Markov-Chain Monte Carlo Score Estimators for Variational Inference with Score Climbing

Abstract

Variational inference (VI) methods that minimize the inclusive Kullback-Leibler (KL) divergence using Markov-chain Monte Carlo (MCMC) have recently been developed. These score climbing VI methods perform stochastic gradient descent by obtaining noisy estimates of the score function using MCMC. In this paper, we compare the different ways to combine MCMC with score climbing VI, including our novel scheme: the *parallel state estimator*. It operates multiple short Markov-chains in parallel, which results in lower variance than the previous schemes. In the traditional MCMC setting, this would come at the cost of higher bias. However, in the score climbing VI setting, we theoretically show that this does not necessarily result in higher bias, making the parallel state estimator better overall.

1 INTRODUCTION

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

Naturally, choosing a discrepancy measure is critical to the problem. This fact had lead to a quest for suitable divergence measures [Salimans et al., 2015, Li and Turner, 2016, Dieng et al., 2017, Wang et al., 2018, Ruiz and Titsias, 2019]. So far, the exclusive KL divergence $D_{\mathrm{KL}}(q_\lambda \parallel \pi)$ (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

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

where the average is taken over π. This is a chicken-and-egg problem as our goal is to obtain π in the first place. Despite this challenge, minimizing (1) has drawn the attention of researchers because it is believed to result in favorable properties [Minka, 2005, MacKay, 2001].

For minimizing the inclusive KL divergence, Naesseth et al. [2020] and Ou and Song [2020] have recently proposed methods that perform stochastic gradient descent (SGD, Robbins and Monro 1951) with the score function estimated using Markov-chain Monte Carlo (MCMC). These MCMC score climbing schemes operate a Markov-chain in conjunction with the VI optimizer. In addition, within the MCMC kernel, they both use Metropolis-Hastings proposals generated from the variational approximation $\mathbf{z}^* \sim q_\lambda(\cdot)$. The MCMC kernel itself benefits from VI, enjoying better proposals over time without the need for computationally expensive proposals as in Hamiltonian Monte Carlo [Duane et al., 1987, Neal, 2011, Betancourt, 2017]. Also, in terms of computational cost, score climbing is efficient compared to other divergences since we do not need to differentiate through the likelihood.

While the convergence of score climbing VI has been astablished by Naesseth et al. [2020], Gu and Kong [1998], the practical performance implications of the design choices have yet to be understood. For example, the methods by Naesseth et al. and Ou and Song [2020] are conceptually similar, but they utilize MCMC kernels in different ways. At each SGD iteration, for estimating the score function, Naesseth et al. use a single sample generated from a relatively expansive MCMC kernel, while Ou and Song average multiple samples generated from a

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

cheaper MCMC kernel. We call the former scheme the *single state estimator* and the later the *sequential state estimator*. Given the two options, it is natural to ask, "which is better? An estimator with multiple cheap samples? or one with a single expensive sample?".

In this paper, we propose a third scheme, the *parallel state estimator*. The parallel state estimator operates N parallel Markov-chains parallel, where only a single state transition is performed on each chain. The variance of this estimator linearly decreases with the computational budget N, unlike the single and sequential state estimators. According to the traditional MCMC intuition, this improvement would come at the cost of increased bias. However, in the example in Section 3.3, we show that this intuition can be wrong in the score climbing setting. In fact, in some cases, it can enjoy *lower* bias than the longer Markov-chains of the sequential state estimator.

To explain the unintuitive performance benefit of the parallel state estimator, we theoretically analyze the variance of the considered estimators. Then, we show that the bias of the parallel state estimator is bounded by the KL objective, meaning that it is affected by the convergence speed of VI. Therefore, the much lower variance of the parallel state estimator results in faster convergence, and thus fast decrease in bias. This suggests that, given a similar computational budget, the parallel state estimator is overall better than the alternative estimators. We also provide experimental evidence on general Bayesian inference problems. Also, within our experiments, score climbing VI with the parallel state estimator shows to be competitive against evidence lower-bound (ELBO) maximization. We further discuss this result in relation with the conclusions of Dhaka et al. [2021] in Section 6.

- ➤ We propose the parallel state estimator for variational score climbing (**Section 3.1**).
- ➤ The parallel state estimator achieves lower variance than the sequential [Ou and Song, 2020] and single state esimators [Naesseth et al., 2020] (**Section 3.4**).
- ➤ We show that the parallel state does not necessarily result in higher bias, making it the best choice in general (**Section 3.5**).
- ➤ We experimentally compare the VI performance of the considered MCMC estimation schemes on general Bayesian inference benchmarks (**Section 4**).

2 BACKGROUND

2.1 INCLUSIVE VARIATIONAL INFERENCE UNTIL NOW

Variational Inference The goal of VI is to find the optimal variational parameter $\boldsymbol{\lambda}$ identifying q_λ that minimizes the discrepancy measure $D(\pi, q_\lambda)$ (in our case, the inclusive KL). A typical way to perform VI is to use stochastic gradient descent (SGD, Robbins and Monro 1951), provided that unbiased gradient estimates of the optimization target $\mathbf{g}(\boldsymbol{\lambda})$ are available. In this case, SGD is performed by repeating the update

$$\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \gamma_t \, \mathbf{g}(\boldsymbol{\lambda}_{t-1}) \tag{2}$$

where $\gamma_1, \ldots, \gamma_T$ is a step-size schedule following the conditions of Robbins and Monro [1951], Bottou [1999]. In the case of inclusive KL divergence minimization, $\mathbf{g}$ estimates

$$\nabla_\lambda D_{\mathrm{KL}}(\pi \parallel q_\lambda) = -\mathbb{E}_\pi [\, \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \,] \approx \mathbf{g}(\boldsymbol{\lambda}) \tag{3}$$

where $s(\boldsymbol{\lambda}; \mathbf{z}) = \nabla_\lambda \log q_\lambda(\mathbf{z})$ is known as the *score function*.

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

$$\mathbb{E}_\pi [\, \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \,] \propto Z \, \mathbb{E}_\pi [\, \mathbf{s}(\mathbf{z}; \boldsymbol{\lambda}) \,] \tag{4}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} w(\mathbf{z}^{(i)}) \, \mathbf{s}(\mathbf{z}^{(i)}; \boldsymbol{\lambda}) \tag{5}$$

$$= \mathbf{g}_{IS}(\boldsymbol{\lambda}) \tag{6}$$

where $w(\mathbf{z}) = p(\mathbf{z}, \mathbf{x}) / q_\lambda(\mathbf{z})$ is known as the *importance weight*, Z is the marginal $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) \, d\mathbf{z}$, and $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ are N independent samples from $q_\lambda(\mathbf{z})$. This scheme is equivalent to adaptive IS methods [Bugallo et al., 2017] since the IS proposal $q_\lambda(\mathbf{z})$ is iteratively optimized based on the current samples. Although IS is unbiased, it is highly numerically unstable because of Z in Equation (4). 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 results in 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 and Casella, 2004].

3 MARKOV-CHAIN MONTE CARLO ESTIMATORS FOR SCORE CLIMBING

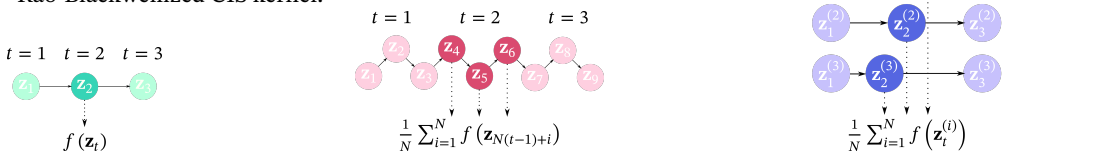
3.1 OVERVIEW OF THE ESTIMATORS

Score Climbing with MCMC Recently, Naesseth et al. [2020] and Ou and Song [2020] proposed two similar but independent score climbing method that minimize the inclusive KL with SGD. Both methods estimate the score function gradient by operating a Markov-chain in parallel with the VI optimization sequence. They notably use MCMC kernels that can effectively utilize the variational approximation $q_{\lambda_t}(\mathbf{z})$. Because of this, both methods are computationally more efficient than previous VI approaches [Ruiz and Titsias, 2019, Hoffman, 2017] that used expensive MCMC kernels such as Hamiltonian Monte Carlo.

Table 1: Computational Costs of Markov-chain Schemes

	Posterior Sampling			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.
Evidence Lower Bound Path Derivative	0	0	N	N	N
Single State Estimator with CIS Kernel (single-CIS)	$N-1$	N	$N-1$	0	1^1 or N^2
Sequential State Estimator with IMH Kernel (seq.-IMH)	N	$N+1$	N	0	N
Parallel State Estimator with IMH Kernel (par.-IMH)	N	$2N$	N	0	N

* We assume that the parameters are cached as much as possible.
* N is the number of samples used in each method.
[1] Vanilla CIS kernel.
[2] Rao-Blackwellized CIS kernel.



(a) Single State Estimator (b) Sequential State Estimator (c) Parallel State Estimator (proposed)

Figure 1: Visualization of the three different ways to utilize MCMC for score climbing VI. The index t denotes the SGD iteration. The dark circles represent the MCMC samples used for estimating the score gradient at $t = 2$.

Single State Estimator In Markovian score climbing (MSC), Naesseth et al. [2020] estimate the score gradient by performing an MCMC transition and estimate the score function gradient as

$$\mathbf{z}_t \sim K_{\lambda_{t-1}}(\mathbf{z}_{t-1}, \cdot), \qquad \mathbf{g}_{\text{single-CIS}}(\lambda) = \mathbf{s}(\lambda; \mathbf{z}_t),$$

where $K_{\lambda_{t-1}}(\mathbf{z}_{t-1}, \cdot)$ is a MCMC kernel leaving $p(\mathbf{z} \mid \mathbf{x})$ invariant and $g_{\text{single}}(\lambda)$ denotes the score estimator. For the MCMC kernel, they propose to use what they call the conditional importance sampling (CIS, previously called iterated sampling importance resampling by Andrieu et al. [2018]). Since the estimator uses *a single state* created by the CIS kernel, we call it the single state estimator with the CIS kernel (**single-CIS**). The CIS kernel internally uses N samples from the $q_{\lambda_{t-1}}(\mathbf{z})$, hence the dependence on λ_{t-1}. When compared to MCMC kernels that only use a single sample from $q_{\lambda_{t-1}}(\mathbf{z})$, it is N times more expensive, but hopefully, statistically superior.

Sequential State Estimator On the other hand, at each SGD iteration t, Ou and Song [2020] perform N sequential Markov-chain transitions and use the average of the intermediate states for estimation. That is, for the transition number $n \in \{1, \dots, N\}$,

$$\mathbf{z}_{T+i} \sim K^i_{\lambda_{t-1}}(\mathbf{z}_T, \cdot), \quad \mathbf{g}_{\text{seq.-IMH}}(\lambda) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{s}(\lambda; \mathbf{z}_{T+n}),$$

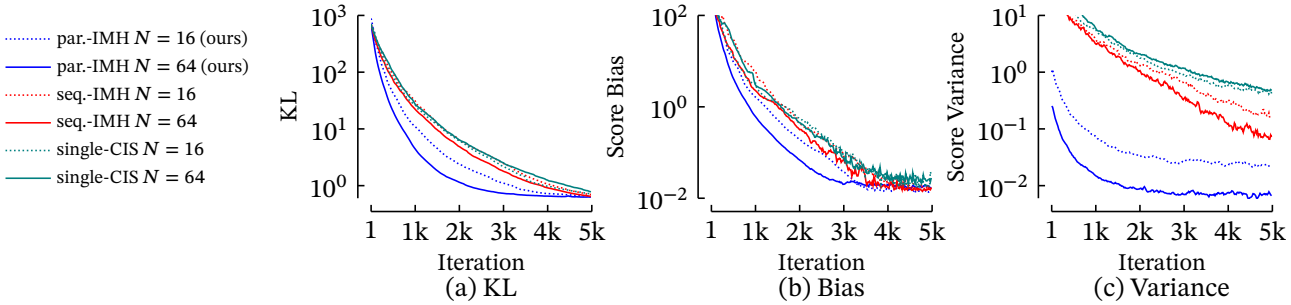
where $\mathbf{z}_T$ is the last Markov-chain state of the previous SGD iteration $t-1$. $K^n_{\lambda_{t-1}}(\mathbf{z}_T, \cdot)$ denotes the MCMC kernel sequentially applied n times. For the MCMC kernel, they use the independent Metropolis-Hastings (IMH, Robert and Casella 2004, Algorithm 25 Hastings 1970) algorithm, which uses only a single sample from $q_{\lambda_{t-1}}(\mathbf{z})$ (notice the dependence on λ_{t-1}). Therefore, the cost of N state transitions with IMH is similar to a single transition with CIS. Since the estimator uses sequential states, we call it the sequential state estimator with the IMH kernel (**seq.-IMH**)

Parallel State Estimator In this work, we propose a new scheme into the mix: *the parallel state estimator*. Like the sequential state estimator, we use the cheaper IMH kernel, but instead of applying the MCMC kernel N times to a single chain, we apply the MCMC kernel a single time to N *parallel Markov-chains* (**par.-IMH**). That is, for each Markov-chain indexed by $i \in \{1, \dots, N\}$,

$$\mathbf{z}_t^{(i)} \sim K_{\lambda_{t-1}}(\mathbf{z}_{t-1}^{(i)}, \cdot), \quad \mathbf{g}_{\text{par.-IMH}}(\lambda) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{s}(\lambda; \mathbf{z}_t^{(i)}),$$

where $\mathbf{z}_{t-1}^{(i)}$ is the state of the ith chain at the previous SGD step. Computationally speaking, we are still applying K N times in total, so the cost is similar to the sequential state estimator. However, the Markov-chains are N times shorter, which, in a traditional MCMC view, might seem to result in worse statistical performance.

Illustration The single and sequential state estimators represent two different ways of using a fixed computational budget for each SGD step. The former uses a single sample generated expensively, while the latter uses multiple samples generated cheaply. On the other hand, the parallel state estimator runs multiple chains where the chains share the budget of the sequential state estimator. An illustration of the three schemes is provided in Figure 1.



(a) KL (b) Bias (c) Variance

Figure 2: Visualization of the bias and variance during score climbing. We can see that the parallel state estimator achieves both the lowest bias and the lowest variance. See the main text for details about the experimental setup.

3.2 COMPUTATIONAL COST

The three schemes using the CIS kernel and the IMH kernel can have different computational costs depending on the parameter N. The computational costs of each scheme are organized in Table 1 while detailed pseudocodes of the considered schemes are provided in the *supplementary material*.

Cost of Sampling Proposals In the CIS kernel, N controls the number of internal proposals sampled from $q_\lambda(\mathbf{z})$. In the sequential and parallel state estimators, the IMH kernel only uses a single sample from $q_\lambda(\mathbf{z})$, but applies the kernel N times. Assuming caching is done as much as possible, the parallel state estimator needs twice the density evaluations of $q_\lambda(\mathbf{z})$ compared to other methods. However, this added cost is minimal since the overall computational cost is dominated by $p(\mathbf{z}, \mathbf{x})$.

Cost of Estimating the Score When estimating the score, the single state estimator computes $\nabla_\lambda \log q_\lambda(\mathbf{z})$ only once, while for the sequential and parallel state estimators compute it N times. However, Naesseth et al. [2020] also discuss a Rao-Blackwellized version of the CIS kernel, which also computes the gradient N times. Lastly, notice that score climbing does not need to differentiate through the likelihood, unlike ELBO maximization, making it's base computational cost significantly cheaper.

3.3 MOTIVATION AND OVERVIEW

Motivating Example According to traditional MCMC theory, multiple short Markov-chains will be more biased than a single long Markov-chain. Therefore, the parallel estimator will surely be more biased than the sequential estimator. However, we show an example where this intuition is wrong: As shown in Figure 2, in this example, the parallel state estimator enjoys not only low variance, but also low bias. We ran score climbing VI with the three different estimators and compared the bias and variance of the estimators. The target distribution was a 10 dimensional multivariate Gaussian where the covariance was sampled from a Wishart distribution with $\nu = 50$ degrees of freedom. The variational family was a mean-field Gaussian. The bias and variance was estimated from 512 independent replications.

Overview of Theoretical Result The traditional MCMC intuition fails because the MCMC kernel depends on the variational approximation q_λ. If we fix q_λ, the traditional MCMC intuition holds: given the same proposal q_λ, the parallel estimator is more biased than the sequential estimator. However, in our case, q_λ is being updated every step. Therefore, if the parallel state estimator somehow results in faster and more stable VI convergence, the intuition obtained with a fixed q_λ completely breaks. In the following sections, we will show that

❶ the parallel state estimator results in substantially low variance (Theorem 2) and that
❷ its bias decreases as VI converges (Theorem 3).
❸ The substantially low variance enables VI to converge faster, leading to a faster decrease of bias.
❹ Even if we compare the bias of the parallel and sequential state estimator with fixed a q_λ, when the KL objective is large, the sequential state estimator does not guarentee lower bias (Theorem 4).

3.4 VARIANCE OF SCORE ESTIMATORS

Stationarity Assumption First, we compare the variance of the three estimators. We generally assume that the Markov-chains have achieved stationarity, which does *not* require $N \to \infty$. Instead, it only requires use to have run SGD sufficiently enough so that the chains have landed near the typical set at least once. This is realistic even if the MCMC kernels are not mixing properly.

In addition, to closely understand the performance of the estimators relying on the IMH kernel, we utilize the following result by Smith and Tierney [1996].

Theorem 1. *(Smith and Tierney 1996, Theorem 1) The t-step marginal IMH transition kernel is given as*

$$K^t(\mathbf{z}, d\mathbf{z}^*) = T_t(w(\mathbf{z}) \vee w(\mathbf{z}^*))\, \pi(\mathbf{z}^*)\, d\mathbf{z}^* + \lambda^t(w(\mathbf{z}))\, \delta_{\mathbf{z}}(d\mathbf{z}^*)$$

where $x \vee y = \max(x, y)$, *and for* $R(w) = \{\mathbf{z}^* \mid w(\mathbf{z}^*) \le w\}$,

$$T_t(w) = \int_w^\infty \frac{t}{v^2} \lambda^{t-1}(v)\, dv, \quad \lambda(w) = \int_{R(w)} \left(1 - \frac{w(\mathbf{z}^*)}{w}\right) \pi(d\mathbf{z}^*).$$

Theorem 2. *Assuming that the Markov-chains have achieved stationarity and $N \geq 2$, the variance of the single state estimator ($\mathbf{g}_{single}$), sequential state estimator with N states of the IMH kernel ($\mathbf{g}_{seq.\text{-}IMH}$), and parallel state estimator ($\mathbf{g}_{par.}$) with N chains are given as*

$$\mathbb{V}\left[\,\mathbf{g}_{single}\,\right] = \sigma^2, \;\; \mathbb{V}\left[\,\mathbf{g}_{seq.\text{-}IMH}\,\right] = \frac{\sigma^2}{N} + C_{gap}, \;\; \mathbb{V}\left[\,\mathbf{g}_{par.}\,\right] = \frac{\sigma^2}{N}$$

where $\sigma^2 = \mathbb{V}_\pi\left[\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\right]$ and $C_{gap} \geq 0$. The variance gap C_{gap} is bounded below as

$$C_{gap} \geq \frac{2}{N}\Big[\, 2\,\sigma^2 + C\big\{\,\exp\left(D_{\mathrm{KL}}(q \parallel \pi)\right) - 3\,\big\} + Cov_q\left(\Delta^2(\mathbf{z}),\, 1/w(\mathbf{z})\right)\Big] + \mathcal{O}\left(N^{-2}\right)$$

where q is the proposal distribution, C is a positive constant, $\Delta(\mathbf{z}) = \mathbf{s}(\boldsymbol{\lambda};\mathbf{z}) - \mathbb{E}_\pi\left[\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\right]$, and, as $N \to \infty$, the bound converges to

$$C_{gap} \geq \frac{2}{N}\Big[\,\sigma^2\big\{\,\exp\left(D_{\mathrm{KL}}(\pi \parallel q)\right) - 1\,\big\} + Cov_\pi\left(\Delta^2(\mathbf{z}),\, w(\mathbf{z})\right)\Big].$$

Proof. See the *full proof* in Appendix D.

As stated in Theorem 2, the variance of the single state estimator is as good as a *single* sample for the posterior. In contrast, the variance of the sequential and parallel state estimator can be reduced by increasing N. Because of the noticable limitation of the single state estimator, from now on, we will restrict our discussion to the other two estimators. For the sequential state estimator, it is difficult to discuss the covariance term in the lower bound of C_{gap} without compromising generality, which is the limitation of our analysis. However, if we are able to assume neglect the covariance term, the KL divergence terms imply that the variance of the sequential estimator will be substantially higher during VI. This is in accordance with the empirical results in Sections 3.3 and 4.

Variance of the Parallel State Estimator The variance of the single and parallel state estimators in Theorem 2 do not assume a specific MCMC kernel. Also, the variance reduction of the parallel state estimator does not require stationarity. Therefore, the parallel state estimator *always* enjoys $\mathcal{O}\left(1/N\right)$ variance reduction. To summarize, the variance of the parallel state estimator will be much lower than both the single and sequential state estimator for $N \geq 2$.

3.5 BIAS OF THE PARALLEL STATE ESTIMATOR

Now, we will formally show that the convergence speed of VI affects the bias of the parallel state estimator.

Relaxing Geometric Ergodicity Previously, the non-asymptotic bias of MCMC kernels have been established throught their geometric convergence rates [Jiang et al., 2021]. For IMH kernels, this requires the rather strong assumption of $\sup_{\mathbf{z}} w(\mathbf{z}) < \infty$ [Mengersen and Tweedie, 1996, Wang, 2020]. We instead use a weaker, more general assumption that uses η, the distribution of the previous states of the parallel chains (such that $\mathbf{z}_{t-1}^{(i)} \sim \eta(\cdot)$). Our key assumption is that $\eta(\mathbf{z}) < M\,\pi(\mathbf{z})$ for some constant $M < \infty$. Since η is the marginal distribution of the MCMC kernel applied $t-1$ times, the following proposition shows a practical condition where our assumption is satisfied.

Proposition 1. *For the initial state $\mathbf{z} \in \mathcal{Z}$, let $w(\mathbf{z}) > \epsilon$ and $\pi(\mathbf{z}) > \epsilon$ for some constant $\epsilon > 0$ and define $w^* = \sup_{\mathbf{z}} w(\mathbf{z})$. Then, the ratio between the t-step marginal kernel and the invariant distribution is bounded as*

$$\frac{K^t(\mathbf{z}, \mathbf{z}^*)}{\pi(\mathbf{z}^*)} \leq \frac{t}{\epsilon}\left(1 - \frac{1}{w^*}\right)^{t-1}.$$

which is finite for all t if $w^* < \infty$.

Proof. See the *full proof* in Appendix D.

Therefore, our assumption can be satisfied for any finite t under mild assumptions, while $w^* < \infty$ extends the guarentee to $t \to \infty$.

Bias Convergence Now, the following bound shows that the bias of the parallel state estimator is bounded by the KL divergence between π and q.

Theorem 3. *Assuming $\eta(\mathbf{z}) < M\,\pi(\mathbf{z})$ for some $M < \infty$ and $\| \mathbf{s}(\boldsymbol{\lambda};\mathbf{z}) \| \leq L$, the bias of the parallel state estimator with an IMH kernel is bounded as*

$$\mathrm{Bias}\left[\mathbf{g}_{par.\text{-}IMH}\right] \leq C\sqrt{D_{\mathrm{KL}}(\pi \parallel q)} + L\left(1 - \frac{1}{w^*}\right)$$

for some positive constant C.

Proof. See the *full proof* in Appendix D.

Therefore, the VI convergence rate determines how quickly the bias goes down. Note that it is also possible to derive a bound without assuming $\| \mathbf{s}(\boldsymbol{\lambda};\mathbf{z}) \| \leq L$, but this unfortunately results in a weaker relationship between the KL and the bias. In addition, this bound is slightly weaker than geometric ergodicity since the bias does not go down to 0 with $w^* < \infty$, but is more practical in the large $D_{\mathrm{KL}}(\pi \parallel q)$ regime.

But what about the sequential state estimator? Despite our result, one might expect that, in some cases, choosing the sequential state estimator with a large N might be beneficial. The following bound shows that this is not likely in general.

Theorem 4. *Assuming $\eta(\mathbf{z}) < M\,\pi(\mathbf{z})$ for some $M < \infty$, when using the IMH kernel, the reduction in bias by using the sequential state estimator with N states instead of the parallel state estimator with N chains is bounded as*

$$\left|\, Bias\left[\mathbf{g}_{seq.\text{-}IMH}\right] - Bias\left[\mathbf{g}_{par.\text{-}IMH}\right] \right|$$

Table 2: Classification Accuracy and Log Predictive Density on Logistic Regression Problems

	`Pima Indians`		`heart disease`		`German credit`	
	Test Accuracy	Test LPD	Test Accuracy	Test LPD	Test Accuracy	Test LPD
ELBO	**0.77 (0.77, 0.77)**	-0.53 (-0.55, -0.51)	0.84 (0.84, 0.84)	**-0.40 (-0.40, -0.39)**	**0.77 (0.77, 0.77)**	-0.54 (-0.58, -0.50)
par.-IMH (ours)	**0.77 (0.77, 0.77)**	**-0.51 (-0.51, -0.50)**	**0.85 (0.84, 0.85)**	**-0.40 (-0.40, -0.40)**	**0.77 (0.77, 0.77)**	**-0.50 (-0.50, -0.50)**
seq.-IMH	0.67 (0.65, 0.69)	-0.71 (-0.76, -0.65)	0.79 (0.78, 0.80)	-0.45 (-0.46, -0.44)	0.76 (0.76, 0.76)	-0.51 (-0.51, -0.51)
single-CIS	0.69 (0.67, 0.71)	-0.68 (-0.73, -0.62)	0.79 (0.78, 0.80)	-0.46 (-0.48, -0.44)	0.76 (0.75, 0.76)	-0.51 (-0.52, -0.51)
single-CISRB	0.71 (0.69, 0.72)	-0.62 (-0.67, -0.58)	0.80 (0.79, 0.81)	-0.44 (-0.45, -0.43)	0.76 (0.76, 0.76)	-0.52 (-0.52, -0.51)
single-HMC	0.75 (0.75, 0.75)	-0.52 (-0.53, -0.52)	0.80 (0.79, 0.81)	-0.45 (-0.46, -0.44)	**0.77 (0.76, 0.77)**	-0.61 (-0.72, -0.49)
SNIS	0.72 (0.71, 0.72)	-0.59 (-0.61, -0.57)	0.78 (0.77, 0.79)	-0.46 (-0.48, -0.45)	0.75 (0.75, 0.76)	-0.52 (-0.52, -0.52)

* LPD denotes the average log predictive density.
* The numbers in the parentheses denote the 80% bootstrap confidence intervals computed from 100 repetitions.

$$\leq C_1 \max\left\{ n\left(1-\frac{1}{w^*}\right)^{n-1} - 1, 1 \right\} + C_2 \left(1-\frac{1}{w^*}\right)$$

for some positive constants C_1 and C_2.

Proof. See the *full proof* in Appendix D.

Proposition 2. *$w^* = \sup_{\mathbf{z}} w(\mathbf{z})$ is bounded below exponentially by the KL divergence such that*

$$\exp\left(D_{\mathrm{KL}}(\pi \parallel q_\lambda)\right) \leq w^*.$$

Proof. See the *full proof* in Appendix D.

Combined with Proposition 2, our bound suggests that, when the KL is large, using the sequential state estimator is not beneficial even with a large N. Even worse, the bias could actually *increase* with N. Therefore, using the sequential state estimator does not guarantee lower bias, especially in the early stages of VI. Given that the parallel state estimator has signifiacantly better variance guarentees, our result suggests that there is close to no practical benefit of using the sequential state estimator.

4 EVALUATIONS

4.1 BASELINES AND IMPLEMENTATION

Implementation For the realistic experiments, we implemented score climbing VI 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 and Ba [2015] with a learning rate of 0.01 in all of the experiments. The computational budget is set to $N = 10$ and $T = 10^4$ for all experiments unless specified.

We compare the following methods.

❶ **par.-IMH**: Score climbing with the parallel state estimator and the IMH kernel.
❷ **seq.-IMH**: Score climbing with the sequential state estimator and the IMH kernel
❸ **single-CIS**: Score climbing with the single state estimator and the CIS kernel [Naesseth et al., 2020].
❹ **single-CISRB**: Rao-Blackwellized version of single-CIS [Naesseth et al., 2020].
❸ **single-HMC**: Score climbing with the single state estimator and the HMC kernel.
❺ **SNIS**: adaptive IS using SNIS (as discussed in Section 2.1).
❻ **ELBO**: evidence lower-bound maximization with automatic differentiation VI [Ranganath et al., 2014, Kucukelbir et al., 2017] and the path derivative estimator [Roeder et al., 2017].

For ELBO, we use only a single sample as originally described by Roeder et al. [2017]. This also ensures a fair comparison against inclusive KL minimization methods since the iteration complexity of computing the ELBO gradient can be easily a few orders of magnitude larger. Also, we only use single-HMC in the logistic regression experiment due to its high computational demands,

4.2 HIERARCHICAL LOGISTIC REGRESSION

Experimental Setup We first perform logistic regression with the `Pima Indians` diabetes ($\mathbf{z} \in \mathbb{R}^{11}$, Smith et al. 1988), `German credit` ($\mathbf{z} \in \mathbb{R}^{27}$), and `heart disease` ($\mathbf{z} \in \mathbb{R}^{16}$, Detrano et al. 1989) datasets obtained from the UCI repository [Dua and 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, we choose a more complex hierarchical logistic regression model

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

where $\mathcal{N}^+(\mu, \sigma)$ is a positive constrained normal distribution with mean μ and standard deviation σ, $\mathbf{x}_i$ and y_i

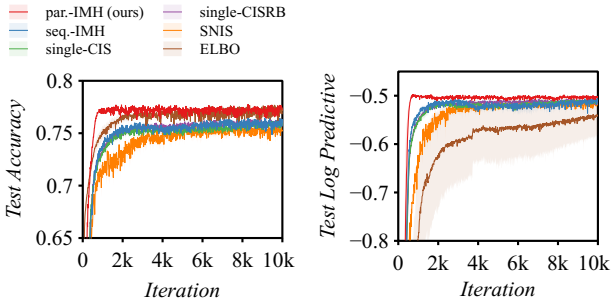
Table 3: Classification Accuracy and Log Predictive Density on Logistic Gaussian Process Problems

	`sonar`		`ionosphere`		`breast`	
	Test Accuracy	Test LPD	Test Accuracy	Test LPD	Test Accuracy	Test LPD
ELBO	0.83 (0.82, 0.84)	-0.39 (-0.40, -0.38)	0.94 (0.94, 0.94)	-0.28 (-0.29, -0.28)	0.96 (0.96, 0.97)	-0.12 (-0.12, -0.11)
par.-IMH	**0.84 (0.83, 0.85)**	**-0.36 (-0.38, -0.35)**	**0.95 (0.94, 0.95)**	**-0.25 (-0.26, -0.24)**	0.96 (0.96, 0.97)	**-0.11 (-0.12, -0.11)**
seq.-IMH	**0.84 (0.83, 0.85)**	-0.37 (-0.38, -0.36)	0.94 (0.94, 0.95)	-0.27 (-0.27, -0.26)	0.96 (0.96, 0.96)	-0.14 (-0.14, -0.13)
single-CIS	0.83 (0.82, 0.84)	-0.38 (-0.39, -0.37)	0.94 (0.94, 0.95)	-0.26 (-0.27, -0.26)	0.96 (0.96, 0.97)	-0.14 (-0.14, -0.13)
single-CISRB	0.83 (0.82, 0.85)	-0.37 (-0.38, -0.36)	0.94 (0.94, 0.95)	-0.26 (-0.27, -0.26)	**0.97 (0.96, 0.97)**	-0.13 (-0.13, -0.12)
SNIS	**0.84 (0.82, 0.85)**	-0.37 (-0.38, -0.36)	0.94 (0.94, 0.94)	**-0.25 (-0.26, -0.25)**	0.96 (0.96, 0.97)	-0.12 (-0.12, -0.12)

* LPD denotes the average log predictive density.
* The numbers in the parentheses denote the 80% bootstrap confidence intervals computed from 100 repetitions.

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 predictive density (Test LPD) results are shown in Table 2. Our proposed parallel state estimator (par.-IMH) achieves the best accuracy and predictive density results. Despite having access to high-quality HMC samples, single-HMC shows poor performance. This supports our analysis that par.-IMH with $N > 1$ superior variance reduction to the single state estimator. Also, seq.-IMH showed poor performance overall due to the correlated samples. Among the two CIS kernel-based methods, single-CISRB performs only marginally better than single-CIS.



(a) Test Accuracy (b) Test LPD

Figure 3: Test accuracy and log predictive density on the `german` dataset. The solid lines and colored regions are the mean and 80% bootstrap confidence interval computed from 100 repetitions.

Inclusive KL v.s. Exclusive KL While both ELBO and par.-IMH showed similar numerical performance, they chose different optimization paths in the parameter space. This is shown in Figure 3. While the test accuracy suggests that ELBO converges quickly around $t = 2000$ (Figure 3a), in terms of uncertainty estimate, it takes much longer to converge (Figure 3b). This shows that inclusive KL minimization chooses a path that has better density coverage as expected.

4.3 GAUSSIAN PROCESS CLASSIFICATION

Experimental Setup For a more challenging problem, we perform classification with latent Gaussian processes [Nguyen and Bonilla, 2014]. The simplified probabilistic model is

$$\begin{aligned} \log \theta &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ f &\sim \mathcal{GP}(\mathbf{0}, \mathbf{\Sigma}_\theta) \\ y_i &\sim \text{Bernoulli-Logit}(f(\mathbf{x}_i)) \end{aligned}$$

where we chose a Matérn 5/2 covariance kernel with automatic relevance determination [Neal, 1996]. For the datasets, we use the `sonar` ($\mathbf{z} \in \mathbb{R}^{249}$, Gorman and Sejnowski 1988), `ionosphere` ($\mathbf{z} \in \mathbb{R}^{351}$, Sigillito et al. 1989), and `breast` ($\mathbf{z} \in \mathbb{R}^{544}$, Wolberg and Mangasarian 1990) datasets. For `breast`, we preprocessed the input features with z-standardization. 10% of the data points were randomly selected in each of the 100 repetitions as test data. For this experiment, the iteration complexity of ELBO is almost two orders of magnitude larger than all inclusive KL minimization methods.

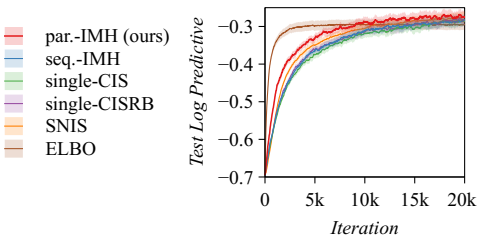


Figure 4: Test log predictive density on the `ionosphere` dataset. The solid lines and colored regions are the medians and 80% percentiles computed from 100 repetitions.

Result The results are shown in Table 3. Again, among inclusive KL minimization, the parallel state estimator (par.-IMH) achieved the best results. Compared to ELBO, its accuracy was lower on `breast`, but the uncertainty estimates were much better. This is better shown in Figure 4, where ELBO quickly converges to a point with poor uncertainty calibration.

4.4 MARGINAL LIKELIHOOD ESTIMATION

Experimental Setup Lastly, we now estimate the marginal log-likelihood of a hierarchical regression model with partial pooling (`radon`, $\mathbf{z} \in \mathbb{R}^{175}$, Gelman and Hill 2007) for modeling radon levels in U.S homes. `radon` contains multiple posterior degeneracies from the hierarchy. We estimated the reference marginal likelihood using *thermodynamic integration* (TI, Gelman and Meng 1998, Neal 2001, Lartillot and Philippe 2006) with HMC implemented by Stan [Carpenter et al., 2017, Betancourt, 2017].

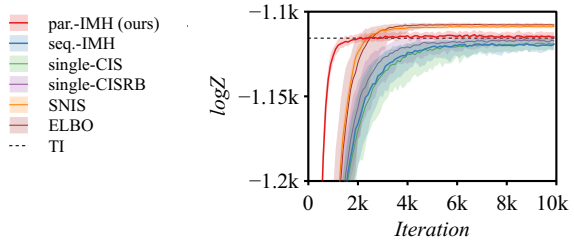


Figure 5: Marginal log-likelihood estimates on the `radon` dataset. The solid lines and colored regions are the medians and 80% percentiles computed from 100 repetitions.

Results The results are shown in Figure 5. par.-IMH converges quickly and provides the most accurate estimate. By contrast, other estimators 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.

5 RELATED WORKS

Inclusive KL minimization Our method directly builds on top of MSC [Naesseth et al., 2020], which minimizes the inclusive KL divergence. Concurrently, in the context of variational autoencoders with discrete latent variables, Ou and Song [2020] proposed JSA. JSA can be viewed as a variant of MSC that takes advantage of models with *i.i.d.* data likelihoods. Meanwhile, Li et al. [2017] used a similar approach to MSC, but their method is slow to achieve stationarity. Other than using MCMC, Bornschein and Bengio [2015], Le et al. [2019] used SNIS while Wu et al. [2020] used sequential Monte Carlo (SMC) for estimating the score, but these methods are restricted to deep generative models. Further away, Jerfel et al. [2021] used boosting instead of SGD to minimize the inclusive KL, which results in a more flexible variational family.

MCMC for VI, VI for MCMC Not restricted to inclusive KL minimization, MCMC has been widely utilized in VI. For example, Salimans et al. [2015], Ruiz and Titsias [2019] construct alternative divergence bounds from samples of an MCMC sampler. More recently, several other methods that apply variational inference for adapting the MCMC kernel have been developed. For adapting the proposals of an IMH sampler, Habib and Barber [2019] minimize the exclusive KL divergence while Neklyudov et al. [2019] minimize the symmetric KL divergence. And for HMC, Zhang et al. [2018], Campbell et al. [2021] have proposed to use score matching, ELBO maximization, and Stein discrepancy minimization.

Adaptive MCMC As pointed out by Ou and Song [2020], using q_λ within the MCMC kernel makes score climbing structurally equivalent to adaptive MCMC. In particular, Andrieu and Thoms [2008], Garthwaite et al. [2016] discuss the use of stochastic approximation in adaptive MCMC. Also, Andrieu and Moulines [2006], Keith et al. [2008], Holden et al. [2009], Giordani and Kohn [2010] specifically discuss adapting the propsosal of IMH kernels. Most similar to score climbing VI is the work of Keith et al. [2008] where they propose to use cross-entropy minimization [Barbakh et al., 2009], which is mathematically identical to inclusive VI.

6 DISCUSSIONS

In this paper, we presented the parallel state estimator for inclusive KL divergence minimization. Compared to previously proposed estimation schemes, our estimator enjoys substantially low variance. We also showed that the parallel state estimator does not result in higher bias in general. We demonstrated empirical evidence of our analysis on Bayesian inference problems.

In our results, minimizing the inclusive KL divergence showed to be competitive against exclusive KL divergence minimization. This is against the conclusions of Dhaka et al. [2021] that the inclusive KL does not work in high-dimensional problems (Dhaka et al. consider few hundreds of dimensions). Theoretically, dimensionality becomes a real challenge when the posterior has complex nonlinear correlations. A practical question would be how correlated our posteriors really are in practice. Also, many of the benefits of alternative divergences are shadowed by the challenges in our inference algorithms. Therefore, our results motivate the development of better inference algorithms for alternative divergence measures, including the inclusive KL.

References

C. Andrieu and É. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16(3), Aug. 2006.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Stat. Comput.*, 18(4):343–373, Dec. 2008.

C. Andrieu, A. Lee, and M. Vihola. Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2), May 2018.

W. A. Barbakh, Y. Wu, and C. Fyfe. *Cross Entropy Methods*, volume 249, pages 151–174. Berlin, Heidelberg, 2009.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *ArXiv170102434 Stat*, Jan. 2017.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *JASA*, 112 (518):859–877, Apr. 2017.

J. Bornschein and Y. Bengio. Reweighted wake-sleep. In *ICLR*, San Diego, California, USA, May 2015.

L. Bottou. On-line learning and stochastic approximations. In *On-Line Learning in Neural Networks*, pages 9–42. First edition, Jan. 1999.

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

A. Campbell, W. Chen, V. Stimper, J. M. Hernandez-Lobato, and Y. Zhang. A gradient based strategy for Hamiltonian Monte Carlo hyperparameter optimization. In *ICML*, volume 139, pages 1238–1248, July 2021.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *J. Stat. Soft.*, 76(1), 2017.

K. S. Chan and C. J. Geyer. Discussion: Markov chains for exploring posterior distributions. *Ann. Stat.*, 22(4): 1747–1758, 1994.

R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *Amer. J. Cardiology*, 64(5):304–310, Aug. 1989.

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

A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. Blei. Variational inference via χ upper bound minimization. In *NIPS*, volume 30, pages 2729–2738, Long Beach, California, USA, 2017.

D. Dua and C. Graff. UCI machine learning repository. 2017.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216–222, 1987.

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, Sept. 2016.

H. Ge, K. Xu, and Z. Ghahramani. Turing: A language for flexible probabilistic inference. In *ICML*, volume 84, pages 1682–1690, 2018.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge; New York, 2007.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.*, 13(2), May 1998.

P. Giordani and R. Kohn. Adaptive independent Metropolis–Hastings by fast estimation of mixtures of normals. *JCGS*, 19(2):243–259, Jan. 2010.

R. Gorman and T. J. Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks*, 1(1):75–89, Jan. 1988.

M. G. Gu and F. H. Kong. A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *PNAS*, 95(13): 7270–7274, June 1998.

R. Habib and D. Barber. Auxiliary variational MCMC. In *ICLR*, 2019.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 (1):97–109, Apr. 1970.

M. D. Hoffman. Learning deep latent Gaussian models with Markov chain Monte Carlo. In *ICML*, volume 70, pages 1510–1519, Aug. 2017.

L. Holden, R. Hauge, and M. Holden. Adaptive independent Metropolis–Hastings. *Ann. Appl. Probab.*, 19(1), Feb. 2009.

G. Jerfel, S. Wang, C. Wong-Fannjiang, K. A. Heller, Y. Ma, and M. I. Jordan. Variational refinement for importance sampling using the forward Kullback-Leibler divergence. In *UAI*, volume 161, pages 1819–1829, July 2021.

Y. H. Jiang, T. Liu, Z. Lou, J. S. Rosenthal, S. Shangguan, F. Wang, and Z. Wu. MCMC confidence intervals and biases. *ArXiv201202816 Math Stat*, June 2021.

J. M. Keith, D. P. Kroese, and G. Y. Sofronov. Adaptive independence samplers. *Stat Comput*, 18(4):409–420, Dec. 2008.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, San Diego, California, USA, 2015.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *JMLR*, 18(14):1–45, 2017.

N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55(2):195–207, Apr. 2006.

T. A. Le, A. R. Kosiorek, N. Siddharth, Y. W. Teh, and F. Wood. Revisiting reweighted wake-sleep for models with stochastic control flow. In *UAI*, Tel Aviv, Israel, July 2019.

Y. Li and R. E. Turner. Rényi divergence variational inference. In *NIPS*, volume 29, 2016.

Y. Li, R. E. Turner, and Q. Liu. Approximate inference with amortised MCMC. *ArXiv170208343 Cs Stat*, May 2017.

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

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

T. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Jan. 2005.

C. Naesseth, F. Lindsten, and D. Blei. Markovian score climbing: Variational inference with KL(p||q). In *NeurIPS*, volume 33, pages 15499–15510, 2020.

R. M. Neal. *Bayesian Learning for Neural Networks*, volume 118. New York, NY, 1996.

R. M. Neal. Annealed importance sampling. *Stat. Comput.*, 11(2):125–139, 2001.

R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. First edition, 2011.

K. Neklyudov, E. Egorov, P. Shvechikov, and D. Vetrov. Metropolis-Hastings view on variational inference and adversarial training. *ArXiv181007151 Cs Stat*, June 2019.

T. V. Nguyen and E. V. Bonilla. Automated variational inference for Gaussian process models. In *NIPS*, volume 27, 2014.

Z. Ou and Y. Song. Joint stochastic approximation and its application to learning discrete latent variable models. In *UAI*, volume 124, pages 929–938, Aug. 2020.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *AISTATS*, volume 33, pages 814–822, Reykjavik, Iceland, Apr. 2014.

H. Robbins and S. Monro. A Stochastic Approximation Method. *Ann. Math. Statist.*, 22(3):400–407, Sept. 1951.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. New York, NY, 2004.

G. Roeder, Y. Wu, and D. K. Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *NIPS*, volume 30, 2017.

F. Ruiz and M. Titsias. A contrastive divergence for combining variational inference and MCMC. In *ICML*, volume 97, pages 5537–5545, June 2019.

T. Salimans, D. Kingma, and M. Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *ICML*, volume 37, pages 1218–1226, Lille, France, July 2015.

V. G. Sigillito, S. Wing, L. V. Hutton, and K. L. Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Tech. Dig.*, 10: 262–266, 1989.

J. 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 *Proc. Annu. Symp. Comput. Appl. Med. Care*, pages 261–265, Nov. 1988.

R. L. Smith and L. Tierney. Exact transition probabilities for the independence metropolis sampler. Technical report, 1996.

Z. Tan. Monte Carlo integration with acceptance-rejection. *JCGS*, 15(3):735–752, Sept. 2006.

D. Wang, H. Liu, and Q. Liu. Variational inference with tail-adaptive f-Divergence. In *NIPS*, volume 31, 2018.

G. Wang. Exact convergence rate analysis of the independent Metropolis-Hastings algorithms. *ArXiv200802455 Math Stat*, Dec. 2020.

W. H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *PNAS*, 87(23):9193–9196, Dec. 1990.

H. Wu, H. Zimmermann, E. Sennesh, T. A. Le, and J.-W. Van De Meent. Amortized population Gibbs samplers with neural sufficient statistics. In *ICML*, volume 119, pages 10421–10431, July 2020.

C. Zhang, B. Shahbaba, and H. Zhao. Variational Hamiltonian Monte Carlo via score matching. *Bayesian Anal.*, 13(2), June 2018.

Markov-Chain Monte Carlo Score Estimators for Variational Inference with Score Climbing
Appendix

A COMPUTATIONAL RESOURCES

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 12 hours on a system with similar computational capabilities.

B PSEUDOCODES OF THE CONSIDERED SCHEMES

Algorithm 1: Score Climbing with the Single State Estimator

Input: MCMC kernel $K_{\boldsymbol{\lambda}}(\mathbf{z}, \cdot)$, 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**
- $\mathbf{z}_t \sim K_{\boldsymbol{\lambda}_{t-1}}(\mathbf{z}_{t-1}, \cdot)$
- $s(\boldsymbol{\lambda}; \mathbf{z}) = \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\lambda}}(\mathbf{z})$
- $g_{\text{single}} = s(\boldsymbol{\lambda}_{t-1}; \mathbf{z}_t)$
- $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \gamma_t \, g_{\text{single}}$

end

Algorithm 2: Score Climbing with the Sequential State Estimator

Input: MCMC kernel $K_{\boldsymbol{\lambda}}(\mathbf{z}, \cdot)$, 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**
- $\tau = N(t-1)$
- **for** $i = 1, 2, \ldots, N$ **do**
  - $\mathbf{z}_t \sim K_{\boldsymbol{\lambda}_{t-1}}(\mathbf{z}_{\tau+i}, \cdot)$
- **end**
- $s(\boldsymbol{\lambda}; \mathbf{z}) = \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\lambda}}(\mathbf{z})$
- $g_{\text{seq.}} = \frac{1}{N} \sum_{i=1}^{N} s(\boldsymbol{\lambda}_{t-1}; \mathbf{z}_{T+i})$
- $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \gamma_t \, g_{\text{seq.}}$

end

Algorithm 3: Score Climbing with the Parallel State Estimator

Input: MCMC kernel $K_{\boldsymbol{\lambda}}(\mathbf{z}, \cdot)$, 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**
  - $\mathbf{z}_t^{(i)} \sim K_{\boldsymbol{\lambda}_{t-1}}(\mathbf{z}_{t-1}^{(i)}, \cdot)$
- **end**
- $s(\boldsymbol{\lambda}; \mathbf{z}) = \nabla_{\boldsymbol{\lambda}} \log q_{\boldsymbol{\lambda}}(\mathbf{z})$
- $g_{\text{par.}} = \frac{1}{N} \sum_{i=1}^{N} s(\boldsymbol{\lambda}_{t-1}; \mathbf{z}_t^{(i)})$
- $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_{t-1} + \gamma_t \, g_{\text{par.}}$

end

Algorithm 4: Conditional Importance Sampling Kernel

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

$\mathbf{z}^{(0)} = \mathbf{z}_{t-1}$

$\mathbf{z}^{(i)} \sim q_{\lambda_{t-1}}(\mathbf{z})$ for $i = 1, 2, \ldots, N$

$w(\mathbf{z}^{(i)}) = p(\mathbf{z}^{(i)}, \mathbf{x}) \,/\, q_{\lambda_{t-1}}(\mathbf{z}^{(i)})$ for $i = 0, 1, \ldots, N$

$\widetilde{w}^{(i)} = w(\mathbf{z}^{(i)}) / \sum_{i=0}^{N} w(\mathbf{z}^{(i)})$ for $i = 0, 1, \ldots, N$

$\mathbf{z}_t \sim \text{Multinomial}(\widetilde{w}^{(0)}, \widetilde{w}^{(1)}, \ldots, \widetilde{w}^{(N)})$

Algorithm 5: Independent Metropolis-Hastings Kernel

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

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

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

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

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

if $u < \alpha$ **then**

  $\mathbf{z}_t = \mathbf{z}^*$

else

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

end

C PROBABILISTIC MODELS CONSIDERED IN ??

C.1 HIERARCHICAL LOGISTIC REGRESSION

The hierarchical logistic regression used in Section 4.2 is

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

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

C.2 GAUSSIAN PROCESS LOGISTIC REGRESSION

The latent Gaussian process model used in Section 4.3 is

$$\begin{aligned}
\log \alpha &\sim \mathcal{N}(0, 1) \\
\log \sigma &\sim \mathcal{N}(0, 1) \\
\log \ell_i &\sim \mathcal{N}(0, 1) \\
f &\sim \mathcal{GP}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\alpha^2, \sigma^2, \ell} + \delta\, \mathbf{I}\right) \\
y_i &\sim \text{Bernoulli-Logit}\left(f\left(\mathbf{x}_i\right)\right).
\end{aligned}$$

The covariance $\boldsymbol{\Sigma}$ is computed using a kernel $k(\cdot, \cdot)$ such that $[\boldsymbol{\Sigma}]_{i,j} = k\left(\mathbf{x}_i, \mathbf{x}_j\right)$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are data points in the dataset. For the kernel, we use the Matern 5/2 kernel with automatic relevance determination [Neal, 1996] defined as

$$k\left(\mathbf{x}, \mathbf{x}';\ \alpha^2, \sigma^2, \ell\right) = \alpha\left(1 + \sqrt{5}r + \frac{5}{3}r^2\right)\exp\left(-\sqrt{5}r\right) \quad \text{where} \quad r = \sum_{i=1}^{D} \frac{\left(\mathbf{x}_i - \mathbf{x}_i'\right)^2}{\ell_i^2}$$

where D is the number of dimensions. The jitter term δ is used for numerical stability. We set a small value of $\delta = 1 \times 10^{-6}$.

D PROOFS

Lemma 1. *For* $w^* = \sup_{\mathbf{z}} w(\mathbf{z})$,

$$T_t(w(\mathbf{z})) \leq \frac{t}{w}\left(1 - \frac{1}{w^*}\right)^{t-1}.$$

Proof. As stated by Smith and Tierney [1996], the essential supremum of $\lambda(w)$ is $1 - \frac{1}{w^*}$. Then, an upper bound of $T_t(w)$ follows as

$$\begin{aligned}
T_t(w) &= \int_w^\infty \frac{t}{v^2}\,\lambda^{t-1}(v)\,dv && (7)\\
&\leq \int_w^\infty \frac{t}{v^2}\left(1 - \frac{1}{w^*}\right)^{t-1} dv && (8)\\
&= t\left(1 - \frac{1}{w^*}\right)^{t-1}\int_w^\infty \frac{1}{v^2}\,dv && (9)\\
&= t\left(1 - \frac{1}{w^*}\right)^{t-1} - \frac{1}{v}\Big|_w^\infty && (10)\\
&= \frac{t}{w}\left(1 - \frac{1}{w^*}\right)^{t-1}. && (11)
\end{aligned}$$

□

Theorem 1. *(Smith and Tierney 1996, Theorem 1) The t-step marginal IMH transition kernel is given as*

$$K^t(\mathbf{z}, d\mathbf{z}^*) = T_t\left(w(\mathbf{z}) \vee w(\mathbf{z}^*)\right)\pi(\mathbf{z}^*)\,d\mathbf{z}^* + \lambda^t(w(\mathbf{z}))\,\delta_{\mathbf{z}}(d\mathbf{z}^*)$$

where $x \vee y = \max(x, y)$, *and for* $R(w) = \{\mathbf{z}^* \mid w(\mathbf{z}^*) \leq w\}$,

$$T_t(w) = \int_w^\infty \frac{t}{v^2}\,\lambda^{t-1}(v)\,dv, \quad \lambda(w) = \int_{R(w)}\left(1 - \frac{w(\mathbf{z}^*)}{w}\right)\pi(d\mathbf{z}^*).$$

Theorem 2. *Assuming that the Markov-chains have achieved stationarity and* $N \geq 2$, *the variance of the single state estimator (*$\mathbf{g}_{single}$*), sequential state estimator with N states of the IMH kernel (*$\mathbf{g}_{seq.\text{-}IMH}$*), and parallel state estimator (*$\mathbf{g}_{par.}$*) with N chains are given as*

$$\mathbb{V}\left[\,\mathbf{g}_{single}\,\right] = \sigma^2, \quad \mathbb{V}\left[\,\mathbf{g}_{seq.\text{-}IMH}\,\right] = \frac{\sigma^2}{N} + C_{gap}, \quad \mathbb{V}\left[\,\mathbf{g}_{par.}\,\right] = \frac{\sigma^2}{N}$$

where $\sigma^2 = \mathbb{V}_\pi\left[\,\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z})\,\right]$ *and* $C_{gap} \geq 0$. *The variance gap* C_{gap} *is bounded below as*

$$\begin{aligned}
C_{gap} \geq \frac{2}{N}\Big[\, 2\sigma^2 + C\big\{\,\exp\left(D_{\mathrm{KL}}(q \parallel \pi)\right) - 3\big\}\\
+ Cov_q\left(\Delta^2(\mathbf{z}),\, 1/w(\mathbf{z})\right)\Big] + \mathcal{O}\left(N^{-2}\right)
\end{aligned}$$

where q *is the proposal distribution,* C *is a positive constant,* $\Delta(\mathbf{z}) = \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) - \mathbb{E}_\pi\left[\,\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z})\,\right]$, *and, as* $N \to \infty$, *the bound converges to*

$$C_{gap} \geq \frac{2}{N}\Big[\,\sigma^2\big\{\,\exp\left(D_{\mathrm{KL}}(\pi \parallel q)\right) - 1\big\} + Cov_\pi\left(\Delta^2(\mathbf{z}),\, w(\mathbf{z})\right)\Big].$$

Proof of ***Theorem*** *2.* ***Variance of the Single State Estimator*** For the single state estimator,

$$\begin{aligned}
\mathbb{V}\left[\,\mathbf{g}_{\text{single}}(\boldsymbol{\lambda})\,\right] = \mathbb{E}\Big[\,\mathbb{V}_{K_{\lambda_{t-1}}(\mathbf{z}_{t-1}, \mathbf{z})}\left[\,\mathbf{s}(\mathbf{z}; \boldsymbol{\lambda}) \mid \mathbf{z}_{t-1}\,\right]\Big]\\
+ \mathbb{V}\Big[\,\mathbb{E}_{K_{\lambda_{t-1}}(\mathbf{z}_{t-1}, \mathbf{z})}\left[\,\mathbf{s}(\mathbf{z}; \boldsymbol{\lambda}) \mid \mathbf{z}_{t-1}\,\right]\Big], \qquad (12)
\end{aligned}$$

and assuming stationarity such that $\mathbf{z}_{t-1} \sim \pi(\mathbf{z})$,

$$= \mathbb{E}_\pi\Big[\,\mathbb{V}_{K_{\lambda_{t-1}}(\mathbf{z}_{t-1}, \mathbf{z})}\left[\,\mathbf{s}(\mathbf{z}; \boldsymbol{\lambda}) \mid \mathbf{z}_{t-1}\,\right]\Big]$$

$$+ \mathbb{V}_\pi \left[\mathbb{E}_{K_{\lambda_{t-1}}(\mathbf{z}_{t-1}, \mathbf{z})} \left[\mathbf{s}(\mathbf{z}; \boldsymbol{\lambda}) \mid \mathbf{z}_{t-1} \right] \right] \tag{13}$$

$$= \mathbb{V}_\pi \left[\mathbf{s}(\mathbf{z}; \boldsymbol{\lambda}) \right] \tag{14}$$

$$= \sigma^2. \tag{15}$$

Variance of the Parallel State Estimator Meanwhile, the parallel state estimator is an average of idenpendent and identically distributed (*i.i.d.*) since the parallel chains are independent. Therefore, from Equation (15), its variance is

$$\mathbb{V}\left[\mathbf{g}_{\text{par.}} \right] = \frac{\sigma^2}{N}. \tag{16}$$

Variance of the Sequential State Estimator Now, for the single state estimator, we first derive an MCMC kernel independent expression. First, remember that the estimator is defined as

$$\mathbf{g}_{\text{seq.}}(\boldsymbol{\lambda}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}), \tag{17}$$

where $\mathbf{z}_{T+i} \sim K^i_{\lambda_{t-1}}(\mathbf{z}_T, \cdot)$ and $\mathbf{z}_T$ is the last Markov-chain state at the previous SGD iteration $t-1$. Then, the variance is given as

$$\mathbb{V}\left[\mathbf{g}_{\text{seq.}} \right] = \mathbb{V}\left[\mathbb{E}_{K(\mathbf{z}_T, \mathbf{z})} \left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}) \,\middle|\, \mathbf{z}_T \right] \right] + \mathbb{E}\left[\mathbb{V}_{K(\mathbf{z}_T, \mathbf{z})} \left[\frac{1}{N} \sum_{i=1}^{N} \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}) \,\middle|\, \mathbf{z}_T \right] \right] \qquad \text{(Total Variance)} \tag{18}$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{V}_{K(\mathbf{z}_T, \mathbf{z})} \left[\mathbb{E}\left[\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}) \mid \mathbf{z}_T \right] \right] + \mathbb{E}\left[\frac{1}{N^2} \sum_{i=1}^{N} \mathbb{V}_{K(\mathbf{z}_T, \mathbf{z})} \left[\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}) \mid \mathbf{z}_T \right] + \frac{2}{N^2} \sum_{i<j}^{N} \operatorname{Cov}\left(s(\boldsymbol{\lambda}; \mathbf{z}_{T+i}), \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}_{T+j}\right) \mid \mathbf{z}_T \right) \right] \tag{19}$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{V}_{K(\mathbf{z}_T, \mathbf{z})} \left[\mathbb{E}\left[\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}) \mid \mathbf{z}_T \right] \right] + \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}_{K(\mathbf{z}_T, \mathbf{z})} \left[\mathbb{V}\left[\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}) \mid \mathbf{z}_T \right] \right] + \frac{2}{N^2} \sum_{i<j} \mathbb{E}\left[\operatorname{Cov}\left(\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}), \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}_{T+j}\right) \mid \mathbf{z}_T \right) \right], \tag{20}$$

where, by assuming stationarity such that $\mathbf{z}_T \sim \pi(\mathbf{z})$,

$$= \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{V}_\pi \left[\mathbb{E}\left[\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}) \mid \mathbf{z}_T \right] \right] + \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}_\pi \left[\mathbb{V}\left[\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}) \mid \mathbf{z}_T \right] \right] + \frac{2}{N^2} \sum_{i<j} \mathbb{E}_\pi \left[\operatorname{Cov}\left(\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}), \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}_{T+j}\right) \mid \mathbf{z}_T \right) \right] \tag{21}$$

$$= \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{V}_\pi \left[\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \right] + \frac{2}{N^2} \sum_{i<j} \mathbb{E}_\pi \left[\operatorname{Cov}\left(\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}), \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}_{T+j}\right) \mid \mathbf{z}_T \right) \right] \qquad \text{(Total Variance)} \tag{22}$$

$$= \frac{1}{N} \mathbb{V}_\pi \left[\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \right] + \frac{2}{N^2} \sum_{i<j} \mathbb{E}_\pi \left[\operatorname{Cov}\left(\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}), \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}_{T+j}\right) \mid \mathbf{z}_T \right) \right] \tag{23}$$

$$= \frac{\sigma^2}{N} + \frac{2}{N^2} \sum_{i<j}^{N} \mathbb{E}_\pi \left[\operatorname{Cov}\left(\mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}_{T+i}), \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}_{T+j}\right) \mid \mathbf{z}_T \right) \right] \tag{24}$$

$$= \frac{\sigma^2}{N} + \frac{2}{N}\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)\gamma_k \tag{25}$$

where γ_k is the k-lag autocovariance.

Specifically for the IMH kernel, Tan [2006] has analyzed γ_k based on the results of Smith and Tierney [1996]. We extend his analysis to obtain our desired conclusion. Let us denote γ_k as the covariance between a state $\mathbf{z}$ and its kth lagged counterpart $\mathbf{z}_k$, and $\Delta(\mathbf{z}) = \mathbf{s}(\mathbf{z}) - \mathbb{E}_\pi[\mathbf{s}]$. Then,

$$\gamma_k = \int \Delta(\mathbf{z})\left(\int K^k(\mathbf{z},\mathbf{z}_k)\,\Delta(\mathbf{z}_k)\,d\mathbf{z}_k\right)\pi(d\mathbf{z}) \tag{26}$$

$$= \int \Delta(\mathbf{z})\left(\int \Delta(\mathbf{z}_k)\,T_k(w(\mathbf{z})\vee w(\mathbf{z}_k))\,\pi(d\mathbf{z}_k) + \Delta(\mathbf{z})\,\lambda^k(w(\mathbf{z}))\right)\pi(d\mathbf{z}) \tag{27}$$

$$= \int\int \Delta(\mathbf{z})\,\Delta(\mathbf{z}_k)\,T_k(w(\mathbf{z})\vee w(\mathbf{z}_k))\,\pi(d\mathbf{z}_k)\,\pi(d\mathbf{z}) + \int \Delta^2(\mathbf{z})\,\lambda^k(w(\mathbf{z}))\,\pi(d\mathbf{z}). \tag{28}$$

For the first term, Tan [2006, Theorem 3] have shown that

$$\int\int \Delta(\mathbf{z})\,\Delta(\mathbf{z}_k)\,T_k(w(\mathbf{z})\vee w(\mathbf{z}_k))\,\pi(d\mathbf{z}_k)\,\pi(d\mathbf{z}) \geq 0. \tag{29}$$

And for the second term,

$$\frac{2}{N}\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)\gamma_k \geq \frac{2}{N}\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)\int \Delta^2(\mathbf{z})\,\lambda^k(w(\mathbf{z}))\,\pi(d\mathbf{z}) \tag{30}$$

$$= \frac{2}{N}\int \Delta^2(\mathbf{z})\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)\lambda^k(w(\mathbf{z}))\,\pi(d\mathbf{z}) \tag{31}$$

$$\geq 0, \tag{32}$$

which proves $\mathbb{V}[\mathbf{g}_{\text{seq.-IMH}}] - \mathbb{V}[\mathbf{g}_{\text{par.}}] = C_{\text{gap}} \geq 0$.

Analyzing the variance gap C_{gap} is a little more involved. Usually, traditional MCMC analysis invokes Cesàro's summability theorem for the sum in Equation (31) by assuming $N \to \infty$ [Chan and Geyer, 1994]. In our case, we avoid this path in order to generalize our result to the small N regime. For clarity, we denote $r = \lambda(w(\mathbf{z}))$. Then,

$$\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)r^k = \sum_{k=1}^{N-1} r^k - \frac{1}{N}\sum_{k=1}^{N-1} k\,r^k \tag{33}$$

$$= \frac{r\left(1-r^{N-1}\right)}{1-r} - \frac{r}{N}\,\frac{1-N\,r^{N-1}+(N-1)\,r^N}{(1-r)^2} \tag{34}$$

$$= \frac{r\left(r^N - N\,r + N - 1\right)}{N(1-r)^2} \tag{35}$$

$$= \frac{r\left(N(1-r)-\left(1-r^N\right)\right)}{N(1-r)^2} \tag{36}$$

$$= \frac{r}{1-r} - \frac{r}{N}\,\frac{\left(1-r^N\right)}{(1-r)^2}, \tag{37}$$

which is monotonically increasing with respect to r.

There are several ways to analyze the behavior of this function. For example, the upper bound

$$\frac{r}{1-r} - \frac{r}{N}\,\frac{\left(1-r^N\right)}{(1-r)^2} \leq \frac{r}{1-r} \tag{38}$$

becomes a good approximation for large N. For small N, it is still accurate for small values of r. In the limit $N \to \infty$, the upper bound becomes exact where we retrieve the result of Tan [2006, Theorem 3].

Lower Bound of C_{gap} On the other hand, it is also possible to derive a lower bound using the formula of geometric sums as

$$-\frac{\left(1-r^N\right)}{(1-r)} = -\frac{(1-r)\left(r^{N-1}+r^{N-2}+\ldots+r+1\right)}{(1-r)} \tag{39}$$

$$= -\left(r^{N-1}+r^{N-2}+\ldots+r^{N-2}+r+1\right) \tag{40}$$

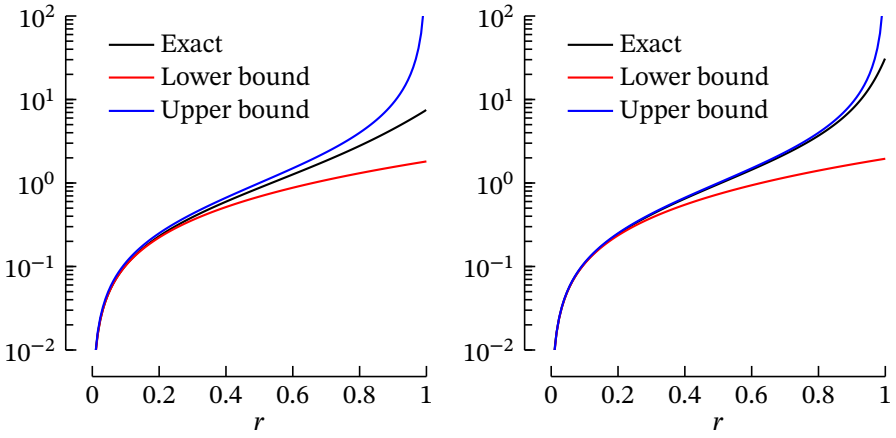
$$\geq -\left(r^2+r^2+\ldots+r^2+r+1\right) \tag{41}$$

$$= -\left((N-2)\,r^2+r+1\right) \tag{42}$$

where we have used the fact that $0 \leq r \leq 1$ and $N \geq 2$. By applying this to Equation (37),

$$\frac{r}{1-r} - \frac{r}{N}\frac{\left(1-r^N\right)}{(1-r)^2} \geq \frac{r}{1-r} - \frac{r}{N}\frac{(N-2)\,r^2+r+1}{(1-r)} \tag{43}$$

$$= \left(1-\frac{1}{N}\right)\frac{r}{1-r} - \frac{1}{N}\frac{r^2}{1-r} - \frac{N-2}{N}\frac{r^3}{1-r}. \tag{44}$$



(a) $N = 16$ (b) $N = 64$

Figure 6: Comparison of Equation (37) against its lower and upper bounds.

.

The quality of the lower and upper bounds are visualized in Figure 6. We can see that the lower bound is quite optimistic in general, but is more accurate for the small N regime.

Now, let us develop our lower bound. Recall that $r = \lambda(w(\mathbf{z}))$. Tan [2006, Proof of Theorem 3] have shown that $1-\frac{1}{u} \leq \lambda(u) \leq 1-\frac{1}{w^*}$. Since the lower bound in Equation (44) is monotonically increasing, applying the bound $r \geq 1-\frac{1}{w(\mathbf{z})}$ results in

$$\begin{aligned}&\left(1-\frac{1}{N}\right)\frac{r}{1-r} - \frac{1}{N}\frac{r^2}{1-r} - \left(1-\frac{2}{N}\right)\frac{r^3}{1-r} \\ &\geq \left(1-\frac{1}{N}\right)w(\mathbf{z})\left(1-\frac{1}{w(\mathbf{z})}\right) - \frac{1}{N}w(\mathbf{z})\left(1-\frac{1}{w(\mathbf{z})}\right)^2 - \left(1-\frac{2}{N}\right)w(\mathbf{z})\left(1-\frac{1}{w(\mathbf{z})}\right)^3.\end{aligned} \tag{45}$$

For clarity, we temporarily set $\alpha = 1-\frac{1}{N}$, $\beta = \frac{1}{N}$, $\gamma = 1-\frac{2}{N}$. Then, Equation (45) becomes

$$\alpha\, w(\mathbf{z})\left(1-\frac{1}{w(\mathbf{z})}\right) - \beta\, w(\mathbf{z})\left(1-\frac{1}{w(\mathbf{z})}\right)^2 - \gamma\, w(\mathbf{z})\left(1-\frac{1}{w(\mathbf{z})}\right)^3$$

$$= \alpha\,(w(\mathbf{z})-1) - \beta\left(w(\mathbf{z}) - 2 + \frac{1}{w(\mathbf{z})}\right) - \gamma\left(w(\mathbf{z}) - 3 + \frac{3}{w(\mathbf{z})} - \frac{1}{w^2(\mathbf{z})}\right) \tag{46}$$

$$= (\alpha - \beta - \gamma)\; w(\mathbf{z}) + (-\alpha + 2\beta + 3\gamma) + (-\beta - 3\gamma)\,\frac{1}{w(\mathbf{z})} + \gamma\frac{1}{w^2(\mathbf{z})} \tag{47}$$

$$= \left(2 - \frac{3}{N}\right) + \left(-3 + \frac{5}{N}\right)\frac{1}{w(\mathbf{z})} + \left(1 - \frac{2}{N}\right)\frac{1}{w^2(\mathbf{z})}. \tag{48}$$

Applying this to Equation (31),

$$\frac{2}{N}\int \Delta^2(\mathbf{z}) \sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\lambda^k(w(\mathbf{z}))\,\pi(d\mathbf{z}) \tag{49}$$

$$\geq \frac{2}{N}\int \Delta^2(\mathbf{z})\left\{\left(2 - \frac{3}{N}\right) + \left(-3 + \frac{5}{N}\right)\frac{1}{w(\mathbf{z})} + \left(1 - \frac{2}{N}\right)\frac{1}{w^2(\mathbf{z})}\right\}\pi(d\mathbf{z}) \tag{50}$$

$$= \frac{2}{N}\left(2 - \frac{3}{N}\right)\sigma^2 + \frac{2}{N}\left(-3 + \frac{5}{N}\right)\int \Delta^2(\mathbf{z})\,\frac{1}{w(\mathbf{z})}\,\pi(d\mathbf{z}) + \frac{2}{N}\left(1 - \frac{2}{N}\right)\int \Delta^2(\mathbf{z})\,\frac{1}{w^2(\mathbf{z})}\,\pi(d\mathbf{z}) \tag{51}$$

$$= \frac{2}{N}\left(2 - \frac{3}{N}\right)\sigma^2 + \frac{2}{N}\left(-3 + \frac{5}{N}\right)\int \Delta^2(\mathbf{z})\,q(\mathbf{z})\,d\mathbf{z} + \frac{2}{N}\left(1 - \frac{2}{N}\right)\int \Delta^2(\mathbf{z})\,\frac{q^2(\mathbf{z})}{\pi(\mathbf{z})}\,d\mathbf{z}. \tag{52}$$

For the last term,

$$\int \Delta^2(\mathbf{z})\,\frac{q^2(\mathbf{z})}{\pi(\mathbf{z})}\,d\mathbf{z} = \mathbb{E}_q\left[\,\Delta^2(\mathbf{z})\,\frac{q(\mathbf{z})}{\pi(\mathbf{z})}\,\right] \tag{53}$$

$$= \mathbb{E}_q\left[\,\Delta^2(\mathbf{z})\,\right]\mathbb{E}_q\left[\,\frac{q(\mathbf{z})}{\pi(\mathbf{z})}\,\right] + \mathrm{Cov}_q\left(\Delta^2(\mathbf{z}),\, w^{-1}(\mathbf{z})\right) \tag{54}$$

$$\geq \mathbb{E}_q\left[\,\Delta^2(\mathbf{z})\,\right]\exp\left(D_{\mathrm{KL}}(q \parallel \pi)\right) + \mathrm{Cov}_q\left(\Delta^2(\mathbf{z}),\, w^{-1}(\mathbf{z})\right). \tag{55}$$

After some algebra, we obtain our result

$$\frac{2}{N}\left(2 - \frac{3}{N}\right)\sigma^2 + \frac{2}{N}\left(-3 + \frac{5}{N}\right)\int \Delta^2(\mathbf{z})\,q(\mathbf{z})\,d\mathbf{z} + \frac{2}{N}\left(1 - \frac{2}{N}\right)\int \Delta^2(\mathbf{z})\,\frac{q^2(\mathbf{z})}{\pi(\mathbf{z})}\,d\mathbf{z} \tag{56}$$

$$\geq \frac{2}{N}\left(2 - \frac{3}{N}\right)\sigma^2 + \frac{2}{N}\left(-3 + \frac{5}{N}\right)\mathbb{E}_q\left[\,\Delta^2\,\right]$$
$$+ \frac{2}{N}\left(1 - \frac{2}{N}\right)\left\{\mathbb{E}_q\left[\,\Delta^2\,\right]\exp\left(D_{\mathrm{KL}}(q \parallel \pi)\right) + \mathrm{Cov}_q\left(\Delta^2(\mathbf{z}),\, w^{-1}(\mathbf{z})\right)\right\} \tag{57}$$

$$\approx \frac{4}{N}\sigma^2 - \frac{6}{N}\mathbb{E}_q\left[\,\Delta^2\,\right] + \frac{2}{N}\left\{\mathbb{E}_q\left[\,\Delta^2\,\right]\exp\left(D_{\mathrm{KL}}(q \parallel \pi)\right) + \mathrm{Cov}_q\left(\Delta^2(\mathbf{z}),\, w^{-1}(\mathbf{z})\right)\right\} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{58}$$

$$= \frac{4}{N}\sigma^2 + \frac{2}{N}\mathbb{E}_q\left[\,\Delta^2\,\right]\left(\exp\left(D_{\mathrm{KL}}(q \parallel \pi)\right) - 3\right) + \frac{2}{N}\,\mathrm{Cov}_q\left(\Delta^2(\mathbf{z}),\, w^{-1}(\mathbf{z})\right) + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{59}$$

Asymptotic Approximation of C_{gap} Before concluding, we discuss the case where N is large such that

$$\frac{r}{1-r} - \frac{r}{N}\,\frac{\left(1 - r^N\right)}{(1-r)^2} \xrightarrow{N\to\infty} \frac{r}{1-r}. \tag{60}$$

Using the lower bound $r \geq 1 - \frac{1}{w(\mathbf{z})}$,

$$\frac{r}{1-r} \geq w(\mathbf{z}) - 1. \tag{61}$$

Therefore,

$$\int \Delta^2(\mathbf{z}) \sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\lambda^k(w(\mathbf{z}))\,\pi(d\mathbf{z}) \to \int \Delta^2(\mathbf{z})\,(w(\mathbf{z}) - 1)\,\pi(d\mathbf{z}) \tag{62}$$

$$= \int \Delta^2(\mathbf{z})\; w(\mathbf{z})\; \pi(d\mathbf{z}) - \sigma^2 \tag{63}$$

$$= \int \Delta^2(\mathbf{z})\; \frac{\pi(\mathbf{z})}{q(\mathbf{z})}\, \pi(d\mathbf{z}) - \sigma^2 \tag{64}$$

$$= \mathbb{E}_\pi\left[\Delta^2(\mathbf{z})\right] \mathbb{E}_\pi\left[\frac{\pi(\mathbf{z})}{q(\mathbf{z})}\right] + \mathrm{Cov}_\pi\left(\Delta^2(\mathbf{z}), w(\mathbf{z})\right) - \sigma^2 \tag{65}$$

$$= \sigma^2\left(\int \pi(\mathbf{z})\; \frac{\pi(\mathbf{z})}{q(\mathbf{z})} d\mathbf{z} - 1\right) + \mathrm{Cov}_\pi\left(\Delta^2(\mathbf{z}), w(\mathbf{z})\right) \tag{66}$$

$$\geq \sigma^2\; (\exp(D_{\mathrm{KL}}(\pi \parallel q)) - 1) + \mathrm{Cov}_\pi\left(\Delta^2(\mathbf{z}), w(\mathbf{z})\right). \tag{67}$$

□

Proposition 1. *For the initial state $\mathbf{z} \in \mathcal{Z}$, let $w(\mathbf{z}) > \epsilon$ and $\pi(\mathbf{z}) > \epsilon$ for some constant $\epsilon > 0$ and define $w^* = \sup_{\mathbf{z}} w(\mathbf{z})$. Then, the ratio between the t-step marginal kernel and the invariant distribution is bounded as*

$$\frac{K^t(\mathbf{z}, \mathbf{z}^*)}{\pi(\mathbf{z}^*)} \leq \frac{t}{\epsilon}\left(1 - \frac{1}{w^*}\right)^{t-1}.$$

which is finite for all t if $w^* < \infty$.

*Proof of **Proposition** 1.* Since the kernel K is a mixture of the accept and reject cases,

$$\frac{K^t(\mathbf{z}, \mathbf{z}^*)}{\pi(\mathbf{z}^*)} \leq \max\left\{ \underbrace{\frac{T_t(w(\mathbf{z}) \vee w(\mathbf{z}^*))}{\pi(\mathbf{z}^*)}}_{\text{accept}}, \underbrace{\frac{\lambda^t(w(\mathbf{z}))}{\pi(\mathbf{z})}}_{\text{accept}} \right\} \qquad \text{(Theorem 1).} \tag{68}$$

For the accept case,

$$\frac{T_t(w(\mathbf{z}) \vee w(\mathbf{z}^*))}{\pi(\mathbf{z}^*)} \leq \frac{t}{w(\mathbf{z}) \vee w(\mathbf{z}^*)}\left(1 - \frac{1}{w^*}\right)^{t-1} \qquad \text{(Lemma 1)} \tag{69}$$

$$\leq \frac{t}{\epsilon}\left(1 - \frac{1}{w^*}\right)^{t-1}. \tag{70}$$

and for the reject case,

$$\frac{\lambda^t(w(\mathbf{z}))}{\pi(\mathbf{z})} \leq \frac{\lambda^t(w(\mathbf{z}))}{\pi(\mathbf{z})} \leq \frac{1}{\pi(\mathbf{z})}\left(1 - \frac{1}{w^*}\right)^t \leq \frac{1}{\epsilon}\left(1 - \frac{1}{w^*}\right)^t. \tag{71}$$

Therefore, Equation (68) becomes

$$\frac{K^t(\mathbf{z}, \mathbf{z}^*)}{\pi(\mathbf{z}^*)} \leq \max\left\{ \frac{t}{\epsilon}\left(1 - \frac{1}{w^*}\right)^{t-1}, \frac{1}{\epsilon}\left(1 - \frac{1}{w^*}\right)^t \right\} \leq \frac{t}{\epsilon}\left(1 - \frac{1}{w^*}\right)^{t-1}, \tag{72}$$

where the last inequality follows from $1 - \frac{1}{w^*} \leq 1$ □

Theorem 3. *Assuming $\eta(\mathbf{z}) < M\,\pi(\mathbf{z})$ for some $M < \infty$ and $\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \| \leq L$, the bias of the parallel state estimator with an IMH kernel is bounded as*

$$\mathrm{Bias}\left[\mathbf{g}_{\textit{par.-IMH}}\right] \leq C\sqrt{D_{\mathrm{KL}}(\pi \parallel q)} + L\left(1 - \frac{1}{w^*}\right)$$

for some positive constant C.

*Proof of **Theorem** 3.* The bias of a the parallel state estimator kernel is bounded as

$$\mathrm{Bias}\left[\mathbf{g}_{\text{par.-IMH}}\right] \tag{73}$$

$$= \left\| \int \eta\left(\mathbf{z}^{(i)}\right) \int \frac{1}{N} \sum_{i=1}^{N} K\left(\mathbf{z}^{(i)}, \mathbf{z}^*\right) \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\; d\mathbf{z}^*\, d\mathbf{z}_{(1:N)} - \int \pi(\mathbf{z}^*)\, \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\, d\mathbf{z}^* \right\| \tag{74}$$

$$= \left\| \frac{1}{N} \int \eta\left(\mathbf{z}^{(i)}\right) \int \sum_{i=1}^{N} \left(K\left(\mathbf{z}^{(i)}, \mathbf{z}^*\right) - \pi\left(\mathbf{z}^*\right) \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \, d\mathbf{z}_{(1:N)} \right\| \tag{75}$$

$$= \left\| \frac{1}{N} \sum_{i=1}^{N} \int \eta\left(\mathbf{z}^{(i)}\right) \int \left(K\left(\mathbf{z}^{(i)}, \mathbf{z}^*\right) - \pi\left(\mathbf{z}^*\right) \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \, d\mathbf{z}_{(1:N)} \right\| \tag{76}$$

$$\leq \frac{1}{N} \sum_{i=1}^{N} \left\| \int \eta\left(\mathbf{z}^{(i)}\right) \int \left(K\left(\mathbf{z}^{(i)}, \mathbf{z}^*\right) - \pi\left(\mathbf{z}^*\right) \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \, d\mathbf{z}_{(1:N)} \right\| \quad \text{(Triangle Inequality)} \tag{77}$$

$$= \left\| \int \eta\left(\mathbf{z}\right) \int \left(K\left(\mathbf{z}, \mathbf{z}^*\right) - \pi\left(\mathbf{z}^*\right) \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \, d\mathbf{z} \right\| \quad \text{(Independence)} \tag{78}$$

$$= \left\| \int \eta\left(\mathbf{z}\right) \int \pi\left(\mathbf{z}^*\right) \left(T_1\left(w\left(\mathbf{z}\right) \vee w\left(\mathbf{z}^*\right)\right) - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \, d\mathbf{z} + \int \eta\left(\mathbf{z}\right) \lambda\left(w\left(\mathbf{z}\right)\right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}\right) \right\| \quad \text{(Theorem 1)} \tag{79}$$

$$\leq \left\| \int \eta\left(\mathbf{z}\right) \int \pi\left(\mathbf{z}^*\right) \left(T_1\left(w\left(\mathbf{z}\right) \vee w\left(\mathbf{z}^*\right)\right) - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \, d\mathbf{z} \right\| + \left\| \int \eta\left(\mathbf{z}\right) \lambda\left(w\left(\mathbf{z}\right)\right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}\right) d\mathbf{z} \right\| \quad \text{(Triangle Inequality)} \tag{80}$$

$$= \underbrace{\left\| \int \eta\left(\mathbf{z}\right) \int \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}\right) \vee w\left(\mathbf{z}^*\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* d\mathbf{z} \right\|}_{\text{T1}} + \underbrace{\left\| \int \eta\left(\mathbf{z}\right) \, \lambda\left(w\left(\mathbf{z}\right)\right) \, \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}\right) d\mathbf{z} \right\|}_{\text{T2}}. \tag{81}$$

For T1, we denote the set of the accepted proposals A as $A\left(w\right) = \left\{ \mathbf{z}^* \mid w\left(\mathbf{z}^*\right) > w \right\}$, and the set of the rejected proposals R as $R\left(w\right) = A^c\left(w\right)$. Then,

$$\left\| \int \eta\left(\mathbf{z}\right) \left(\int \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}\right) \vee w\left(\mathbf{z}^*\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \right) d\mathbf{z} \right\| \tag{82}$$

$$= \left\| \int \eta\left(\mathbf{z}\right) \left\{ \int_{A(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}^*\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* + \int_{R(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \right\} d\mathbf{z} \right\| \tag{83}$$

$$\leq \underbrace{\left\| \int \eta\left(\mathbf{z}\right) \int_{A(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}^*\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* d\mathbf{z} \right\|}_{\text{T3}} + \underbrace{\left\| \int \eta\left(\mathbf{z}\right) \int_{R(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* d\mathbf{z} \right\|}_{\text{T4}} \tag{84}$$

For T3,

$$\left\| \int \eta\left(\mathbf{z}\right) \int_{A(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}^*\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* d\mathbf{z} \right\| = \left\| \int \eta\left(\mathbf{z}\right) \, d\mathbf{z} \int_{A(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}^*\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \right\| \tag{85}$$

$$= \left\| \int_{A(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}^*\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* \right\| \tag{86}$$

$$\leq \int_{A(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \left| \frac{1}{w\left(\mathbf{z}^*\right)} - 1 \right| \left\| \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) \right\| d\mathbf{z}^* \tag{87}$$

$$\leq \int \pi\left(\mathbf{z}^*\right) \left| \frac{1}{w\left(\mathbf{z}^*\right)} - 1 \right| \left\| \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) \right\| d\mathbf{z}^* \tag{88}$$

$$\leq \int \left| \pi\left(\mathbf{z}^*\right) - q\left(\mathbf{z}^*\right) \right| \left\| \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) \right\| d\mathbf{z}^* \tag{89}$$

$$\leq L \int \left| \pi\left(\mathbf{z}^*\right) - q\left(\mathbf{z}^*\right) \right| d\mathbf{z}^* \tag{90}$$

$$\leq 2\, L \left\| \pi - q \right\|_{\text{TV}}. \tag{91}$$

Now, for T4,

$$\left\| \int \eta\left(\mathbf{z}\right) \int_{R(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \left(\frac{1}{w\left(\mathbf{z}\right)} - 1 \right) \mathbf{s}\left(\boldsymbol{\lambda}; \mathbf{z}^*\right) d\mathbf{z}^* d\mathbf{z} \right\| = \left\| \int \eta\left(\mathbf{z}\right) \left(\frac{1}{w\left(\mathbf{z}\right)} - 1 \right) d\mathbf{z} \int_{R(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \mathbf{s}\left(\mathbf{z}^*\right) d\mathbf{z}^* \right\| \tag{92}$$

$$= \underbrace{\left\| \int \eta\left(\mathbf{z}\right) \left(\frac{1}{w\left(\mathbf{z}\right)} - 1 \right) d\mathbf{z} \right\|}_{\text{T5}} \underbrace{\left\| \int_{R(w(\mathbf{z}))} \pi\left(\mathbf{z}^*\right) \mathbf{s}\left(\mathbf{z}^*\right) d\mathbf{z}^*, \right\|}_{\text{T6}}, \tag{93}$$

while T5 is bounded as

$$\left\| \int \eta(\mathbf{z}) \left(\frac{1}{w(\mathbf{z})} - 1 \right) d\mathbf{z} \right\| = \int \eta(\mathbf{z}) \left| \frac{q(\mathbf{z})}{\pi(\mathbf{z})} - 1 \right| d\mathbf{z} \tag{94}$$

$$= \int \frac{\eta(\mathbf{z})}{\pi(\mathbf{z})} \left| \left(q(\mathbf{z}) - \pi(\mathbf{z}) \right) \right| d\mathbf{z} \tag{95}$$

$$\leq M \int \left| q(\mathbf{z}) - \pi(\mathbf{z}) \right| d\mathbf{z} \tag{96}$$

$$= 2M \left\| \pi - q \right\|_{\mathrm{TV}} \tag{97}$$

and T6 is bounded as

$$\left\| \int_{R(w(\mathbf{z}))} \pi(\mathbf{z}^*) \, \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \, d\mathbf{z}^* \right\| \leq \int_{R(w(\mathbf{z}))} \pi(\mathbf{z}^*) \left\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \right\| d\mathbf{z}^* \tag{98}$$

$$\leq \int \pi(\mathbf{z}^*) \left\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \right\| d\mathbf{z}^* \tag{99}$$

$$\leq L \int \pi(\mathbf{z}^*) \, d\mathbf{z}^* \tag{100}$$

$$\leq L. \tag{101}$$

Finally, for T2,

$$\left\| \int \eta(\mathbf{z}) \, \lambda(w(\mathbf{z})) \, \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \, d\mathbf{z} \right\| \leq \int \eta(\mathbf{z}) \, \lambda(w(\mathbf{z})) \left\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \right\| d\mathbf{z} \tag{102}$$

$$\leq \int \eta(\mathbf{z}) \left(1 - \frac{1}{w^*} \right) \left\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \right\| d\mathbf{z} \tag{103}$$

$$= \left(1 - \frac{1}{w^*} \right) \int \eta(\mathbf{z}) \left\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \right\| d\mathbf{z} \tag{104}$$

$$\leq L \left(1 - \frac{1}{w^*} \right) \int \eta(\mathbf{z}) \, d\mathbf{z} \tag{105}$$

$$= L \left(1 - \frac{1}{w^*} \right) \tag{106}$$

By combining T1, T2, T3, and T4, we conclude

$$\mathrm{Bias}\left[\mathbf{g}_{\text{par.-IMH}} \right] \leq 2L \left\| \pi - q \right\|_{\mathrm{TV}} + 2LM \left\| \pi - q \right\|_{\mathrm{TV}} + L \left(1 - \frac{1}{w^*} \right) \tag{107}$$

$$= 2L(1 + M) \left\| \pi - q \right\|_{\mathrm{TV}} + L \left(1 - \frac{1}{w^*} \right) \tag{108}$$

$$\leq 2L(1 + M) \sqrt{D_{\mathrm{KL}}(\pi \parallel q)} + L \left(1 - \frac{1}{w^*} \right). \qquad \text{(Pinsker's Inequality)} \tag{109}$$

□

Theorem 4. *Assuming $\eta(\mathbf{z}) < M \pi(\mathbf{z})$ for some $M < \infty$, when using the IMH kernel, the reduction in bias by using the sequential state estimator with N states instead of the parallel state estimator with N chains is bounded as*

$$\left| \mathit{Bias}\left[\mathbf{g}_{\mathit{seq.-IMH}} \right] - \mathit{Bias}\left[\mathbf{g}_{\mathit{par.-IMH}} \right] \right| \leq C_1 \max\left\{ n \left(1 - \frac{1}{w^*} \right)^{n-1} - 1, 1 \right\} + C_2 \left(1 - \frac{1}{w^*} \right)$$

for some positive constants C_1 and C_2.

*Proof of **Theorem** 4.*

$$\left| \mathrm{Bias}\left[\mathbf{g}_{\text{seq.-IMH}} \right] - \mathrm{Bias}\left[\mathbf{g}_{\text{par.-IMH}} \right] \right| \tag{110}$$

$$\leq \left\| \int \eta(\mathbf{z}) \int \frac{1}{N} \sum_{n=1}^{N} K^n(\mathbf{z}, \mathbf{z}^*)\, \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\, d\mathbf{z}^*\, d\mathbf{z} - \int \eta(\mathbf{z}) \int \frac{1}{N} \sum_{n=1}^{N} K\left(\mathbf{z}^{(n)}, \mathbf{z}^*\right) \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\, d\mathbf{z}^*\, d\mathbf{z}_{(1:N)} \right\| \quad \text{(Triangle Inequality)} \tag{111}$$

$$= \left\| \int \eta(\mathbf{z}) \int \frac{1}{N} \sum_{n=1}^{N} K^n(\mathbf{z}, \mathbf{z}^*)\, \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\, d\mathbf{z}^*\, d\mathbf{z} - \int \eta(\mathbf{z}) \int \frac{1}{N} \sum_{n=1}^{N} K(\mathbf{z}, \mathbf{z}^*)\, \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\, d\mathbf{z}^*\, d\mathbf{z} \right\| \quad \text{(Independence)} \tag{112}$$

$$= \left\| \int \eta(\mathbf{z}) \int \frac{1}{N} \sum_{n=1}^{N} \left(K^n(\mathbf{z}, \mathbf{z}^*) - K(\mathbf{z}, \mathbf{z}^*)\right) \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\, d\mathbf{z}^*\, d\mathbf{z} \right\| \tag{113}$$

$$\leq \frac{1}{N} \sum_{n=1}^{N} \left\| \int \int \eta(\mathbf{z}) \left(K^n(\mathbf{z}, \mathbf{z}^*) - K(\mathbf{z}, \mathbf{z}^*)\right) \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\, d\mathbf{z}^*\, d\mathbf{z} \right\| \quad \text{(Triangle Inequality)} \tag{114}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \Bigg\| \int \eta(\mathbf{z}) \int \pi(\mathbf{z}^*) \left(T_n(w(\mathbf{z}) \vee w(\mathbf{z}^*)) - T_1(w(\mathbf{z}) \vee w(\mathbf{z}^*))\right) \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\, d\mathbf{z}^*\, d\mathbf{z} + \int \eta(\mathbf{z}) \left(\lambda^n(w(\mathbf{z})) - \lambda(w(\mathbf{z}))\right) \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z})\, d\mathbf{z} \Bigg\| \quad \text{(Theorem 1)} \tag{115}$$

$$\leq \frac{1}{N} \sum_{n=1}^{N} \left\| \int \eta(\mathbf{z}) \int \pi(\mathbf{z}^*) \left(T_n(w(\mathbf{z}) \vee w(\mathbf{z}^*)) - T_1(w(\mathbf{z}) \vee w(\mathbf{z}^*))\right) \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*)\, d\mathbf{z}^*\, d\mathbf{z} \right\| + \frac{1}{N} \sum_{n=1}^{N} \left\| \int \eta(\mathbf{z}) \left(\lambda^n(w(\mathbf{z})) - \lambda(w(\mathbf{z}))\right) \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z})\, d\mathbf{z} \right\| \quad \text{(Triangle Inequality)} \tag{116}$$

$$\leq \frac{1}{N} \sum_{n=1}^{N} \underbrace{\int \eta(\mathbf{z}) \int \pi(\mathbf{z}^*) \left| T_n(w(\mathbf{z}) \vee w(\mathbf{z}^*)) - T_1(w(\mathbf{z}) \vee w(\mathbf{z}^*)) \right| \left\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \right\| d\mathbf{z}^*\, d\mathbf{z}}_{T1} + \frac{1}{N} \sum_{n=1}^{N} \underbrace{\int \eta(\mathbf{z}) \left| \lambda^n(w(\mathbf{z})) - \lambda(w(\mathbf{z})) \right| \left\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}) \right\| d\mathbf{z}}_{T2} \tag{117}$$

Now, denoting $w' = w(\mathbf{z}) \vee w(\mathbf{z}^*)$,

$$\left| T_n(w(\mathbf{z}) \vee w(\mathbf{z}^*)) - T_1(w(\mathbf{z}) \vee w(\mathbf{z}^*)) \right| = \left| \int_{w'}^{\infty} \frac{1}{v^2}\, n\, \lambda^{n-1}(w(\mathbf{z}))\, dv - \int_{w'}^{\infty} \frac{1}{v^2} dv \right| \tag{118}$$

$$= \left| \int_{w'}^{\infty} \frac{1}{v^2} \left(n\, \lambda^{n-1}(w(\mathbf{z})) - 1\right) dv \right| \tag{119}$$

$$\leq \int_{w'}^{\infty} \frac{1}{v^2} \left| n\, \lambda^{n-1}(w(\mathbf{z})) - 1 \right| dv. \tag{120}$$

Since

$$0 \leq \lambda(w(\mathbf{z})) \leq 1 - \frac{1}{w^*} \leq 1,$$

if follows that

$$-1 \leq n\, \lambda^{n-1}(w(\mathbf{z})) - 1 \leq n\left(1 - \frac{1}{w^*}\right)^{n-1} - 1,$$

and thus

$$\left| n\, \lambda^{n-1}(w(\mathbf{z})) - 1 \right| \leq \max\left\{ n\left(1 - \frac{1}{w^*}\right)^{n-1} - 1, 1 \right\}.$$

Then, we can bound Equation (120) as

$$\int_{w'}^{\infty} \frac{1}{v^2} \left| n\, \lambda^{n-1}(w(\mathbf{z})) - 1 \right| dv \leq \int_{w'}^{\infty} \frac{1}{v^2}\, \max\left\{ n\left(1 - \frac{1}{w^*}\right)^{n-1} - 1, 1 \right\} dv \tag{121}$$

$$= \max\left\{ n\left(1-\frac{1}{w^*}\right)^{n-1} - 1, 1 \right\} \int_{w'}^{\infty} \frac{1}{v^2} \tag{122}$$

$$= \max\left\{ n\left(1-\frac{1}{w^*}\right)^{n-1} - 1, 1 \right\} \frac{1}{w'} \tag{123}$$

$$= \max\left\{ n\left(1-\frac{1}{w^*}\right)^{n-1} - 1, 1 \right\} \frac{1}{w(\mathbf{z}) \vee w(\mathbf{z}^*)}. \tag{124}$$

Now, back to Equation (117), T1 can be bounded as

$$\int \eta(\mathbf{z}) \int \pi(\mathbf{z}^*) \mid T_n(w(\mathbf{z}) \vee w(\mathbf{z}^*)) - T_1(w(\mathbf{z}) \vee w(\mathbf{z}^*)) \mid \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \, d\mathbf{z} \tag{125}$$

$$\leq \int \eta(\mathbf{z}) \int \pi(\mathbf{z}^*) \max\left\{ n\left(1-\frac{1}{w^*}\right)^{n-1} - 1, 1 \right\} \frac{1}{w(\mathbf{z}) \vee w(\mathbf{z}^*)} \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \, d\mathbf{z} \tag{126}$$

$$= \max\left\{ n\left(1-\frac{1}{w^*}\right)^{n-1} - 1, 1 \right\}$$
$$\times \int \eta(\mathbf{z}) \left\{ \int_{A(w(\mathbf{z}))} \pi(\mathbf{z}^*) \frac{1}{w(\mathbf{z}^*)} \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* + \int_{R(w(\mathbf{z}))} \pi(\mathbf{z}^*) \frac{1}{w(\mathbf{z})} \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \right\} d\mathbf{z} \tag{127}$$

$$= \max\left\{ n\left(1-\frac{1}{w^*}\right)^{n-1} - 1, 1 \right\}$$
$$\times \left\{ \underbrace{\int \eta(\mathbf{z}) \int_{A(w(\mathbf{z}))} \pi(\mathbf{z}^*) \frac{1}{w(\mathbf{z}^*)} \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \, d\mathbf{z}}_{T3} + \underbrace{\int \eta(\mathbf{z}) \frac{1}{w(\mathbf{z})} d\mathbf{z} \int_{R(w(\mathbf{z}))} \pi(\mathbf{z}^*) \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \, d\mathbf{z}}_{T4} \right\}. \tag{128}$$

T3 reduces to

$$\int \eta(\mathbf{z}) \int_{A(w(\mathbf{z}))} \pi(\mathbf{z}^*) \frac{1}{w(\mathbf{z}^*)} \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \, d\mathbf{z} \leq \int \eta(\mathbf{z}) \int \pi(\mathbf{z}^*) \frac{1}{w(\mathbf{z}^*)} \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \, d\mathbf{z} \tag{129}$$

$$= \int \eta(\mathbf{z}) \, d\mathbf{z} \int \pi(\mathbf{z}^*) \frac{1}{w(\mathbf{z}^*)} \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \tag{130}$$

$$= \int \pi(\mathbf{z}^*) \frac{1}{w(\mathbf{z}^*)} \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \tag{131}$$

$$= \int q(\mathbf{z}^*) \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \tag{132}$$

$$= \mathbb{E}_q [\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \|], \tag{133}$$

which is a positive constant.

On the other hand, T4 becomes

$$\int \eta(\mathbf{z}) \frac{1}{w(\mathbf{z})} d\mathbf{z} \int_{R(w(\mathbf{z}))} \pi(\mathbf{z}^*) \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \leq \int \eta(\mathbf{z}) \frac{q(\mathbf{z})}{\pi(\mathbf{z})} d\mathbf{z} \int \pi(\mathbf{z}^*) \| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \| \, d\mathbf{z}^* \tag{134}$$

$$= \mathbb{E}_\pi [\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \|] \int \eta(\mathbf{z}) \frac{q(\mathbf{z})}{\pi(\mathbf{z})} d\mathbf{z} \tag{135}$$

$$= \mathbb{E}_\pi [\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \|] \int q(\mathbf{z}) \frac{\eta(\mathbf{z})}{\pi(\mathbf{z})} d\mathbf{z} \tag{136}$$

$$\leq M \, \mathbb{E}_\pi [\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \|] \int q(\mathbf{z}) \, d\mathbf{z} \tag{137}$$

$$= M \, \mathbb{E}_\pi [\| \mathbf{s}(\boldsymbol{\lambda}; \mathbf{z}^*) \|]. \tag{138}$$

Meanwhile, for T2,

$$\frac{1}{N}\sum_{n=1}^{N}\int \eta(\mathbf{z})\,|\,\lambda^{n}(w(\mathbf{z}))-\lambda(w(\mathbf{z}))\,|\,\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\|\,d\mathbf{z} = \frac{1}{N}\sum_{n=1}^{N}\int \eta(\mathbf{z})\,\lambda(w(\mathbf{z}))\,|\,\lambda^{n-1}(w(\mathbf{z}))-1\,|\,\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\|\,d\mathbf{z} \tag{139}$$

$$\leq \frac{1}{N}\sum_{n=1}^{N}\int \eta(\mathbf{z})\,\lambda(w(\mathbf{z}))\,\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\|\,d\mathbf{z} \tag{140}$$

$$= \int \eta(\mathbf{z})\,\lambda(w(\mathbf{z}))\,\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\|\,d\mathbf{z} \tag{141}$$

$$\leq \left(1-\frac{1}{w^{*}}\right)\int \eta(\mathbf{z})\,\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\|\,d\mathbf{z} \tag{142}$$

$$\leq \left(1-\frac{1}{w^{*}}\right)\mathbb{E}_{\eta}\left[\,\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\|\,\right]. \tag{143}$$

Combining T1, T2, T3, and T4, we conclude by

$$\left|\,\mathrm{Bias}\left[\mathbf{g}_{\text{seq.-IMH}}\right]-\mathrm{Bias}\left[\mathbf{g}_{\text{par.-IMH}}\right]\,\right| \tag{144}$$

$$\leq \max\left\{n\left(1-\frac{1}{w^{*}}\right)^{n-1}-1,1\right\}\left(\mathbb{E}_{q}\left[\,\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\|\,\right]+M\,\mathbb{E}_{\pi}\left[\,\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z}^{*})\,\|\,\right]\right)+\left(1-\frac{1}{w^{*}}\right)\mathbb{E}_{\eta}\left[\,\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{z})\,\|\,\right]. \tag{145}$$

□

Proposition 2. *$w^{*}=\sup_{\mathbf{z}} w(\mathbf{z})$ is bounded below exponentially by the KL divergence such that*

$$\exp\left(D_{\mathrm{KL}}(\pi\,\|\,q_{\lambda})\right)\leq w^{*}.$$

*Proof of **Proposition** 2.*

$$D_{\mathrm{KL}}(\pi\,\|\,q_{\lambda})=\int \pi(\mathbf{z})\log\frac{\pi(\mathbf{z})}{q_{\lambda}(\mathbf{z})}\,d\mathbf{z}\leq\int \pi(\mathbf{z})\log w^{*}\,d\mathbf{z}=\log w^{*}$$

□