A Relationship Between Sampling Methods

We organize the sampling methods described in this work in Table 1.

Table 1: Comparison of Sampling Method Designs

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

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

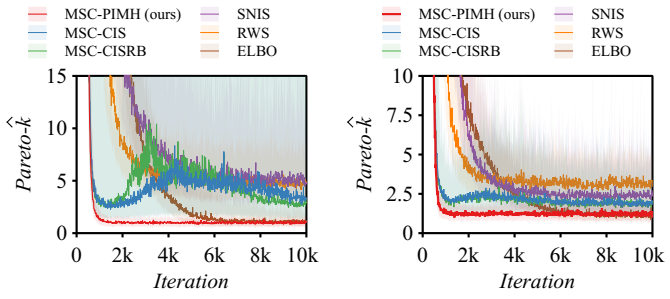
In this paper, we designated kernels that use independent proposals and perform a Metropolis-Hastings (M-H) test as "IMH type" kernels. While the original paper of CIS does not mention it as an IMH type, we have shown in Section 3.1 that it is indeed an IMH type kernel that uses Barker's acceptance ratio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

B Additional Experimental Results

B.1 Experimental Environment

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

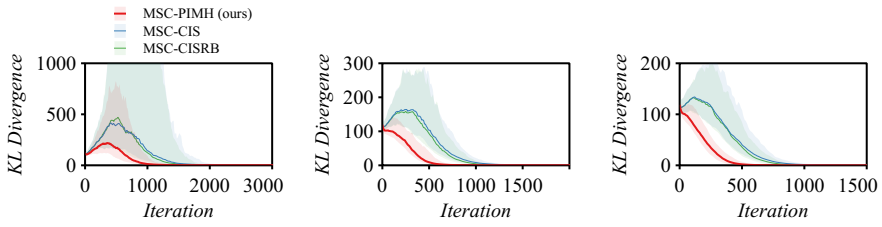
B.2 Additional Results of Logistic Regression Experiments



(a) Pima (b) Heart

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

B.3 Isotropic Gaussian Experiments



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

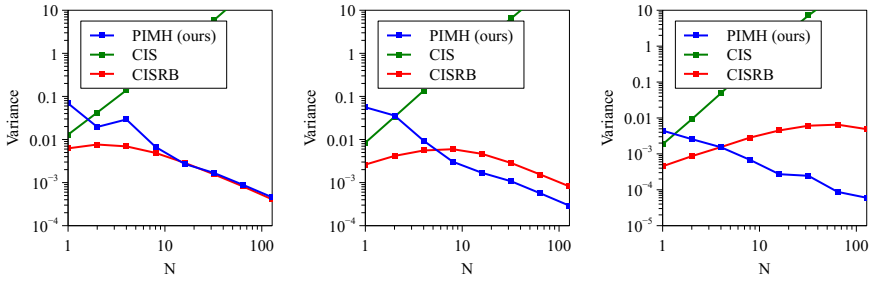
Figure 6: 100-D isotropic Gaussian example with a varying computational budget N. MSC-PIMH converges faster than MSC-CIS and MSC-CISRB regardless of N. Also, the convergence of MSC-PIMH becomes more stable/monotonic as N increases. The solid lines and colored regions are the medians and 80% percentiles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

We perform experiments with a 100-D isotropic multivariate Gaussian distribution. With Gaussian distributions, convergence can be evaluated exactly since their KL divergence is available in a closed form. We compare the performance of MSC-PIMH, MSC-CIS, and MSC-CISRB with respect to the N (number of proposals for MSC-CIS, MSC-CISRB; number of parallel chains for MSC-PIMH). The results are shown in Figure 6. While MSC-PIMH shows some level of overshoot wih $N = 4$, it shows monotonic convergence with larger N. On the other hand, both MSC-CIS and MSC-CISRB overshoots even with $N = 64$. This clearly shows that PIMH enjoys better gradient estimates compared to the CIS kernel.

C Numerical Simulation

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

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



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

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

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

D Probabilistic Models Considered in Section 4

D.1 Hierarchical Logistic Regression

The hierarchical logistic regression used in Section 4.2 is

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

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

D.2 Stochastic Volatility

The stochastic volatility model used in Section 4.3 is

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

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

D.3 Radon Hierarchical Regression

The partially pooled linear regression model used in Section 4.3 is

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

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

E Proofs

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

Proof of ***Proposition*** *1.* We evaluate the variance by approximating the rejection probability as an independent constant. First,

$$\mathbb{V}_{q_\lambda}\left[\, \mathbb{E}\left[\, f \mid \mathbf{z}_{t-1}, \mathbf{z}^{(1:N)} \,\right] \right]. \tag{25}$$

Applying (22),

$$= \mathbb{V}_{q_\lambda}\left[\left(1 - r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right)\right) f_{\text{IS}} + r\left(\mathbf{z}_{t-1} \mid \mathbf{z}^{(1:N)}\right) f(\mathbf{z}_{t-1}) \mid \mathbf{z}_{t-1}\right] \tag{26}$$

$$\approx \mathbb{V}_{q_\lambda}\left[\left(1 - r\left(\mathbf{z}_{t-1}\right)\right) f_{\text{IS}} + r\left(\mathbf{z}_{t-1}\right) f(\mathbf{z}_{t-1}) \mid \mathbf{z}_{t-1}\right] \tag{27}$$

$$= \mathbb{V}_{q_\lambda}\left[\left(1 - r\left(\mathbf{z}_{t-1}\right)\right) f_{\text{IS}} \mid \mathbf{z}_{t-1}\right] \tag{28}$$

$$= \left(1 - r\left(\mathbf{z}_{t-1}\right)\right)^2 \mathbb{V}_{q_\lambda}\left[f_{\text{IS}} \mid \mathbf{z}_{t-1}\right]. \tag{29}$$

The equality of (28) follows from the fact that $r\left(\mathbf{z}_{t-1}\right) f(\mathbf{z}_{t-1})$ is a constant. □

Proposition 2. *The rejection rate $r\left(\mathbf{z}_{t-1}\right)$ of a CIS sampler with N proposals is bounded below such that*

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

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

*Proof of **Proposition** 2.* The rejection rate $r(\mathbf{z}_{t-1})$ is given by

$$r(\mathbf{z}_{t-1}) = \mathbb{E}_{q_\lambda}\left[\frac{w(\mathbf{z}_{t-1})}{\sum_{k=1}^{N} w(\mathbf{z}^{(k)}) + w(\mathbf{z}_{t-1})}\right] \tag{30}$$

$$= \mathbb{E}_{q_\lambda}\left[\left(\frac{\sum_{k=1}^{N} w(\mathbf{z}^{(k)})}{w(\mathbf{z}_{t-1})} + 1\right)^{-1}\right]. \tag{31}$$

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

$$\geq \frac{1}{1 + \mathbb{E}_{q_\lambda}\left[\frac{\sum_{k=1}^{N} w(\mathbf{z}^{(k)})}{w(\mathbf{z}_{t-1})}\right]} \tag{32}$$

$$= \frac{1}{1 + \frac{1}{w(\mathbf{z}_{t-1})}\mathbb{E}_{q_\lambda}\left[\sum_{k=1}^{N} w(\mathbf{z}^{(k)})\right]}. \tag{33}$$

From the independence of the N proposals, we obtain

$$= \frac{1}{1 + \frac{1}{w(\mathbf{z}_{t-1})} N \mathbb{E}_{q_\lambda}\left[w(\mathbf{z})\right]} \tag{34}$$

$$= \frac{1}{1 + \frac{1}{w(\mathbf{z}_{t-1})} NZ}. \tag{35}$$

□

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

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

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

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

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

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

Proof. By the definition of the inclusive KL divergence,

$$D_{\mathrm{KL}}(p \parallel q_\lambda) = \int p(\mathbf{z} \mid \mathbf{x}) \log \frac{p(\mathbf{z} \mid \mathbf{x})}{q_\lambda(\mathbf{z})}\, d\mathbf{z} \le \log \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\, \frac{p(\mathbf{z} \mid \mathbf{x})}{q_\lambda(\mathbf{z})} \,\right] \tag{36}$$

$$= \log \mathbb{E}_{p(\mathbf{z}|\mathbf{x})}\left[\, \frac{w(\mathbf{z})}{Z} \,\right] \tag{37}$$

where the right-hand side follows from Jensen's inequality. By a simple change of notation, we relate (37) with the rejection states $\mathbf{z}_{t-1}$ such as

$$D_{\mathrm{KL}}(p \parallel q_\lambda) \le \log \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, \frac{w(\mathbf{z}_{t-1})}{Z} \,\right]. \tag{38}$$

Then,

$$\exp\big(D_{\mathrm{KL}}(p \parallel q_\lambda)\big) \le \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, \frac{w(\mathbf{z}_{t-1})}{Z} \,\right] \tag{39}$$

$$Z\exp\big(D_{\mathrm{KL}}(p \parallel q_\lambda)\big) \le \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, w(\mathbf{z}_{t-1}) \,\right]. \tag{40}$$

□

Now, from the result of Proposition 2,

$$\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, r(\mathbf{z}_{t-1}) \,\right] \ge \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, \frac{w(\mathbf{z}_{t-1})}{w(\mathbf{z}_{t-1}) + N\,Z} \,\right] = \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, \varphi\big(w(\mathbf{z}_{t-1})\big) \,\right], \tag{41}$$

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

$$\varphi\big(\,\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, w(\mathbf{z}_{t-1}) \,\right]\big) \ge \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, \varphi\,(w(\mathbf{z}_{t-1})) \,\right] \tag{42}$$

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

$$\delta = \varphi\big(\,\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, w(\mathbf{z}_{t-1}) \,\right]\big) - \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, \varphi\big(w(\mathbf{z}_{t-1})\big) \,\right], \tag{43}$$

where $\delta \ge 0$. Then, by applying (43) to (41),

$$\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, r(\mathbf{z}_{t-1}) \,\right] \ge \mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, \varphi\big(w(\mathbf{z}_{t-1})\big) \,\right] \tag{44}$$

$$= \varphi\big(\,\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}\left[\, w(\mathbf{z}_{t-1}) \,\right]\big) - \delta, \tag{45}$$

and by the monotonicity of φ and Lemma 1,

$$\ge \varphi\big(\,Z\,\exp\big(D_{\mathrm{KL}}(p \parallel q_\lambda)\big)\big) - \delta \tag{46}$$

$$= \frac{Z\,\exp\big(D_{\mathrm{KL}}(p \parallel q_\lambda)\big)}{Z\,\exp\big(D_{\mathrm{KL}}(p \parallel q_\lambda)\big) + N\,Z} - \delta \tag{47}$$

$$= \frac{\exp\big(D_{\mathrm{KL}}(p \parallel q_\lambda)\big)}{\exp\big(D_{\mathrm{KL}}(p \parallel q_\lambda)\big) + N} - \delta \tag{48}$$

$$= \frac{1}{1 + \frac{N}{\exp\left(D_{\mathrm{KL}}(p\|q_\lambda)\right)}} - \delta. \tag{49}$$

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

$$\inf_{x \in (a,b)} h(x;\mu)\,\sigma^2 \le \mathbb{E}\left[\, \widetilde{\varphi}(X) \,\right] - \widetilde{\varphi}\big(\mathbb{E}\left[\, X \,\right]\big), \quad \text{where} \quad h(x;\nu) = \frac{\widetilde{\varphi}(x) - \widetilde{\varphi}(\nu)}{(x-\nu)^2} - \frac{\widetilde{\varphi}\,'(\nu)}{x-\nu}, \tag{50}$$

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

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

$$\lim_{x\to 0} h(x;\mu) = \lim_{x\to 0} \frac{1}{(x-\mu)^2}(\varphi(x) - \varphi(\mu)) - \frac{1}{x-\mu}\varphi'(\mu) \tag{51}$$

$$= \lim_{x\to 0} \frac{1}{(x-\mu)^2}\left(\frac{x}{x+NZ} - \frac{\mu}{\mu+NZ}\right) - \frac{1}{x-\mu}\left(\frac{NZ}{(\mu+NZ)^2}\right) \tag{52}$$

$$= -\frac{1}{\mu^2}\left(\frac{\mu}{\mu+NZ}\right) + \frac{1}{\mu}\left(\frac{NZ}{(\mu+NZ)^2}\right) \tag{53}$$

$$= -\frac{1}{\mu(\mu+NZ)} + \frac{NZ}{\mu(\mu+NZ)^2} \tag{54}$$

$$> -\frac{1}{\mu^2}. \tag{55}$$

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

We finally discuss the variance term σ^2 in (50). Since we assume $\sup {}^{p(\mathbf{z}|\mathbf{x})}/_{q_\lambda(\mathbf{z})} = M < \infty$, $0 < \frac{p(\mathbf{z}|\mathbf{x})}{q_\lambda(\mathbf{z})} < M$ for all $\mathbf{z} \in \mathcal{Z}$. Then,

$$\sigma^2 = \mathbb{E}\left[\, w^2(\mathbf{z})\,\right] - \mathbb{E}\left[\, w(\mathbf{z})\,\right]^2 \tag{56}$$

$$= \mathbb{E}\left[\left(\frac{Z\, p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right)^2\right] - \mathbb{E}\left[\frac{Z\, p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right]^2 \tag{57}$$

$$= Z^2\left(\mathbb{E}\left[\left(\frac{p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right)^2\right] - \mathbb{E}\left[\frac{p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right]^2\right) \tag{58}$$

$$= Z^2\,\mathbb{V}\left[\frac{p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right], \tag{59}$$

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

$$0 \le \sigma^2 = Z^2\,\mathbb{V}\left[\frac{p(\mathbf{z}\mid\mathbf{x})}{q_\lambda(\mathbf{z})}\right] \le Z^2(M-\mu)\,\mu. \tag{60}$$

By combining the results, we obtain

$$0 \le \delta \le -\inf_{x\in(a,b)} h(x;\mu)\,\sigma^2 = -\sigma^2 \lim_{x\to 0} h(x;\mu) < \frac{\sigma^2}{\mu^2} < \frac{Z^2 M}{\mu^2} = \frac{Z^2 M}{\mathbb{E}_{p(\mathbf{z}_{t-1}|\mathbf{x})}[\, w(\mathbf{z}_{t-1})\,]^2}, \tag{61}$$

and by Lemma 1,

$$0 \le \delta < \frac{M}{\exp^2\left(D_{\mathrm{KL}}(p \parallel q_\lambda)\right)}. \tag{62}$$

□

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

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

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

$$r(\mathbf{z}_{t-1}) = 1 - \int \alpha(\mathbf{z},\mathbf{z}_{t-1})\; q_\lambda(\mathbf{z})\, d\mathbf{z}. \tag{63}$$

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

$$
\begin{aligned}
r\left(\mathbf{z}_{t-1}\right) &= 1 - \int \min\left(\frac{w\left(\mathbf{z}\right)}{w\left(\mathbf{z}_{t-1}\right)}, 1\right) q_{\lambda}(\mathbf{z})\, d\mathbf{z} && (64)\\
&= 1 - \frac{1}{w\left(\mathbf{z}_{t-1}\right)} \int \min\left(w\left(\mathbf{z}\right), w\left(\mathbf{z}_{t-1}\right)\right) q_{\lambda}(\mathbf{z})\, d\mathbf{z} && (65)\\
&= 1 - \frac{1}{w\left(\mathbf{z}_{t-1}\right)} \int \min\left(\frac{p\left(\mathbf{z}, \mathbf{x}\right)}{q_{\lambda}(\mathbf{z})}, w\left(\mathbf{z}_{t-1}\right)\right) q_{\lambda}(\mathbf{z})\, d\mathbf{z} && (66)\\
&= 1 - \frac{1}{w\left(\mathbf{z}_{t-1}\right)} \int \min\left(p\left(\mathbf{z}, \mathbf{x}\right), w\left(\mathbf{z}_{t-1}\right) q_{\lambda}(\mathbf{z})\right) d\mathbf{z} && (67)\\
&\geq 1 - \frac{1}{w\left(\mathbf{z}_{t-1}\right)} \int p\left(\mathbf{z}, \mathbf{x}\right) d\mathbf{z} && (68)\\
&= 1 - \frac{Z}{w\left(\mathbf{z}_{t-1}\right)} && (69)
\end{aligned}
$$

The inequality in Equation (68) follows from $\min\left(p\left(\mathbf{z}, \mathbf{x}\right), \cdot\right) \leq p\left(\mathbf{z}, \mathbf{x}\right)$ for $\forall \mathbf{z} \in \mathcal{Z}$. □