Massively Parallel Reweighted Wake-Sleep (Supplementary Material)

Thomas Heap[1] **Gavin Leech**[1] **Laurence Aitchison**[1]

[1]Department of Computer Science, University of Bristol, Bristol

1 PROOF OF EQUIVALENCE OF THE DIFFERENT FORMS OF THE GLOBAL RWS UPDATES

We start with the RWS P update (Eq. 8a), then use $\nabla_\theta \log \mathcal{P}_{\text{global}}(z) = \left(\nabla_\theta \mathcal{P}_{\text{global}}(z)\right) / \mathcal{P}_{\text{global}}(z)$,

$$\mathrm{E}\left[\Delta\theta_{\text{RWS}}\right] = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\nabla_\theta \mathcal{P}_{\text{global}}(z)}{\mathcal{P}_{\text{global}}(z)}\right]. \tag{1}$$

Using the definition of $\mathcal{P}_{\text{global}}(z)$ (Eq. 4),

$$\mathrm{E}\left[\Delta\theta_{\text{RWS}}\right] = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\nabla_\theta \frac{1}{K}\sum_k r_k(z)}{\mathcal{P}_{\text{global}}(z)}\right] \tag{2}$$

Substituting for $r_k(z)$ (Eq. 5) in the numerator,

$$\mathrm{E}\left[\Delta\theta_{\text{RWS}}\right] = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\frac{1}{K}\sum_k \frac{\nabla_\theta \mathrm{P}_\theta\left(z^k, x\right)}{\mathrm{Q}_\phi(z^k|x)}}{\mathcal{P}_{\text{global}}(z)}\right] \tag{3}$$

substituting $\nabla_\theta \mathrm{P}_\theta\left(z^k, x\right) = \mathrm{P}_\theta\left(z^k, x\right) \nabla_\theta \log \mathrm{P}_\theta\left(z^k, x\right)$,

$$\mathrm{E}\left[\Delta\theta_{\text{RWS}}\right] = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{1}{K}\sum_k \frac{\frac{\mathrm{P}_\theta\left(z^k, x\right)}{\mathrm{Q}_\phi(z^k|x)}}{\mathcal{P}_{\text{global}}(z)} \nabla_\theta \log \mathrm{P}_\theta\left(z^k, x\right)\right] \tag{4}$$

Noticing that the ratio of $\mathrm{P}_\theta\left(z^k, x\right)$ and $\mathrm{Q}_\phi\left(z^k|x\right)$ in the numerator is equal to $r_k(z)$ (Eq. 5), we get back to Eq. (7a), as required.

The RWS Q update is very similar. Again, we start with Eq. (8b), then use $\nabla_\theta \log \mathcal{P}_{\text{global}}(z) = \left(\nabla_\theta \mathcal{P}_{\text{global}}(z)\right) / \mathcal{P}_{\text{global}}(z)$,

$$\mathrm{E}\left[\Delta\phi_{\text{RWS}}\right] = -\mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\nabla_\phi \mathcal{P}_{\text{global}}(z)}{\mathcal{P}_{\text{global}}(z)}\right] \tag{5}$$

Using the definition of $\mathcal{P}_{\text{global}}(z)$ (Eq. 4),

$$\mathrm{E}\left[\Delta\phi_{\text{RWS}}\right] = -\mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\nabla_\phi \frac{1}{K}\sum_k r_k(z)}{\mathcal{P}_{\text{global}}(z)}\right] \tag{6}$$

Substituting for $r_k(z)$ (Eq. 5) in the numerator,

$$\mathrm{E}\left[\Delta\phi_{\text{RWS}}\right] = -\mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\frac{1}{K}\sum_k \nabla_\phi \frac{\mathrm{P}_\theta\left(z^k, x\right)}{\mathrm{Q}_\phi(z^k|x)}}{\mathcal{P}_{\text{global}}(z)}\right] \tag{7}$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Computing the derivative,

$$\mathrm{E}\left[\Delta\phi_{\text{RWS}}\right] = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\frac{1}{K}\sum_k \frac{\mathrm{P}_\theta\left(z^k,x\right)}{\left(\mathrm{Q}_\phi(z^k|x)\right)^2}\nabla_\phi \mathrm{Q}_\phi\left(z^k|x\right)}{\mathcal{P}_{\text{global}}(z)}\right]. \tag{8}$$

Noticing that $\left(\nabla_\phi \mathrm{Q}_\phi\left(z^k|x\right)\right)/\mathrm{Q}_\phi\left(z^k|x\right) = \nabla_\phi \log \mathrm{Q}_\phi\left(z^k|x\right)$,

$$\mathrm{E}\left[\Delta\phi_{\text{RWS}}\right] = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\frac{1}{K}\sum_k \frac{\mathrm{P}_\theta\left(z^k,x\right)}{\mathrm{Q}_\phi(z^k|x)}\nabla_\phi \log \mathrm{Q}_\phi\left(z^k|x\right)}{\mathcal{P}_{\text{global}}(z)}\right]. \tag{9}$$

Finally, noticing that the ratio of $\mathrm{P}_\theta\left(z^k,x\right)$ and $\mathrm{Q}_\phi\left(z^k|x\right)$ in the numerator is equal to $r_k(z)$ (Eq. 5), we get back to Eq. (7b), as required.

Both of these derivations may be straightforwardly repeated for the massively parallel setting, simply by replacing $k \in \mathcal{K}$ with $\mathbf{k} \in \mathcal{K}^n$, and by replacing $1/K$ with $1/K^n$.

2 TMC VS MASSIVELY PARALLEL APPROXIMATE POSTERIORS

TMC approximate posteriors draw the K samples of the ith latent variable IID,

$$\mathrm{Q}_{\text{TMC}}\left(z_i|z_j \text{ for all } j \in \text{qa}\left(i\right)\right) = \prod_{k_i \in \mathcal{K}} \mathrm{Q}_{\text{TMC}}\left(z_i^{k_i}\middle|z_j \text{ for all } j \in \text{qa}\left(i\right)\right) \tag{10}$$

Specifically, TMC draws each sample from an equally weighted mixture over all parent particles,

$$\mathrm{Q}_{\text{TMC}}\left(z_i^{k_i}\middle|z_j \text{ for all } j \in \text{qa}\left(i\right)\right) = \frac{1}{K^{|\text{qa}(i)|}}\sum_{\mathbf{k}_{\text{qa}(i)}} \mathrm{Q}_{\text{global}}\left(z_i^{k_i}\middle|z_j^{k_j} \text{ for all } j \in \text{qa}\left(i\right)\right). \tag{11}$$

In contrast, massively parallel methods do not force us to sample particles IID. The key issue with IID sampling is that it introduces the risk of particle degeneracy [Carpenter et al., 1999, Li et al., 2012, 2014, Zhou et al., 2016, Wang et al., 2017]. In particle degeneracy, some of the parent samples (e.g. z_j^1 where $j \in \text{qa}\left(i\right)$) might have multiple children, in the sense that multiple z_i^k are sampled from the mixture component arising from z_j^1. At the same time, some of the parents, (e.g. z_j^2) might have no children, in the sense that no z_i^k are sampled from a mixture component arising from z_j^2. This is problematic because it reduces diversity in the population of samples, $z_i = (z_i^1, \ldots, z_i^K)$, and this reduction in diversity can be especially problematic in models with long chains of latent variables, such as timeseries models. To reduce the risk of particle degeneracy, the massively parallel methods considered here couple the distribution over each of the K particles,

$$\mathrm{Q}_{\text{MP}}\left(z_i|z_j \text{ for all } j \in \text{qa}\left(i\right)\right) \neq \prod_{k_i \in \mathcal{K}} \mathrm{Q}_{\text{TMC}}\left(z_i^{k_i}\middle|z_j \text{ for all } j \in \text{qa}\left(i\right)\right). \tag{12}$$

However, we do ensure that the marginal for a single particle is the same as for TMC,

$$\mathrm{Q}_{\text{MP}}\left(z_i^{k_i}\middle|z_j \text{ for all } j \in \text{qa}\left(i\right)\right) = \frac{1}{K^{|\text{qa}(i)|}}\sum_{\mathbf{k}_{\text{qa}(i)}} \mathrm{Q}_{\text{global}}\left(z_i^{k_i}\middle|z_j^{k_j} \text{ for all } j \in \text{qa}\left(i\right)\right). \tag{13}$$

To achieve this, we sample a permutation, π for each latent variable, and the permutation tells us which parent particle to consider. To give an example for one parent,

$$\mathrm{Q}_{\text{MP}}\left(z_i|\pi, z_j\right) = \prod_{k_i} \mathrm{Q}_{\text{MP}}\left(z_i^{k_i}\middle|\pi, z_j\right) \tag{14}$$

$$\mathrm{Q}_{\text{MP}}\left(z_i^{k_i}\middle|\pi, z_j\right) = \frac{1}{K}\sum_{k_i} \mathrm{Q}_\phi\left(z_i^{k_i}\middle|z_j^{\pi_{k_i}}\right) \tag{15}$$

Critically, if we marginalise over the permutation, the distribution over a single $z_i^{k_i}$ has the same density as that from a uniform mixture,

$$\mathrm{Q}_{\mathrm{MP}}\left(z_i^{k_i}\middle|z_j\right) = \sum_{\pi} \mathrm{Q}_{\mathrm{MP}}\left(z_i^{k_i}\middle|\pi, z_j\right) \tag{16}$$

$$\mathrm{Q}_{\mathrm{MP}}\left(z_i^{k_i}\middle|z_j\right) = \tfrac{1}{K}\sum_{k_j} \mathrm{Q}_{\phi}\left(z_i^{k_i}\middle|z_j^{k_j}\right). \tag{17}$$

Finally, if we have multiple parent latent variables, we independently sample a permutation for each latent variable.

3 MASSIVELY PARALLEL IWAE AND RWS

Before getting started, it will prove useful to define some briefer notation than that used in the main text. Specifically, we use,

$$z_{\mathrm{qa}(i)} = \left\{z_j \text{ for all } j \in \mathrm{qa}\left(i\right)\right\}, \tag{18}$$

$$z_{\mathrm{qa}(i)}^{k_i} = \left\{z_j^{k_i} \text{ for all } j \in \mathrm{qa}\left(i\right)\right\}, \tag{19}$$

$$z_{\mathrm{pa}(i)}^{\mathbf{k}_{\mathrm{pa}(i)}} = \left\{z_j^{k_j} \text{ for all } j \in \mathrm{pa}\left(i\right)\right\}, \tag{20}$$

so,

$$\mathrm{Q}\left(z_i^{k_i}\middle|x, z_{\mathrm{qa}(i)}\right) = \mathrm{Q}\left(z_i^{k_i}\middle|x, z_j \text{ for all } j \in \mathrm{qa}\left(i\right)\right), \tag{21}$$

$$\mathrm{P}_{\theta}\left(z_i^{k_i}\middle|z_{\mathrm{pa}(i)}^{\mathbf{k}_{\mathrm{pa}(i)}}\right) = \mathrm{P}_{\theta}\left(z_i^{k_i}\middle|z_j^{k_j} \text{ for all } j \in \mathrm{pa}\left(i\right)\right) \tag{22}$$

Note that in Eq. (21), we allow for the possibility of a slightly more general form for the approximate posterior, where the distribution over $z_i^{k_i}$ may depend on any of the parent samples. This generalisation ensures that the subsequent derivations generalise to other possible forms for the approximate posterior, such as those for TMC (Eq. 11).

In addition, it is useful to introduce notation to describe the "non-indexed" latent variables (i.e. everything in z that is not $z^{\mathbf{k}}$). The ith non-indexed latents are, $z_i^{/\mathbf{k}_i}$,

$$z_i^{/k_i} = \left(z_i^1, \ldots, z_i^{k_i-1}, z_i^{k_i+1}, \ldots, z_i^K\right) \in \mathcal{Z}_i^{K-1}. \tag{23}$$

and $z^{/\mathbf{k}}$ are all non-indexed latents,

$$z^{/\mathbf{k}} = \left(z_1^{/k_1}, z_2^{/k_2}, \ldots, z_n^{/k_n}\right) \in \mathcal{Z}^{K-1}. \tag{24}$$

3.1 IWAE

3.1.1 Single-Sample VI

We begin by building intuition by looking at the derivation for the ELBO in the standard single-sample VAE. We start by writing the marginal likelihood as an integral,

$$\mathrm{P}_{\theta}\left(x\right) = \int dz' \, \mathrm{P}_{\theta}\left(x, z'\right). \tag{25}$$

Here, we use $z' \in \mathcal{Z}$ to denote a single sample from the full joint state space; we use z' instead of z because z is reserved for K samples (Eq. 2). Next, we divide and multiply by the approximate posterior probability, $\mathrm{Q}_{\phi}\left(z'|x\right)$,

$$\mathrm{P}_{\theta}\left(x\right) = \int dz' \mathrm{Q}_{\phi}\left(z'|x\right) \frac{\mathrm{P}_{\theta}\left(x, z'\right)}{\mathrm{Q}_{\phi}\left(z'|x\right)}. \tag{26}$$

Now, we can rewrite the integral as an expectation under the approximate posterior,

$$\mathrm{P}_\theta(x) = \mathrm{E}_{\mathrm{Q}_\phi(z'|x)}\left[\frac{\mathrm{P}_\theta(x,z')}{\mathrm{Q}_\phi(z'|x)}\right]. \tag{27}$$

Now we take the logarithm on both sides and apply Jensen's inequality,

$$\log \mathrm{P}_\theta(x) = \log \mathrm{E}_{\mathrm{Q}_\phi(z'|x)}\left[\frac{\mathrm{P}_\theta(x,z')}{\mathrm{Q}_\phi(z'|x)}\right] \geq \mathrm{E}_{\mathrm{Q}_\phi(z'|x)}\left[\log \frac{\mathrm{P}_\theta(x,z')}{\mathrm{Q}_\phi(z'|x)}\right] = \mathcal{L}_{\text{VAE}} \tag{28}$$

Of course, this derivation is specific to the single-sample VAE. But we can pull out an underlying strategy that generalises to the multi-sample setting. In particular, we first come up with an unbiased estimator of the marginal likelihood. In our VAE, this is,

$$\mathcal{P}_{\text{VAE}}(z') = \frac{\mathrm{P}_\theta(x,z')}{\mathrm{Q}_\phi(z'|x)} \tag{29}$$

Following Eq. (27) we can see that this quantity is an unbiased estimator of the marginal likelihood if z' is sampled from $\mathrm{Q}_\phi(z'|x)$,

$$\mathrm{P}_\theta(x) = \mathrm{E}_{\mathrm{Q}_\phi(z'|x)}\left[\mathcal{P}_{\text{VAE}}(z')\right] \tag{30}$$

Then we apply Jensen's inequality (Eq. 28),

$$\log \mathrm{P}_\theta(x) \geq \mathcal{L}_{\text{VAE}} = \mathrm{E}_{\mathrm{Q}_\phi(z'|x)}\left[\log \mathcal{P}_{\text{VAE}}(z')\right]. \tag{31}$$

However, this approach highlights key issues with the usual single-sample bound. In particular, the single-sample estimator, $\mathcal{P}_{\text{VAE}}(z')$ can be very high-variance, and variance in the unbiased estimator causes the Jensen bound to become looser.

3.1.2 Global IWAE

To reduce variance in the unbiased estimator, a natural approach is to average K independent samples, and this is exactly what global IWAE does,

$$\mathcal{P}_{\text{global}}(z) = \frac{1}{K}\sum_{k=1}^{K} r_k(z) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{P}_{\text{VAE}}(z^k) \tag{32}$$

This is of course an unbiased estimator, as it is the average of K unbiased estimators,

$$\mathrm{P}_\theta(x) = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\mathcal{P}_{\text{global}}(z)\right]. \tag{33}$$

Therefore, applying Jensen's inequality gives a new lower-bound on the log-marginal likelihood,

$$\log \mathrm{P}_\theta(x) = \log \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\mathcal{P}_{\text{MP}}(z)\right] \geq \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\log \mathcal{P}_{\text{global}}(z)\right] = \mathcal{L}_{\text{IWAE}} \tag{34}$$

which is tighter than the usual single-sample ELBO [Burda et al., 2015], and which matches Eq. (6) in the main text.

3.1.3 Massively Parallel IWAE

Our proposed $\mathcal{P}_{\text{MP}}(z)$ (Eq. 15) is the average of K^n terms, rather than K terms in global IWAE. To prove that our massively parallel strategy is valid, our strategy is to show that every term in this average is an unbiased estimator of $\log \mathrm{P}_\theta(x)$, in which case the average is also an unbiased estimator, and we can again apply Jensen.

Each term in the average $\mathcal{P}_{\text{MP}}(z)$ (Eq. 15) is of the form $r_{\mathbf{k}}(z)$ (Eq. 16). The expectation of each term is,

$$\mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[r_{\mathbf{k}}(z)\right] = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\mathrm{P}_\theta\left(x, z^{\mathbf{k}}\right)}{\prod_i \mathrm{Q}_\phi\left(z_i^{k_i}|x, z_{\text{qa}(i)}\right)}\right]. \tag{35}$$

We can rewrite the expectation as an integral,

$$\mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[r_{\mathbf{k}}(z)\right] = \int dz\, \mathrm{P}_\theta\left(x, z^{\mathbf{k}}\right) \prod_i \frac{\mathrm{Q}_\phi\left(z_i|x, z_{\mathrm{qa}(i)}\right)}{\mathrm{Q}_\phi\left(z_i^{k_i}|x, z_{\mathrm{qa}(i)}\right)}. \tag{36}$$

Bayes theorem tells us,

$$\frac{\mathrm{Q}_\phi\left(z_i|x, z_{\mathrm{qa}(i)}\right)}{\mathrm{Q}_\phi\left(z_i^{k_i}|x, z_{\mathrm{qa}(i)}\right)} = \frac{\mathrm{Q}_\phi\left(z_i^{k_i}, z_i^{/k_i}|x, z_{\mathrm{qa}(i)}\right)}{\mathrm{Q}_\phi\left(z_i^{k_i}|x, z_{\mathrm{qa}(i)}\right)} = \mathrm{Q}\left(z_i^{/k_i}|x, z_i^{k_i}, z_{\mathrm{qa}(i)}\right), \tag{37}$$

Applying Bayes theorem,

$$\mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[r_{\mathbf{k}}(z)\right] = \int dz\, \mathrm{P}_\theta\left(x, z^{\mathbf{k}}\right) \prod_i \mathrm{Q}\left(z_i^{/k_i}|x, z_i^{k_i}, z_{\mathrm{qa}(i)}\right). \tag{38}$$

Importantly, the integrand is a valid joint distribution over x and z, or equivalently over x, $z^{\mathbf{k}}$ and $z^{/\mathbf{k}}$. Thus, integrating over $z^{/\mathbf{k}}$ then $z^{\mathbf{k}}$, we find,

$$\mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[r_{\mathbf{k}}(z)\right] = \mathrm{P}_\theta\left(x\right). \tag{39}$$

As such, each of the $r_{\mathbf{k}}(z)$ terms is an unbiased estimator of the marginal likelihood. As $\mathcal{P}_{\mathrm{MP}}(z)$ (Eq. 15) is just an average of K^n $r_{\mathbf{k}}(z)$ terms, it is also an unbiased estimator. Applying Jensen's inequality to this unbiased estimator,

$$\log \mathrm{P}_\theta\left(x\right) \geq \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\log \mathcal{P}_{\mathrm{MP}}(z)\right] = \mathcal{L}_{\mathrm{MP}}, \tag{40}$$

which mirrors Eq. (17) in the main text.

3.2 RWS

3.2.1 Global RWS

To build intuition, we first give a derivation of the standard RWS updates. Ideally the updates would use samples drawn from the true posterior, $\mathrm{P}_\theta\left(z|x\right)$,

$$\Delta\theta_{\mathrm{post}} = \mathrm{E}_{\mathrm{P}_\theta(z^k|x)}\left[\nabla_\theta \log \mathrm{P}_\theta\left(z, x\right)\right] \tag{41a}$$

$$\Delta\phi_{\mathrm{post}} = \mathrm{E}_{\mathrm{P}_\theta(z^k|x)}\left[\nabla_\phi \log \mathrm{Q}_\phi\left(z|x\right)\right] \tag{41b}$$

The P update is exactly the M-step in EM, and the Q step trains $\mathrm{Q}_\phi\left(z|x\right)$ using maximum likelihood based on samples from the true posterior. To simplify the derivations, we note that both of these updates can be understood as computing a moment under the true posterior,

$$\Delta_{\mathrm{post}} = \mathrm{E}_{\mathrm{P}_\theta(z^k|x)}\left[\Delta(z^k)\right]. \tag{42}$$

For the P update, we have $\Delta_{\mathrm{post}} = \Delta\theta_{\mathrm{post}}$ and $\Delta(z^k) = \nabla_\theta \log \mathrm{P}_\theta\left(z, x\right)$. For the Q update, we have $\Delta_{\mathrm{post}} = \Delta\phi_{\mathrm{post}}$ and $\Delta(z^k) = \nabla_\theta \log \mathrm{Q}_\phi\left(z, x\right)$. Of course, in practice, the true posterior is intractable, so instead we must use some form of importance weighting. We begin by writing the generic form for the updates as an integral,

$$\Delta_{\mathrm{post}} = \int dz^k\, \mathrm{P}\left(z^k|x\right) \Delta(z^k). \tag{43}$$

We then multiply and divide by an approximate posterior, $\mathrm{Q}\left(z^k|x\right)$,

$$\Delta_{\mathrm{post}} = \int dz^k \mathrm{Q}\left(z^k|x\right) \frac{\mathrm{P}\left(z^k|x\right)}{\mathrm{Q}\left(z^k|x\right)} \Delta(z^k). \tag{44}$$

We can rewrite the integral as expectation over the approximate posterior, $\mathrm{Q}\left(z^k|x\right)$,

$$\Delta_{\text{post}} = \mathrm{E}_{\mathrm{Q}(z^k|x)}\left[\frac{\mathrm{P}\left(z^k|x\right)}{\mathrm{Q}\left(z^k|x\right)}\Delta(z^k)\right]. \tag{45}$$

This quantity is difficult to use directly, because computing the posterior, $\mathrm{P}\left(z^k|x\right)$ involves the marginal likelihood, $\mathrm{P}_\theta\left(x\right)$, which is intractable,

$$\mathrm{P}_\theta\left(z^k|x\right) = \frac{\mathrm{P}_\theta\left(z^k,x\right)}{\mathrm{P}_\theta(x)} \qquad\qquad \mathrm{P}_\theta\left(x\right) = \int dz^k\, \mathrm{P}_\theta\left(z^k, x\right). \tag{46}$$

As the true marginal likelihood is intractable, we instead use $\mathcal{P}_{\text{global}}(z)$ (Eq. 4), which is an unbiased estimator of $\mathrm{P}\left(x\right)$, and is correct in the limit as $K \to \infty$ [Burda et al., 2015]. This gives updates of the form,

$$\Delta_{\text{global}} = \mathrm{E}_{\mathrm{Q}(z^k|x)}\left[\frac{\frac{\mathrm{P}\left(z^k,x\right)}{\mathrm{Q}(z^k|x)}}{\mathcal{P}_{\text{global}}(z)}\Delta(z^k)\right]. \tag{47}$$

Remembering the definition of $r_k(z)$ (Eq. 5), this can be written,

$$\Delta_{\text{global}} = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{r_k(z)}{\mathcal{P}_{\text{global}}(z)}\Delta(z^k)\right]. \tag{48}$$

Finally as the expectation is the same for all k, we can average over k, which gives the expression in the main text (Eq. 7)

3.2.2 Massively Parallel RWS

Now, we can move on to massively parallel RWS. In the previous derivation for global RWS, we showed that each sample, z^k, individually constituted an unbiased estimator. In the massively parallel setting, the key difference is that instead of having K samples z^k, we have K^n samples, $z^{\mathbf{k}}$. In particular,

$$\Delta_{\text{post}} = \mathrm{E}_{\mathrm{P}(z^{\mathbf{k}}|x)}\left[\Delta(z^{\mathbf{k}})\right] = \int dz^{\mathbf{k}}\, \mathrm{P}\left(z^{\mathbf{k}}|x\right)\Delta(z^{\mathbf{k}}). \tag{49}$$

Now, we multiply and divide by $\prod_i \mathrm{Q}\left(z_i^{k_i}\middle|x, z_{\text{qa}(i)}\right)$,

$$\Delta_{\text{post}} = \int dz^{\mathbf{k}}\left(\prod_i \mathrm{Q}\left(z_i^{k_i}\middle|x, z_{\text{qa}(i)}\right)\right)\frac{\mathrm{P}\left(z^{\mathbf{k}}|x\right)}{\prod_i \mathrm{Q}\left(z_i^{k_i}\middle|x, z_{\text{qa}(i)}\right)}\Delta(z^{\mathbf{k}}). \tag{50}$$

Now, we introduce and integrate out a distribution over the non-indexed latent variables, $\prod_i \mathrm{Q}\left(z_i^{/k_i}\middle|x, z_i^{k_i}, z_{\text{qa}(i)}\right)$

$$1 = \int dz^{/\mathbf{k}}\prod_i \mathrm{Q}\left(z_i^{/k_i}\middle|x, z_i^{k_i}, z_{\text{qa}(i)}\right), \tag{51}$$

Multiplying Eq. (50) by 1 (Eq. 51),

$$\Delta_{\text{post}} = \int dz^{\mathbf{k}}\left(\prod_i \mathrm{Q}\left(z_i^{k_i}\middle|x, z_{\text{qa}(i)}\right)\right)\frac{\mathrm{P}\left(z^{\mathbf{k}}|x\right)}{\prod_i \mathrm{Q}\left(z_i^{k_i}\middle|x, z_{\text{qa}(i)}\right)}\Delta(z^{\mathbf{k}})\int dz^{/\mathbf{k}}\prod_i \mathrm{Q}\left(z_i^{/k_i}\middle|x, z_i^{k_i}, z_{\text{qa}(i)}\right). \tag{52}$$

Combining the integrals over $z^{\mathbf{k}}$ and $z^{/\mathbf{k}}$ into a single integral over z,

$$\Delta_{\text{post}} = \int dz\mathrm{Q}\left(z|x\right)\frac{\mathrm{P}\left(z^{\mathbf{k}}|x\right)}{\prod_i \mathrm{Q}\left(z_i^{k_i}\middle|x, z_{\text{qa}(i)}\right)}\Delta(z^{\mathbf{k}}). \tag{53}$$

Writing the integral as an expectation,

$$\Delta_{\text{post}} = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\mathrm{P}\left(z^{\mathbf{k}}|x\right)}{\prod_i \mathrm{Q}\left(z_i^{k_i}|x, z_{\text{pa}(i)}\right)}\Delta(z^{\mathbf{k}})\right]. \tag{54}$$

Again, the posterior can be written,

$$\mathrm{P}_\theta\left(z^{\mathbf{k}}|x\right) = \frac{\mathrm{P}_\theta\left(z^{\mathbf{k}}, x\right)}{\mathrm{P}_\theta(x)} \qquad \mathrm{P}_\theta(x) = \int dz^{\mathbf{k}}\, \mathrm{P}_\theta\left(z^{\mathbf{k}}, x\right). \tag{55}$$

Again, the marginal likelihood, $\mathrm{P}_\theta(x)$ is intractable. Instead, we use the massively parallel estimate of the marginal likelihood, which was shown to be unbiased in Sec. 3.1.3,

$$\Delta_{\text{MP}} = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{\frac{\mathrm{P}\left(z^{\mathbf{k}}, x\right)}{\prod_i \mathrm{Q}\left(z_i^{k_i}|x, z_{\text{pa}(i)}\right)}}{\mathcal{P}_{\text{MP}}(z)}\Delta(z^{\mathbf{k}})\right]. \tag{56}$$

Remembering the definition of $r_{\mathbf{k}}(z)$ (Eq. 16), this can be written,

$$\Delta_{\text{MP}} = \mathrm{E}_{\mathrm{Q}_\phi(z|x)}\left[\frac{r_{\mathbf{k}}(z)}{\mathcal{P}_{\text{MP}}(z)}\Delta(z^{\mathbf{k}})\right]. \tag{57}$$

Finally, note that the expectation is the same for every value of $\mathbf{k}$. Averaging over all K^n values of $\mathbf{k}$, we get the form in the main text (Eq. 18).

3.3 MOVIELENS GRAPHICAL MODEL

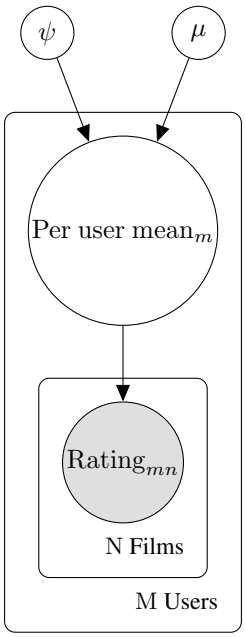

Figure 1: Graphical model for the MovieLens dataset

3.4 BUS DELAY MODEL SPECIFICATION

$$
\begin{aligned}
\text{YearVariance} &\sim \text{Cat}([0.1, 0.5, 0.4, 0.05, 0.05]) \\
\text{YearMean} &\sim \mathcal{N}(0, 10^{-4}) \\
\text{BoroughMean}_m &\sim \mathcal{N}(\text{YearMean}, \exp(\text{YearVariance})),\ m = 1, ..., \text{M} \\
\text{BoroughVariance}_j &\sim \text{Cat}([0.1, 0.4, 0.05, 0.5, 0.05]), j = 1, ..., \text{J} \\
\text{IdMean}_{mj} &\sim \mathcal{N}(\text{BoroughMean}_m, \text{BoroughVariance}_j),\ j = 1, ..., \text{J},\ m = 1, ..., \text{M} \\
\text{WeightVariance}_i &\sim \text{Cat}([0.1, 0.4, 0.5, 0.05, 0.05]),\ i = 1, ..., \text{I} \\
\mathbf{C}_i &\sim \mathcal{N}(\mathbf{0}_{\#\text{BusCo.s}}, \text{WeightVariance}_i),\ i = 1, ..., \text{I} \\
\mathbf{J}_i &\sim \mathcal{N}(\mathbf{0}_{\#\text{JourneyTypes}}, \text{WeightVariance}_i),\ i = 1, ..., \text{I} \\
\text{logits}_{mji} &= \text{IdMean}_{mj} + \mathbf{C}_i * \text{Bus company name}_{mji} + \mathbf{J}_i * \text{Journey type}_{mji} \\
\text{Delay}_{mji} &\sim \text{NegativeBinomial}(\text{total count} = 130, \text{logits}_{mji}),\ i = 1, ..., \text{I},\ j = 1, ..., \text{J}, \boldsymbol{\mu} = 1, ..., \text{M}
\end{aligned}
\tag{58}
$$

Where $\text{Bus company name}_{mji}$ is a one-hot encoded indicator variable indicating which bus company was running that route, and $\text{Journey type}_{mji}$ similarly indicates which kind of bus journey was being undertaken. A total county of 130 is chosen as this is the largest recorded delay in the dataset.

3.5 BUS BREAKDOWN GRAPHICAL MODEL

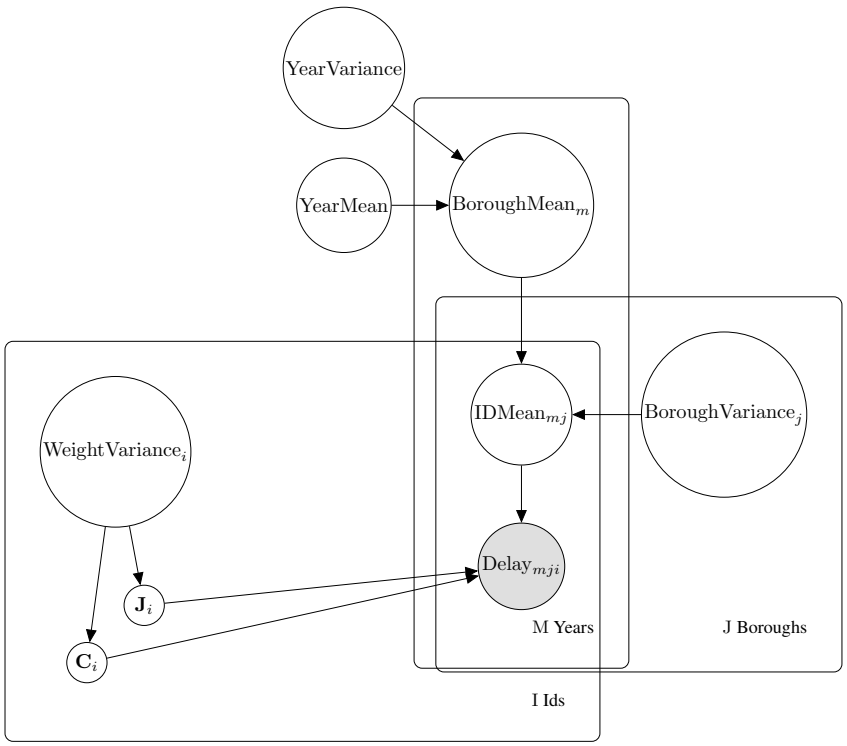

Figure 2: Graphical model for the bus breakdown dataset