Doubly Adaptive Importance Sampling

Willem van den Boom[*]
Institute for Human Development and Potential,
Agency for Science, Technology and Research, Singapore
and
Andrea Cremaschi
School of Science and Technology, IE University, Madrid, Spain
and
Alexandre H. Thiery
Department of Statistics and Data Science,
National University of Singapore

Abstract

We propose an adaptive importance sampling scheme for Gaussian approximations of intractable posteriors. Optimization-based approximations like variational inference can be too inaccurate while existing Monte Carlo methods can be too slow. Therefore, we propose a hybrid where, at each iteration, the Monte Carlo effective sample size can be guaranteed at a fixed computational cost by interpolating between natural-gradient variational inference and importance sampling. The amount of damping in the updates adapts to the posterior and guarantees the effective sample size. Gaussianity enables the use of Stein's lemma to obtain gradient-based optimization in the highly damped variational inference regime and a reduction of Monte Carlo error for undamped adaptive importance sampling. The result is a generic, embarrassingly parallel and adaptive posterior approximation method. Numerical studies on simulated and real data show its competitiveness with other, less general methods.

Keywords: Adaptive Monte Carlo; Embarrassingly parallel computing; Gaussian posterior approximation; Natural-gradient variational inference; Stein's lemma

[*] Email: willem@wvdboom.nl. This work was partially supported by the Singapore Ministry of Education Academic Research Fund Tier 2 grant MOE2016-T2-2-135.

1 Introduction

It is common in several modeling settings to encounter distributions that are only known up to proportionality. Some examples are models characterized by an intractable likelihood, complex prior specifications or general non-conjugate Bayesian models. The goal of this work is to propose a novel approach to approximate a $\mathbb{R}^d$-valued probability distribution π for which the normalizing constant is unknown in closed form. Although the distribution π can be approximated by numerical integration in low-dimensional settings, the problem is typically challenging in higher dimensions due to the curse of dimensionality. Popular approximation schemes include optimization-based methods such as Laplace approximation, variational inference (VI, Blei et al., 2017) and expectation propagation (EP, Minka, 2001), as well as Monte Carlo (MC) approaches such as Markov chain MC, (adaptive) importance sampling (IS, Bugallo et al., 2017) or sequential Monte Carlo (SMC, Del Moral et al., 2006). However, optimization-based approximations are often too inaccurate while MC can be computationally too expensive. This motivates hybrid approaches that blend features of optimization and MC methods (Jerfel et al., 2021). We propose a novel hybrid method that adaptively interpolates between the computational speed and approximate nature of VI (Khan and Nielsen, 2018) and the accuracy and computational cost of IS.

The task of approximating a density known only up to proportionality typically arises in Bayesian statistics, where intractable posterior distributions are commonly encountered. Let $x \in \mathbb{R}^d$ indicate a d-dimensional parameter vector and $y \in \mathbb{R}^n$ be the data, so that $\ell(y \,|\, x)$ provides the model likelihood. Let p_0 be the prior distribution of the parameter x and p_y the marginal density of y. Then, by Bayes' rule, the posterior distribution of x given y is obtained as $\pi(x) = \ell(y \,|\, x)\, p_0(x)/p_y(y)$, where we indicate the distributions and the corresponding densities with the same symbols. The normalizing constant $p_y(y)$ of the posterior density π is often intractable and cumbersome to approximate in high-dimensional settings, such as Bayesian inverse problems (Stuart, 2010).

We devise an adaptive IS (AIS) scheme that iteratively adapts a proposal distribution q by matching its sufficient statistics with an annealed version q_γ of the target distribution

π. Specifically, q_γ is obtained through an adaptive *damping*[1] mechanism that guarantees a target effective sample size (ESS) used for quantifying the MC error. In particular, $q_\gamma(x) \propto q^{1-\gamma}(x)\,\pi^\gamma(x)$ where the damping parameter γ, which specifies the annealing at each iteration of the algorithm, is obtained by numerically solving a fixed lower bound on the ESS. The adaptation of both the proposal q and the damping γ motivates the name *doubly adaptive importance sampling* (DAIS). The proposed approach is based on IS and can consequently easily leverage parallel computations and modern compute environments. We highlight two contributions to the AIS literature:

1. Differently from previous AIS approaches, we allow the iterative scheme to converge to (and thus terminate at) an annealed target distribution were the amount of annealing (if any) depends on the ESS bound, sample size and shape of the target.

2. Working with Gaussian proposals,[2] we use Stein's lemma to build a variance reduction scheme that significantly enhances the robustness of the proposed method.

We set the proposed methodology in the context of variational inference (VI). In its most common form, VI determines an approximating distribution q by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}(\,q\,\|\,\pi\,)$ of π from q over a tractable family of distributions. The quantity $\mathrm{KL}(\,q\,\|\,\pi\,)$, often referred to as the *reverse* KL divergence, involves an expectation with respect to the tractable approximating distribution q. While it is relatively straightforward to implement VI, the use of the reverse KL divergence is known to often yield an approximating distribution q with lower variance than π (Minka, 2005; Li and Turner, 2016; Jerfel et al., 2021) and thus overconfident Bayesian inference. Minimizing the *forward* KL divergence $\mathrm{KL}(\,\pi\,\|\,q\,)$ can mitigate these issues but is typically computationally challenging since it involves an expectation with respect to the intractable distribution π. Linking both extremes, the α-divergence $\mathrm{K}_\alpha(\,\pi\,\|\,q\,)$ (Minka, 2005; Li and Turner, 2016) interpolates between the reverse and the forward KL divergences using the parameter α.

[1] We borrow the term 'damping' from work on expectation propagation (Vehtari et al., 2020, Section 5.2).

[2] See Section 6 for adapting DAIS to more general proposals.

Recall that damping is based on ESS, which closely relates to the χ^2-divergence and thus MC error (Sanz-Alonso, 2018). Still, we find that the fixed points of DAIS' iterative procedure can be described as a stationary point of the functional $q \mapsto \mathrm{K}_\alpha(\, \pi \,\|\, q \,)$ for a parameter $0 < \alpha < 1$ related to the amount of damping used within our method. The proposed adaptive damping scheme thus produces an automatic trade-off between minimizing the computationally more convenient reverse KL divergence and the forward KL divergence, which yields more accurate approximations. Finally, we establish that, in the limit of maximally damped updates (i.e. slow updates), our method corresponds to minimizing the reverse KL divergence $\mathrm{KL}(\, q \,\|\, \pi \,)$ with a natural-gradient descent scheme.

The rest of the article is organized as follows. Section 2 introduces DAIS. Section 3 discusses related work. Section 4 analyses the link with natural-gradient VI and α-divergence. Section 5 presents empirical results. Section 6 concludes.

2 Doubly Adaptive Importance Sampling

2.1 Algorithm Setting and Notation

We consider a target distribution on $\mathbb{R}^d$ with strictly positive and continuously differentiable density π with respect to the Lebesgue measure. We constrain ourselves to building Gaussian approximations of π, although parts of our development also apply to other families of approximating distributions, as discussed in Section 6. DAIS iteratively builds a sequence of Gaussian approximations to the target distribution π. At iteration $t \geq 1$, the current Gaussian approximation is denoted by $q_t(x) \equiv \mathcal{N}(x \,|\, \mu_t, \Gamma_t)$ for a mean vector $\mu_t \in \mathbb{R}^d$ and positive-definite covariance matrix Γ_t. DAIS requires the log-target density $\log \pi(x)$ to be known up to an additive constant and its gradient $\nabla_x \log \pi(x)$ to be efficiently evaluated: if a computer program exists for the evaluation of $\log \pi(x)$, then algorithmic differentiation can provide its gradient at a similar computational cost as evaluation of the target density (Griewank and Walther, 2008, Section 3.3).

Throughout the article, we denote expectations with respect to a distribution η by $\mathbb{E}_\eta[\varphi(X)] = \int \varphi(x)\, \eta(x)\, dx$. For two vectors $u, v \in \mathbb{R}^d$, we indicate their inner product by $\langle u, v \rangle = \sum_{i=1}^{d} u_i v_i$, and their outer product by $u \otimes v \equiv u\,v^\top \in \mathbb{R}^{d \times d}$. For two vector-valued functions $U : \mathbb{R}^d \to \mathbb{R}^{d_U}$ and $V : \mathbb{R}^d \to \mathbb{R}^{d_V}$, and an $\mathbb{R}^d$-valued random variable X with distribution η, the covariance matrix between the random variables $U(X)$ and $V(X)$ is denoted as $\mathrm{cov}_\eta[U(X), V(X)] \in \mathbb{R}^{d_U \times d_V}$. We denote the covariance matrix by $\mathrm{cov}_\eta[X] = \mathrm{cov}_\eta[X, X]$. Finally, the KL divergence between two densities $p \ll q$ is defined as $\mathrm{KL}(\, p \,\|\, q \,) = \mathbb{E}_p[\log\{p(X)/q(X)\}]$.

2.2 Adaptive Gaussian Approximations

We can assess the quality of the approximation to the target distribution $\pi(x)$ by measuring the closeness of the two probability distributions. This can be done in different ways, such as minimizing the forward or reverse KL divergences. These divergences measure closeness differently. DAIS builds a Gaussian approximation $q_\star$ to the target distribution whose first two moments are matched to those of π. This corresponds to minimizing the forward KL (Bishop, 2006, Equation (10.187)),

$$q_\star \;=\; \mathrm{argmin}\, \{\mathrm{KL}(\, \pi \,\|\, q \,) \;:\; q \in \mathcal{Q}_{\mathrm{gauss}}\},$$

where $\mathcal{Q}_{\mathrm{gauss}}$ denotes the exponential family of Gaussian distributions. This objective is desirable in a Bayesian setting as the posterior mean and covariance are often of interest.

The DAIS procedure is initialized from a user-specified Gaussian approximation $q_1(x) \equiv \mathcal{N}(x \,|\, \mu_1, \Gamma_1)$. Approaches for setting the initial approximation include using (a Gaussian approximation to) the prior distribution or a Laplace approximation to the target distribution π, which we use in Section 5.2. Also, more sophisticated procedures are possible such as running DAIS multiple times with random initializations. The quality of the initial approximation can greatly influence the number of iterations required for convergence and might affect quality of the final approximation, e.g. if π is multimodal.

Given the current approximation q_t, an improved Gaussian approximation q_{t+1} is obtained by matching its moments with an annealed distribution q_{t,γ_t} that interpolates between the current Gaussian distribution q_t and the target distribution π. The annealing is used since directly targeting π might result in too high MC error of the moment estimates if q_t is too different from π. The intermediate target, defined as $q_{t,\gamma_t}(x) \propto q_t^{1-\gamma_t}(x)\,\pi^{\gamma_t}(x)$, is coined the *damped target density* in analogy with damped expectation propagation (EP) (Vehtari et al., 2020, Section 5.2). The parameter $0 < \gamma_t \leq 1$ is referred to as the *damping parameter* with smaller γ_t corresponding to more damping of the update from q_t to q_{t+1}. Furthermore,

$$q_{t,\gamma_t}(x) \propto q_t(x)\, e^{\gamma_t \Phi_t(x)} \qquad \text{for} \qquad \Phi_t(x) = \log \pi(x) - \log q_t(x). \tag{1}$$

The function $\Phi_t : \mathbb{R}^d \to \mathbb{R}$ captures the discrepancy between the current approximation q_t and the target distribution π. Since DAIS only requires the gradient of Φ_t, the target distribution can be specified up to a multiplicative constant.

The mean μ_{t,γ_t} and covariance Γ_{t,γ_t} of the damped target density q_{t,γ_t} can be expressed as perturbations of the current parameters μ_t and Γ_t:

$$\begin{cases} \mu_{t,\gamma_t} = \mathbb{E}_{q_{t,\gamma_t}}[X] = \mu_t + \gamma_t\, \mathrm{g}_\mu(q_{t,\gamma_t}) \\ \Gamma_{t,\gamma_t} = \mathrm{cov}_{q_{t,\gamma_t}}[X] = \Gamma_t + \gamma_t\, \mathrm{G}_\Gamma(q_{t,\gamma_t}) \end{cases} \quad \text{where} \quad \begin{cases} \mathrm{g}_\mu(q_{t,\gamma_t}) = \mathbb{E}_{q_{t,\gamma_t}}[\Gamma_t \nabla \Phi_t(X)] \\ \mathrm{G}_\Gamma(q_{t,\gamma_t}) = \mathrm{cov}_{q_{t,\gamma_t}}[\Gamma_t \nabla \Phi_t(X), X]. \end{cases} \tag{2}$$

The proof of the identities in (2) is presented in Section 2.3 and Appendix A. Furthermore, Section 4 connects the quantities $\mathrm{g}_\mu(q_{t,\gamma_t})$ and $\mathrm{G}_\Gamma(q_{t,\gamma_t})$ to the (negative) natural gradients of both the functionals $q_t \mapsto \mathrm{KL}(\, q_t \,\|\, \pi\,)$ and $q_t \mapsto \mathrm{KL}(\, \pi \,\|\, q_t\,)$. Since both $\mathrm{g}_\mu(q_{t,\gamma_t})$ and $\mathrm{G}_\Gamma(q_{t,\gamma_t})$ are expressed as expectations with respect to the damped target density q_{t,γ_t}, these quantities can be estimated as $\widehat{\mathrm{g}}_\mu(q_{t,\gamma_t})$ and $\widehat{\mathrm{G}}_\Gamma(q_{t,\gamma_t})$ with self-normalized IS with $S \geq 1$ samples generated from q_t.

The damping parameter $0 < \gamma_t \leq 1$, that controls the closeness of the intermediate distribution q_{t,γ_t} to the target distribution π, is chosen adaptively so that the ESS is above a user-specified threshold $1 < N_{\mathrm{ESS}} < S$. Here, we condition on S samples $x_{t,1:S} = (x_{t,1}, \ldots, x_{t,S})$ from the proposal distribution q_t. Then, the associated ESS is computed as:

$$\mathrm{ESS}\{w_{t,1:S}(\gamma)\} = \frac{\big\{ \sum_s w_{t,s}(\gamma) \big\}^2}{\sum_s w_{t,s}^2(\gamma)} \qquad \text{with} \qquad w_{t,s}(\gamma) = \exp\{\gamma\, \Phi_t(x_s)\} \propto \frac{q_{t,\gamma}(x_{t,s})}{q_t(x_{t,s})}.$$

Since $\gamma \mapsto ESS\{w_{t,1:S}(\gamma)\}$ is a continuous and decreasing function of $0 < \gamma \leq 1$ (Beskos et al., 2016, Lemma 3.1), then the least amount of damping given the particles $x_{t,1:S}$ at iteration t can efficiently be computed with a standard root-finding method such as the bisection method, solving:

$$\gamma_t = \max \Big\{ \gamma \in (0, 1] \;:\; \mathrm{ESS}\{w_{t,1:S}(\gamma)\} \geq N_{\mathrm{ESS}} \Big\}. \tag{3}$$

Algorithm 1 Doubly adaptive importance sampling

```
1. Initialize the algorithm with $q_1(x) \equiv \mathcal{N}(x \,|\, \mu_1, \Gamma_1)$.

2. For $t = 1, 2, \dots$ until the ELBO no longer improves:

    (a) Sample $x_{t,s} \sim q_t(x) \equiv \mathcal{N}\left(x \,|\, \mu_t, \Gamma_t\right)$ independently for $s = 1, \dots, S$.

    (b) For a damping parameter $\gamma$, the unnormalized importance weights are:
        $w_{t,s}(\gamma) = \exp\{\gamma\, \Phi_t(x_{t,s})\} \propto q_{t,\gamma}(x_{t,s})/q_t(x_{t,s})$
        and the ESS is then:
        $\mathrm{ESS}_t(\gamma) = \big\{ \sum_{s=1}^S w_{t,s}(\gamma) \big\}^2 / \sum_{s=1}^S w_{t,s}^2(\gamma)$
        For a given threshold $1 < N_{\mathrm{ESS}} < S$, update the damping parameter $\gamma$:
         $\gamma_t = 1,$                                                          if  $\mathrm{ESS}_t(1) \geq N_{\mathrm{ESS}}$
         $\gamma_t = \max\left\{\gamma \in (0,1) : \mathrm{ESS}_t(\gamma) \geq N_{\mathrm{ESS}}\right\},$   if  $\mathrm{ESS}_t(1) < N_{\mathrm{ESS}}$

    (c) Compute $q_{t+1}(x) \equiv \mathcal{N}(x \,|\, \mu_{t+1}, \Gamma_{t+1})$ using the samples $x_{t,s}$ weighted by $w_{t,s}(\gamma_t)$:
        set $\mu_{t+1}$ and $\Gamma_{t+1}$ equal to an IS estimate of the right-hand sides of (2).

3. Return the final Gaussian $q_{t+1}$ as the approximation to the target distribution $\pi$.
```

The resulting scheme is summarized in Algorithm 1.

As explained in Section 4, the quantities $\mathrm{g}_\mu(q_{t,\gamma_t})$ and $\mathrm{G}_\Gamma(q_{t,\gamma_t})$ can heuristically be thought of as (natural) gradients. This motivates the updates

$$\begin{cases} \mu_{t+1} = \mu_t + \zeta_t \, \widehat{\mathrm{g}}_\mu(q_{t,\gamma_t}) \\ \Gamma_{t+1} = \Gamma_t + \zeta_t \, \widehat{\mathrm{G}}_\Gamma(q_{t,\gamma_t}) \end{cases}$$

for a sequence of learning rates $\zeta_t > 0$. The choice $\zeta_t = \gamma_t$ corresponds to matching the first two moments of q_{t+1} to the (estimate of) the first two moments of the damped target distribution q_{t,γ_t}. Since $\widehat{\mathrm{g}}_\mu(q_{t,\gamma_t})$ and $\widehat{\mathrm{G}}_\Gamma(q_{t,\gamma_t})$ are only stochastic estimates, for improved stability, we advocate choosing $\zeta_t = c\,\gamma_t$ for a *robustness parameter* $0 < c \leq 1$. We monitor convergence and decide when to terminate Algorithm 1 using the Evidence Lower BOund (ELBO) as discussed in Appendix C.

2.3 Control Variate for Gaussian Perturbations

Since q_{t,γ_t} is a perturbation of the current Gaussian approximation q_t, estimating the moments μ_{t,γ_t} and Γ_{t,γ_t} from scratch is statistically suboptimal. Instead, we derive update equations using Stein's (1972) identity: for a probability density $p(x)$ on $\mathbb{R}^d$ and a continuously differentiable test function $\varphi : \mathbb{R}^d \to \mathbb{R}^d$, we have that (Oates et al., 2017, Proposition 2)

$$\mathbb{E}_p[(\langle \nabla \log p, \varphi \rangle + \mathrm{div}\varphi)(X)] = 0 \tag{4}$$

Equation (4) follows from an integration by parts that is justified under mild growth and regularity assumptions (Mira et al., 2013; Oates et al., 2017). As derived in Appendix A, applying (4) to the annealed density q_{t,γ_t} and appropriate test functions gives Equation (2).[3] The identities in (2) show that, given knowledge of μ_t and Γ_t, the first two moments of q_{t,γ_t} can be estimated with a root-mean-square error (RMSE) of order $\gamma_t/\sqrt{S}$ when using S samples. A standard IS procedure that estimates these two quantities from scratch (i.e. without exploiting knowledge of μ_t and Γ_t) would typically lead to an RMSE of order $1/\sqrt{S}$, i.e. the MC error does not vanish as $\gamma_t \to 0$. Specifically, using results and regularity conditions from Agapiou et al. (2017), we have the following (see Appendix B for a proof).

Proposition 1. *Assume that* $\mathbb{E}_{q_t}[e^{2\gamma_t \Phi_t(X)}] < \infty$.

(i) *If* $\mathbb{E}_{q_t}[X_i^2\, X_j^2\, e^{2\gamma_t \Phi_t(X)}] < \infty$ *for all* i, j, *then the self-normalized IS estimators based on the standard moments in the left-hand side of* (2) *have an RMSE of order* $1/\sqrt{S}$ *as* $S \to \infty$.

(ii) *If* $\mathbb{E}_{q_t}[\{\nabla_i \Phi_t(X)\}^2\, e^{2\gamma_t \Phi_t(X)}] < \infty$ *and* $\mathbb{E}_{q_t}[\{\nabla_i \Phi_t(X)\}^2 X_j^2\, e^{2\gamma_t \Phi_t(X)}] < \infty$ *for all* i, j, *then the self-normalized IS estimators based on the gradient-based expressions in the right-hand side of* (2) *have an RMSE of order* $\gamma_t/\sqrt{S}$ *as* $S \to \infty$.

Furthermore, when q_t is a good approximation to the target distribution π, which is expected as the DAIS procedure progresses, the discrepancy function Φ_t and its gradient typically become small, leading to improved robustness. We conclude this section by illustrating the

[3] A similar use of Stein's lemma for a gradient flow in VI appears in Section 4.2.2 of Chen et al. (2023).

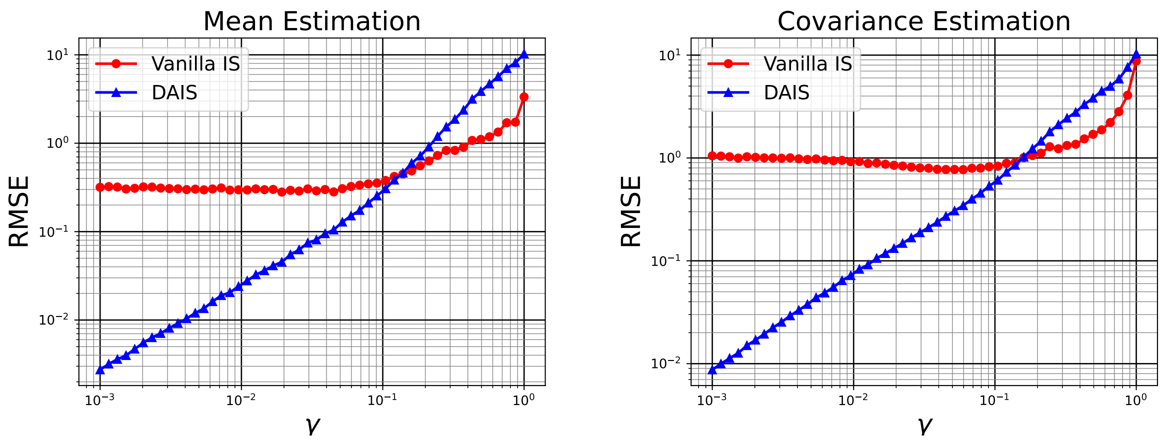

Figure 1: Estimation of $\widehat{\mu}_{t,\gamma_t}$ and $\widehat{\Gamma}_{t,\gamma_t}$ as a function of γ_t with standard self-normalized IS and with DAIS based on (2) with $S = 10^2$ particles. The proposal $q_t(x) = \mathcal{N}(x \,|\, 0, \mathbf{1}_D)$ is a standard isotropic distribution in dimension $D = 10$ and the target distribution is also Gaussian $\pi(x) = \mathcal{N}(x \,|\, m, \Sigma)$ with mean $m = (1, \ldots, 1) \in \mathbb{R}^D$ and covariance Σ with $\Sigma_{i,j} = 0.9 + 0.1\, \delta(i = j)$.

statistical advantages of estimating the first two moments μ_{t,γ_t} and Γ_{t,γ_t} of q_{t,γ_t} through (2) when compared to a naive IS estimation of these two quantities.

Example. *Consider a tractable setting where $q_t(x) = \mathcal{N}(x \,|\, 0, \mathrm{I}_d)$ is a standard isotropic distribution of dimension $d = 10$ and the target distribution is also Gaussian $\pi(x) = \mathcal{N}\,(x \,|\, m, \Sigma)$ with mean $m = (1, \ldots, 1)$ and covariance Σ with $\Sigma_{i,j} = 0.9 + 0.1\, \delta(i = j)$. Figure 1 reports, as a function of γ_t, the RMSE quantities $\mathbb{E}[\|\widehat{\mu}_{t,\gamma_t} - \mu_{t,\gamma_t}\|^2]^{1/2}$ and $\mathbb{E}[\|\widehat{\Gamma}_{t,\gamma_t} - \Gamma_{t,\gamma_t}\|_{\mathrm{F}}^2]^{1/2}$, where $\|M\|_{\mathrm{F}}^2 = \sum M_{i,j}^2$ is the squared Frobenius norm of the matrix M. The RMSEs are approximated with 10^2 independent experiments and the IS estimates use $S = 10^2$ particles.*

3 Related Work

We now provide an overview of existing posterior approximation approaches to contextualize the proposed approach DAIS and highlight relevant connections.

3.1 Methods Based on Importance Sampling

We start by reviewing AIS (Bugallo et al., 2017) with Gaussian proposal distributions to approximate the target π. AIS is firstly initialized to q_1, a Gaussian distribution with mean μ_1 and covariance Γ_1, for instance obtained from the prior distribution or from an initial approximation to the target π. Then, the proposal q_{t+1} at each iteration is determined via the IS approximation to π from the previous iteration. That is, μ_{t+1} and Γ_{t+1} are chosen so that q_{t+1} is closer to π, resulting in a more accurate IS approximation.

AIS (Bugallo et al., 2017) and DAIS iteratively improve q_t via moment matching. A novelty of DAIS is in the choice of the new parameters μ_{t+1} and Γ_{t+1} in (2), which is based on an application of Stein's lemma. Another difference from previous AIS schemes is the adaptation of the IS target in the ultimate iteration to guarantee ESS. Ryu and Boyd (2015), Schuster (2015), Akyildiz and Míguez (2021), Elvira and Chouzenoux (2022), and Elvira et al. (2023) propose AIS approaches that, like ours, use the gradient of the target π to update the proposal. Though, their gradients appear as part of optimization algorithms while ours derives from moment matching with Stein's lemma.

Smoothing of IS weights, of which the damping in DAIS is a special case, has been employed in IS (e.g. Koblents and Míguez, 2013; Vehtari et al., 2024) and in AIS (Paananen et al., 2021). Most similarly to DAIS, Koblents and Míguez (2013) base their decision whether to temper the weights on ESS, though they do not adapt the amount of tempering to the target. Like DAIS but with a different smoothing method, Paananen et al. (2021) use a stopping criterion based on the regularity of the weights to determine the number of AIS iterations. Their proposal distributions, arising from specific tasks such as Bayesian cross-validation, are complicated while DAIS considers Gaussian proposals with a focus on approximating the posterior mean and covariance of the target distribution. As we do in Section 4, Guilmeau et al. (2024a) link an AIS scheme to α-divergence minimization. There, α is adapted through a one-to-one correspondence with the tail behaviour of the IS proposal distribution.

Similarly to DAIS, SMC (Del Moral et al., 2006) and annealed IS (Neal, 2001) adapt the target and proposal across iterations. Moreover, automatic tempering using ESS is also used in adaptive SMC (e.g. Chopin and Papaspiliopoulos, 2020, Algorithm 17.3). In these methods, proposals are discrete distributions based on reweighted, resampled or rejuvenated particles while DAIS adapts a Gaussian proposal. Typically, proposal adaptation is based on geometric averaging which coincides with the damping in (1) and entropic mirror descent on $\mathrm{KL}(\, q_t \,\|\, \pi \,)$ (Chopin et al., 2024), though moment matching has been explored as well (Grosse et al., 2013). DAIS uses both damping and moment matching simultaneously.

3.2 Optimization-based Methods

Moment matching is fundamental to EP (Minka, 2001) and expectation consistent approximate inference (Opper and Winther, 2005). These methods usually entail such matching iteratively across factors of the target density π. That is, expectations are propagated across a Bayesian network. This contrasts with our matching, which uses all of π at once, e.g. $\mu_{t+1} = \mathbb{E}_{q_{t,\gamma_t}}[X]$. Disregarding this discrepancy, the damped moment matching of DAIS is equivalent to damping in EP (Vehtari et al., 2020, Section 5.2) and to the α-divergence minimization scheme in Equation (18) of Minka (2005). Like DAIS, the EP methods by Wiegerinck and Heskes (2003), Minka (2004) and Hernández-Lobato et al. (2016) minimize the α-divergence. The updates in (2) involve taking the expectation, or smoothing, of gradients. Dehaene (2016) links smoothed gradients to EP and minimization of α-divergence.

Prangle and Viscardi (2022) consider the same damped target q_{t,γ_t}, also based on ESS, while iteratively updating an IS proposal q_t as in DAIS. Differences include that q_t is not a Gaussian distribution but a normalizing flow and that q_t is updated via gradient descent for the objective $\mathrm{KL}(\, q_{t,\gamma_t} \,\|\, q_{t+1} \,)$. DAIS updates q_t through moment matching, e.g. $\mu_{t+1} = \mathbb{E}_{q_{t,\gamma_t}}[X]$, which directly targets the minimizer of $\mathrm{KL}(\, q_{t,\gamma_t} \,\|\, q_{t+1} \,)$.

There is a VI literature on improving variational objectives and approximating families via MC (e.g. Li and Turner, 2016; Ruiz and Titsias, 2019) including IS (e.g. Domke and Sheldon, 2018; Wang et al., 2018) and AIS (Han and Liu, 2017; Jerfel et al., 2021). DAIS constitutes a substantially different hybrid between VI and MC as it performs AIS that happens to recover natural-gradient VI via damping as $\gamma_t \to 0$ (see Section 4). Some VI methods (e.g. Li and Turner, 2016) replace the reverse KL divergence by α-divergence, which DAIS effectively also minimizes (see Section 4). Moreover, Daudel et al. (2023) and Guilmeau et al. (2024b) obtain damped moment matching updates as in DAIS. However, DAIS, derived as AIS instead of a change in VI objective, yields principled adaption of the damping parameter $\alpha = \gamma_t$. In this context, Wang et al. (2018) consider adaptation of the divergence based on tail probabilities of importance weights. Yao et al. (2018) evaluate the accuracy of VI using IS. Liu and Wang (2016) use Stein's lemma for VI. They minimize the reverse KL divergence via a functional gradient descent derived from the Stein discrepancy. Han and Liu (2017) expand that Stein VI method to use AIS.

4 Analysis of DAIS

Implementing the DAIS algorithm with learning rate $\zeta_t = \gamma_t$ corresponds to iteratively matching, up to MC variability, the first two moments of the damped target density $q_{t,\gamma}$ to the ones of the next Gaussian approximation q_{t+1}. This demonstrates that the algorithm does not depend on the mean-covariance parametrization of the Gaussian family, and that any other parametrization would lead to exactly the same sequence of Gaussian approximations. This remark motivates the connections described in this section between DAIS and the natural-gradient descent method (Amari, 1998). We defer derivations to Appendix D.

While the standard gradient is the steepest descent direction when the usual Euclidean distance is used, the natural gradient is the steepest descent direction in the space of distributions where distance is measured by the KL divergence (Martens, 2020). In particular, natural-gradient flows are parametrization invariant. In the Gaussian setting of this article, the natural-gradient flow for minimizing a loss function $\mathcal{L}(\mu, \Gamma)$ over the space of Gaussian distributions $q \in \mathcal{Q}_{\text{gauss}}$ is

$$\begin{cases} \dfrac{d\mu}{dt} = -\Gamma\,\nabla_\mu \mathcal{L} \\ \dfrac{d\Gamma^{-1}}{dt} = 2\,\nabla_\Gamma \mathcal{L} \end{cases} \qquad \Longleftrightarrow \qquad \begin{cases} \dfrac{d\mu}{dt} = -\Gamma\,\nabla_\mu \mathcal{L} \equiv -\widetilde{\nabla}_\mu \mathcal{L} \\ \dfrac{d\Gamma}{dt} = -2\,\Gamma(\nabla_\Gamma \mathcal{L})\Gamma \equiv -\widetilde{\nabla}_\Gamma \mathcal{L}, \end{cases} \tag{5}$$

where $\widetilde{\nabla}_\mu \mathcal{L} = \Gamma\,\nabla_\mu \mathcal{L}$ and $\widetilde{\nabla}_\Gamma \mathcal{L} = 2\,\Gamma(\nabla_\Gamma \mathcal{L})\Gamma$ denote the natural gradient with respect to the mean and covariance parameters, respectively. For $\mathrm{g}_\mu(q_{t,\gamma_t})$ and $\mathrm{G}_\Gamma(q_{t,\gamma_t})$ defined in (2),

$$\begin{cases} \lim\limits_{\gamma_t \to 0} \mathrm{g}_\mu(q_{t,\gamma_t}) = \Gamma \mathbb{E}_{q_t}[\, \nabla_x \log \pi(X)] = -\widetilde{\nabla}_\mu \mathrm{KL}(\, q_t \,\|\, \pi \,) \\ \lim\limits_{\gamma_t \to 0} \mathrm{G}_\Gamma(q_{t,\gamma_t}) = \Gamma \, \mathbb{E}_{q_t}[\nabla^2_{xx} \log \pi(X)] \, \Gamma + \Gamma = -\widetilde{\nabla}_\Gamma \mathrm{KL}(\, q_t \,\|\, \pi \,). \end{cases} \tag{6}$$

Equation (6) shows that, in the limit of small damping parameter $\gamma_t \to 0$, DAIS can be understood as a natural-gradient descent for minimizing the reverse KL. Furthermore,

$$\begin{cases} \lim\limits_{\gamma_t \to 1} \mathrm{g}_\mu(q_{t,\gamma_t}) = \mu_\pi - \mu = -\widetilde{\nabla}_\mu \mathrm{KL}(\, \pi \,\|\, q_t \,) \\ \lim\limits_{\gamma_t \to 1} \mathrm{G}_\Gamma(q_{t,\gamma_t}) = \Gamma_\pi - \Gamma = -\widetilde{\nabla}_\Gamma \mathrm{KL}(\, \pi \,\|\, q_t \,) - (\mu_\pi - \mu) \otimes (\mu_\pi - \mu) \end{cases} \tag{7}$$

where $\mu_\pi = \mathbb{E}_\pi[X]$ and $\Gamma_\pi = \mathrm{cov}_\pi[X]$. Equation (7) establishes a connection between DAIS and the natural-gradient flow for minimizing the forward KL, whose global minimizer is indeed given by the Gaussian distribution with first two moments matching those of the target distribution π.

To conclude this section, we characterize the limiting distribution obtained by the DAIS methodology. For this purpose, assume that the DAIS algorithm has converged towards an approximating distribution $q_\infty(x) = \mathcal{N}(x \mid \mu_\infty, \Gamma_\infty)$ with final damping parameter $0 < \gamma_\infty < 1$. The moment matching conditions mean that

$$\mathbb{E}_{q_\infty}[T(X)] \;=\; \mathbb{E}_{q_{\infty,\gamma_\infty}}[T(X)] \tag{8}$$

where $q_{\infty,\gamma_\infty}(x) \propto q_\infty^{1-\gamma_\infty}(x) \, \pi^{\gamma_\infty}(x)$. In the identity above, $T : \mathbb{R}^d \to \mathbb{R}^{d+d(d+1)/2}$ equals $T(x) = (x_i, x_i x_j)_{i \leq j}$, representing the sufficient statistic vector for a d-dimensional Gaussian distribution in its natural parametrization. Recall that the Gaussian family can be parametrized as $q_\lambda(x) = \exp(\langle \lambda, T(x) \rangle)/Z(\lambda)$ for natural parameter $\lambda \in \Lambda \subset \mathbb{R}^{d+d(d+1)/2}$ and associated normalizing constant $Z(\lambda) > 0$. We remark that the following can be generalized to any natural exponential family. Condition (8) describes the stationary points of the α-divergence functional $\lambda \mapsto \mathrm{K}_\alpha(\, \pi \,\|\, q_\lambda \,)$ (Hernández-Lobato et al., 2016, Equation (7)). Consequently, (8) shows that the limiting Gaussian distribution q_∞ is a stationary point of the α-divergence functional $\lambda \mapsto \mathrm{K}_\alpha(\, \pi \,\|\, q_\lambda \,)$ when choosing $\alpha = \gamma_\infty$. Since $\mathrm{K}_\alpha(\, \pi \,\|\, q \,) \to \mathrm{KL}(\, \pi \,\|\, q \,)$ as $\alpha \to 1$ and $\mathrm{K}_\alpha(\, \pi \,\|\, q \,) \to \mathrm{KL}(\, \pi \,\|\, q \,)$ as $\alpha \to 0$, this result further indicates that large update parameters γ_t are to be favored since minimizing $\mathrm{KL}(\, \pi \,\|\, q_t \,)$ is preferred over minimizing $\mathrm{KL}(\, q_t \,\|\, \pi \,)$.

5 Applications

This section compares the performance of DAIS with other approximations. Additionally, Appendix H considers an inverse problem where, without any problem-specific adjustments or reduced approximation accuracy, DAIS is faster than an approximation that exploits the structure of the problem. In Algorithm 1, We set the ESS threshold N_{ESS} to 10^3, the importance sample size S to 10^5 and the robustness parameter c to 0.5 unless otherwise specified. We use the Python package JAX (Bradbury et al., 2018) for automatic differentiation to obtain $\nabla_x \Phi_t(x)$ and for parallelization of importance samples across CPU cores.

5.1 Two-dimensional Synthetic Examples

As a first example, we consider two bivariate distributions from Ruiz and Titsias (2019) as their low dimensionality allows for easy inspection of approximations. Specifically, we consider the banana-shaped target distribution

$$\pi(x) \propto \mathcal{N}\left\{ \left. \begin{pmatrix} x_1 \\ x_2 + x_1^2 + 1 \end{pmatrix} \right| 0, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \right\}$$

and the mixture of two Gaussian distributions

$$\pi(x) = 0.3\mathcal{N}\left\{ x \left| \begin{pmatrix} 0.8 \\ 0.8 \end{pmatrix}, \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \right. \right\} + 0.7\mathcal{N}\left\{ x \left| \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \begin{pmatrix} 1 & -0.6 \\ -0.6 & 1 \end{pmatrix} \right. \right\}$$

visualized in Figure 2.

We approximate the distributions using Gaussian proposals. Algorithm 1, i.e. DAIS, reaches $\gamma_t = 1$ in 2 and 3 iterations for the banana-shaped and the mixture distribution, respectively. Appendix E considers a lower sample size S resulting in a final γ_t less than one. For comparison, we run VI with the same full covariance Gaussian distribution as DAIS uses. Specifically, we minimize the reverse KL divergence $\text{KL}(\, q_t \,\|\, \pi \,)$ via gradient descent. Additionally, Appendix F compares with adaptive IS methods.

Figure 2 summarizes the results. DAIS captures the mean and covariance of the target distribution π more accurately than VI. In particular, the DAIS approximation for the mixture distribution is virtually indistinguishable from the exact mean and covariance.

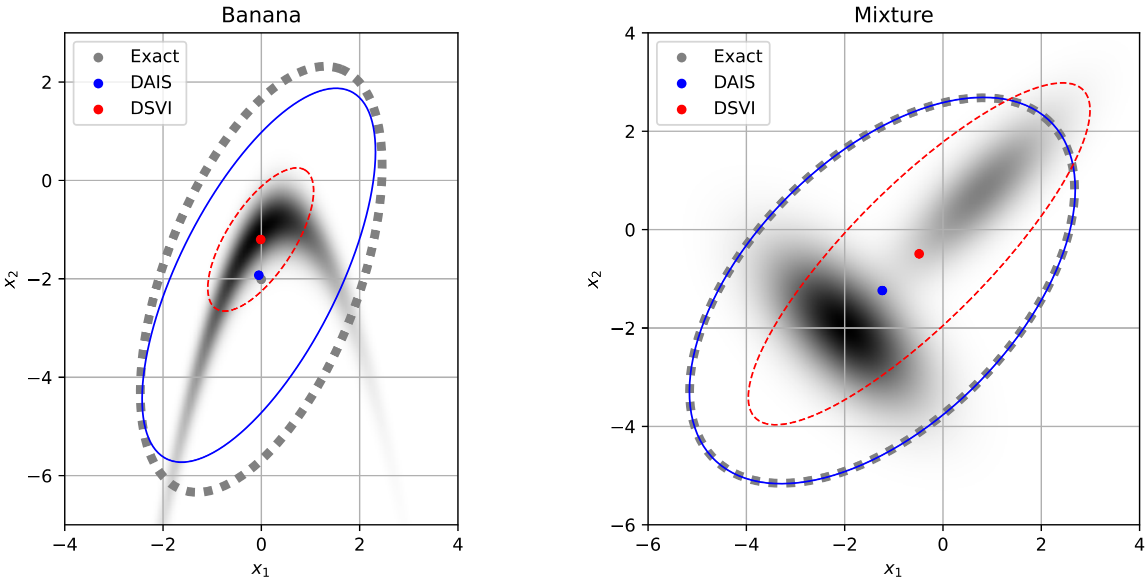

Figure 2: The banana-shaped and mixture densities in grayscale with mean (dot) and 95% credible regions (ellipse) of the corresponding Gaussian approximations overlaid. The Gaussian approximations have their moments equal to the exact moments (dotted), the DAIS estimates (solid) and the VI estimates (dashed).

This shows the benefit of minimizing $\mathrm{KL}(\,\pi \,\|\, q_t\,)$ instead of $\mathrm{KL}(\,q_t \,\|\, \pi\,)$. The covariance underestimation by DAIS for the banana-shaped distribution is likely due to the ESS estimator in Step 2b of Algorithm 1 underestimating MC error (Martino et al., 2017; Elvira et al., 2022). Additionally, the results in Appendix E where $\gamma_t < 1$ are in line with the fact that the adaptation of γ_t interpolates between moment matching and VI.

5.2 Logistic Regression

Lastly, we apply DAIS to the four logistic regression examples from Section 4.2.1 of Ong et al. (2018). Each data set consists of a binary response $y_i \in \{-1, 1\}$ and a d-dimensional feature vector $a_i \in \mathbb{R}^d$ for $1 \leq i \leq n$ where n is the number of cases. Then, the likelihood is $\ell(y \,|\, x) = \prod_{i=1}^n 1/\{1 + \exp(-y_i \langle a_i, x\rangle)\}$ where x is the coefficient vector. The prior on x is $p_0(x) = \mathcal{N}(x \,|\, 0, 10\, \mathrm{I}_d)$ such that the posterior follows as $\pi(x) \propto \mathcal{N}\,(x \,|\, 0, 10\, \mathrm{I}_d)\, \ell(y \,|\, x)$.

The data involved are binarized versions of the spam, krkp, ionosphere and mushroom

Table 1: Number of cases, number of categorical and continuous attributes, resulting number of predictors d, number of iterations and computation time in seconds of DAIS for the logistic regression examples.

Data name	Cases	Categorical	Continuous	d	Iterations	Time
Spam	4,601	57	0	105	7	24.3s
Krkp	3,196	0	36	38	6	14.0s
Ionosphere	351	32	0	111	12	11.3s
Mushroom	8,124	0	22	96	10	58.5s

data from the UCI Machine Learning Repository (Dua and Graff, 2017). The binarization follows Gelman et al. (2008, Section 5.1). First, any continuous attributes are discretized using the method from Fayyad and Irani (1993). Then, the resulting set of categorical attributes are encoded using dummy variables with the most frequent category as baseline. These dummy variables plus an intercept constitute the d predictors considered. The resulting problem dimensionalities are summarized in Table 1.

Algorithm 1 approximates the mean and covariance of π with as initial approximation $q_1(x) = \mathcal{N}(x \,|\, \mu_1, \Gamma_1)$ the posterior mode $\mu_1 = \arg\max_x \pi(x)$ and the inverse Hessian of the negative log-density $\Gamma_1 = \{-\nabla^2 U(\mu_1)\}^{-1}$. Table 1 also lists computation times, using an Intel i5-10600 CPU with six cores, and number of DAIS iterations. To assess approximation accuracy, we run Hamiltonian MC using the Python package BlackJAX (Cabezas et al., 2024) for 100,000 iterations, of which 10,000 are burn-in iterations.

Figure 3 shows that DAIS provides highly accurate estimates of posterior moments. In Appendix G, the estimates are compared with variational inference with mean-field and full-covariance structure, as well as with adaptive multiple importance sampling (AMIS, Cornuet et al., 2012). To explore the effect of reduced MC error from the application of Stein's identity (Section 2.3) on the approximation of π, Figure S10 in Supplementary Material is the same as Figure 3 except that it uses standard moment matching instead of the gradient-based updates in (2). Comparing these figures reveals that the methodology

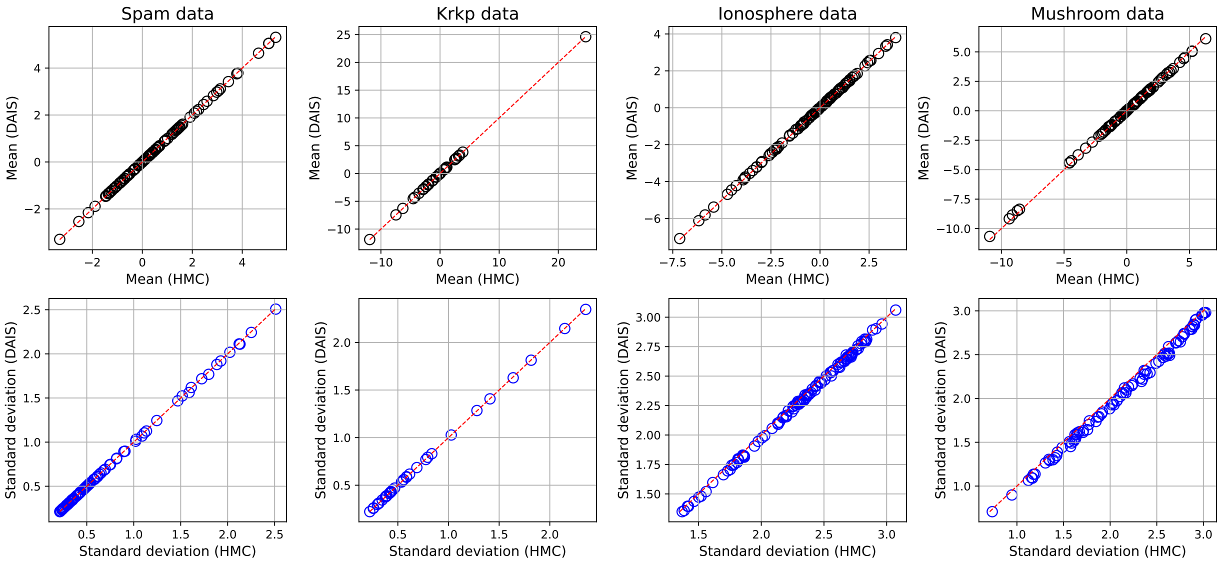

Figure 3: Scatter plots of the DAIS estimates versus the Hamiltonian MC estimates of the posterior means and standard deviations for the logistic regression examples.

from Section 2.3 indeed improves approximation accuracy.

6 Discussion

We propose a novel iterative approach to the approximation of the posterior distribution in general Bayesian models. The methodology is based on producing a sequence of Gaussian distributions whose moments match those of a damped target distribution, thus adapting to the target. The sequence is identified by exploiting Stein's lemma, which provides an updating rule for two consecutive sets of moments. The moments are computed via importance sampling while damping of the target is used to control the effective sample size (ESS) of the samples in the importance sampling. The adaptation guarantees that the ESS is above a pre-specified threshold, which controls MC error, and provides a trade-off between minimizing the reverse and forward KL divergences based on computational constraints. We call the method *doubly adaptive importance sampling* (DAIS). DAIS is a general methodology and competitive with methods that are more tailored to a problem-specific posterior.

DAIS inherits certain limitations from IS. Firstly, high dimensionality typically results in too low ESS for IS to be feasible and translates to exceedingly high damping in DAIS. Also, a large number of samples S requires considerable computer memory. At the same time, the standard estimate for ESS used here can underestimate MC error as alluded to in Section 5.1, especially if the proposal q_t is far from the target π. See for instance Martino et al. (2017), who also explore alternative measures of ESS with markedly different behavior such as the inverse of the maximum normalized IS weight: they find it to result in lower MC error at the same number of resampling steps in an SMC context. Such alternative ESS measures are readily incorporated in DAIS. Similarly, schemes that minimize the MC error resulting from the self-normalization in IS could be explored (Branchini and Elvira, 2024): we have focused on self-normalized IS for simplicity. Furthermore, methods for regularizing IS weights can be applied to smoothen the weights beyond what is achieved by damping (Vehtari et al., 2024).

In the event that the IS estimate for the covariance Γ_{t+1} in Algorithm 1 is not positive-definite, standard post-processing methods can be used for transforming the estimate into a positive-definite version of it. Possible approaches include setting the negative eigenvalues to small positive numbers. A more principled approach consists in reducing the damping parameter γ_t. Since the computational bottleneck generally lies in the evaluation of the target density π, recomputing the mean and covariance estimates for a reduced damping parameter γ_t is generally computationally straightforward since no additional evaluation of the target density π is necessary.

The fact that q_t is Gaussian limits how well it can approximate the target π. To go beyond this limitation, π can instead be approximated by the importance-weighted samples from the last iteration of DAIS. The approximation can be made arbitrarily accurate by increasing the number of samples S in this last iteration. Additionally, samples from multiple iterations of DAIS can be combined to approximate π as in Equation (17) of Bugallo et al. (2017), especially if the amount of damping γ_t is nearly constant over these iterations, or by treating the proposals across iterations as components of a mixture proposal distribution as in Cornuet et al. (2012).

Another avenue for increasing accuracy is going beyond Gaussianity for q_t. The Gaussian constraint enables the MC error reduction in (2). Other aspects of DAIS such as its adaptation and the analysis in Section 4 do not require Gaussianity. Moreover, Lin et al. (2019) extend Stein's lemma beyond Gaussian distributions to mixtures of an exponential family, potentially enabling MC error reduction similar to (2) for more general q_t. As such, ideas behind DAIS can be used with non-Gaussian approximating distributions.

In IS, it is typically desirable that the proposal has heavy tails to improve robustness. Therefore, one might consider a distribution $\overline{q}_t \neq q_t$ as proposal such as the multivariate t-distribution with location μ_t and scale Γ_t. However, we have found that the choice can result in exceedingly small ESS for a high dimensionality d, and thus large amounts of damping. After all, if π is a posterior, it will often have Gaussian-like tails, e.g. because of Bernstein-von Mises. Similarly, π is then also likely unimodal. Thus, we do not consider the multivariate t or mixture distributions further.

SUPPLEMENTARY MATERIAL

Supplement: Derivations, discussion on monitoring convergence of Algorithm 1 and additional empirical results. (DAIS_supplement.pdf, PDF file)

Code: Scripts that produce the empirical results along with a readme file are available at https://github.com/thiery-lab/dais. (DAIS_code.zip, GitHub repository)

The authors report there are no competing interests to declare.

References

Agapiou, S., O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science 32*(3), 405–431.

Akyildiz, O. D. and J. Míguez (2021). Convergence rates for optimised adaptive importance samplers. *Statistics and Computing 31*(2), 12.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation 10*(2), 251–276.

Beskos, A., A. Jasra, N. Kantas, and A. Thiery (2016). On the convergence of adaptive sequential Monte Carlo methods. *The Annals of Applied Probability 26*(2), 1111–1146.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York.

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

Bradbury, J., R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, and S. Wanderman-Milne (2018). JAX: composable transformations of Python+NumPy programs. http://github.com/google/jax.

Branchini, N. and V. Elvira (2024). Generalizing self-normalized importance sampling with couplings. arXiv:2406.19974v1.

Bugallo, M. F., V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine 34*(4), 60–79.

Cabezas, A., A. Corenflos, J. Lao, R. Louf, A. Carnec, K. Chaudhari, et al. (2024). BlackJAX: Composable Bayesian inference in JAX. arXiv:2402.10797v2.

Chen, Y., D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart (2023). Gradient flows for sampling: Mean-field models, Gaussian approximations and affine invariance. arXiv:2302.11024v7.

Chopin, N., F. R. Crucinio, and A. Korba (2024). A connection between tempering and entropic mirror descent. arXiv:2310.11914v3.

Chopin, N. and O. Papaspiliopoulos (2020). *An Introduction to Sequential Monte Carlo*. Springer Series in Statistics. Springer Nature Switzerland.

Cornuet, J.-M., J.-M. Marin, A. Mira, and C. P. Robert (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics 39*(4), 798–812.

Daudel, K., R. Douc, and F. Roueff (2023). Monotonic alpha-divergence minimisation for variational inference. *Journal of Machine Learning Research 24*, 62.

Dehaene, G. P. (2016). Expectation propagation performs a smoothed gradient descent. arXiv:1612.05053v1.

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(3), 411–436.

Domke, J. and D. R. Sheldon (2018). Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

Dua, D. and C. Graff (2017). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml.

Elvira, V. and E. Chouzenoux (2022). Optimized population Monte Carlo. *IEEE Transactions on Signal Processing 70*, 2489–2501.

Elvira, V., E. Chouzenoux, O. D. Akyildiz, and L. Martino (2023). Gradient-based adaptive importance samplers. *Journal of the Franklin Institute 360*(13), 9490–9514.

Elvira, V., L. Martino, and C. P. Robert (2022). Rethinking the effective sample size. *International Statistical Review 90*(3), 525–550.

Fayyad, U. M. and K. B. Irani (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, Vol. 2*, pp. 1022–1027.

Gelman, A., A. Jakulin, M. G. Pittau, and Y. Su (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics 2*(4), 1360–1383.

Griewank, A. and A. Walther (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation* (2nd ed.). Society for Industrial and Applied Mathematics, Philadelphia, PA.

Grosse, R. B., C. J. Maddison, and R. R. Salakhutdinov (2013). Annealing between distributions by averaging moments. In *Advances in Neural Information Processing Systems*, Volume 26. Curran Associates, Inc.

Guilmeau, T., N. Branchini, E. Chouzenoux, and V. Elvira (2024a). Adaptive importance sampling for heavy-tailed distributions via α-divergence minimization. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, Volume 238 of *Proceedings of Machine Learning Research*, pp. 3871–3879. PMLR.

Guilmeau, T., E. Chouzenoux, and V. Elvira (2024b). Regularized Rényi divergence minimization through Bregman proximal gradient algorithms. arXiv:2211.04776v4.

Han, J. and Q. Liu (2017). Stein variational adaptive importance sampling. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press.

Hernández-Lobato, J., Y. Li, M. Rowland, T. Bui, D. Hernández-Lobato, and R. Turner (2016). Black-box α-divergence minimization. In *Proceedings of The 33rd International Conference on Machine Learning*, Volume 48 of *Proceedings of Machine Learning Research*, New York, NY, pp. 1511–1520. PMLR.

Jerfel, G., S. Wang, C. Wong-Fannjiang, K. A. Heller, Y. Ma, and M. I. Jordan (2021). Variational refinement for importance sampling using the forward Kullback-Leibler divergence. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, Volume 161 of *Proceedings of Machine Learning Research*, pp. 1819–1829. PMLR.

Khan, M. E. and D. Nielsen (2018). Fast yet simple natural-gradient descent for variational inference in complex models. In *2018 International Symposium on Information Theory and Its Applications (ISITA)*, pp. 31–35. IEEE.

Koblents, E. and J. Míguez (2013). A population Monte Carlo scheme with transformed weights and its application to stochastic kinetic models. *Statistics and Computing 25*(2), 407–425.

Li, Y. and R. E. Turner (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, Volume 29. Curran Associates, Inc.

Lin, W., M. E. Khan, and M. Schmidt (2019). Stein's lemma for the reparameterization trick with exponential family mixtures. arXiv:1910.13398v1.

Liu, Q. and D. Wang (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc.

Martens, J. (2020). New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research 21*, 146.

Martino, L., V. Elvira, and F. Louzada (2017). Effective sample size for importance sampling based on discrepancy measures. *Signal Processing 131*, 386–401.

Minka, T. (2004). Power EP. Technical Report MSR-TR-2004-149, Microsoft Research Ltd., Cambridge, UK.

Minka, T. (2005). Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd., Cambridge, UK.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 362–369.

Mira, A., R. Solgi, and D. Imparato (2013). Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing 23*(5), 653–662.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing 11*(2), 125–139.

Oates, C. J., M. Girolami, and N. Chopin (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(3), 695–718.

Ong, V. M.-H., D. J. Nott, and M. S. Smith (2018). Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics 27*(3), 465–478.

Opper, M. and O. Winther (2005). Expectation consistent approximate inference. *Journal of Machine Learning Research 6*, 2177–2204.

Paananen, T., J. Piironen, P.-C. Bürkner, and A. Vehtari (2021). Implicitly adaptive importance sampling. *Statistics and Computing 31*(2), 16.

Prangle, D. and C. Viscardi (2022). Distilling importance sampling. arXiv:1910.03632v4.

Ruiz, F. and M. Titsias (2019). A contrastive divergence for combining variational inference and MCMC. In *Proceedings of the 36th International Conference on Machine Learning*, Volume 97 of *Proceedings of Machine Learning Research*, pp. 5537–5545. PMLR.

Ryu, E. K. and S. P. Boyd (2015). Adaptive importance sampling via stochastic convex programming. arXiv:1412.4845v2.

Sanz-Alonso, D. (2018). Importance sampling and necessary sample size: An information theory approach. *SIAM/ASA Journal on Uncertainty Quantification 6*(2), 867–879.

Schuster, I. (2015). Gradient importance sampling.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, Volume 6, pp. 583–603. University of California Press.

Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numerica 19*, 451–559.

Vehtari, A., A. Gelman, T. Sivula, P. Jylänki, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, and C. P. Robert (2020). Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *Journal of Machine Learning Research 21*, 17.

Vehtari, A., D. Simpson, A. Gelman, Y. Yao, and J. Gabry (2024). Pareto smoothed importance sampling. arXiv:1507.02646v9.

Wang, D., H. Liu, and Q. Liu (2018). Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc.

Wiegerinck, W. and T. Heskes (2003). Fractional belief propagation. In *Advances in Neural Information Processing Systems 15*. MIT Press.

Yao, Y., A. Vehtari, D. Simpson, and A. Gelman (2018). Yes, but did it work?: Evaluating variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80 of *Proceedings of Machine Learning Research*, pp. 5581–5590. PMLR.

Supplement to “Doubly Adaptive Importance Sampling” published in the Journal of Computational and Graphical Statistics

Willem van den Boom, Andrea Cremaschi and Alexandre H. Thiery

A Derivation of Equation (2)

For the standard orthonormal basis $(e_1, \ldots, e_d)$ of $\mathbb{R}^d$, consider the constant test functions $\varphi^{[i]}(x) = e_i$ for $1 \leq i \leq d$. An application of Stein's identity (4) to these functions $\varphi^{[i]}$ gives $\mathbb{E}_{q_{t,\gamma_t}}[\nabla \log q_{t,\gamma_t}(X)] = 0$. The first line of Equation (2) follows now from

$$\nabla \log q_{t,\gamma_t}(x) = -\Gamma_t^{-1}(x - \mu_t) + \gamma_t \nabla \Phi_t(x). \tag{S1}$$

Similar manipulations of (4) show that for a function $F : \mathbb{R}^d \to \mathbb{R}^d$ with Jacobian matrix $\mathbb{J}_F(x)_{ij} = \partial_{x_j} F_i(x)$ and using test functions $\varphi^{[ij]}(x) = e_i F_j(x)$,

$$\mathbb{E}_{q_{t,\gamma_t}}\left[\nabla \log q_{t,\gamma_t}(X) \otimes F(X) + \mathbb{J}_F^\top(X)\right] = 0_{d\times d}.$$

Inserting (S1) and $F(X) = X - \mu_{t,\gamma_t}$ where $\mu_{t,\gamma_t} = \mathbb{E}_{q_{t,\gamma_t}}[X]$ yields that

$$\mathbb{E}_{q_{t,\gamma_t}}[(X - \mu_t) \otimes (X - \mu_{t,\gamma_t})] = \Gamma_t + \gamma_t \Gamma_t \, \mathbb{E}_{q_{t,\gamma_t}}[\nabla \Phi_t(X) \otimes (X - \mu_{t,\gamma_t})]$$

from which the second line of (2) follows.

B Proof of Proposition 1

By Equation (2.3) of Agapiou et al. (2017), we have that self-normalized IS estimation of $\mathbb{E}_{q_{t,\gamma_t}}[\varphi(X)]$ with S samples from q_t has an RMSE of $1/\sqrt{S}$ as $S \to \infty$ under the following two conditions:

$$\begin{gathered}\mathbb{E}_{q_t}[\{q_{t,\gamma_t}(X)/q_t(X)\}^2] \propto \mathbb{E}_{q_t}[e^{2\gamma_t \Phi_t(X)}] < \infty \\ \mathbb{E}_{q_t}[\{\varphi(X)\, q_{t,\gamma_t}(X)/q_t(X)\}^2] \propto \mathbb{E}_{q_t}[\varphi^2(X)\, e^{2\gamma_t \Phi_t(X)}] < \infty\end{gathered} \tag{S2}$$

These conditions are satisfied in part (ii) of Proposition 1 by assumption. Furthermore, the presence of the factor γ_t in the right-hand side of (2) results in the asymptotic RMSE of order $\gamma_t/\sqrt{S}$.

For part (i), consider the unnormalized density $f(x) = q_t(x)\, e^{2\gamma_t \Phi_t(x)}$. Then,

$$\mathbb{E}_{q_t}[X_i^2\, e^{2\gamma_t\Phi_t(X)}]^2 = \left\{\int x_i^2\, q_t(x)\, e^{2\gamma_t\Phi_t(x)} dx\right\}^2 = \mathbb{E}_f[X_i^2]^2$$

$$< \mathbb{E}_f[X_i^4] = \int x_i^4\, q_t(x)\, e^{2\gamma_t\Phi_t(x)} dx = \mathbb{E}_{q_t}[X_i^4\, e^{2\gamma_t\Phi_t(X)}]$$

where the inequality follows from Jensen's inequality and $\mathbb{E}_{q_t}[X_i^4\, e^{2\gamma_t\Phi_t(X)}] < \infty$ by assumption. Thus, (S2) is also satisfied when considering self-normalized IS estimation of the left-hand side of the first line of (2). Furthermore, (S2) is satisfied for the left-hand side of the second line of (2) by assumption such that the required result follows.

C Monitoring Convergence

In challenging settings where the target distribution departs significantly from Gaussianity, running DAIS with a fixed number of IS particles S per iteration produces a sequence of damping parameters γ_t that does not eventually converge to one. Furthermore, it is typically not feasible to reliably estimate the forward KL divergence $\mathrm{KL}(\,\pi \,\|\, q_t\,)$ with MC methods. Although the trajectory $t \mapsto \gamma_t$ is typically noisy and not necessarily increasing, we observe that the damping parameter eventually (and often rapidly) reaches a stationary regime. For further monitoring of convergence, we track the Evidence Lower BOund

$$\mathrm{ELBO}(q_t) = \int_x \log\left\{\frac{\overline{\pi}(x)}{q_t(x)}\right\} q_t(x)\, dx \tag{S3}$$

where $\pi(x) = \overline{\pi}(x)/Z$ for an unknown normalization constant $Z > 0$. Producing an estimate $\widehat{\mathrm{ELBO}}$ of (S3) with importance sampling is straightforward since all quantities necessary for its evaluation would have typically already been evaluated while running the DAIS algorithm.

Figure S1 displays the trajectories of $t \mapsto \gamma_t$ and $t \mapsto -\widehat{\mathrm{ELBO}}(q_t)$ when DAIS is used for approximating the following two target densities:

(i) a $d = 2$ dimensional density $\pi(x_1, x_2)$ defined as

$$\frac{\pi(x_1, x_2)}{p_0(x_0, x_1)} \propto \exp\left[-\frac{\{x_2 - \mathcal{F}(x_1)\}^2}{2\sigma^2}\right]$$

for $\sigma = 0.1$, non-linear function $\mathcal{F}(x) = 1 + \sin(2x)$ and standard Gaussian prior density $p_0(x_0, x_1)$;

(ii) a $d = 100$ dimensional Gaussian distribution with mean $\mu = (1, \dots, 1) \in \mathbb{R}^d$ and covariance $\Gamma_{i,j} = 0.9 + 0.1\, \delta(i = j)$.

In both cases, DAIS is started from a standard multivariate Gaussian distribution, i.e. $\mu_1 = 0_{d\times 1}$ and $\Gamma_1 = \mathrm{I}_d$. We use the robustness parameters of $c = 0.1$ and $c = 0.3$ for the 2-dimensional and the 100-dimensional target, respectively. Figure S1 shows that monitoring either the damping parameter $\gamma_t \in (0, 1]$ or the ELBO leads to roughly the same conclusion.

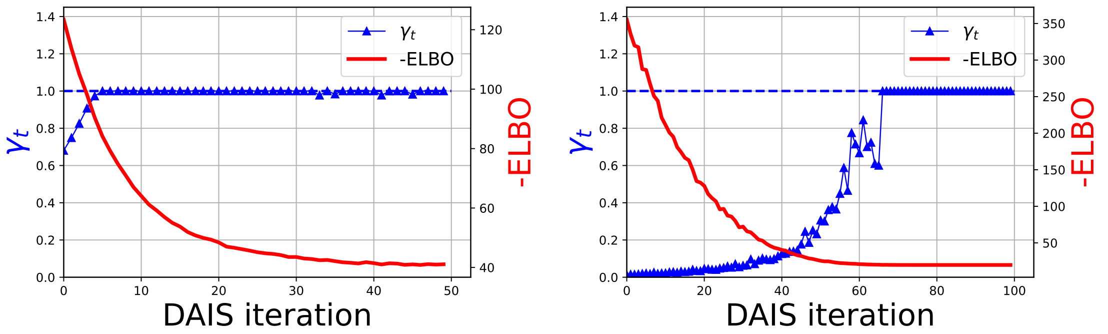

Figure S1: Tracking of the damping parameter γ_t and (estimate of) the negative ELBO for monitoring convergence. **(Left)** 2-dimensional target $\pi(x) \propto \exp[-\{x_2 - \mathcal{F}(x_1)\}^2/(2\sigma^2)]\, p_0(x_1, x_2)$ with $\mathcal{F}(x) = 1 + \sin(2x)$ and $p_0(x_1, x_2)$ the density of a standard Gaussian density in $\mathbb{R}^2$. **(Right)** $d = 100$ dimensional Gaussian target distribution with mean $\mu = (1, \ldots, 1) \in \mathbb{R}^d$ and covariance $\Gamma_{i,j} = 0.9 + 0.1\,\delta(i = j)$.

D Derivations for Section 4

A derivation of (5) can be found in Khan et al. (2017) and the equivalence between the two formulations follows from the chain rule. Since $\nabla_\Gamma \mathbb{E}_q[\varphi(X)] = \frac{1}{2}\,\mathbb{E}_q[\nabla_{xx}\varphi(X)]$ for any test function $\varphi : \mathbb{R}^d \to \mathbb{R}$ (Opper and Archambeau, 2009, Equation (A.3)), standard algebraic manipulations show that the forward (Fwd) and reverse (Rev) KL divergences satisfy:

$$\text{(Rev)} : \begin{cases} \nabla_\mu \mathrm{KL}(\, q \,\|\, \pi \,) = -\mathbb{E}_q[\nabla_x \log \pi(X)] \\ \nabla_\Gamma \mathrm{KL}(\, q \,\|\, \pi \,) = \dfrac{1}{2}\left(-\mathbb{E}_q[\nabla^2_{xx} \log \pi(X)] - \Gamma^{-1}\right) \end{cases}$$

$$\text{(Fwd)} : \begin{cases} \nabla_\mu \mathrm{KL}(\, \pi \,\|\, q \,) = -\Gamma^{-1}\, \mathbb{E}_\pi[(X - \mu)] \\ \nabla_\Gamma \mathrm{KL}(\, \pi \,\|\, q \,) = -\dfrac{1}{2}\, \mathbb{E}_\pi[\Gamma^{-1}(X - \mu) \otimes (X - \mu)\Gamma^{-1} - \Gamma^{-1}]. \end{cases}$$

It follows that the natural-gradient flow for minimizing the forward and reverse KL divergences are given by

$$\text{(Rev)} : \begin{cases} \dfrac{d\mu}{dt} = \Gamma\, \mathbb{E}_q[\nabla_x \log \pi(X)] \\ \dfrac{d\Gamma}{dt} = \Gamma\, \mathbb{E}_q[\nabla^2_{xx} \log \pi(X)]\, \Gamma + \Gamma \end{cases} \qquad \text{(Fwd)} : \begin{cases} \dfrac{d\mu}{dt} = \mathbb{E}_\pi[(X - \mu)] \\ \dfrac{d\Gamma}{dt} = \mathbb{E}_\pi[(X - \mu) \otimes (X - \mu)] - \Gamma. \end{cases}$$

Since q_{t,γ_t} converges to $q_t(x) = \mathcal{N}(x \,|\, \mu, \Gamma)$ as $\gamma_t \to 0$ and $\nabla \Phi_t(x) = \nabla \log \pi(x) + \Gamma^{-1}(x - \mu)$, the quantities $\mathrm{g}_\mu(q_{t,\gamma_t}) = \mathbb{E}_{q_{t,\gamma_t}}[\Gamma\, \nabla \Phi_t(X)]$ and $\mathrm{G}_\Gamma(q_{t,\gamma_t}) = \mathrm{cov}_{q_{t,\gamma_t}}[\Gamma\, \nabla \Phi_t(X), X]$ satisfy (6): the second equality in (6) follows from an integration by parts (or Stein's lemma). Furthermore, since q_{t,γ_t} converges to π as $\gamma_t \to 1$, the definitions $\mu_{t,\gamma_t} = \mu_t + \gamma_t\, \mathrm{g}_\mu(q_{t,\gamma_t})$ and $\Gamma_{t,\gamma_t} = \Gamma_t + \gamma_t\, \mathrm{G}_\Gamma(q_{t,\gamma_t})$ show (7).

Amari (1985) defines the α-divergence as

$$\mathrm{K}_\alpha(\, \pi \,\|\, q_\lambda \,) = \frac{1}{\alpha(1 - \alpha)} \left[1 - \int_{\mathcal{X}} \left\{ \frac{\pi(x)}{q_\lambda(x)} \right\}^\alpha q_\lambda(x)\, dx \right].$$

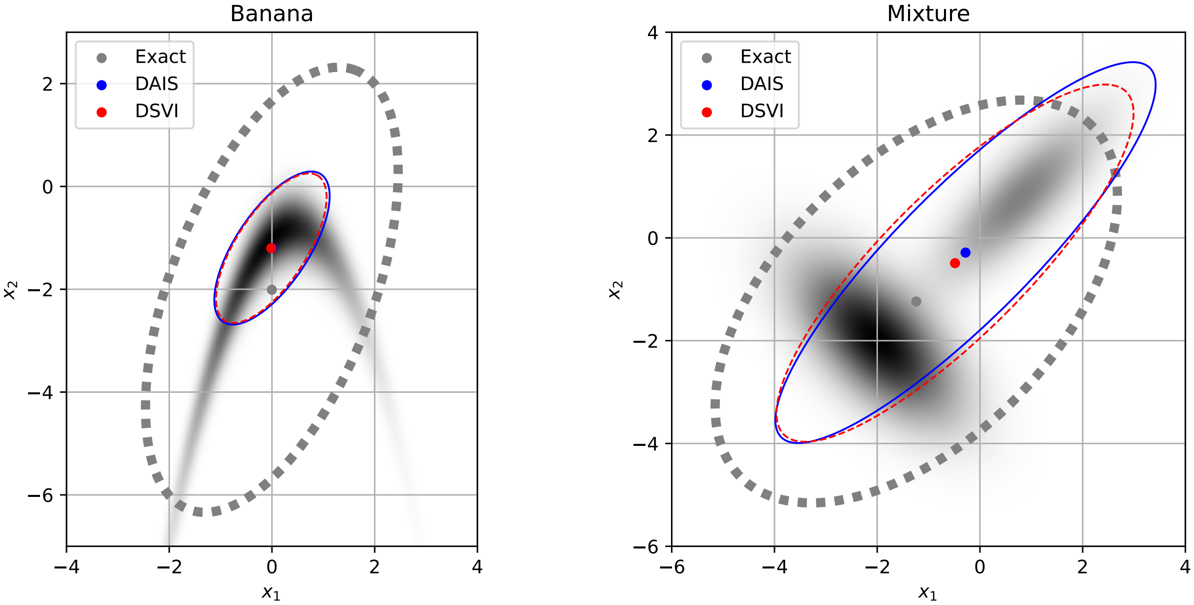

Figure S2: The banana-shaped and mixture densities in grayscale with the mean (dot) and the 95% credible regions (ellipse) of their Gaussian approximations overlaid. The Gaussian approximations have their moments equal to the exact moments (gray, dotted), the DAIS estimates using $S = 1010$ resulting in $\gamma_t < 1$ (blue, solid) and the VI estimates (red, dashed).

Since $\nabla_\lambda q_\lambda(x) = q_\lambda(x)\{T(x) - \mathbb{E}_{q_\lambda}[T(X)]\}$, we obtain (Hernández-Lobato et al., 2016, Equation (7))

$$\nabla_\lambda \mathrm{K}_\alpha(\,\pi \,\|\, q_\lambda\,) = \frac{Z(\lambda,\alpha)}{\alpha}\big(\mathbb{E}_{q_\lambda}[T(X)] - \mathbb{E}_{q_{\lambda,\alpha}}[T(X)]\big)$$

where $Z(\lambda,\alpha) = \int \pi(x)^\alpha\, q_\lambda^{1-\alpha}(x)\,dx$. Thus, condition (8) describes the stationary points of the α-divergence functional $\lambda \mapsto \mathrm{K}_\alpha(\,\pi \,\|\, q_\lambda\,)$.

E Two-dimensional Examples Converging to $\gamma_t < 1$

The set-up of Section 5.1 uses importance sample size $S = 10^5$ such that DAIS finishes with $\gamma_t = 1$. This appendix instead considers $S = 1010$, which is only slightly higher than the effective sample size threshold $N_{\text{ESS}} = 10^3$. Then, DAIS converges to $\gamma_t \approx 0.14$ and $\gamma_t \approx 0.10$ for the banana-shaped and mixture distributions, respectively. Comparing Figure 2 in the main text with Figure S2 confirms that the adaptation of γ_t indeed interpolates between moment matching and variational inference (VI).

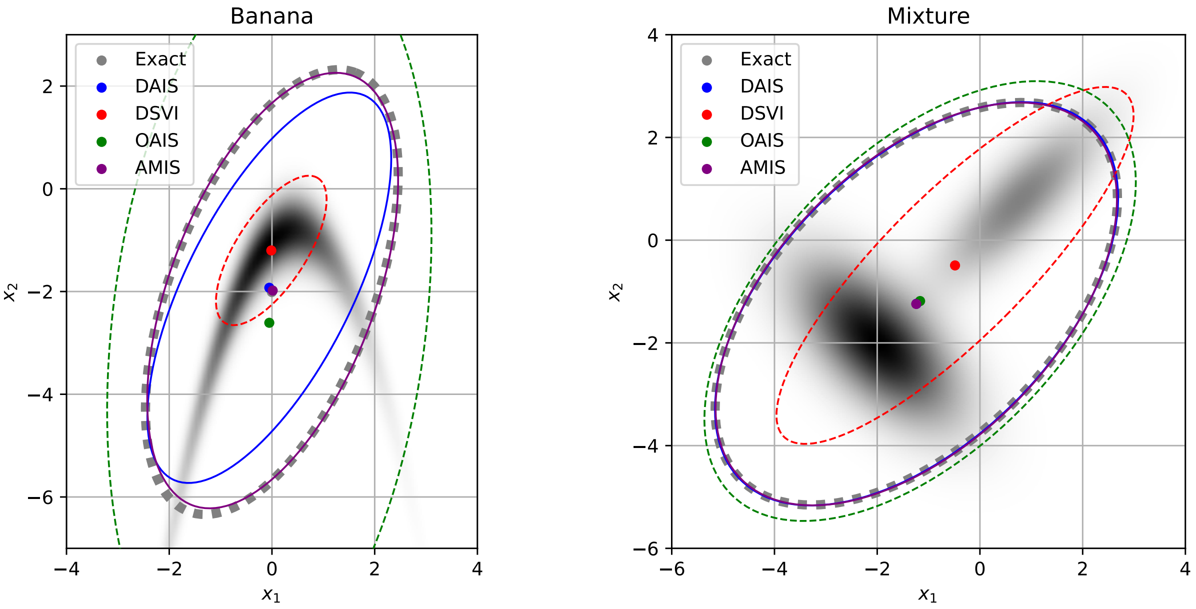

Figure S3: The banana-shaped and mixture densities are shown in grayscale, with the mean (dot) and the 95% credible regions (ellipse) of their Gaussian approximations overlaid. In these non-Gaussian examples, DSVI (red, dashed) fails to accurately approximate the second moments. In contrast, for the mixture density example, DAIS, OAIS and AMIS all provide good approximations of the second moments, with AMIS being indistinguishable from the ground truth. The severely non-Gaussian example of the banana-shaped density presents a greater challenge, with the methods approximating the second moments with varying degrees of accuracy.

F Comparisons to AMIS and OAIS: Two-dimensional Synthetic Examples

In the two-dimensional examples presented in Section 5.1, we compare DAIS with adaptive multiple importance sampling (AMIS, Cornuet et al., 2012) and optimized adaptive importance sampling (OAIS, Akyildiz and Míguez, 2021), as well as Gaussian variational inference fitted through the doubly stochastic variational inference (DSVI) approach of Titsias and Lázaro-Gredilla (2014). The resulting approximations are reported and discussed in Figure S3.

AMIS is a foundational method in the field of multiple importance sampling and has inspired numerous subsequent works (Bugallo et al., 2017). For this comparison, AMIS is implemented using a Student's t-distribution with three degrees of freedom as the proposal family. At each of the $R = 20$ rounds of adaptation, $S = 10^5$ samples were generated.

The DAIS method is also compared to OAIS (Akyildiz and Míguez, 2021), which optimizes the upper bound $R(\theta) = \mathbb{E}_{q_\theta}[\Pi^2(X)/q_\theta^2(X)]$ to minimize the mean square error (MSE) in self-normalized importance sampling. This approach employs a proposal density $q_\theta(x)$ to approximate expectations under the target density $\pi(x) = \Pi(x)/\mathcal{Z}$. While we successfully implemented OAIS for these low-dimensional examples, extending the method to higher-dimensional settings, such as the logistic regression examples considered in the next section, proved challenging. The primary difficulty lies in the ratio $\Pi^2(x)/q_\theta^2(x)$, which can span several orders of magnitude in high-dimensional scenarios, thereby complicating the adaptive procedure and its dependence on the normalization constant $\mathcal{Z}$. OAIS was implemented using $S = 10^3$ samples and 10^4 iterations of the ADAM optimizer with a learning rate of 10^{-2}. We experimented with a range of learning rates and significantly increased the number of importance samples, but the results were relatively insensitive to these choices.

In contrast, we have found that VI methods, which optimize $\mathbb{E}_{q_\theta}[\log\{q_\theta(X)/\pi(X)\}]$, are both straightforward and stable to implement. Notably, the gradient $\nabla_\theta \mathbb{E}_{q_\theta}[\log\{q_\theta(X)/\pi(X)\}]$ does not depend on the normalizing constant $\mathcal{Z}$, and the logarithm stabilizes the optimization. Conversely, we have found it challenging to reliably minimize quantities such as $R(\theta) = \mathbb{E}_{q_\theta}[\Pi^2(X)/q_\theta^2(X)]$. While our proposed method also relies on quantities of the type $\Pi(x)/q(x)$, the adaptive damping procedure is indeed one key mechanism for stabilizing the adaptation procedure.

G Comparisons to AMIS and Variational Inference: Logistic Regression

In the four logistic regression examples presented in Section 5.2, we compare DAIS with adaptive multiple importance sampling (AMIS, Cornuet et al., 2012), as well as with variational inference using both a diagonal covariance matrix and a full covariance matrix. The VI methods were implemented using the doubly stochastic variational inference framework of Titsias and Lázaro-Gredilla (2014). As detailed in Section F, we were unable to successfully apply the optimized adaptive importance sampling method of Akyildiz and Míguez (2021) in these high-dimensional settings. We report estimates of the marginal means and standard deviations, comparing them to the ground truth obtained from a Hamiltonian Monte Carlo (HMC) run at convergence.

For DAIS (Figure S4, same as Figure 3 in the main text), we used $S = 10^5$ importance samples with a target effective sample size of $N_{\text{ESS}} = 10^3$ and a robustness parameter of $c = 0.5$ as mentioned in Section 5. The AMIS method (Figure S5) was implemented using a Student's t-distribution with three degrees of freedom as the proposal family. At each of the $R = 20$ adaptation rounds, $S = 10^5$ samples were generated. The VI methods (Figures S6 and S7) were run for 10^3 iterations, with 10^3 Monte Carlo samples per iteration and an ADAM learning rate of 10^{-2}. All methods were initialized using a Laplace approximation of the target distribution.

For the spam and krkp data, both DAIS and AMIS provide highly accurate estimates. VI with full covariance is accurate as well though slightly less so. Finally, VI with a diagonal covariance has lower accuracy in terms of the posterior standard deviations. For the ionosphere and mushroom data, DAIS is most accurate. VI with full covariance performs well too while AMIS and VI with a diagonal covariance result in notable estimation error including underestimation of posterior standard deviations.

Figure S8 summarizes the computation times of the various algorithms. We used a computer with 8 CPU cores and 32 GB of RAM, specifically the machine type `e2-standard-2` from Google Cloud Platform. The results are mixed across the four datasets with DAIS

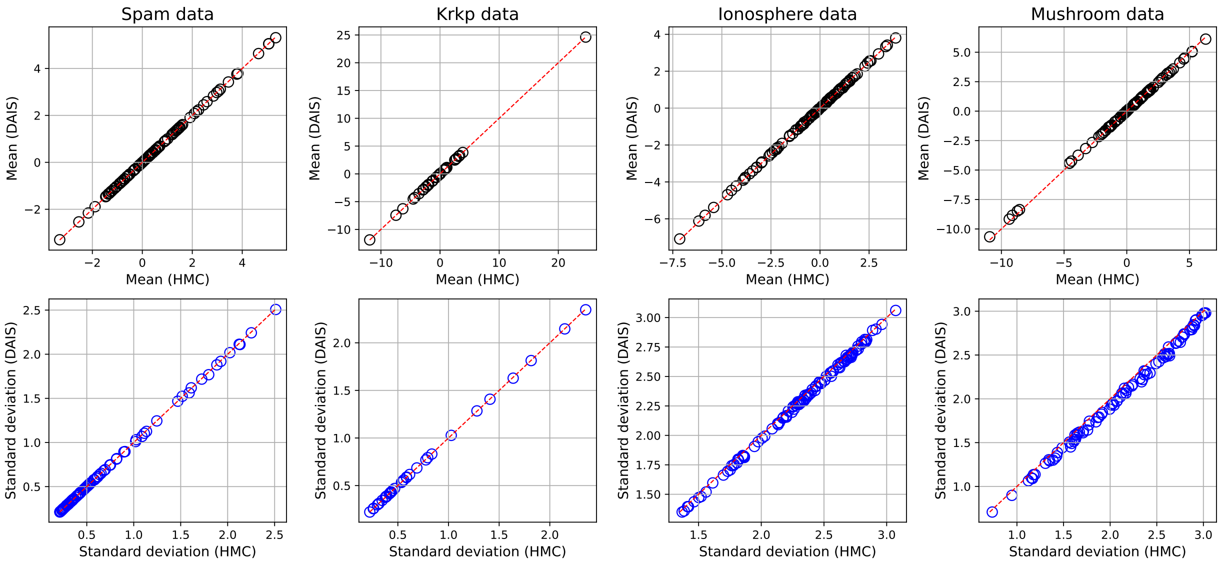

Figure S4: Scatter plots of the DAIS estimates versus the Hamiltonian MC estimates of the posterior means and standard deviations for the logistic regression examples.

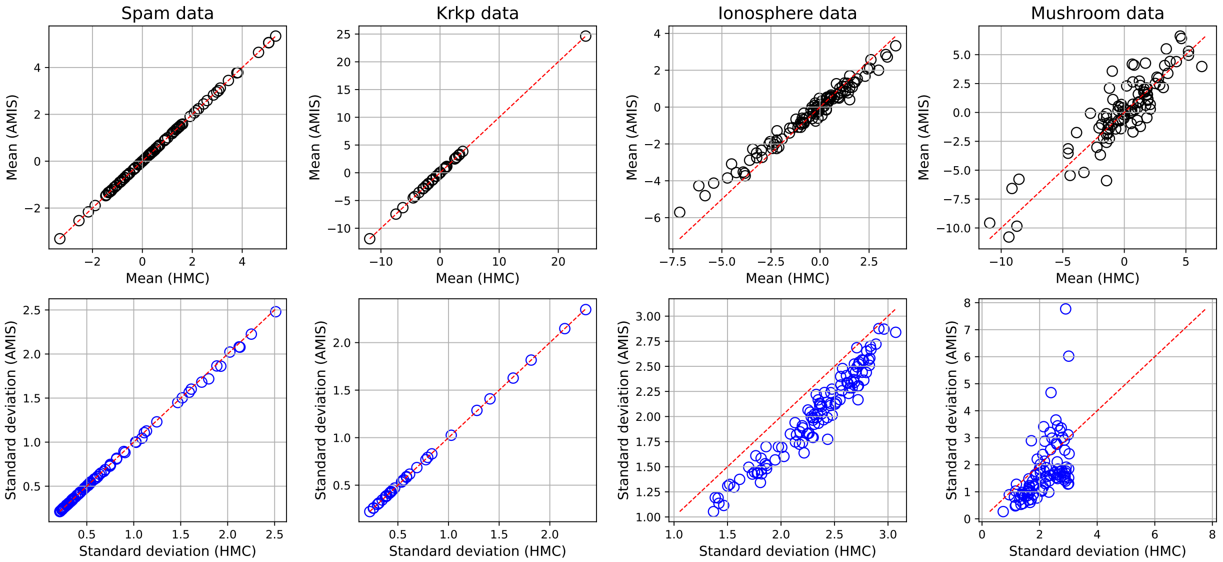

Figure S5: Scatter plots of the AMIS estimates versus the Hamiltonian MC estimates of the posterior means and standard deviations for the logistic regression examples.

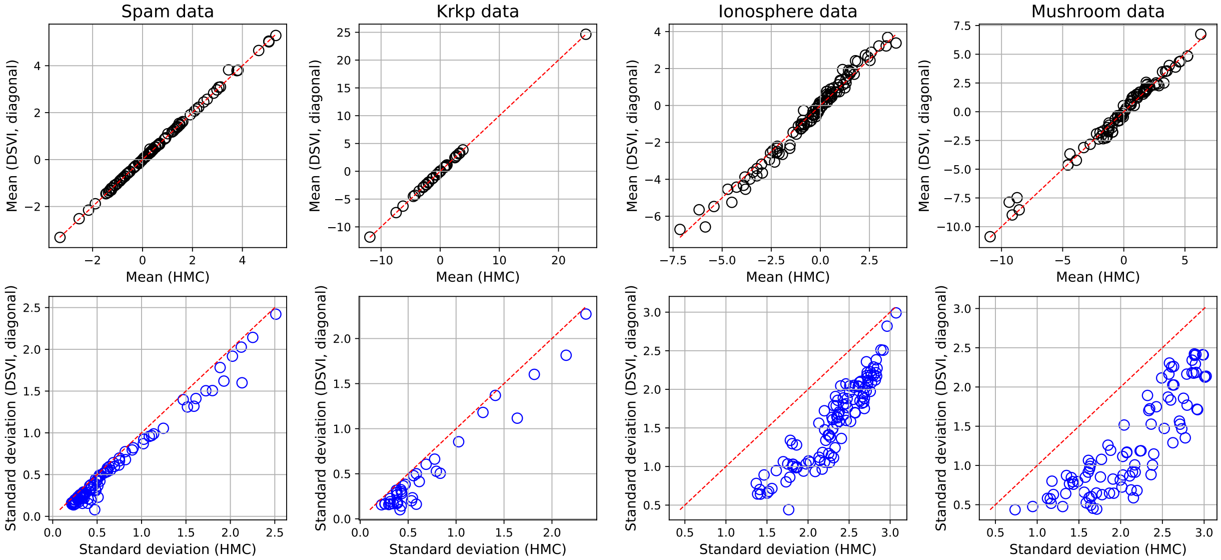

Figure S6: Scatter plots of the estimates from VI with diagonal covariance matrix versus the Hamiltonian MC estimates of the posterior means and standard deviations for the logistic regression examples.

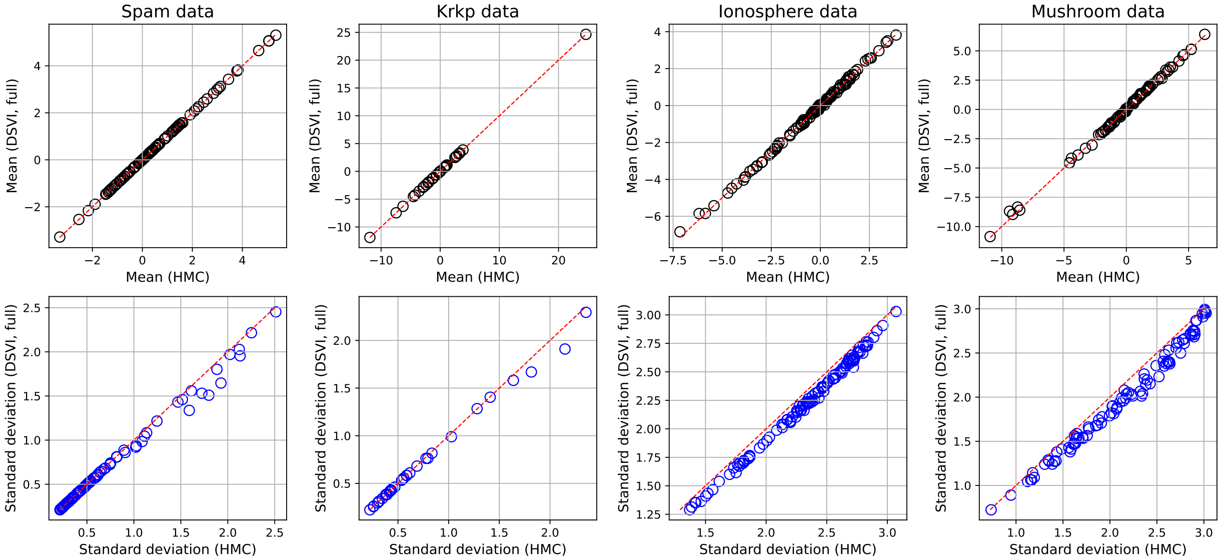

Figure S7: Scatter plots of the estimates from VI with full covariance matrix versus the Hamiltonian MC estimates of the posterior means and standard deviations for the logistic regression examples.

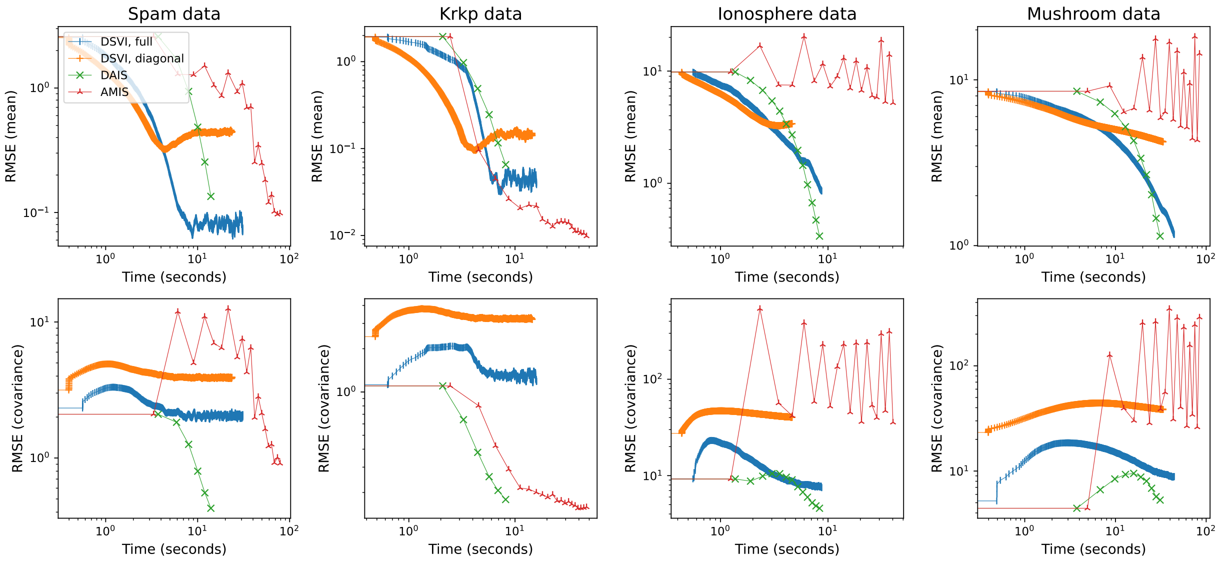

Figure S8: Approximation accuracy across iterations versus computation time. The accuracy is characterized by the RMSE quantities $\|\widehat{\mu}_t - \mu\|$ and $\|\widehat{\Gamma}_t - \Gamma\|_{\mathrm{F}}$, where μ (resp. Γ) is the posterior mean (resp. covariance) from HMC, μ_t and Γ_t are the estimates at the tth iteration of DSVI/DAIS/AMIS, and $\|M\|_{\mathrm{F}} = (\sum M_{i,j}^2)^{1/2}$ is the Frobenius norm of the matrix M.

adaptive stopping rule resulting in a trade-off between computation time and accuracy that is competitive. We note that computation times are highly dependent on implementation and tuning parameters such as number of particles, and Figure S8 should be considered with this limitation in mind.

H Synthetic Inverse Problem

We consider the inverse problem from van den Boom and Thiery (2019) as a synthetic problem that is higher-dimensional than the two-dimensional examples in Section 5.1. Consider the function $f : \mathbb{R} \to \mathbb{R}$ distributed as a zero-mean Gaussian process with covariance function $k(t,t') = \exp\{-400\,(t-t')^2\}$. Then, $x \equiv \{f\left(\frac{t-1}{d-1}\right), t = 1, \ldots, d = 100\}^\top$ is a discretization of f. Define the $d \times d$ blurring matrix G by first setting $G_{ij} = \exp\{-\min(i+j,\, d-i-j)^2/25\}$ for $1 \le i, j \le d$ and then scaling its rows to sum to one. Consider the d-dimensional vector $\overline{H}(x) = G\, x^{\odot 3}$ obtained by multiplying the matrix G by the vector $x^{\odot 3} = \{x_i^3,\, i = 1, \ldots, d\}^\top$. Then, generate a 30-dimensional vector $H(x)$ by sampling 30 elements at random with replacement from the elements of $\overline{H}(x)$ with odd indices, as a type of subsampling. Finally, we generate 30-dimensional data according to $y \sim \mathcal{N}\{H(x), \mathrm{I}_{30}\}$ with x fixed to a prior draw. The target density follows as the posterior defined by this likelihood and the Gaussian prior p_0 induced by the Gaussian process, $\pi(x) \propto p_0(x)\, \mathcal{N}\{y \,|\, H(x), I_{30}\}$. Computing π is a Bayesian inverse problem where H is the forward map that maps x to a distribution on y. The goal is to “invert” H by inferring x from y.

We compare three Gaussian approximations of π. The first is the proposed Algorithm 1 with importance sample size $S = 10^4$, $N_{\mathrm{ESS}} = 100$ and robustness parameter $c = 1$ with

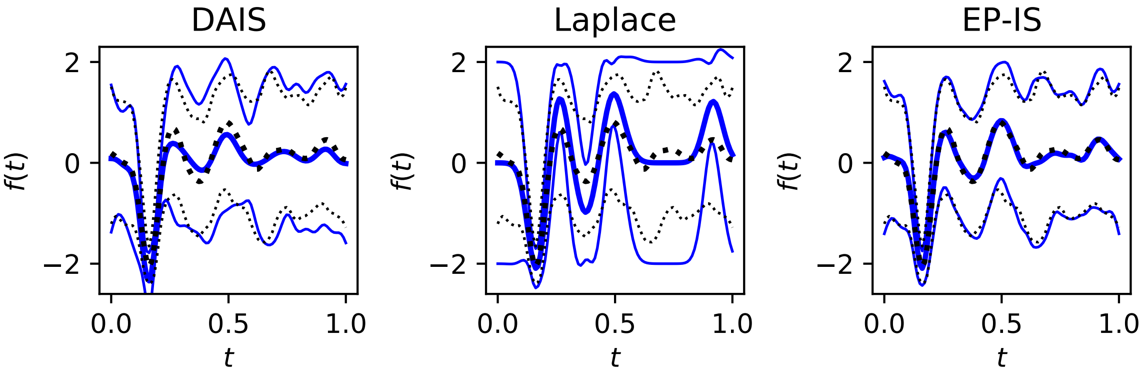

Figure S9: Posterior mean (thick line) and 2.5% and 97.5% quantiles (thin lines) of π from the inverse problem as estimated by the preconditioned Crank–Nicolson algorithm (dotted), and compared with estimates from DAIS, the Laplace approximation and EP-IS (solid).

which DAIS finishes in 20 iterations with $\gamma_t = 0.30$.[1] The second is a Laplace approximation from Steinberg and Bonilla (2014) based on a Taylor series linearization of the forward map H. Lastly, we run EP-IS with covariance matrix tapering from van den Boom and Thiery (2019), which exploits that the data are independent. As an arbitrarily accurate MC baseline, we run a preconditioned Crank–Nicolson algorithm (Cotter et al., 2013) with the prior p_0 as reference measure for 100,000 iterations, which is substantially slower than the Gaussian approximations.

Figure S9 summarizes the results of the different approximations. The Laplace approximation is both the least accurate and the fastest approximation, taking only 1.0 seconds. DAIS and EP-IS are similarly accurate. DAIS is faster than EP-IS (13 versus 16 seconds). Thus, DAIS is faster than the approximation that exploits the structure of the inverse problem without any problem-specific adjustments or reduced approximation accuracy.

[1] Here, we use the stopping criterion detailed in van den Boom et al. (2024).

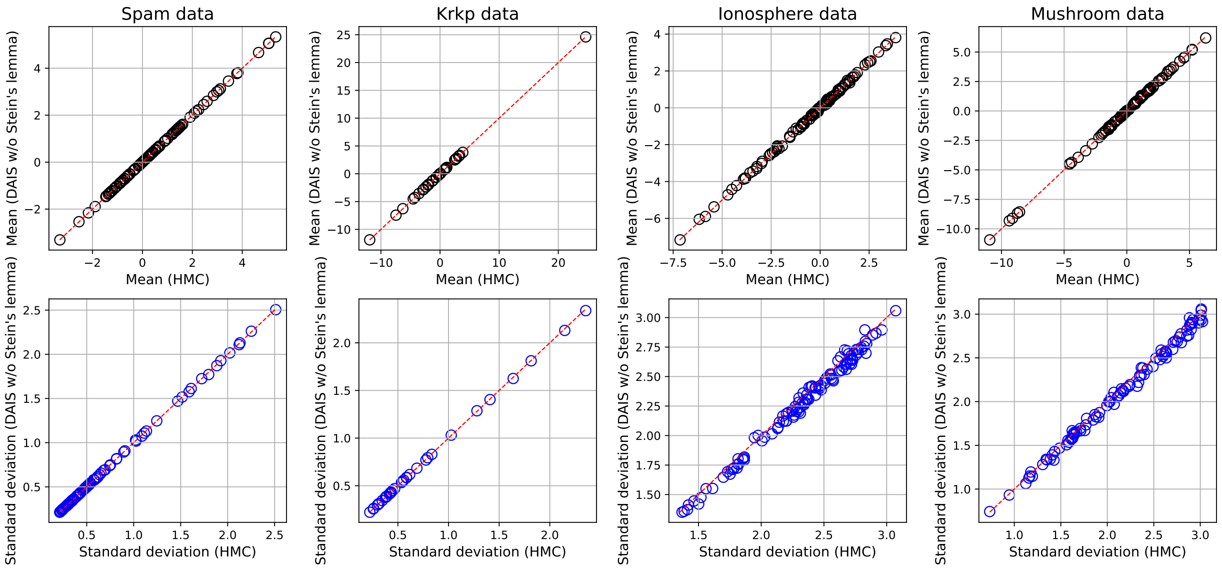

Figure S10: Scatter plots of the estimates from DAIS without using Stein's identity to reduce MC error through the identities in (2) (see Section 2.3) versus the Hamiltonian MC estimates of the posterior means and standard deviations for the logistic regression examples.

References

Agapiou, S., O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science 32*(3), 405–431.

Akyildiz, O. D. and J. Míguez (2021). Convergence rates for optimised adaptive importance samplers. *Statistics and Computing 31*(2), 12.

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics. New York, NY: Springer. Chapter 3.

Bugallo, M. F., V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine 34*(4), 60–79.

Cornuet, J.-M., J.-M. Marin, A. Mira, and C. P. Robert (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics 39*(4), 798–812.

Cotter, S. L., G. O. Roberts, A. M. Stuart, and D. White (2013). MCMC methods for functions: Modifying old algorithms to make them faster. *Statistical Science 28*(3), 424–446.

Hernández-Lobato, J., Y. Li, M. Rowland, T. Bui, D. Hernández-Lobato, and R. Turner (2016). Black-box α-divergence minimization. In *Proceedings of The 33rd International Conference on Machine Learning*, Volume 48 of *Proceedings of Machine Learning Research*, New York, NY, pp. 1511–1520. PMLR.

Khan, M. E., W. Lin, V. Tangkaratt, Z. Liu, and D. Nielsen (2017). Variational adaptive-Newton method for explorative learning. In *Advances in Approximate Bayesian Inference. NIPS 2017 Workshop.*

Opper, M. and C. Archambeau (2009). The variational Gaussian approximation revisited. *Neural Computation 21*(3), 786–792.

Steinberg, D. M. and E. V. Bonilla (2014). Extended and unscented Gaussian processes. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc.

Titsias, M. and M. Lázaro-Gredilla (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning*, Volume 32 of *Proceedings of Machine Learning Research*, Bejing, China, pp. 1971–1979. PMLR.

van den Boom, W., A. Cremaschi, and A. H. Thiery (2024). Doubly adaptive importance sampling. arXiv:2404.18556v1.

van den Boom, W. and A. H. Thiery (2019). EP-IS: Combining expectation propagation and importance sampling for Bayesian nonlinear inverse problems. In *Proceeding of 62nd ISI World Statistics Congress 2019. Contributed Paper Session: Volume 2*, pp. 145–152. Department of Statistics Malaysia. https://2019.isiproceedings.org/Files/8.Contributed-Paper-Session(CPS)-Volume-2.pdf.