$	$1.23 \cdot 10^{-2}_{\pm 1.4 \cdot 10^{-3}}$	$1.44 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$9.15 \cdot 10^{-3}_{\pm 8.4 \cdot 10^{-4}}$

Table 17. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression. Details can be found in Section 5.3. Results are averaged over 3 runs.

	d	MAE	IWVI$_{\text{SIR}}$	DAIS (2)	DAIS (4)	DAIS (8)	DAIS (16)
					$N = 2$		
sonar	61	mean	$9.03 \cdot 10^{-2}_{\pm 1.4 \cdot 10^{-3}}$	$3.53 \cdot 10^{-1}_{\pm 1.8 \cdot 10^{-2}}$	$1.85 \cdot 10^{-1}_{\pm 2.0 \cdot 10^{-2}}$	$1.35 \cdot 10^{-1}_{\pm 1.1 \cdot 10^{-2}}$	$1.58 \cdot 10^{-1}_{\pm 7.7 \cdot 10^{-3}}$
		std.	$2.13 \cdot 10^{-1}_{\pm 1.0 \cdot 10^{-3}}$	$2.59 \cdot 10^{-1}_{\pm 1.6 \cdot 10^{-2}}$	$1.17 \cdot 10^{-1}_{\pm 1.3 \cdot 10^{-2}}$	$8.64 \cdot 10^{-2}_{\pm 3.2 \cdot 10^{-3}}$	$9.04 \cdot 10^{-2}_{\pm 7.9 \cdot 10^{-3}}$
ionosphere	35	mean	$7.94 \cdot 10^{-2}_{\pm 9.8 \cdot 10^{-4}}$	$9.14 \cdot 10^{-2}_{\pm 6.5 \cdot 10^{-3}}$	$1.39 \cdot 10^{-1}_{\pm 3.1 \cdot 10^{-2}}$	$1.19 \cdot 10^{-1}_{\pm 3.2 \cdot 10^{-3}}$	$9.99 \cdot 10^{-2}_{\pm 1.7 \cdot 10^{-2}}$
		std.	$1.81 \cdot 10^{-1}_{\pm 6.5 \cdot 10^{-4}}$	$1.20 \cdot 10^{-1}_{\pm 5.8 \cdot 10^{-3}}$	$1.26 \cdot 10^{-1}_{\pm 2.5 \cdot 10^{-2}}$	$7.04 \cdot 10^{-2}_{\pm 3.7 \cdot 10^{-3}}$	$6.06 \cdot 10^{-2}_{\pm 3.9 \cdot 10^{-3}}$
heart-disease	16	mean	$1.63 \cdot 10^{-2}_{\pm 1.4 \cdot 10^{-3}}$	$1.84 \cdot 10^{-2}_{\pm 1.7 \cdot 10^{-3}}$	$1.84 \cdot 10^{-2}_{\pm 5.8 \cdot 10^{-4}}$	$1.86 \cdot 10^{-2}_{\pm 1.5 \cdot 10^{-3}}$	$1.98 \cdot 10^{-2}_{\pm 4.7 \cdot 10^{-3}}$
		std.	$1.28 \cdot 10^{-1}_{\pm 6.2 \cdot 10^{-4}}$	$7.61 \cdot 10^{-2}_{\pm 8.5 \cdot 10^{-4}}$	$3.00 \cdot 10^{-2}_{\pm 2.1 \cdot 10^{-3}}$	$3.18 \cdot 10^{-2}_{\pm 1.4 \cdot 10^{-3}}$	$2.08 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$
heart-attack	14	mean	$4.71 \cdot 10^{-2}_{\pm 4.3 \cdot 10^{-3}}$	$4.85 \cdot 10^{-2}_{\pm 2.8 \cdot 10^{-3}}$	$4.79 \cdot 10^{-2}_{\pm 1.8 \cdot 10^{-3}}$	$4.67 \cdot 10^{-2}_{\pm 2.9 \cdot 10^{-3}}$	$5.08 \cdot 10^{-2}_{\pm 1.8 \cdot 10^{-3}}$
		std.	$1.45 \cdot 10^{-1}_{\pm 1.8 \cdot 10^{-3}}$	$6.93 \cdot 10^{-2}_{\pm 1.4 \cdot 10^{-3}}$	$2.66 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$	$2.42 \cdot 10^{-2}_{\pm 1.4 \cdot 10^{-3}}$	$2.77 \cdot 10^{-2}_{\pm 9.6 \cdot 10^{-4}}$
loan	12	mean	$1.64 \cdot 10^{-2}_{\pm 7.5 \cdot 10^{-4}}$	$1.44 \cdot 10^{-2}_{\pm 2.4 \cdot 10^{-3}}$	$1.62 \cdot 10^{-2}_{\pm 2.7 \cdot 10^{-3}}$	$8.57 \cdot 10^{-3}_{\pm 5.4 \cdot 10^{-4}}$	$1.52 \cdot 10^{-2}_{\pm 1.3 \cdot 10^{-3}}$
		std.	$9.39 \cdot 10^{-2}_{\pm 1.8 \cdot 10^{-3}}$	$5.31 \cdot 10^{-2}_{\pm 1.7 \cdot 10^{-3}}$	$9.09 \cdot 10^{-3}_{\pm 1.2 \cdot 10^{-3}}$	$1.26 \cdot 10^{-2}_{\pm 1.5 \cdot 10^{-3}}$	$8.07 \cdot 10^{-3}_{\pm 4.3 \cdot 10^{-4}}$

Table 18. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression. Details can be found in Section 5.3. Results are averaged over 3 runs.

	d	MAE	IWVI	DAIS_0 (2)	DAIS_0 (4)	DAIS_0 (8)	DAIS_0 (16)
					$N = 4$		
sonar	61	mean	$8.68 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-3}}$	$8.45 \cdot 10^{-2}{}_{\pm 8.5 \cdot 10^{-4}}$	$8.64 \cdot 10^{-2}{}_{\pm 9.0 \cdot 10^{-4}}$	$8.89 \cdot 10^{-2}{}_{\pm 2.3 \cdot 10^{-3}}$	$8.42 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-3}}$
		std.	$1.48 \cdot 10^{-1}{}_{\pm 1.1 \cdot 10^{-3}}$	$9.98 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-3}}$	$6.92 \cdot 10^{-2}{}_{\pm 1.9 \cdot 10^{-3}}$	$6.95 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-3}}$	$6.63 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$
ionosphere	35	mean	$4.61 \cdot 10^{-2}{}_{\pm 8.7 \cdot 10^{-4}}$	$4.48 \cdot 10^{-2}{}_{\pm 4.0 \cdot 10^{-4}}$	$4.41 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-3}}$	$4.42 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$	$4.59 \cdot 10^{-2}{}_{\pm 9.7 \cdot 10^{-4}}$
		std.	$7.97 \cdot 10^{-2}{}_{\pm 5.7 \cdot 10^{-4}}$	$6.88 \cdot 10^{-2}{}_{\pm 3.2 \cdot 10^{-4}}$	$5.63 \cdot 10^{-2}{}_{\pm 6.3 \cdot 10^{-4}}$	$5.65 \cdot 10^{-2}{}_{\pm 6.7 \cdot 10^{-4}}$	$4.26 \cdot 10^{-2}{}_{\pm 3.0 \cdot 10^{-3}}$
heart-disease	16	mean	$1.59 \cdot 10^{-2}{}_{\pm 6.4 \cdot 10^{-4}}$	$1.69 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-3}}$	$2.21 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-3}}$	$1.84 \cdot 10^{-2}{}_{\pm 3.3 \cdot 10^{-4}}$	$2.05 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-3}}$
		std.	$8.56 \cdot 10^{-2}{}_{\pm 9.9 \cdot 10^{-4}}$	$5.08 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$	$2.00 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-3}}$	$1.68 \cdot 10^{-2}{}_{\pm 6.5 \cdot 10^{-4}}$	$1.34 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-3}}$
heart-attack	14	mean	$4.95 \cdot 10^{-2}{}_{\pm 3.4 \cdot 10^{-3}}$	$4.74 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-3}}$	$4.72 \cdot 10^{-2}{}_{\pm 4.0 \cdot 10^{-3}}$	$4.85 \cdot 10^{-2}{}_{\pm 3.5 \cdot 10^{-3}}$	$4.99 \cdot 10^{-2}{}_{\pm 4.4 \cdot 10^{-3}}$
		std.	$7.97 \cdot 10^{-2}{}_{\pm 6.6 \cdot 10^{-4}}$	$3.46 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-3}}$	$3.49 \cdot 10^{-2}{}_{\pm 5.2 \cdot 10^{-3}}$	$2.65 \cdot 10^{-2}{}_{\pm 4.2 \cdot 10^{-3}}$	$2.74 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-3}}$
loan	12	mean	$1.31 \cdot 10^{-2}{}_{\pm 5.4 \cdot 10^{-3}}$	$1.09 \cdot 10^{-2}{}_{\pm 3.5 \cdot 10^{-4}}$	$1.42 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-3}}$	$1.31 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-3}}$	$1.85 \cdot 10^{-2}{}_{\pm 2.3 \cdot 10^{-3}}$
		std.	$5.52 \cdot 10^{-2}{}_{\pm 4.5 \cdot 10^{-4}}$	$2.79 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-3}}$	$1.16 \cdot 10^{-2}{}_{\pm 5.1 \cdot 10^{-4}}$	$7.90 \cdot 10^{-3}{}_{\pm 1.7 \cdot 10^{-3}}$	$5.57 \cdot 10^{-3}{}_{\pm 6.3 \cdot 10^{-4}}$

Table 19. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression. Details can be found in Section 5.3. Results are averaged over 3 runs.

	d	MAE	IWVI_{SIR}	DAIS (2)	DAIS (4)	DAIS (8)	DAIS (16)
					$N = 4$		
sonar	61	mean	$8.86 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-3}}$	$2.45 \cdot 10^{-1}{}_{\pm 4.3 \cdot 10^{-2}}$	$1.54 \cdot 10^{-1}{}_{\pm 1.6 \cdot 10^{-2}}$	$1.21 \cdot 10^{-1}{}_{\pm 6.7 \cdot 10^{-3}}$	$1.40 \cdot 10^{-1}{}_{\pm 9.4 \cdot 10^{-3}}$
		std.	$1.52 \cdot 10^{-1}{}_{\pm 1.1 \cdot 10^{-3}}$	$1.76 \cdot 10^{-1}{}_{\pm 2.8 \cdot 10^{-2}}$	$9.18 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-2}}$	$7.43 \cdot 10^{-2}{}_{\pm 6.3 \cdot 10^{-3}}$	$7.80 \cdot 10^{-2}{}_{\pm 3.1 \cdot 10^{-3}}$
ionosphere	35	mean	$7.67 \cdot 10^{-2}{}_{\pm 7.0 \cdot 10^{-4}}$	$1.29 \cdot 10^{-1}{}_{\pm 3.8 \cdot 10^{-2}}$	$2.27 \cdot 10^{-1}{}_{\pm 2.3 \cdot 10^{-2}}$	$1.01 \cdot 10^{-1}{}_{\pm 5.4 \cdot 10^{-3}}$	$8.49 \cdot 10^{-2}{}_{\pm 2.8 \cdot 10^{-3}}$
		std.	$1.39 \cdot 10^{-1}{}_{\pm 9.9 \cdot 10^{-4}}$	$1.25 \cdot 10^{-1}{}_{\pm 3.0 \cdot 10^{-2}}$	$1.68 \cdot 10^{-1}{}_{\pm 1.9 \cdot 10^{-2}}$	$7.11 \cdot 10^{-2}{}_{\pm 1.7 \cdot 10^{-3}}$	$5.82 \cdot 10^{-2}{}_{\pm 4.7 \cdot 10^{-3}}$
heart-disease	16	mean	$1.70 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-4}}$	$1.63 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$	$2.20 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-3}}$	$1.82 \cdot 10^{-2}{}_{\pm 3.7 \cdot 10^{-4}}$	$1.98 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-3}}$
		std.	$8.54 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-3}}$	$5.05 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$	$1.92 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-3}}$	$1.71 \cdot 10^{-2}{}_{\pm 7.7 \cdot 10^{-4}}$	$1.43 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-3}}$
heart-attack	14	mean	$4.70 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-3}}$	$4.79 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$	$4.62 \cdot 10^{-2}{}_{\pm 5.6 \cdot 10^{-3}}$	$4.87 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-3}}$	$5.10 \cdot 10^{-2}{}_{\pm 4.7 \cdot 10^{-3}}$
		std.	$8.13 \cdot 10^{-2}{}_{\pm 3.5 \cdot 10^{-4}}$	$3.47 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$	$3.14 \cdot 10^{-2}{}_{\pm 5.1 \cdot 10^{-3}}$	$2.48 \cdot 10^{-2}{}_{\pm 3.8 \cdot 10^{-3}}$	$2.69 \cdot 10^{-2}{}_{\pm 1.2 \cdot 10^{-3}}$
loan	12	mean	$2.23 \cdot 10^{-2}{}_{\pm 5.5 \cdot 10^{-3}}$	$1.16 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-3}}$	$1.23 \cdot 10^{-2}{}_{\pm 2.1 \cdot 10^{-3}}$	$1.37 \cdot 10^{-2}{}_{\pm 2.8 \cdot 10^{-4}}$	$1.85 \cdot 10^{-2}{}_{\pm 3.6 \cdot 10^{-3}}$
		std.	$5.22 \cdot 10^{-2}{}_{\pm 3.9 \cdot 10^{-4}}$	$2.79 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$	$9.15 \cdot 10^{-3}{}_{\pm 9.6 \cdot 10^{-4}}$	$6.22 \cdot 10^{-3}{}_{\pm 1.7 \cdot 10^{-3}}$	$5.36 \cdot 10^{-3}{}_{\pm 2.8 \cdot 10^{-5}}$

Table 20. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression. Details can be found in Section 5.3. Results are averaged over 3 runs.

	d	MAE	DAIS_0 (2)	DAIS_0 (4)	DAIS_0 (8)
				$N = 16$	
sonar	61	mean	$8.79 \cdot 10^{-2}{}_{\pm 1.9 \cdot 10^{-3}}$	$8.82 \cdot 10^{-2}{}_{\pm 2.0 \cdot 10^{-3}}$	$8.37 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$
		std.	$6.11 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-3}}$	$4.43 \cdot 10^{-2}{}_{\pm 5.0 \cdot 10^{-4}}$	$4.57 \cdot 10^{-2}{}_{\pm 4.8 \cdot 10^{-4}}$
ionosphere	35	mean	$4.30 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-3}}$	$4.53 \cdot 10^{-2}{}_{\pm 8.9 \cdot 10^{-4}}$	$4.49 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-3}}$
		std.	$4.27 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-4}}$	$3.41 \cdot 10^{-2}{}_{\pm 6.1 \cdot 10^{-4}}$	$3.46 \cdot 10^{-2}{}_{\pm 4.5 \cdot 10^{-4}}$
heart-disease	16	mean	$2.03 \cdot 10^{-2}{}_{\pm 2.1 \cdot 10^{-3}}$	$1.94 \cdot 10^{-2}{}_{\pm 3.0 \cdot 10^{-3}}$	$2.21 \cdot 10^{-2}{}_{\pm 4.6 \cdot 10^{-3}}$
		std.	$1.41 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$	$9.70 \cdot 10^{-3}{}_{\pm 1.4 \cdot 10^{-3}}$	$7.85 \cdot 10^{-3}{}_{\pm 4.2 \cdot 10^{-4}}$
heart-attack	14	mean	$4.89 \cdot 10^{-2}{}_{\pm 3.4 \cdot 10^{-3}}$	$5.48 \cdot 10^{-2}{}_{\pm 6.4 \cdot 10^{-3}}$	$4.81 \cdot 10^{-2}{}_{\pm 2.3 \cdot 10^{-3}}$
		std.	$3.16 \cdot 10^{-2}{}_{\pm 3.3 \cdot 10^{-3}}$	$3.99 \cdot 10^{-2}{}_{\pm 5.8 \cdot 10^{-4}}$	$3.50 \cdot 10^{-2}{}_{\pm 2.8 \cdot 10^{-3}}$
loan	12	mean	$1.73 \cdot 10^{-2}{}_{\pm 3.7 \cdot 10^{-3}}$	$2.03 \cdot 10^{-2}{}_{\pm 3.3 \cdot 10^{-3}}$	$1.50 \cdot 10^{-2}{}_{\pm 2.7 \cdot 10^{-3}}$
		std.	$8.10 \cdot 10^{-3}{}_{\pm 1.3 \cdot 10^{-3}}$	$1.65 \cdot 10^{-2}{}_{\pm 2.5 \cdot 10^{-3}}$	$1.22 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-3}}$

Table 21. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression. Details can be found in Section 5.3. Results are averaged over 3 runs.

	d	MAE	DAIS (2)	DAIS (4)	DAIS (8)
				$N = 16$	
sonar	61	mean	$1.87 \cdot 10^{-1}{}_{\pm 6.0 \cdot 10^{-2}}$	$1.20 \cdot 10^{-1}{}_{\pm 7.9 \cdot 10^{-3}}$	$1.07 \cdot 10^{-1}{}_{\pm 5.7 \cdot 10^{-3}}$
		std.	$1.25 \cdot 10^{-1}{}_{\pm 2.5 \cdot 10^{-2}}$	$7.26 \cdot 10^{-2}{}_{\pm 6.0 \cdot 10^{-3}}$	$6.06 \cdot 10^{-2}{}_{\pm 5.0 \cdot 10^{-3}}$
ionosphere	35	mean	$1.36 \cdot 10^{-1}{}_{\pm 4.3 \cdot 10^{-2}}$	$2.05 \cdot 10^{-1}{}_{\pm 5.2 \cdot 10^{-2}}$	$1.15 \cdot 10^{-1}{}_{\pm 9.7 \cdot 10^{-3}}$
		std.	$1.08 \cdot 10^{-1}{}_{\pm 2.1 \cdot 10^{-2}}$	$1.27 \cdot 10^{-1}{}_{\pm 2.7 \cdot 10^{-2}}$	$7.53 \cdot 10^{-2}{}_{\pm 2.0 \cdot 10^{-3}}$
heart-disease	16	mean	$1.99 \cdot 10^{-2}{}_{\pm 2.0 \cdot 10^{-3}}$	$1.94 \cdot 10^{-2}{}_{\pm 3.0 \cdot 10^{-3}}$	$2.21 \cdot 10^{-2}{}_{\pm 4.2 \cdot 10^{-3}}$
		std.	$1.44 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-3}}$	$1.07 \cdot 10^{-2}{}_{\pm 1.3 \cdot 10^{-3}}$	$8.96 \cdot 10^{-3}{}_{\pm 1.5 \cdot 10^{-4}}$
heart-attack	14	mean	$4.85 \cdot 10^{-2}{}_{\pm 5.1 \cdot 10^{-3}}$	$5.49 \cdot 10^{-2}{}_{\pm 6.4 \cdot 10^{-3}}$	$4.87 \cdot 10^{-2}{}_{\pm 2.2 \cdot 10^{-3}}$
		std.	$3.15 \cdot 10^{-2}{}_{\pm 3.4 \cdot 10^{-3}}$	$3.65 \cdot 10^{-2}{}_{\pm 8.4 \cdot 10^{-4}}$	$3.35 \cdot 10^{-2}{}_{\pm 2.6 \cdot 10^{-3}}$
loan	12	mean	$1.66 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-3}}$	$1.95 \cdot 10^{-2}{}_{\pm 2.9 \cdot 10^{-3}}$	$1.33 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$
		std.	$9.12 \cdot 10^{-3}{}_{\pm 1.6 \cdot 10^{-3}}$	$1.52 \cdot 10^{-2}{}_{\pm 2.6 \cdot 10^{-3}}$	$1.21 \cdot 10^{-2}{}_{\pm 2.1 \cdot 10^{-3}}$

Table 22. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression. Details can be found in Section 5.3. Results are averaged over 3 runs.

	d	MAE	MSC_{1c} ($N = 1$)	MSC_{1c} ($N = 4$)	MSC_{1c} ($N = 8$)	MSC_{1c} ($N = 16$)
sonar	61	mean	$4.71 \cdot 10^{-1}{}_{\pm 7.1 \cdot 10^{-3}}$	$2.30 \cdot 10^{-1}{}_{\pm 1.9 \cdot 10^{-2}}$	$1.57 \cdot 10^{-1}{}_{\pm 9.5 \cdot 10^{-3}}$	$1.39 \cdot 10^{-1}{}_{\pm 5.9 \cdot 10^{-3}}$
		std.	$8.85 \cdot 10^{-1}{}_{\pm 8.7 \cdot 10^{-7}}$	$2.10 \cdot 10^{-1}{}_{\pm 2.1 \cdot 10^{-2}}$	$1.43 \cdot 10^{-1}{}_{\pm 1.4 \cdot 10^{-3}}$	$9.43 \cdot 10^{-2}{}_{\pm 1.1 \cdot 10^{-2}}$
ionosphere	35	mean	$7.62 \cdot 10^{-1}{}_{\pm 1.2 \cdot 10^{-2}}$	$2.68 \cdot 10^{-1}{}_{\pm 2.4 \cdot 10^{-2}}$	$2.35 \cdot 10^{-1}{}_{\pm 1.1 \cdot 10^{-2}}$	$2.10 \cdot 10^{-1}{}_{\pm 9.4 \cdot 10^{-3}}$
		std.	$5.87 \cdot 10^{-1}{}_{\pm 4.4 \cdot 10^{-7}}$	$1.68 \cdot 10^{-1}{}_{\pm 1.5 \cdot 10^{-3}}$	$1.32 \cdot 10^{-1}{}_{\pm 1.2 \cdot 10^{-2}}$	$1.02 \cdot 10^{-1}{}_{\pm 1.8 \cdot 10^{-3}}$
heart-disease	16	mean	$6.72 \cdot 10^{-1}{}_{\pm 1.3 \cdot 10^{-2}}$	$3.63 \cdot 10^{-1}{}_{\pm 1.4 \cdot 10^{-2}}$	$3.50 \cdot 10^{-1}{}_{\pm 1.1 \cdot 10^{-2}}$	$3.40 \cdot 10^{-1}{}_{\pm 5.7 \cdot 10^{-3}}$
		std.	$3.47 \cdot 10^{-1}{}_{\pm 3.6 \cdot 10^{-11}}$	$7.21 \cdot 10^{-2}{}_{\pm 2.3 \cdot 10^{-3}}$	$4.44 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-3}}$	$3.45 \cdot 10^{-2}{}_{\pm 1.5 \cdot 10^{-3}}$
heart-attack	14	mean	$9.63 \cdot 10^{-1}{}_{\pm 2.3 \cdot 10^{-2}}$	$1.87 \cdot 10^{-1}{}_{\pm 1.9 \cdot 10^{-2}}$	$1.83 \cdot 10^{-1}{}_{\pm 6.2 \cdot 10^{-3}}$	$1.69 \cdot 10^{-1}{}_{\pm 1.5 \cdot 10^{-3}}$
		std.	$5.47 \cdot 10^{-1}{}_{\pm 1.0 \cdot 10^{-6}}$	$8.66 \cdot 10^{-2}{}_{\pm 3.2 \cdot 10^{-3}}$	$7.91 \cdot 10^{-2}{}_{\pm 2.0 \cdot 10^{-3}}$	$7.15 \cdot 10^{-2}{}_{\pm 3.2 \cdot 10^{-3}}$
loan	12	mean	$4.83 \cdot 10^{-1}{}_{\pm 1.6 \cdot 10^{-2}}$	$1.97 \cdot 10^{-1}{}_{\pm 1.2 \cdot 10^{-2}}$	$1.80 \cdot 10^{-1}{}_{\pm 9.3 \cdot 10^{-3}}$	$1.76 \cdot 10^{-1}{}_{\pm 5.3 \cdot 10^{-3}}$
		std.	$4.55 \cdot 10^{-1}{}_{\pm 2.1 \cdot 10^{-15}}$	$7.94 \cdot 10^{-2}{}_{\pm 7.0 \cdot 10^{-3}}$	$6.32 \cdot 10^{-2}{}_{\pm 1.8 \cdot 10^{-3}}$	$5.28. 10^{-2}{}_{\pm 1.1 \cdot 10^{-3}}$

Table 23. MAE calculated between learned mean and standard deviation to HMC samples for Bayesian logistic regression with MSC with 8 parallel chains (MSC_{8c}). Results are averaged over 3 runs.

	d	MAE	MSC_{8c} ($N = 1$)	MSC_{8c} ($N = 4$)	MSC_{8c} ($N = 2$)	MSC_{8c} ($N = 16$)
sonar	61	mean	$4.54 \cdot 10^{-1}{}_{\pm 2.3 \cdot 10^{-3}}$	$1.70 \cdot 10^{-1}{}_{\pm 8.2 \cdot 10^{-3}}$	$1.37 \cdot 10^{-1}{}_{\pm 6.9 \cdot 10^{-3}}$	$1.33 \cdot 10^{-1}{}_{\pm 8.1 \cdot 10^{-3}}$
		std.	$7.96 \cdot 10^{-1}{}_{\pm 7.3 \cdot 10^{-4}}$	$1.40 \cdot 10^{-1}{}_{\pm 2.9 \cdot 10^{-3}}$	$1.06 \cdot 10^{-1}{}_{\pm 4.7 \cdot 10^{-3}}$	$7.26 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-3}}$
ionosphere	35	mean	$7.61 \cdot 10^{-1}{}_{\pm 9.7 \cdot 10^{-3}}$	$2.11 \cdot 10^{-1}{}_{\pm 1.2 \cdot 10^{-2}}$	$2.15 \cdot 10^{-1}{}_{\pm 4.6 \cdot 10^{-3}}$	$2.14 \cdot 10^{-1}{}_{\pm 5.0 \cdot 10^{-3}}$
		std.	$5.01 \cdot 10^{-1}{}_{\pm 2.7 \cdot 10^{-3}}$	$1.30 \cdot 10^{-1}{}_{\pm 1.5 \cdot 10^{-3}}$	$1.09 \cdot 10^{-1}{}_{\pm 4.6 \cdot 10^{-3}}$	$8.55 \cdot 10^{-2}{}_{\pm 2.4 \cdot 10^{-4}}$
heart-disease	16	mean	$6.73 \cdot 10^{-1}{}_{\pm 5.9 \cdot 10^{-3}}$	$3.43 \cdot 10^{-1}{}_{\pm 1.2 \cdot 10^{-2}}$	$3.46 \cdot 10^{-1}{}_{\pm 5.2 \cdot 10^{-4}}$	$3.37 \cdot 10^{-1}{}_{\pm 3.2 \cdot 10^{-3}}$
		std.	$2.67 \cdot 10^{-1}{}_{\pm 6.8 \cdot 10^{-3}}$	$5.33 \cdot 10^{-2}{}_{\pm 6.4 \cdot 10^{-3}}$	$3.77 \cdot 10^{-2}{}_{\pm 4.0 \cdot 10^{-3}}$	$2.81 \cdot 10^{-2}{}_{\pm 1.0 \cdot 10^{-3}}$
heart-attack	14	mean	$9.87 \cdot 10^{-1}{}_{\pm 5.1 \cdot 10^{-3}}$	$1.79 \cdot 10^{-1}{}_{\pm 6.2 \cdot 10^{-3}}$	$1.75 \cdot 10^{-1}{}_{\pm 4.7 \cdot 10^{-3}}$	$1.75 \cdot 10^{-1}{}_{\pm 1.1 \cdot 10^{-3}}$
		std.	$4.66 \cdot 10^{-1}{}_{\pm 3.4 \cdot 10^{-3}}$	$8.18 \cdot 10^{-2}{}_{\pm 6.5 \cdot 10^{-3}}$	$7.91 \cdot 10^{-2}{}_{\pm 1.4 \cdot 10^{-3}}$	$7.41 \cdot 10^{-2}{}_{\pm 3.5 \cdot 10^{-3}}$
loan	12	mean	$4.88 \cdot 10^{-1}{}_{\pm 8.3 \cdot 10^{-3}}$	$1.83 \cdot 10^{-1}{}_{\pm 7.9 \cdot 10^{-3}}$	$1.72 \cdot 10^{-1}{}_{\pm 4.8 \cdot 10^{-3}}$	$1.76 \cdot 10^{-1}{}_{\pm 8.9 \cdot 10^{-4}}$
		std.	$3.71 \cdot 10^{-1}{}_{\pm 7.5 \cdot 10^{-3}}$	$6.09 \cdot 10^{-2}{}_{\pm 7.1 \cdot 10^{-3}}$	$5.45 \cdot 10^{-2}{}_{\pm 5.5 \cdot 10^{-4}}$	$5.08 \cdot 10^{-2}{}_{\pm 1.6 \cdot 10^{-3}}$