Inclusive Variational Inference with Independent Metropolis-Hastings

Anonymous authors
Paper under double-blind review

Abstract

Markovian score climbing (MSC) is a recently proposed variational inference (VI) method based on Markov-chain Monte Carlo (MCMC) for minimizing the inclusive Kullback-Leibler (KL) divergence. This paper shows that independent Metropolis-Hastings (IMH) type kernels can automatically trade off bias and variance when used for MSC. In addition, we also show that the variance of the conditional importance sampling kernel, which was originally proposed for MSC, may increase with an additional computational budget. To fix this, we propose to use parallel IMH (PIMH) chains for obtaining stochastic gradients. We find that MSC with PIMH achieves better performance compared to inclusive as well as exclusive VI methods.

1 Introduction

Given an observed data $\mathbf{x}$ and a latent variable $\mathbf{z}$, Bayesian inference aims to analyze $p(\mathbf{z} \mid \mathbf{x})$ given an unnormalized joint density $p(\mathbf{z}, \mathbf{x})$ where the relationship is given by Bayes' rule such that $p(\mathbf{z} \mid \mathbf{x}) = p(\mathbf{z}, \mathbf{x}) / p(\mathbf{x}) \propto p(\mathbf{z}, \mathbf{x})$. Instead of working directly with the target distribution $p(\mathbf{z} \mid \mathbf{x})$, variational inference (VI, Jordan et al. 1999; Blei et al. 2017; Zhang et al. 2019) searches for a variational approximation $q_{\lambda}(\mathbf{z})$ that is similar to $p(\mathbf{z} \mid \mathbf{x})$ according to a discrepancy measure $D(p, q_{\lambda})$.

Naturally, choosing a good discrepancy measure, or objective function, is a critical part of the problem. This fact had lead to a quest for good divergence measures (Li & Turner, 2016; Dieng et al., 2017; Wang et al., 2018; Ruiz & Titsias, 2019). So far, the exclusive KL divergence $D_{\mathrm{KL}}(q_{\lambda} \parallel p)$ (or reverse KL divergence) has been used "exclusively" among various discrepancy measures. This is partly because the exclusive KL is defined as an average over $q_{\lambda}(\mathbf{z})$, which can be estimated efficiently. By contrast, the inclusive KL is defined as

$$D_{\mathrm{KL}}(p \parallel q_{\lambda}) = \int p(\mathbf{z} \mid \mathbf{x}) \log \frac{p(\mathbf{z} \mid \mathbf{x})}{q_{\lambda}(\mathbf{z})} \, d\mathbf{z} = \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{z} \mid \mathbf{x})}{q_{\lambda}(\mathbf{z})} \right] \tag{1}$$

where the average is taken over $p(\mathbf{z} \mid \mathbf{x})$. Interestingly, this is a chicken-and-egg problem as our goal is to obtain $p(\mathbf{z} \mid \mathbf{x})$ in the first place. Despite this challenge, minimizing (1) has drawn the attention of researchers because it can overcome some known limitations of the exclusive KL (Minka, 2005; MacKay, 2001).

For performing inclusive VI, Naesseth et al. (2020); Ou & Song (2020) recently proposed *Markovian score climbing* (MSC), which is a blend of Markov-chain Monte Carlo (MCMC) and variational inference. In MSC, stochastic gradients of the inclusive KL are obtained by operating a Markov-chain in parallel with the VI optimizer. In this paper, we find an interesting property of MSC when it is combined with specific types of MCMC kernels. Specifically, we show that *independent Metropolis-Hastings* (IMH, Robert & Casella 2004) type kernels can automatically trade off bias and variance when used for MSC. This family of kernels includes the *condition importance sampling* (CIS, Naesseth et al. 2020) kernel, which was originally proposed for MSC. Surprisingly, this automatic tradeoff property is unique to IMH type kernels and does not occur in MCMC kernels with state-dependent proposals such as Hamiltonian Monte Carlo (HMC, Duane et al. 1987; Neal 2011a; Betancourt 2017).

Following our analysis of the CIS kernel, we also show that its performance can degrade with the number of proposals (which is equivalent to the *per-transition computational budget*) used in each Markov-chain transition. As a simple solution to this, we propose to use parallel IMH (MSC-PIMH) chains, which reduce variance given the same amount of computation. We evaluate the performance of MSC with PIMH against other inclusive VI (Bornschein & Bengio, 2015; Naesseth et al., 2020) and exclusive VI (Ranganath et al., 2014; Kucukelbir et al., 2017) methods.

Contribution Summary **(i)** We show that IMH type kernels (which include the CIS kernel originally used in MSC; **Section 3.2**) automatically perform bias-variance tradeoff (**Section 3.3**). **(ii)** We show that increasing the computation budget of the CIS kernel may *increase* its variance. To overcome this limication, we propose to use parallel IMH (PIMH) (**Section 3.4**). **(iii)** We evaluate the performance of MSC with PIMH against other inclusive and exclusive VI methods (**Section 4**).

2 Background and Motivation

2.1 Inclusive Variational Inference Until Now

A typical way to perform VI is to use stochastic gradient descent (SGD, Robbins & Monro 1951; Bottou 1999), which requires unbiased gradient estimates of the optimization target. In the case of inclusive variational inference, this corresponds to estimating

$$\nabla_{\boldsymbol{\lambda}} D_{\mathrm{KL}}(p \parallel q_{\lambda}) = \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\, -\nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\lambda}}(\mathbf{z}) \,\right] = -\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\, s\,(\mathbf{z}; \boldsymbol{\lambda}) \,\right] \tag{2}$$

where $s\,(\mathbf{z}; \boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\lambda}}(\mathbf{z})$ is known as the *score function*. Evidently, estimating $\nabla_{\boldsymbol{\lambda}} D_{\mathrm{KL}}(p \parallel q_{\lambda})$ requires integrating the score function over $p(\mathbf{z} \mid \mathbf{x})$, which is prohibitive.

Importance Sampling When it is easy to sample from the variational approximation $q_{\lambda}(\mathbf{z})$, one can use importance sampling (IS, Robert & Casella 2004; Owen 2013) since

$$\mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\, s\,(\mathbf{z}; \boldsymbol{\lambda}) \,\right] \propto \mathbb{E}_{q_{\boldsymbol{\lambda}}}\left[\, w\,(\mathbf{z})\, s\,(\mathbf{z}; \boldsymbol{\lambda}) \,\right] \approx \frac{1}{N} \sum_{i=1}^{N} w\,(\mathbf{z}^{(i)})\, s\,(\mathbf{z}^{(i)}; \boldsymbol{\lambda}) \tag{3}$$

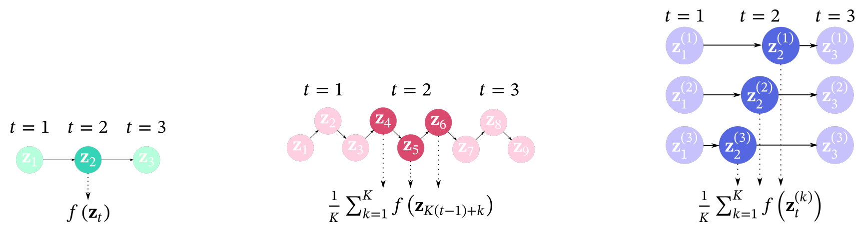
where $w\,(\mathbf{z}) = p\,(\mathbf{z}, \mathbf{x}) / q_{\boldsymbol{\lambda}}(\mathbf{z})$ is known as the *importance weight*, and $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ are N independent samples from $q_{\boldsymbol{\lambda}}(\mathbf{z})$. This scheme is equivalent to adaptive IS methods (Cappé et al., 2008; Bugallo et al., 2017) since the IS proposal $q_{\boldsymbol{\lambda}}(\mathbf{z})$ is iteratively optimized based on the current samples. Though IS is unbiased, it is highly unstable in practice. A more stable alternative is to use the *normalized weight* $\widetilde{w}^{(i)} = w\,(\mathbf{z}^{(i)}) / \sum_{i=1}^{N} w\,(\mathbf{z}^{(i)})$, which is known as the self-normalized IS (SNIS) approximation. Unfortunately, SNIS still fails to converge even on moderate dimensional objectives and unlike IS, it is no longer unbiased (Robert & Casella, 2004; Owen, 2013).

2.2 Stochastic Approximation with Markov-Chain Monte Carlo

MSC and JSA Recently, Naesseth et al. and Ou & Song proposed two similar but independent methods for performing inclusive variational inference. Both methods operate a Markov-chain in parallel with the VI optimization sequence, but the formulation is slightly different. Naesseth et al. proposed *Markovian score climbing* (MSC) which operates a MCMC kernel leaving $p\,(\mathbf{z} \mid \mathbf{x})$ invariant. Also, for the MCMC kernel, they propose conditional importance sampling (CIS), which is inspired by the particle MCMC method by Andrieu et al. (2010).

On the other hand, Ou & Song (2020) propose *joint stochastic approximation* (JSA) which operats a MCMC kernel leaving the likelihoods of the independent data points $p\,(\mathbf{z}_i \mid \mathbf{x}_i)$ invariant. For the MCMC kernel, unlike Naesseth et al., they use the independence Metropolis Hastings (IMH) sampler. Also, they operate multiple

They showed that MSC achieves better and more robust performance compared to methods such as SNIS and expectation propagation (EP, Minka 2001). MSC is described in Algorithm 4. It obtains stochastic gradients from a Markov-chain $\{\, \mathbf{z}_1, \mathbf{z}_2, \ldots \mathbf{z}_T \,\}$ generated from a π-invariant transition operator $K(\mathbf{z}, \cdot)$ running in parallel with the VI optimizer (represented by the sequence $\{\, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_T \,\}$). Naesseth et al. specifically proposed to use the *conditional importance sampling* (CIS) kernel for MSC.



(a) Single mode (b) Sequential mode (c) Parallel mode (proposed)

Figure 1: Visualization of different ways of combining MCMC with stochastic approximation variational inference. The index t denotes the stochastic approximation iteration. The dark circles denote the MCMC samples used for estimating the score gradient for $t = 2$.

Table 1: Computational Cost of Markov-chain Schemes

	Estimation			Stochastic gradient	
	$p(\mathbf{z},\mathbf{x})$ # Eval.	$q_\lambda(\mathbf{z})$ # Eval.	$q_\lambda(\mathbf{z})$ # Samples	$p(\mathbf{z},\mathbf{x})$ # Grad.	$q_\lambda(\mathbf{z})$ # Grad.
ADVI	0	0	K	K	0
Single Mode (with CIS)	$K-1$	K	$K-1$	0	1[1]or K[2]
Sequetial Mode (with IMH)	K	K	K	0	K
Parallal Mode (with IMH)	K	K	K	0	K

[*] K is the number of samples used in each method.
[1] Vanilla CIS kernel.
[2] Rao-Blackwellized CIS kernel.

3 Inclusive Variational Inference with Independent Metropolis-Hastings

3.1 Overview of Markov-Chain Monte Carlo Schemes

First, the different modes of operating Markov-chains are illustrated in Figure 1. The three different methods all have a similar computational cost, but their statistical performances are very different. The computational costs of each scemes are organized in Table 1.

Theorem 1. *Assuming* $w^* = \sup_{\mathbf{z}} p(\mathbf{z}|\mathbf{x})/q_\lambda(\mathbf{z}) < \infty$ *for* $\forall\lambda$*, for a bounded functional* $f : \mathcal{Z} \to [\text{-}L/2, L/2]$*, the bias of the sequential mode estimator with an IMH kernel at iteration t is bounded as*

$$\mathrm{Bias}\left[g_{\text{seq.},t}\right] \leq L\, C^{N(t-1)} \frac{1}{N} \sum_{i=1}^{N} C^i$$

where $C = (1 - 1/w^*) < 1$*.*

Proof. The proof is in the *supplementary material.*

Theorem 2. *Assuming* $w^* = \sup_{\mathbf{z}} p(\mathbf{z}|\mathbf{x})/q_{\lambda_\tau}(\mathbf{z}) < \infty$ *for* $\forall\lambda$*, for a bounded functional* $f : \mathcal{Z} \to [\text{-}L/2, L/2]$*, the bias of the parallel mode estimator with an IMH kernel at iteration t is bounded as*

$$\mathrm{Bias}\left[g_{\text{par.},t}\right] \leq L\, C^t.$$

where $C = 1 - 1/w^*$*.*

Proof. The proof is in the *supplementary material.*

Proposition 1. *variance of sequential estimator*

Proof. The proof is in the *supplementary material.*

Single Mode First, the "single mode" (see Figure 1a) used by Naesseth et al. (2020) performs a "single" Markov-chain transition at each stochastic approximation step, and estimates the score using the single state such as $g(\boldsymbol{\lambda}) = \nabla \log q_{\boldsymbol{\lambda}}(\mathbf{z}_t)$. The single state is generated by a computationally expensive MCMC kernel such as the CIS kernel proposed by Naesseth et al..

Sequential Mode On the other hand, (Ou & Song, 2020) perform *multiple* "sequental" state transitions such that $\{\mathbf{z}_i, \mathbf{z}_{i+1}, \ldots, \mathbf{z}_{i+K}\}$ at each stochastic approximation step. The score is estimated by averaging the multiple intermediate states such that $\frac{1}{K}\sum_{k=1}^{K} \nabla \log q_{\boldsymbol{\lambda}}(\mathbf{z}_{t+k})$. When using the independent Metropolis Hastings (IMH) kernel, the computational cost is comparable to the single mode with the CIS kernel. We call this scheme the "sequential mode" in contrast with the parallel mode described below.

Parallel Mode The "parallel mode" shown in Figure 1c is another possible scheme that has been underappreciated. Unlike both the single and sequential modes that operate only a single Markov-chain, the parallel mode operates *multiple chains* in parallel. This scheme differs from the traditional way of running MCMC, but has multiple benefits for the specific purpose of VI.

MCMC uses marginal estimates. To explain the counterintuitive observation in **??**, we first discuss a subtle difference between the usual MCMC setting and the MSC setting. In MCMC, what matters is the estimate provided by the average of *all* the states. This *marginal* estimate of f is

$$\frac{1}{T}\sum_{t=1}^{T} f(\mathbf{z}_t) = \mathbb{E}\left[\, \mathbb{E}_{\mathbf{z}_t \sim K(\mathbf{z}_{t-1},\cdot)}\left[\, f(\mathbf{z}_t) \mid \mathbf{z}_{t-1} \,\right]\right] \tag{4}$$

where $\mathbb{E}_{\mathbf{z}_t \sim K(\mathbf{z}_{t-1},\cdot)}\left[\, f(\mathbf{z}_t) \mid \mathbf{z}_{t-1} \,\right]$ is the *conditional estimate* of $K(\mathbf{z}_{t-1}, \cdot)$ given the previous state $\mathbf{z}_{t-1}$.

MSC uses conditional estimates. On the other hand, in MSC, the stochastic gradient is estimated using a *single* state of the Markov-chain denoted as $\mathbf{z}_t \sim K(\mathbf{z}_{t-1}, \cdot)$. This means that one needs to estimate the gradients using the conditional estimate such as

$$\nabla_{\lambda} D_{\mathrm{KL}}(p \parallel q_{\lambda}) \approx \mathbb{E}_{\mathbf{z}_t \sim K(\mathbf{z}_{t-1},\cdot)}\left[\, s(\mathbf{z}_t; \boldsymbol{\lambda}) \mid \mathbf{z}_{t-1} \,\right]. \tag{5}$$

This seemingly subtle difference of using either the marginal or conditional estimate reveals the following facts about MSC: (i) The usual central limit theorem guarantees of MCMC do not apply to MSC. (ii) In addition, ergodicity guarantees work differently since $q_{\boldsymbol{\lambda}}(\mathbf{z})$ is updated at each iteration. The most important difference is, however, the way we analyze the variance of the estimates.

3.2 Interpreting Conditional Importance Sampling as Independent Metropolis-Hastings

Many popular MCMC kernels are based on the Metropolis-Hastings test where a random proposal $\mathbf{z}^*$ is either accepted into the Markov-chain ($\mathbf{z}_t = \mathbf{z}^*$) or rejected ($\mathbf{z}_t = \mathbf{z}_{t-1}$). Among these, independent Metropolis-Hastings (IMH) kernels generate proposals independently of $\mathbf{z}_{t-1}$ such as $\mathbf{z}^* \sim q(\mathbf{z})$ instead of $\mathbf{z}^* \sim q(\mathbf{z} \mid \mathbf{z}_{t-1})$.We show that the CIS kernel proposed by Naesseth et al. (2020) turns out to be a type of IMH kernel that uses Barker's acceptance ratio (Barker, 1965) for the Metropolis-Hastings test. This interpretation enables the analysis of the *rejection rate* of the CIS kernel.

Algorithm 1: Conditional Importance Sampling

Input: previous sample $\mathbf{z}_{t-1}$, previous parameter $\boldsymbol{\lambda}_{t-1}$, number of proposals N
$\mathbf{z}^{(0)} = \mathbf{z}_{t-1}$
Propose $\mathbf{z}^{(i)} \sim q_{\boldsymbol{\lambda}_{t-1}}(\mathbf{z})$ for $i = 1, 2, \ldots, N$
Weight $w(\mathbf{z}^{(i)}) = p(\mathbf{z}^{(i)}, \mathbf{x}) \,/\, q_{\boldsymbol{\lambda}_{t-1}}(\mathbf{z}^{(i)})$ for $i = 0, 1, \ldots, N$
Normalize $\widetilde{w}^{(i)} = w(\mathbf{z}^{(i)}) / \sum_{i=0}^{N} w(\mathbf{z}^{(i)})$ for $i = 0, 1, \ldots, N$
$\mathbf{z}_t \sim \text{Multinomial}(\widetilde{w}^{(0)}, \widetilde{w}^{(1)}, \ldots, \widetilde{w}^{(N)})$

Conditional Importance Sampling A pseudocode of the CIS kernel is shown in Algorithm 1. The original algorithmic description of CIS is to (i) obtain N samples from $q_\lambda(\mathbf{z})$, (ii) compute the importance weight including the previous Markov-chain state $\mathbf{z}_{t-1}$, and (iii) resample $\mathbf{z}_{t-1}$ from the multinomial distribution of $N+1$ proposals. While particle MCMC (Andrieu et al., 2010) originally inspired the CIS kernel, it is possible to find connections in multiple-try MCMC methods (Martino, 2018). In particular, the CIS kernel is identical to the previously proposed *ensemble MCMC sampler* (Austad, 2007; Neal, 2011b) with independent proposals, which is an instance of multiple-try MCMC (Martino, 2018, Table 12).

CIS as a Metropolis-Hastings Kernel Now we show that the CIS kernel is an accept-reject type kernel with Barker's acceptance ratio. First, by defining $\mathbf{z}_t = \mathbf{z}_{t-1}$ as "reject" and $\mathbf{z}_t \neq \mathbf{z}_{t-1}$ as "accept", the CIS kernel can be understood as an accept-reject type kernel. By denoting the N parallel proposals as an *ensemble state* $\mathbf{z}^{(1:N)} = (\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)})$ (Neal, 2011b), the CIS kernel conditional estimate can be written as

$$\mathbb{E}_{K(\mathbf{z}_{t-1},\mathbf{z}_t)}\left[\, f(\mathbf{z}_t) \mid \mathbf{z}_{t-1} \,\right] = \mathbb{E}_{q_\lambda}\left[\alpha\left(\mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right) \frac{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)})}{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right)} \right] + r\left(\mathbf{z}_{t-1}\right) f(\mathbf{z}_{t-1}) \tag{6}$$

where $q_\lambda(\mathbf{z}^{(1:N)}) = \prod_{i=1}^{N} q_\lambda(\mathbf{z}^{(i)})$, the acceptance ratio $\alpha(\mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}) = \sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) / \sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)$ is the probability of accepting the ensemble state $\mathbf{z}^{(1:N)}$, and

$$r\left(\mathbf{z}_{t-1}\right) = \mathbb{E}_{q_\lambda(\mathbf{z}^{(1:N)})}\left[\, r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right) \right] = \mathbb{E}_{q_\lambda(\mathbf{z}^{(1:N)})}\left[\left(1 - \alpha\left(\mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right)\right) \right] \tag{7}$$

is the probability of staying on $\mathbf{z}_{t-1}$ by rejecting *any* ensemble state, $r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right)$ is the rejection rate given $\mathbf{z}^{(1:N)}$. The expression of $\alpha\left(\mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right)$ is known as Barker's acceptance ratio (Barker, 1965), which is a special case of the original Metropolis ratio (Metropolis et al., 1953). (A detailed derivation is in the *supplementary material.*)

3.3 Bias-Variance Tradeoff of Conditional Importance Sampling

Variance of Conditional Importance Sampling The IMH (or accept-reject) view in Section 3.2 now enables us to discuss the rejection rate of the CIS kernel. As discussed in Section 3.1, MSC obtains gradients using the conditional estimates of MCMC. The variance of the conditional estimate is closely related to the rejection rate such as

$$\begin{aligned}
\mathbb{V}_{K(\mathbf{z}_{t-1},\cdot)}\left[\, f \mid \mathbf{z}_{t-1} \,\right] &= \mathbb{V}_{q_\lambda}\left[\, \mathbb{E}\left[\, f \mid \mathbf{z}_{t-1},\, \mathbf{z}^{(1:N)} \,\right] \right] + \underbrace{\mathbb{E}_{q_\lambda}\left[\, \mathbb{V}\left[\, f \mid \mathbf{z}_{t-1},\, \mathbf{z}^{(1:N)} \,\right] \right]}_{\text{Rao-Blackwellization gain}} && (8)\\
&\geq \mathbb{V}_{q_\lambda}\left[\, \mathbb{E}\left[\, f \mid \mathbf{z}_{t-1},\, \mathbf{z}^{(1:N)} \,\right] \right] && (9)\\
&= \mathbb{V}_{q_\lambda}\left[\left(1 - r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right)\right) f_{\mathrm{IS}} + r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right) f(\mathbf{z}_{t-1}) \,\middle|\, \mathbf{z}_{t-1} \right] && (10)\\
&\text{where} \;\; f_{\mathrm{IS}} = \sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)}) \,/\, \sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right).
\end{aligned}$$

The bound in (9) becomes exact if we use Rao-Blackwellization (that is, if we use the SNIS estimator $\sum_{i=1}^{N} \widetilde{w}^{(i)} f\left(\mathbf{z}^{(i)}\right)$ instead of the resampled $\mathbf{z}_t$) as mentioned by Naesseth et al. (2020). The expansion in (8) follows from the law of total variance where the right-hand term is the variance reduction we gain from using Rao-Blackwellization (Bernton et al., 2015), and the equality in (10) follows from (6).

Low rejection rate means high conditional variance. Because of the dependence of $r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right)$ on $\mathbf{z}^{(1:N)}$, it is in general difficult to interpret the result of (10). Nonetheless, when $w\left(\mathbf{z}_{t-1}\right) \gg w\left(\mathbf{z}^{(i)}\right)$, $r(\mathbf{z} \mid \mathbf{z}^{(1:N)})$ is close to 1 almost independently of $q_\lambda(\mathbf{z})$. The intuition is that if the *rejection weight* $w\left(\mathbf{z}_{t-1}\right)$ is large, most proposals will be rejected regardless of their values. By translating this intuition into an approximation, we obtain the following result.

Proposition 2. *Assuming $w\left(\mathbf{z}_{t-1}\right)$ is large enough to make $r\left(\mathbf{z} \mid \mathbf{z}^{(1:N)}\right)$ independent of $\mathbf{z}^{(1:N)}$, the variance can be approximated by*

$$\mathbb{V}_{q_\lambda}\left[\, \mathbb{E}\left[\, f \mid \mathbf{z}_{t-1}, \mathbf{z}^{(1:N)} \,\right] \right] \approx \left(1 - r\left(\mathbf{z}_{t-1}\right)\right)^2 \mathbb{V}_{q_\lambda}\left[\, f_{IS} \,|\, \mathbf{z}_{t-1} \,\right]. \tag{11}$$

Proof. The proof is in the *supplementary material*.

The statement of Proposition 2 is intuitive; if we reject all the states, there is no conditional variance. However, this obvious fact becomes more interesting combined with the followings.

CIS has a high rejection rate until MSC converges. The following bound provides a condition for the rejection rate of a CIS kernel to be high.

Proposition 3. *The rejection rate $r(\mathbf{z}_{t-1})$ of a CIS sampler with N proposals is bounded below such that*

$$r(\mathbf{z}_{t-1}) \geq \frac{1}{1 + \frac{N Z}{w(\mathbf{z}_{t-1})}}$$

where $Z = \mathbb{E}_{q_\lambda(\mathbf{z})}\left[p(\mathbf{z},\mathbf{x}) / q_\lambda(\mathbf{z}) \right] = \int p(\mathbf{z}, \mathbf{x})\, d\mathbf{z}$ is the normalizing constant.

Proof. The proof is in the *supplementary material*.

When $w(\mathbf{z}_{t-1})$ is large, the lower bound of $r(\mathbf{z}_{t-1})$ becomes close to one. In this case, according to Proposition 2, the conditional variance becomes minimal.

In the context of VI, the following shows that $r(\mathbf{z}_{t-1})$ is huge when the KL divergence is large.

Theorem 3. *Assuming $\sup_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}) / q_\lambda(\mathbf{z}) = M < \infty$, the average rejection rate $r = \int r(\mathbf{z}_{t-1})\, p(\mathbf{z}_{t-1} \mid \mathbf{x})\, d\mathbf{z}_{t-1}$ of a CIS kernel with N proposals is bounded below such that*

$$r \geq \frac{1}{1 + \frac{N}{\exp\left(D_{\mathrm{KL}}(p \| q_\lambda)\right)}} - \delta,$$

where the sharpness of the bound is given as $0 \leq \delta \leq \frac{M}{\exp^2\left(D_{\mathrm{KL}}(p \| q_\lambda)\right)}$.

Proof. The proof is in the *supplementary material*.

This result states that in the ideal case when the Markov-chain has achieved stationarity and $\mathbf{z}_{t-1}$ closely follows $p(\mathbf{z} \mid \mathbf{x})$, the average rejection weight is bounded below exponentially by the KL divergence. The bound is tight as long as $D_{\mathrm{KL}}(p \parallel q_\lambda)$ is large. In practical conditions, the rejection rate cannot be improved by increasing N since (i) the iteration complexity also increases, and (ii) $w(\mathbf{z}_t)$ can easily be larger by many orders of magnitude.

The results of Theorem 3 signify that, until MSC converges such that $D_{\mathrm{KL}}(p \parallel q_\lambda)$ is small, the rejection rate will be very high. And by Proposition 2, the variance will be small, improving the convergence of SGD. This explains why the Markov-chain in **??** does not move until MSC converges and why MSC works well with the CIS kernel. Note that these properties hold similarly in any IMH type kernels.

Is bias guaranteed to decrease? The bias of the conditional estimate is closely related to the total variation (TV) distance $\| K(\mathbf{z}_{t-1}, \cdot) - p(\cdot \mid \mathbf{x}) \|_{\mathrm{TV}}$. For bounded functions, the TV distance provides an upper bound of the bias. Unfortunately, it is in general difficult to specify the TV distance (and hence the bias) with respect to $p(\mathbf{z} \mid \mathbf{x})$ and $q_\lambda(\mathbf{z})$. Neverthless, Wang (2020) recently showed that the rejection rate is related with the TV distance such that $r(\mathbf{z}_{t-1}) \leq \| K(\mathbf{z}_{t-1}, \cdot) - p(\cdot \mid \mathbf{x}) \|_{\mathrm{TV}}$. Therefore, a low rejection rate is *necessary* for the bias to decrease.

Automatic Bias-Variance Tradeoff of IMH Type Kernels To summarize, IMH type kernels (including the CIS kernel) have an automatic bias-variance tradeoff mechanism. In the initial steps where the inclusive KL divergence is large, variance is suppressed by rejecting most proposals. However, as MSC converges, the bound on the rejection rate becomes loose, admitting a lower rejection rate, which enables bias to decrease. This mechanism provides an interesting case where rejections in MCMC can actually be beneficial. Lastly, we note that the automatic tradeoff mechanism is a unique property of IMH type kernels; it does not exist in kernels with state-dependent proposals such as random-walk Metropolis-Hastings or HMC.

3.4 Reducing Variance with Parallel Independent Metropolis-Hastings Markov-Chains

Recall that the CIS kernel uses N multiple-try type proposals. Thus, it is natural to expect the variance to decrease as the (per-transition) computational budget N increases. However, under specific conditions, we find that the variance actually *increases* with N.

The variance of CIS can increase with N. The bound in Proposition 3 can be reinterpreted as a bound on the acceptance rate

$$1 - r(\mathbf{z}_{t-1}) \leq \frac{N Z}{w(\mathbf{z}_{t-1}) + N Z}, \tag{12}$$

which is, in general, very tight. More importantly, when $w(\mathbf{z}_{t-1}) \gg N Z$, the acceptance rate grows like $\mathcal{O}(N)$. On the other hand, the variance of an SNIS estimator is known to decrease approximately at a rate of $\mathcal{O}(1/N)$ (Kong et al., 1994; Robert & Casella, 2004; Elvira et al., 2018). That is,

$$\mathbb{V}_{q_\lambda}\left[\mathbb{E}\left[f \mid \mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right]\right] \approx \underbrace{\left(1 - r(\mathbf{z}_{t-1})\right)^2}_{\text{approx. } \mathcal{O}(N^2)} \underbrace{\mathbb{V}_{q_\lambda}\left[f_{\mathrm{IS}}\right]}_{\text{approx. } \mathcal{O}(1/N)}. \tag{13}$$

Thus, when $w(\mathbf{z}_{t-1}) \gg N Z$, the conditional variance of CIS approximately grows as $\mathcal{O}(N)$. We provide numerical simulations that support our analysis in the *supplementary material*.

Algorithm 2: Markovian Score Climbing with Parallel Chains

Input: initial samples $\mathbf{z}_0^{(1)}, \ldots, \mathbf{z}_0^{(N)}$, initial parameter $\boldsymbol{\lambda}_0$, number of iterations T, stepsize schedule γ_t

for $t = 1, 2, \ldots, T$ **do**
- **for** $i = 1, 2, \ldots, N$ **do**
  - $\mathbf{z}_t^{(i)} \sim K(\mathbf{z}_{t-1}^{(i)}, \cdot)$
- **end**
- $s(\mathbf{z}_t^{(i)}; \boldsymbol{\lambda}) = \nabla_\lambda \log q_\lambda(\mathbf{z}_t^{(i)})$
- $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \gamma_t \, \frac{1}{N} \sum_{i=1}^{N} s(\mathbf{z}_t^{(i)}; \boldsymbol{\lambda}_{t-1})$

end

Algorithm 3: Independent Metropolis-Hastings

Input: previous sample $\mathbf{z}_{t-1}$, previous parameter $\boldsymbol{\lambda}_{t-1}$,

$\mathbf{z}^* \sim q_{\lambda_{t-1}}(\mathbf{z})$

$w(\mathbf{z}) = p(\mathbf{z}, \mathbf{x}) / q_{\lambda_{t-1}}(\mathbf{z})$

$\alpha = \min\left(w(\mathbf{z}^*)/w(\mathbf{z}_{t-1}), 1\right)$

$u \sim \text{Uniform}(0, 1)$

if $u < \alpha$ **then**
- $\mathbf{z}_t = \mathbf{z}^*$

else
- $\mathbf{z}_t = \mathbf{z}_{t-1}$

end

Variance Reduction with Paralel IMH Chains To resolve the aforementioned limitation of the CIS kernel, we propose a simple but effective remedy: running N parallel IMH (PIMH) Markov-chains $\{\mathbf{z}_t^{(1)}\}, \{\mathbf{z}_t^{(2)}\}, \ldots, \{\mathbf{z}_t^{(N)}\}$ where each of the chains performs a Metropolis-Hastings test with only a *single* proposal each. The modified MSC algorithm incorporating parallel chains is shown in Algorithm 6 (the modified parts are highlighted in blue and purple), while the IMH kernel is described in Algorithm 3. Since the parallel chains generate an *independent* conditional estimate each, the gradient estimate $1/N \sum_{i=1}^{N} s(\mathbf{z}_t^{(i)}; \boldsymbol{\lambda})$ is an average of N independent and identical estimators. This obviously reduces the variance of a single conditional estimate as $\mathcal{O}(1/N)$. In the best case, relative to CIS, PIMH will have a variance reduction close to $\mathcal{O}(1/N^2)$.

Lower bound of rejection rate in IMH IMH also enjoys variance control properties similar to the CIS kernel. That is, a lower bound similar to Proposition 3 can be shown.

Proposition 4. *The rejection rate $r(\mathbf{z}_{t-1})$ of a IMH sampler is bounded below such that*

$$r(\mathbf{z}_{t-1}) \geq 1 - \frac{Z}{w(\mathbf{z}_{t-1})} \quad \textit{where} \quad Z = \int p(\mathbf{z}, \mathbf{x}) \, d\mathbf{z}.$$

Proof. The proof is in the *supplementary material*.

Unlike CIS, we can see that the bound does not depend on N. This means the rejection rate will *not* prematurely increase. We thus obtain all the benefits of CIS except its limitations.

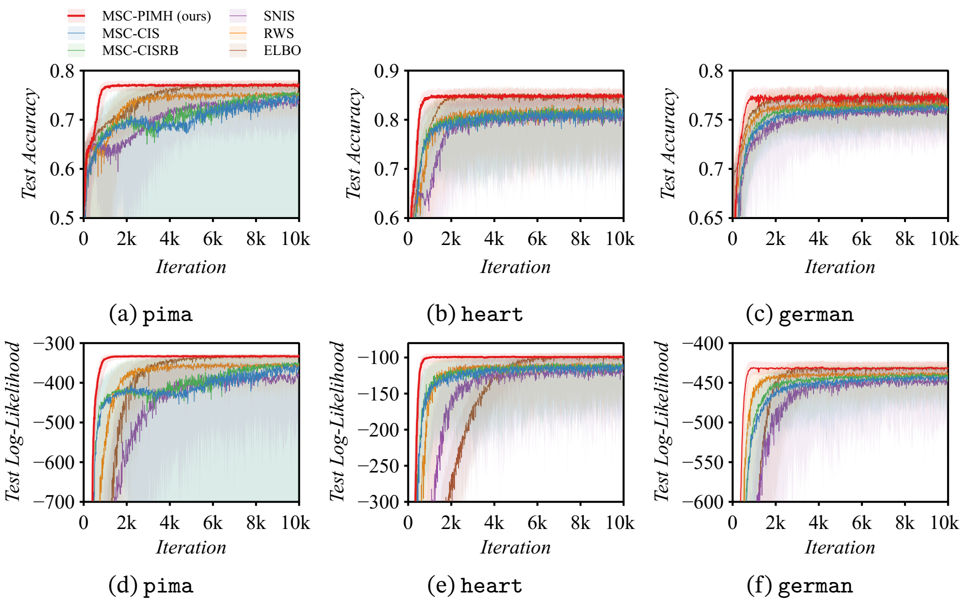


Figure 2: Test accuracy and log-likelihood of logistic regression problems. The solid lines and colored regions are the medians and 80% percentiles computed from 100 repetitions.

Costs and Limitation of PIMH The computational cost of sampling $\mathbf{z}_t^{(i)}$ (blue region in Algorithm 6) for N chains is equal to a CIS kernel with N proposals. On the other hand, the cost of estimating the stochastic gradient (purple region in Algorithm 6) is now $\mathcal{O}(N)$ instead of $\mathcal{O}(1)$ of the CIS kernel. This cost is, however, also imposed on the CIS kernel if we use Rao-Blackwellization. Thus, the overall computational cost of PIMH is more or less equal to that of the CIS kernel. The only downside of PIMH is that the Metropolis-Hastings acceptance ratio that we use has a slightly larger acceptance rate than Barker's (Peskun, 1973; Minh & Minh, 2015). In the small N regime, PIMH has a slightly larger conditional variance, but it can be easily fixed by using Barker's ratio.

4 Evaluations

4.1 Experimental Setup

Implementation We implemented MSC with PIMH on top of the Turing (Ge et al., 2018) probabilistic programming framework. Our implementation works with any model described in Turing, which automatically handles distributions with constrained support (Kucukelbir et al., 2017). We use the ADAM optimizer by Kingma & Ba (2015) with a learning rate of 0.01 in all of the experiments. We set the computational budget $N = 10$ and $T = 10^4$ for all experiments unless specified.

Considered Baselines We compare MSC-PIMH with **(i)** MSC using the CIS kernel (**MSC-CIS**, Naesseth et al. 2020), **(ii)** MSC using the CIS kernel with Rao-Blackwellization (**MSC-CISRB**, Naesseth et al. 2020), **(iii)** the adaptive IS method using SNIS as introduced in Section 2.1 (**SNIS**), **(iv)** the reweighted wake-sleep algorithm (**RWS**, Bornschein & Bengio 2015), and **(v)** evidence lower-bound maximization (**ELBO**, Ranganath et al. 2014). Specifically, we use automatic differentiation VI (ADVI, Kucukelbir et al. 2017) implemented by Turing.

Reinterpreting RWS The original RWS algorithm assumes that independent samples from $p(\mathbf{z} \mid \mathbf{x})$ are available, possibly with an additional cost. Since this is not the case in our setting, we reinterpret RWS as alternating between cheap (*sleep update*) and expensive (*wake update*) estimates. We respectively use SNIS and HMC for the sleep and wake updates, and perform the wake update every $K = 5$ steps as originally recommended by Bornschein & Bengio (2015).

4.2 Hierarchical Logistic Regression

Experimental Setup We evaluate MSC-PIMH on logistic regression with the Pima Indians diabetes (`pima`, $\mathbf{z} \in \mathbb{R}^{11}$, Smith et al. 1988), German credit (`german`, $\mathbf{z} \in \mathbb{R}^{27}$), and heart disease (`heart`, $\mathbf{z} \in \mathbb{R}^{16}$, Detrano et al. 1989) datasets obtained from the UCI repository (Dua & Graff, 2017). 10% of the data points were randomly selected in each of the 100 repetitions as test data.

Probabilistic Model Instead of the usual single-level probit/logistic regression models used in VI, we choose a more complex hierarchical logistic regression model

$$y_i \sim \text{Bernoulli-Logit}\,(p),\;\; p \sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta} + \alpha,\, \sigma_\alpha^2),\;\; \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0},\, \sigma_\beta^2 \mathbf{I}),\;\; \sigma_\beta,\, \sigma_\alpha \sim \mathcal{N}^+(0, 1.0) \qquad (14)$$

where $\mathcal{N}^+(\mu, \sigma)$ is a positive constrained normal distribution with mean μ and standard deviation σ, $\mathbf{x}_i$ and y_i are the feature vector and target variable of the ith datapoint. The extra degrees of freedom σ_β and σ_α make this model relatively more challenging.

Results The test accuracy and test log-likelihood results are shown in Figure 2. Our proposed MSC-PIMH is the fastest to converge on all the datasets. Despite having access to high-quality HMC samples, RWS fails to achieve a similar level of performance to MSC-PIMH. However, RWS converges faster than MSC-CIS and MSC-CISRB. Among the two, MSC-CISRB performs only marginally better than MSC-CIS. Meanwhile, SNIS converges the most slowly among inclusive VI methods. Although much slower to converge, ELBO achieves competitive results.

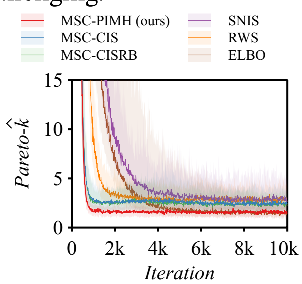
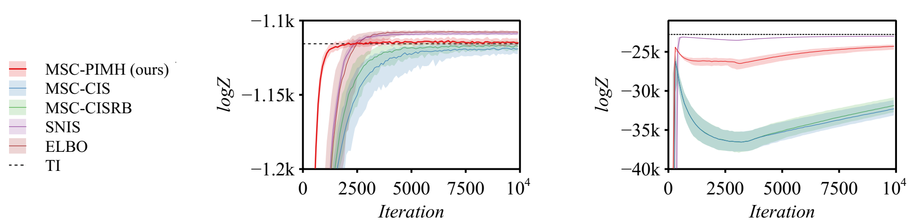


Figure 3: Pareto-$\hat{k}$ statistics result on `german`. The solid lines and colored regions are the medians and 80% percentiles computed from 100 repetitions.

Inclusive VI v.s. Exclusive VI The results of Figure 2 might be misleading to conclude that inclusive and exclusive VI deliver similar results. However, in the parameter space, they choose different optimization paths. This is shown in Figure 3 through the Pareto-$\hat{k}$ diagnostic (Dhaka et al., 2020; Vehtari et al., 2021), which determines how reliable the importance weights are when computed using $q_{\boldsymbol{\lambda}}(\mathbf{z})$. While the test accuracy suggests that ELBO converges around $t = 2000$, in terms of Pareto-$\hat{k}$, it takes much longer to converge (about $t = 5000$). This shows that, even if their predictive performance is similar, the inclusive VI chooses paths that have better density coverage as expected.

4.3 Marginal Likelihood Estimation



(a) `radon` (b) `stock`

Figure 4: Marginal log-likelihood ($\log Z$) estimates of considered methods. ELBO is omitted in Figure 4b as it failed to deliver reasonable estimates. The solid lines and colored regions are the medians and 80% percentiles computed from 100 repetitions.

Experimental Setup We now estimate the marginal log-likelihood $\log Z$ of a stochastic volatility model (`stock`, $\mathbf{z} \in \mathbb{R}^{2613}$, Kim et al. 1998) and a hierarchical regression model with partial pooling (`radon`, $\mathbf{z} \in \mathbb{R}^{175}$, Gelman & Hill 2007) for modeling radon levels in U.S homes. For `stock`, we use 10 years of the S&P index daily closing price (May 3, 2007, to May 3, 2017). `stock` is highly challenging as it is both high dimensional and strongly correlated. We estimated the reference marginal likelihood using *thermodynamic integration* (TI, Gelman & Meng 1998; Neal 2001; Lartillot & Philippe 2006) with HMC implemented by Stan (Carpenter et al., 2017; Betancourt, 2017).

Results The results are shown in Figure 4. On `radon`, MSC-PIMH converges quickly and provides the most accurate estimate. By contrast, MSC-CIS and MSC-CISRB converge much slowly. SNIS and ELBO, on the other hand, overestimate $\log Z$, which can be attributed to the mode-seeking behavior of ELBO and the small sample bias of SNIS. On `stock`, SNIS is unexpectedly the most accurate. Unfortunately, MSC-PIMH, MSC-CIS, MSC-CISRB all underestimate $\log Z$. Nevertheless, MSC-PIMH provides much better estimates than the latter two. Lastly, we observe that only ELBO fails to converge given the same amount of SGD steps.

5 Related Works

Inclusive VI with SGD Our method directly builds on top of MSC (Naesseth et al., 2020), which is a method for minizing the inclusive KL divergence. While many works minimizing the inclusive KL have emerged (Bornschein & Bengio, 2015; Li et al., 2017; Minka, 2001; Ou & Song, 2020; Kim et al., 2021), only a few have been proposed for general VI based on SGD. Notably, Bornschein & Bengio (2015) use SNIS for estimating the stochastic gradients, while Li et al. (2017) use an MCMC kernel to refine samples from $q_{\boldsymbol{\lambda}}(\mathbf{z})$ to better resemble samples from $p(\mathbf{z} \mid \mathbf{x})$. Meanwhile, a synonymous method to MSC, *general stochastic approximation* (GSA) by Ou & Song (2020, Algorithm 1) has been proposed concurrently in the context of discrete latent variables. Kim et al. (2021) recently proposed a method that essentially blends GSA/MSC with RWS.

Adaptive MCMC As pointed out by Ou & Song (2020), MSC is structurally equivalent to adaptive MCMC methods. Strong resemblence can be found in methods using stochastic approximation for adapting the proposal distribution used inside the MCMC kernel. In particular, Andrieu & Thoms (2008); Garthwaite et al. (2016) discuss the use of stochastic approximation in adaptive MCMC.

Adaptive IMH Among adaptive MCMC methods, those that use independent proposals (Andrieu & Moulines, 2006; Keith et al., 2008; Holden et al., 2009; Giordani & Kohn, 2010) are the most related to our work. Keith et al. (2008) propose to use *cross-entroy minimization* (Barbakh et al., 2009), which is mathematically identical to inclusive VI, for adaptation. Our work, on the other hand, contrasts with previous adaptive IMH algorithms in that we use SGD for adapting $q_{\boldsymbol{\lambda}}(\mathbf{z})$. This enables VI methods such as ADVI to consider proposals that are much more complex (Kucukelbir et al., 2017).

Ergodicity and Inclusive VI Meanwhile, in the context of MCMC, Mengersen & Tweedie (1996) showed that it is necessary to ensure $\sup_{\mathbf{z}} w(\mathbf{z}) = M < \infty$ (finite weight condition) for an IMH kernel to be geometrically ergodic. While this might seem less relevant for inclusive VI, the bound

$$D_{\mathrm{KL}}(p \parallel q_{\boldsymbol{\lambda}}) = \int p(\mathbf{z} \mid \mathbf{x}) \log w(\mathbf{z})\, d\mathbf{z} \leq \int p(\mathbf{z} \mid \mathbf{x}) \log M \, d\mathbf{z} = \log M. \tag{15}$$

suggests that it is in fact a sufficient condition for the KL divergence to be finite. This condition can easily be violated as shown by Andrieu & Thoms (2008). To ensure this does not happen, Giordani & Kohn (2010); Holden et al. (2009) use proposal distributions of the form of $w\, q_0(\mathbf{z}) + (1 - w)\, q_{\boldsymbol{\lambda}}(\mathbf{z})$ for some $0 < w < 1$ for their adaptive IMH sampler. Here, q_0 is supposed to be a heavy tailed distribution in the spirit of defensive mixtures (Hesterberg, 1995). A research direction in the interest of both adaptive MCMC and inclusive VI would be to investigate whether such precaution is actually necessary for convergence. If that is the case, it would be beneficial to consider variational families of heavy-tailed distributions as proposed by Domke & Sheldon (2018) for exclusive VI.

6 Conclusions

In this paper, we investigated the properties of Markovian score climbing (MSC) with independent Metropolis-Hastings (IMH) type Markov-chain Monte Carlo (MCMC) kernels. We proved that IMH type kernels are able to automatically perform bias-variance tradeoff using their accept-reject mechanism. We also analyzed the limitation of the conditional importance sampling (CIS) kernel originally used in MSC. We then proposed parallel IMH (PIMH) as an alternative that enjoys the benefits of CIS without its limitations. Our experiments verify that

MSC combined with PIMH performs well on the considered Bayesian inference problems, even compared to exclusive variational inference methods.

ACKNOWLEDGMENTS

This work initially started in the process of understanding the performance of the Markovian score climbing algorithm by Naesseth et al. (2020). We sincerely thank Hongseok Yang for pointing us to a relevant related work by Kim et al. (2021), Guanyang Wang for insightful discussions about the independent Metropolis-Hastings algorithm, Geon Park and Kwanghee Choi for constructive comments that enriched this paper. We also aknowledge the Computer Science Department of Sogang University for providing computational resources.

REFERENCES

Christophe Andrieu and Éric Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3), August 2006.

Christophe Andrieu and Johannes Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, December 2008.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72 (3):269–342, June 2010.

Haakon Michael Austad. *Parallel Multiple Proposal MCMC Algorithms*. Master thesis, Norwegian University of Science and Technology, June 2007.

Wesam Ashour Barbakh, Ying Wu, and Colin Fyfe. *Cross Entropy Methods*, volume 249, pp. 151–174. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Anthony A. Barker. Monte Carlo calculations of the radial distribution functions for a proton electron plasma. *Australian Journal of Physics*, 18(2):119, 1965.

Espen Bernton, Shihao Yang, Yang Chen, Neil Shephard, and Jun S. Liu. Locally weighted Markov chain Monte Carlo. *arXiv:1506.08852 [stat]*, June 2015.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*, January 2017.

Michael Betancourt. Hierarchical Modeling, November 2020.

Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, April 2000.

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

Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR)*, San Diego, California, USA, May 2015.

Léon Bottou. On-line learning and stochastic approximations. In *On-Line Learning in Neural Networks*, pp. 9–42. Cambridge University Press, first edition, January 1999.

Monica F. Bugallo, Victor Elvira, Luca Martino, David Luengo, Joaquin Miguez, and Petar M. Djuric. Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, July 2017.

Olivier Cappé, Randal Douc, Arnaud Guillin, Jean-Michel Marin, and Christian P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4): 447–459, December 2008.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310, August 1989.

Akash Kumar Dhaka, Alejandro Catalina, Michael R Andersen, Måns Magnusson, Jonathan Huggins, and Aki Vehtari. Robust, accurate stochastic optimization for variational inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 10961–10973. Curran Associates, Inc., 2020.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pp. 2729–2738, Long Beach, California, USA, 2017. Curran Associates, Inc.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Dheeru Dua and Casey Graff. UCI machine learning repository. 2017.

Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

Víctor Elvira, Luca Martino, and Christian P. Robert. Rethinking the Effective Sample Size. *arXiv:1809.04129 [stat]*, September 2018.

P. H. Garthwaite, Y. Fan, and S. A. Sisson. Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process. *Communications in Statistics - Theory and Methods*, 45(17):5098–5111, September 2016.

Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: A language for flexible probabilistic inference. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1682–1690. ML Research Press, 2018.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, Cambridge; New York, 2007.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2), May 1998.

Paolo Giordani and Robert Kohn. Adaptive independent Metropolis–Hastings by fast estimation of mixtures of normals. *Journal of Computational and Graphical Statistics*, 19(2):243–259, January 2010.

Tim Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, May 1995.

Lars Holden, Ragnar Hauge, and Marit Holden. Adaptive independent Metropolis–Hastings. *The Annals of Applied Probability*, 19(1), February 2009.

Yu Hang Jiang, Tong Liu, Zhiya Lou, Jeffrey S. Rosenthal, Shanshan Shangguan, Fei Wang, and Zixuan Wu. MCMC confidence intervals and biases. *arXiv:2012.02816 [math, stat]*, June 2021.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

Jonathan M. Keith, Dirk P. Kroese, and George Y. Sofronov. Adaptive independence samplers. *Statistics and Computing*, 18(4):409–420, December 2008.

Hyunsu Kim, Juho Lee, and Hongseok Yang. Adaptive strategy for resetting a non-stationary markov chain during learning via joint stochastic approximation. In *Proceedings of the 3rd Symposium on Advances in Approximate Bayesian, to Appear*, 2021.

Sangjoon Kim, Neil Shepherd, and Siddhartha Chib. Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies*, 65(3):361–393, July 1998.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*, San Diego, California, USA, 2015.

Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American Statistical Association*, 89(425):278–288, March 1994.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14): 1–45, 2017.

Nicolas Lartillot and Hervé Philippe. Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207, April 2006.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems (NIPS)*, volume 29. Curran Associates, Inc., 2016.

Yingzhen Li, Richard E. Turner, and Qiang Liu. Approximate inference with amortised MCMC. *arXiv:1702.08343 [cs, stat]*, May 2017.

J. G. Liao and Arthur Berg. Sharpening Jensen's inequality. *The American Statistician*, 73(3): 278–281, July 2019.

David J.C. MacKay. Local minima, symmetry-breaking, and model pruning in variational free energy minimization. Technical Report, June 2001.

Luca Martino. A review of multiple try MCMC algorithms for signal processing. *Digital Signal Processing*, 75:134–152, April 2018.

K. L. Mengersen and R. L. Tweedie. Rates of convergence of the hastings and metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 1996.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, June 1953.

David D. L. Minh and Do Le (Paul) Minh. Understanding the Hastings Algorithm. *Communications in Statistics - Simulation and Computation*, 44(2):332–349, February 2015.

Thomas P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

Christian Naesseth, Fredrik Lindsten, and David Blei. Markovian score climbing: Variational inference with KL(p||q). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 15499–15510. Curran Associates, Inc., 2020.

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

Radford M Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, pp. 113–162. Chapman and Hall/CRC, first edition, 2011a.

Radford M. Neal. MCMC Using Ensembles of States for Problems with Fast and Slow Variables such as Gaussian Process Regression. Dept. of Statistics Technical Report 1011, University of Toronto, January 2011b.

Zhijian Ou and Yunfu Song. Joint stochastic approximation and its application to learning discrete latent variable models. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 929–938. ML Research Press, August 2020.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013.

P. H. Peskun. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612, 1973.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 33 of *Proceedings of Machine Learning Research*, pp. 814–822, Reykjavik, Iceland, April 2014. ML Research Press.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York, New York, NY, 2004.

Francisco Ruiz and Michalis Titsias. A contrastive divergence for combining variational inference and MCMC. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5537–5545. ML Research Press, June 2019.

Jack W Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp. 261–265, November 1988.

Stan Development Team. Stan modeling language users guide and reference manual, version 2.23.0. 2020.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv:1507.02646 [stat]*, February 2021.

Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-Divergence. In *Advances in Neural Information Processing Systems (NIPS)*, volume 31. Curran Associates, Inc., 2018.

Guanyang Wang. Exact convergence rate analysis of the independent Metropolis-Hastings algorithms. *arXiv:2008.02455 [math, stat]*, December 2020.

Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, August 2019.

A Relationship Between Sampling Methods

We organize the sampling methods described in this work in Table 2.

Table 2: Comparison of Sampling Method Designs

Algorithm	Origin	Proposal	M-H Test	Acceptance Ratio	Multiple Proposals	Reference
RWMH[1]	MCMC	Dependent	✓	M-H	✗	
HMC	MCMC	Dependent	✓	M-H	✗	Duane et al. (1987)
SNIS	IS	Independent	✗		✓	
IMH	MCMC	Independent	✓	M-H	✗	
CIS	PMCMC[4]	Independent	✓	Barker	✓	Naesseth et al. 2020
En. MCMC[2]	MCMC	Both	✓	Barker	✓	Neal 2011b
PMP MCMC[3]	MCMC	Dependent	✓	Barker	✓	Austad 2007

[1] Random-walk Metropolis-Hastings
[2] Ensemble MCMC
[3] Parallel multiple proposals MCMC
[4] Particle MCMC

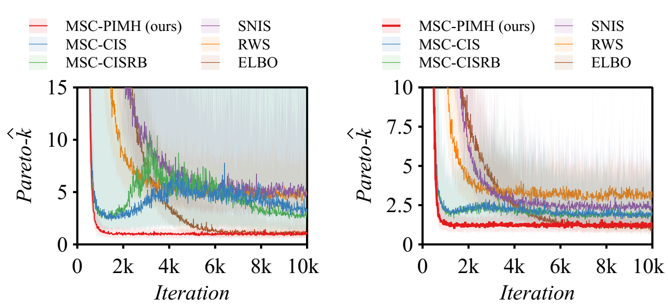
In this paper, we designated kernels that use independent proposals and perform a Metropolis-Hastings (M-H) test as "IMH type" kernels. While the original paper of CIS does not mention it as an IMH type, we have shown in Section 3.1 that it is indeed an IMH type kernel that uses Barker's acceptance ratio and multiple proposals per transition. This, in turn, reveals close connections with ensemble MCMC by Neal (2011b). While parallel multiple proposals MCMC by Austad (2007) also uses Barker's acceptance ratio and multiple proposals, it only considers dependent proposals, unlike ensemble MCMC. Although in principle, it should work with independent proposals without modification.

B Additional Experimental Results

B.1 Experimental Environment

All of our experiments presented in this paper were executed on a server with 20 Intel Xeon E5–2640 CPUs and 64GB RAM. Each of the CPUs has 20 logical threads with 32k L1 cache, 256k L2 cache, and 25MB L3 cache. All of our experiments can be executed within a few days on a system with similar computational capabilities.

B.2 Additional Results of Logistic Regression Experiments



(a) Pima (b) Heart

Figure 5: Pareto-$\hat{k}$ results of logistic regression problems. The solid lines are the median of 100 repetitions while the colored regions are the 80% empirical percentiles.

C Pseudocodes of the Considered Schemes

Algorithm 4: Single Mode

Input: MCMC kernel $K(\mathbf{z}, \cdot)$, initial sample $\mathbf{z}_0$, initial parameter $\boldsymbol{\lambda}_0$, number of iterations T, stepsize schedule γ_t

for $t = 1, 2, \ldots, T$ **do**
- $\mathbf{z}_t \sim K(\mathbf{z}_{t-1}, \cdot)$
- $s(\mathbf{z}; \boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\lambda}}(\mathbf{z}_t)$
- $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \gamma_t \, s(\mathbf{z}_t; \boldsymbol{\lambda}_{t-1})$

end

Algorithm 5: Sequential Mode

Input: initial sample $\mathbf{z}_0$, initial parameter $\boldsymbol{\lambda}_0$, number of iterations T, stepsize schedule γ_t

for $t = 1, 2, \ldots, T$ **do**
- $j = N(t-1)$
- **for** $i = 1, 2, \ldots, N$ **do**
  - $\mathbf{z}_t \sim K(\mathbf{z}_{j+i}, \cdot)$
- **end**
- $s(\mathbf{z}_t^{(i)}; \boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\lambda}}(\mathbf{z}_t^{(i)})$
- $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \gamma_t \, \frac{1}{N} \sum_{i=1}^{N} s(\mathbf{z}_{j+i}; \boldsymbol{\lambda}_{t-1})$

end

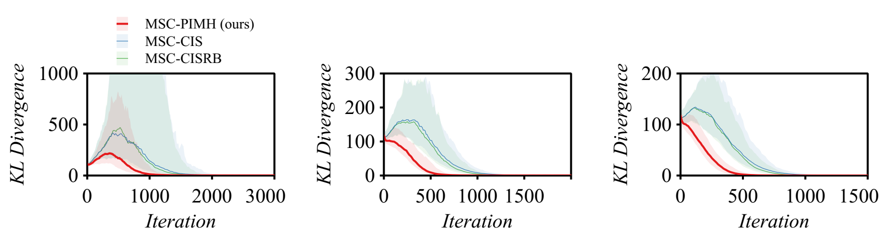
Algorithm 6: parallel Mode

Input: initial samples $\mathbf{z}_0^{(1)}, \ldots, \mathbf{z}_0^{(N)}$, initial parameter $\boldsymbol{\lambda}_0$, number of iterations T, stepsize schedule γ_t

for $t = 1, 2, \ldots, T$ **do**
- **for** $i = 1, 2, \ldots, N$ **do**
  - $\mathbf{z}_t^{(i)} \sim K(\mathbf{z}_{t-1}^{(i)}, \cdot)$
- **end**
- $s(\mathbf{z}_t^{(i)}; \boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\lambda}}(\mathbf{z}_t^{(i)})$
- $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \gamma_t \, \frac{1}{N} \sum_{i=1}^{N} s(\mathbf{z}_t^{(i)}; \boldsymbol{\lambda}_{t-1})$

end

C.1 Isotropic Gaussian Experiments



(a) $N = 2^2$ (b) $N = 2^4$ (c) $N = 2^6$

Figure 6: 100-D isotropic Gaussian example with a varying computational budget N. MSC-PIMH converges faster than MSC-CIS and MSC-CISRB regardless of N. Also, the convergence of MSC-PIMH becomes more stable/monotonic as N increases. The solid lines and colored regions are the medians and 80% percentiles computed from 100 repetitions.

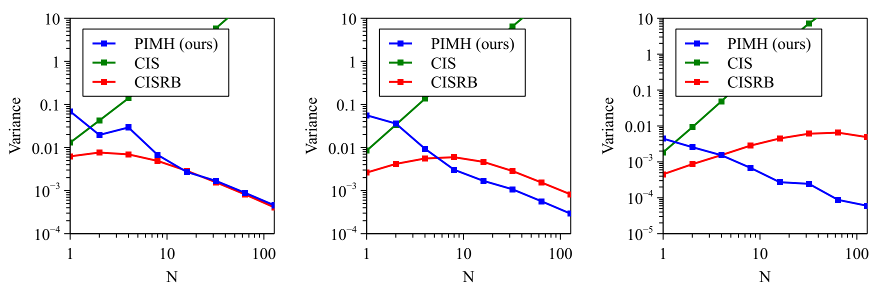
We perform experiments with a 100-D isotropic multivariate Gaussian distribution. With Gaussian distributions, convergence can be evaluated exactly since their KL divergence is available in a closed form. We compare the performance of MSC-PIMH, MSC-CIS, and MSC-CISRB with respect to the N (number of proposals for MSC-CIS, MSC-CISRB; number of parallel chains for MSC-PIMH). The results are shown in Figure 6. While MSC-PIMH shows some level of overshoot wih $N = 4$, it shows monotonic convergence with larger N. On the other hand, both

MSC-CIS and MSC-CISRB overshoots even with $N = 64$. This clearly shows that PIMH enjoys better gradient estimates compared to the CIS kernel.

D Numerical Simulation

We present numerical simulations of our analyses in Section 3.3 and Section 3.4. In particular, we visualize the fact that the variance of the CIS kernel can increase with the number of proposals N when the KL divergence is large, as described in (13).

Experimental Setup We first set the target distribution as $p(z \mid x) = \mathcal{N}(0, 1)$ and the proposal distribution as $q(z; \mu) = \mathcal{N}(\mu, 2)$ with varying mean. We measure the variance of estimating the score function $s(z, \mu) = \frac{\partial q(z;\mu)}{\partial \mu}$ using the CIS, CISRB, and PIMH kernels, given the previous Markov-chain denoted bystate z_{t-1} and computational budget N. For CIS and CISRB, we set a fixed z_{t-1}, while for PIMH, we randomly sample N samples from $\mathbf{z}_{t-1} \sim p(z \mid z)$ (we obtained similar trends regardless of the distribution of z_{t-1}). The variance is estimated using 2^{14} samples from $K(\mathbf{z}_{t-1}, \cdot)$. We report the variance across varying N and varying KL divergence between $q_\lambda(z)$ and $p(\mathbf{z} \mid \mathbf{x})$. The latter is performed by varying the difference between the mean of the proposal and the target distributions denoted by $\Delta\mu = \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}[\, z\,] - \mathbb{E}_{q_\lambda}[\, z\,]$.



(a) $\Delta\mu = 0$ (b) $\Delta\mu = 2$ (c) $\Delta\mu = 4$

Figure 7: Conditional variance of different MCMC kernels with varying N and varying difference between the mean of the target and proposal distributions.

Results Summary The results are presented in Figure 7. We can see that, when the difference of the mean of the p and q is large, the variance of CISRB *increases* with N. This increasing trend becomes stronger as the KL divergence between p and q increases. While this simulation suggests that CISRB has much smaller variance compared to CIS, our realistic experiments in Section 4 did not reveal such levels of performance gains. Is also visible that PIMH has a slightly larger variance compared to CIS in the small N regime. This is due to the higher acceptance rate of the Metropolis-Hastings acceptance ratio used by PIMH compared to Barker's acceptance ratio used by CIS (Peskun, 1973; Minh & Minh, 2015).

E Probabilistic Models Considered in Section 4

E.1 Hierarchical Logistic Regression

The hierarchical logistic regression used in Section 4.2 is

$$\begin{aligned}
\sigma_\beta &\sim \mathcal{N}^+(0, 1.0) \\
\sigma_\alpha &\sim \mathcal{N}^+(0, 1.0) \\
\boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_\beta^2 \mathbf{I}\right) \\
\alpha &\sim \mathcal{N}\left(0, \sigma_\alpha^2\right) \\
p &\sim \mathcal{N}\left(\mathbf{x}_i^\top \boldsymbol{\beta} + \alpha,\, \sigma_\alpha^2\right) \\
y_i &\sim \text{Bernoulli-Logit}(p)
\end{aligned}$$

where $\mathbf{x}_i$ and y_i are the predictors and binary target variable of the ith datapoints.

E.2 Stochastic Volatility

The stochastic volatility model used in Section 4.3 is

$$\begin{aligned}
\mu &\sim \text{Cauchy}(0, 10) \\
\phi &\sim \text{Uniform}(-1, 1) \\
\sigma &\sim \text{Cauchy}^{+}(0, 5) \\
h_1 &\sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right) \\
h_{t+1} &\sim \mathcal{N}\left(\mu + \phi\,(h_t - \mu), \sigma^2\right) \\
y_t &\sim \mathcal{N}(0, \exp(h_t))
\end{aligned}$$

where y_t is the stock price at the tth point in time. We used the reparameterized version where h_t is sampled from a white multivariate Gaussian described by the Stan Development Team (2020).

E.3 Radon Hierarchical Regression

The partially pooled linear regression model used in Section 4.3 is

$$\begin{aligned}
\sigma_{a_1} &\sim \text{Gamma}(\alpha = 1, \beta = 0.02) \\
\sigma_{a_2} &\sim \text{Gamma}(\alpha = 1, \beta = 0.02) \\
\sigma_{y} &\sim \text{Gamma}(\alpha = 1, \beta = 0.02) \\
\mu_{a_1} &\sim \mathcal{N}(0, 1) \\
\mu_{a_2} &\sim \mathcal{N}(0, 1) \\
a_{1,c} &\sim \mathcal{N}\left(\mu_{a_1}, \sigma_{a_1}^2\right) \\
a_{2,c} &\sim \mathcal{N}\left(\mu_{a_2}, \sigma_{a_2}^2\right) \\
y_i &\sim \mathcal{N}\left(a_{1,c_i} + a_{2,c_i}\, x_i,\, \sigma_y^2\right)
\end{aligned}$$

where $a_{1,c}$ is the intercept at the county c, $a_{2,c}$ is the slope at the county c, c_i is the county of the ith datapoint, x_i and y_i are the floor predictor of the measurement and the measured radon level of the $i$the datapoint, respectively. The model pools the datapoints into their respective counties, which complicates the posterior geometry (Betancourt, 2020).

F Proofs

Detailed derivation of ***Equation*** **(6)**

$$\mathbb{E}_{K(\mathbf{z}_{t-1}, \mathbf{z})}\left[\, f(\mathbf{z})\,\right] \tag{16}$$

$$= \mathbb{E}_{q_\lambda}\left[\sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)}) \,/\, \sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)\right] \tag{17}$$

(with a slight abuse of notation, $\mathbf{z}_{t-1} = \mathbf{z}^{(0)}$)

$$= \mathbb{E}_{q_\lambda}\left[\Big(\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)}) + w\left(\mathbf{z}_{t-1}\right) f(\mathbf{z}_{t-1}) \Big) / \sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right) \right] \tag{18}$$

$$= \mathbb{E}_{q_\lambda}\left[\frac{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)})}{\sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)} + \frac{w\left(\mathbf{z}_{t-1}\right)}{\sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)} f(\mathbf{z}_{t-1}) \right] \tag{19}$$

$$= \mathbb{E}_{q_\lambda}\left[\frac{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right)}{\sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)} \frac{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)})}{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right)} + \frac{w\left(\mathbf{z}_{t-1}\right)}{\sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)} f(\mathbf{z}_{t-1}) \right] \tag{20}$$

$$= \mathbb{E}_{q_\lambda}\left[\left(1 - \frac{w\left(\mathbf{z}_{t-1}\right)}{\sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)}\right) \frac{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)})}{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right)} + \frac{w\left(\mathbf{z}_{t-1}\right)}{\sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)} f(\mathbf{z}_{t-1}) \right] \tag{21}$$

$$= \mathbb{E}_{q_\lambda}\left[\alpha\left(\mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right) \frac{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)})}{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right)} + r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right) f(\mathbf{z}_{t-1}) \right] \tag{22}$$

where we denote $\alpha\left(\mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right) = \frac{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right)}{\sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)}$, $r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right) = \frac{w\left(\mathbf{z}_{t-1}\right)}{\sum_{i=0}^{N} w\left(\mathbf{z}^{(i)}\right)}$, and thus,

$$= \mathbb{E}_{q_\lambda}\left[\alpha\left(\mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right) \frac{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)})}{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right)} \right] + \mathbb{E}_{q_\lambda}\left[r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right) \right] f(\mathbf{z}_{t-1}) \tag{23}$$

$$= \mathbb{E}_{q_\lambda}\left[\alpha\left(\mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right) \frac{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right) f(\mathbf{z}^{(i)})}{\sum_{i=1}^{N} w\left(\mathbf{z}^{(i)}\right)} \right] + r\left(\mathbf{z}_{t-1}\right) f(\mathbf{z}_{t-1}) \tag{24}$$

$\square$

Theorem 1. *Assuming $w^* = \sup_{\mathbf{z}} {}^{p(\mathbf{z}|\mathbf{x})}/_{q_\lambda(\mathbf{z})} < \infty$ for $\forall \boldsymbol{\lambda}$, for a bounded functional $f : \mathcal{Z} \rightarrow [{}^{-L}/_2, {}^{L}/_2]$, the bias of the sequential mode estimator with an IMH kernel at iteration t is bounded as*

$$\mathrm{Bias}\left[g_{\mathrm{seq.,t}}\right] \leq L\, C^{N(t-1)} \frac{1}{N} \sum_{i=1}^{N} C^i$$

where $C = (1 - {}^1/_{w^*}) < 1$.

*Proof of **Theorem** 1.* We employ a similar proof strategy with the works of Jiang et al. (2021, Theorem 4). We denote the empirical distribution of the Markov-chain states at iteration t as

$$\eta_{seq.,t}(\mathbf{z}) = \frac{1}{N} \sum_{i=1}^{N} K^{N(t-1)+i}(\mathbf{z}_0, \mathbf{z}) \tag{25}$$

$$\left\| \eta_{seq.,t}(\cdot) - p\left(\cdot \mid \mathbf{x}\right) \right\|_{\mathrm{TV}} = \left\| \frac{1}{N} \sum_{i=1}^{N} K^{N(t-1)+i}(\mathbf{z}_0, \cdot) - p\left(\cdot \mid \mathbf{x}\right) \right\|_{\mathrm{TV}} \tag{26}$$

$$\leq \frac{1}{N} \sum_{i=1}^{N} \left\| K^{N(t-1)+i}(\mathbf{z}_0, \cdot) - p\left(\cdot \mid \mathbf{x}\right) \right\|_{\mathrm{TV}} \quad \text{(Triangle inequality)} \tag{27}$$

$$\tag{28}$$

For an IMH kernel with $w^* < \infty$, the geometric ergodicity of the IMH kernel (Mengersen & Tweedie, 1996, Theorem 2.1) gives the bound

$$\left\| K^t(\mathbf{z}_0, \cdot) - p(\cdot \mid \mathbf{x}) \right\|_{\mathrm{TV}} \leq \left(1 - \frac{1}{w^*}\right)^t. \tag{29}$$

But, since in our case w^* is dependent on $\boldsymbol{\lambda}$, our bound is given as

$$\left\| K^t(\mathbf{z}_0, \cdot) - p(\cdot \mid \mathbf{x}) \right\|_{\mathrm{TV}} \leq \prod_{\tau=1}^{t} \Big(1 - \frac{1}{w^*(\boldsymbol{\lambda}_\tau)}\Big). \tag{30}$$

Thus,

$$\begin{aligned}
\left\| \eta_{seq.,t}(\cdot) - p\left(\cdot \mid \mathbf{x}\right) \right\|_{\mathrm{TV}} &\leq \prod_{\tau=1}^{t-1} \Big(1 - \frac{1}{w^*(\boldsymbol{\lambda}_\tau)}\Big)^N \frac{1}{N} \sum_{i=1}^{N} \Big(1 - \frac{1}{w^*(\boldsymbol{\lambda}_t)}\Big)^i & (31)\\
&\leq \prod_{\tau=1}^{t-1} \Big(1 - \frac{1}{w^*}\Big)^N \frac{1}{N} \sum_{i=1}^{N} \Big(1 - \frac{1}{w^*}\Big)^i & (32)\\
&\leq C^{N(t-1)} \frac{1}{N} \sum_{i=1}^{N} C^i & (33)
\end{aligned}$$

where $C = 1 - \frac{1}{w^*}$.

Finally, by the definition of the total-variation distance,

$$\begin{aligned}
\mathrm{bias}\left[g_{\mathrm{seq.,t}}\right] &\leq L \left\| \eta_{seq.,t}(\cdot) - p(\cdot \mid \mathbf{x}) \right\|_{\mathrm{TV}} & (34)\\
&\leq L \sup_{h : \mathcal{Z} \to [\text{-}L/2, L/2]} \left| \mathbb{E}_{\eta_{seq.,t}(\cdot)}\left[\, h \,\right] - \mathbb{E}_{p(\cdot \mid \mathbf{x})}\left[\, h \,\right] \right| & (35)\\
&\leq L\, C_1^{N(t-1)} \frac{1}{N} \sum_{i=1}^{N} C_2^i. & (36)
\end{aligned}$$

□

Theorem 2. *Assuming $w^* = \sup_{\mathbf{z}} p(\mathbf{z}|\mathbf{x})/q_{\lambda_\tau}(\mathbf{z}) < \infty$ for $\forall \boldsymbol{\lambda}$, for a bounded functional $f : \mathcal{Z} \to [\text{-}L/2, L/2]$, the bias of the parallel mode estimator with an IMH kernel at iteration t is bounded as*

$$\mathrm{Bias}\left[g_{\mathrm{par.,t}}\right] \leq L\, C^t.$$

where $C = 1 - 1/w^*$.

*Proof of **Theorem** 2.* We denote the empirical distribution of the Markov-chain states at iteration t as

$$\eta_{par.,t}(\mathbf{z}) = \frac{1}{N} \sum_{i}^{N} K^t\left(\mathbf{z}_0^{(i)}, \mathbf{z}\right). \tag{37}$$

Similarly with Theorem 1,

$$\begin{aligned}
\left\| \eta_{par.,t}(\mathbf{z}) - p\left(\cdot \mid \mathbf{x}\right) \right\|_{\mathrm{TV}} &= \left\| \frac{1}{N} \sum_{i=1}^{N} K^t\left(\mathbf{z}_0^{(i)}, \mathbf{z}\right) - p\left(\cdot \mid \mathbf{x}\right) \right\|_{\mathrm{TV}} & (38)\\
&\leq \frac{1}{N} \sum_{i=1}^{N} \left\| K^t(\mathbf{z}_0^{(i)}, \cdot) - p\left(\cdot \mid \mathbf{x}\right) \right\|_{\mathrm{TV}} \qquad \text{(Triangle inequality)} & (39)\\
&\leq \left\| K^t(\mathbf{z}_0, \cdot) - p\left(\cdot \mid \mathbf{x}\right) \right\|_{\mathrm{TV}} \qquad \text{(Independence of the chains)} & (40)\\
&\leq \prod_{\tau=1}^{t} \Big(1 - \frac{1}{w^*(\boldsymbol{\lambda}_\tau)}\Big) & (41)\\
&\leq \prod_{\tau=1}^{t} \Big(1 - \frac{1}{w^*}\Big) & (42)\\
&\leq C^t & (43)
\end{aligned}$$

where $w^*(\boldsymbol{\lambda}_\tau) = \sup_{\mathbf{z}} {p(\mathbf{z}|\mathbf{x})}/{q_{\lambda_\tau}(\mathbf{z})}$ and $C = 1 - 1/w^*$. And, finally the bias is given as

$$\begin{aligned}
\text{bias}\left[g_{\text{seq.,t}}\right] &\leq L \left\| \eta_{seq.,t}(\cdot) - p(\cdot \mid \mathbf{x}) \right\|_{\text{TV}} && (44) \\
&\leq L \sup_{h:\mathcal{Z}\to[-L/2,L/2]} \left| \mathbb{E}_{\eta_{seq.,t}(\cdot)}\left[h\right] - \mathbb{E}_{p(\cdot|\mathbf{x})}\left[h\right] \right| && (45) \\
&\leq L\, C^t. && (46)
\end{aligned}$$

□

Proposition 1. *variance of sequential estimator*

*Proof of **Proposition** 1.*

$$\begin{aligned}
\mathbb{V}\left[g_{\text{seq.,t}}\right] &= \mathbb{V}\left[\frac{1}{N}\sum_{i=1}^{N} f\left(\mathbf{z}_{T+i}\right)\right] && (47) \\
&= \frac{1}{N^2}\sum_{i=1}^{N} \mathbb{V}\left[f\left(\mathbf{z}_{T+i}\right)\right] + \frac{1}{N^2}\sum_{i=1,\, i\leq j}^{N} \mathrm{Cov}\left(f\left(\mathbf{z}_{T+i}\right), f\left(\mathbf{z}_{T+j}\right)\right) && (48) \\
&= \frac{1}{N}\mathbb{V}\left[f\left(\mathbf{z}\right)\right] + \frac{1}{N^2}\sum_{i=1,\, i\leq j}^{N} \mathrm{Cov}\left(f\left(\mathbf{z}_{T+i}\right), f\left(\mathbf{z}_{T+j}\right)\right) \quad \text{(By stationarity)} && (49) \\
& && (50)
\end{aligned}$$

□

Proposition 2. *Assuming $w\left(\mathbf{z}_{t-1}\right)$ is large enough to make $r\left(\mathbf{z} \mid \mathbf{z}^{(1:N)}\right)$ independent of $\mathbf{z}^{(1:N)}$, the variance can be approximated by*

$$\mathbb{V}_{q_\lambda}\left[\mathbb{E}\left[f \mid \mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right]\right] \approx \left(1 - r\left(\mathbf{z}_{t-1}\right)\right)^2 \mathbb{V}_{q_\lambda}\left[f_{IS} \mid \mathbf{z}_{t-1}\right]. \tag{11}$$

*Proof of **Proposition** 2.* We evaluate the variance by approximating the rejection probability as an independent constant. First,

$$\mathbb{V}_{q_\lambda}\left[\mathbb{E}\left[f \mid \mathbf{z}_{t-1}, \mathbf{z}^{(1:N)}\right]\right]. \tag{51}$$

Applying (22),

$$\begin{aligned}
&= \mathbb{V}_{q_\lambda}\left[\left(1 - r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right)\right) f_{\text{IS}} + r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right) f(\mathbf{z}_{t-1}) \,\middle|\, \mathbf{z}_{t-1}\right] && (52) \\
&\approx \mathbb{V}_{q_\lambda}\left[\left(1 - r\left(\mathbf{z}_{t-1}\right)\right) f_{\text{IS}} + r\left(\mathbf{z}_{t-1}\right) f(\mathbf{z}_{t-1}) \,\middle|\, \mathbf{z}_{t-1}\right] && (53) \\
&= \mathbb{V}_{q_\lambda}\left[\left(1 - r\left(\mathbf{z}_{t-1}\right)\right) f_{\text{IS}} \mid \mathbf{z}_{t-1}\right] && (54) \\
&= \left(1 - r\left(\mathbf{z}_{t-1}\right)\right)^2 \mathbb{V}_{q_\lambda}\left[f_{\text{IS}} \mid \mathbf{z}_{t-1}\right]. && (55)
\end{aligned}$$

The equality of (54) follows from the fact that $r\left(\mathbf{z}_{t-1}\right) f(\mathbf{z}_{t-1})$ is a constant. □

Proposition 3. *The rejection rate $r\left(\mathbf{z}_{t-1}\right)$ of a CIS sampler with N proposals is bounded below such that*

$$r\left(\mathbf{z}_{t-1}\right) \geq \frac{1}{1 + \frac{N Z}{w(\mathbf{z}_{t-1})}}$$

where $Z = \mathbb{E}_{q_\lambda(\mathbf{z})}\left[{p(\mathbf{z},\mathbf{x})}/{q_\lambda(\mathbf{z})}\right] = \int p\left(\mathbf{z}, \mathbf{x}\right) d\mathbf{z}$ is the normalizing constant.

*Proof of **Proposition** 3.* The rejection rate $r(\mathbf{z}_{t-1})$ is given by

$$\begin{aligned}
r(\mathbf{z}_{t-1}) &= \mathbb{E}_{q_\lambda}\left[\frac{w(\mathbf{z}_{t-1})}{\sum_{k=1}^{N} w(\mathbf{z}^{(k)}) + w(\mathbf{z}_{t-1})}\right] && (56) \\
&= \mathbb{E}_{q_\lambda}\left[\left(\frac{\sum_{k=1}^{N} w(\mathbf{z}^{(k)})}{w(\mathbf{z}_{t-1})} + 1\right)^{-1}\right]. && (57)
\end{aligned}$$

At this point, we apply Jensen's inequality subject to the convex function $f(x) = 1/(1+x)$,

$$\geq \frac{1}{1 + \mathbb{E}_{q_\lambda}\left[\frac{\sum_{k=1}^{N} w(\mathbf{z}^{(k)})}{w(\mathbf{z}_{t-1})}\right]} \tag{58}$$

$$= \frac{1}{1 + \frac{1}{w(\mathbf{z}_{t-1})}\mathbb{E}_{q_\lambda}\left[\sum_{k=1}^{N} w(\mathbf{z}^{(k)})\right]}. \tag{59}$$

From the independence of the N proposals, we obtain

$$= \frac{1}{1 + \frac{1}{w(\mathbf{z}_{t-1})} N \mathbb{E}_{q_\lambda}\left[\, w(\mathbf{z})\,\right]} \tag{60}$$

$$= \frac{1}{1 + \frac{1}{w(\mathbf{z}_{t-1})} N Z}. \tag{61}$$

□

Theorem 3. *Assuming* $\sup_{\mathbf{z}} p(\mathbf{z}|\mathbf{x})/q_\lambda(\mathbf{z}) = M < \infty$, *the average rejection rate* $r = \int r(\mathbf{z}_{t-1})\, p(\mathbf{z}_{t-1} \mid \mathbf{x})\, d\mathbf{z}_{t-1}$ *of a CIS kernel with N proposals is bounded below such that*

$$r \geq \frac{1}{1 + \frac{N}{\exp\left(D_{\mathrm{KL}}(p \| q_\lambda)\right)}} - \delta,$$

where the sharpness of the bound is given as $0 \leq \delta \leq \frac{M}{\exp^2\left(D_{\mathrm{KL}}(p \| q_\lambda)\right)}$.

*Proof of **Theorem** 3.* We first show a simple Lemma that relates the rejection weight $w(\mathbf{z}_{t-1})$ with the KL divergence.

Lemma 1. *The average unnormalized weight of the rejection states is bounded below by the KL divergence such as*

$$Z \exp\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right) \leq \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, w(\mathbf{z}_{t-1})\,\right].$$

Proof. By the definition of the inclusive KL divergence,

$$D_{\mathrm{KL}}(p \parallel q_\lambda) = \int p(\mathbf{z} \mid \mathbf{x}) \log \frac{p(\mathbf{z} \mid \mathbf{x})}{q_\lambda(\mathbf{z})}\, d\mathbf{z} \leq \log \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\frac{p(\mathbf{z} \mid \mathbf{x})}{q_\lambda(\mathbf{z})}\right] \tag{62}$$

$$= \log \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\frac{w(\mathbf{z})}{Z}\right] \tag{63}$$

where the right-hand side follows from Jensen's inequality. By a simple change of notation, we relate (63) with the rejection states $\mathbf{z}_{t-1}$ such as

$$D_{\mathrm{KL}}(p \parallel q_\lambda) \leq \log \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\frac{w(\mathbf{z}_{t-1})}{Z}\right]. \tag{64}$$

Then,

$$\exp\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right) \leq \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\frac{w(\mathbf{z}_{t-1})}{Z}\right] \tag{65}$$

$$Z \exp\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right) \leq \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, w(\mathbf{z}_{t-1})\,\right]. \tag{66}$$

□

Now, from the result of Proposition 3,

$$\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, r(\mathbf{z}_{t-1})\,\right] \geq \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\frac{w(\mathbf{z}_{t-1})}{w(\mathbf{z}_{t-1}) + N Z}\right] = \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, \varphi\left(w(\mathbf{z}_{t-1})\right)\right], \tag{67}$$

where $\varphi(x) = x/(x+NZ)$. The lower bound has the following relationship

$$\varphi\left(\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[w(\mathbf{z}_{t-1})\right]\right) \geq \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\varphi\left(w(\mathbf{z}_{t-1})\right)\right] \tag{68}$$

by the concavity of φ and Jensen's inequality. From this, we denote the *Jensen gap*

$$\delta = \varphi\left(\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[w(\mathbf{z}_{t-1})\right]\right) - \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\varphi\left(w(\mathbf{z}_{t-1})\right)\right], \tag{69}$$

where $\delta \geq 0$. Then, by applying (69) to (67),

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[r(\mathbf{z}_{t-1})\right] &\geq \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\varphi\left(w(\mathbf{z}_{t-1})\right)\right] && (70)\\
&= \varphi\left(\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[w(\mathbf{z}_{t-1})\right]\right) - \delta, && (71)
\end{aligned}$$

and by the monotonicity of φ and Lemma 1,

$$\begin{aligned}
&\geq \varphi\left(Z\exp\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right)\right) - \delta && (72)\\
&= \frac{Z\exp\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right)}{Z\exp\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right) + NZ} - \delta && (73)\\
&= \frac{\exp\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right)}{\exp\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right) + N} - \delta && (74)\\
&= \frac{1}{1 + \frac{N}{\exp\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right)}} - \delta. && (75)
\end{aligned}$$

Now we discuss the Jensen gap δ, which directly gives the sharpness of our lower bound. Liao & Berg (2019, Theorem 1) have shown that, for a random variable X satisfying $P(X \in (a,b)) = 1$, where $-\infty \leq a < b \leq \infty$, and a differentiable function $\widetilde{\varphi}(x)$, the following inequality holds:

$$\inf_{x\in(a,b)} h(x;\mu)\,\sigma^2 \leq \mathbb{E}\left[\widetilde{\varphi}(X)\right] - \widetilde{\varphi}(\mathbb{E}[X]), \quad \text{where} \quad h(x;\nu) = \frac{\widetilde{\varphi}(x) - \widetilde{\varphi}(\nu)}{(x-\nu)^2} - \frac{\widetilde{\varphi}'(\nu)}{x-\nu}, \tag{76}$$

μ and σ^2 are the mean and variance of X, respectively. Also, Liao & Berg (2019, Lemma 1) have shown that, if $\widetilde{\varphi}'(x)$ is convex, then $\inf_{x\in(a,b)} h(x;\mu) = \lim_{x\to a} h(x;\mu)$.

In our case, the domain is $(a,b) = (0,\infty)$ since $w(\mathbf{z}_{t-1}) > 0$. Since $\varphi'(x) = NZ/(x+NZ)^2$ is convex, we have

$$\begin{aligned}
\lim_{x\to 0} h(x;\mu) &= \lim_{x\to 0} \frac{1}{(x-\mu)^2}\left(\varphi(x) - \varphi(\mu)\right) - \frac{1}{x-\mu}\varphi'(\mu) && (77)\\
&= \lim_{x\to 0} \frac{1}{(x-\mu)^2}\left(\frac{x}{x+NZ} - \frac{\mu}{\mu+NZ}\right) - \frac{1}{x-\mu}\left(\frac{NZ}{(\mu+NZ)^2}\right) && (78)\\
&= -\frac{1}{\mu^2}\left(\frac{\mu}{\mu+NZ}\right) + \frac{1}{\mu}\left(\frac{NZ}{(\mu+NZ)^2}\right) && (79)\\
&= -\frac{1}{\mu(\mu+NZ)} + \frac{NZ}{\mu(\mu+NZ)^2} && (80)\\
&> -\frac{1}{\mu^2}. && (81)
\end{aligned}$$

Notice that in the context of the original problem, $\mu = \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[w(\mathbf{z}_{t-1})\right]$.

We finally discuss the variance term σ^2 in (76). Since we assume $\sup p(\mathbf{z}|\mathbf{x})/q_\lambda(\mathbf{z}) = M < \infty$, $0 < \frac{p(\mathbf{z}|\mathbf{x})}{q_\lambda(\mathbf{z})} < M$ for all $\mathbf{z} \in \mathcal{Z}$. Then,

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left[w^2(\mathbf{z})\right] - \mathbb{E}\left[w(\mathbf{z})\right]^2 && (82)\\
&= \mathbb{E}\left[\left(\frac{Z\,p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right)^2\right] - \mathbb{E}\left[\frac{Z\,p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right]^2 && (83)\\
&= Z^2\left(\mathbb{E}\left[\left(\frac{p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right)^2\right] - \mathbb{E}\left[\frac{p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right]^2\right) && (84)\\
&= Z^2\,\mathbb{V}\left[\frac{p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right], && (85)
\end{aligned}$$

and by Bhatia & Davis (2000)'s inequality,

$$0 \leq \sigma^2 = Z^2\, \mathbb{V}\left[\, \frac{p\,(\mathbf{z} \mid \mathbf{x})}{q_\lambda(\mathbf{z})} \,\right] \leq Z^2\,(M-\mu)\,\mu. \tag{86}$$

By combining the results, we obtain

$$0 \leq \delta \leq -\inf_{x \in (a,b)} h\,(x;\mu)\,\sigma^2 = -\sigma^2 \lim_{x \to 0} h\,(x;\mu) < \frac{\sigma^2}{\mu^2} < \frac{Z^2 M}{\mu^2} = \frac{Z^2\, M}{\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, w\,(\mathbf{z}_{t-1})\,\right]^2}\,, \tag{87}$$

and by Lemma 1,

$$0 \leq \delta < \frac{M}{\exp^2\,\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right)}. \tag{88}$$

□

Proposition 4. *The rejection rate $r\,(\mathbf{z}_{t-1})$ of a IMH sampler is bounded below such that*

$$r\,(\mathbf{z}_{t-1}) \geq 1 - \frac{Z}{w\,(\mathbf{z}_{t-1})} \quad \textit{where} \quad Z = \int p\,(\mathbf{z}, \mathbf{x})\, d\mathbf{z}.$$

*Proof of **Proposition** 4.* The rejection rate $r(\mathbf{z}_{t-1})$ is given by

$$r\,(\mathbf{z}_{t-1}) = 1 - \int \alpha\,(\mathbf{z}, \mathbf{z}_{t-1})\; q_\lambda(\mathbf{z})\, d\mathbf{z}. \tag{89}$$

For an IMH sampler with the Metropolis-Hastings acceptance function and independent proposals, the rejection rate is bounded such that

$$\begin{aligned}
r\,(\mathbf{z}_{t-1}) &= 1 - \int \min\left(\frac{w\,(\mathbf{z})}{w\,(\mathbf{z}_{t-1})}, 1\right) q_\lambda(\mathbf{z})\, d\mathbf{z} && (90)\\
&= 1 - \frac{1}{w\,(\mathbf{z}_{t-1})} \int \min\Big(w\,(\mathbf{z}), w\,(\mathbf{z}_{t-1})\Big)\, q_\lambda(\mathbf{z})\, d\mathbf{z} && (91)\\
&= 1 - \frac{1}{w\,(\mathbf{z}_{t-1})} \int \min\left(\frac{p\,(\mathbf{z}, \mathbf{x})}{q_\lambda(\mathbf{z})}, w\,(\mathbf{z}_{t-1})\right) q_\lambda(\mathbf{z})\, d\mathbf{z} && (92)\\
&= 1 - \frac{1}{w\,(\mathbf{z}_{t-1})} \int \min\Big(p\,(\mathbf{z}, \mathbf{x}), w\,(\mathbf{z}_{t-1})\, q_\lambda(\mathbf{z})\Big)\, d\mathbf{z} && (93)\\
&\geq 1 - \frac{1}{w\,(\mathbf{z}_{t-1})} \int p\,(\mathbf{z}, \mathbf{x})\, d\mathbf{z} && (94)\\
&= 1 - \frac{Z}{w\,(\mathbf{z}_{t-1})} && (95)
\end{aligned}$$

The inequality in Equation (94) follows from $\min\big(p\,(\mathbf{z}, \mathbf{x}),\, \cdot\,\big) \leq p\,(\mathbf{z}, \mathbf{x})$ for $\forall \mathbf{z} \in \mathcal{Z}$. □