Massively Parallel Reweighted Wake-Sleep

Thomas Heap, Gavin Leech, Laurence Aitchison

Department of Computer Science, University of Bristol, UK

Premise

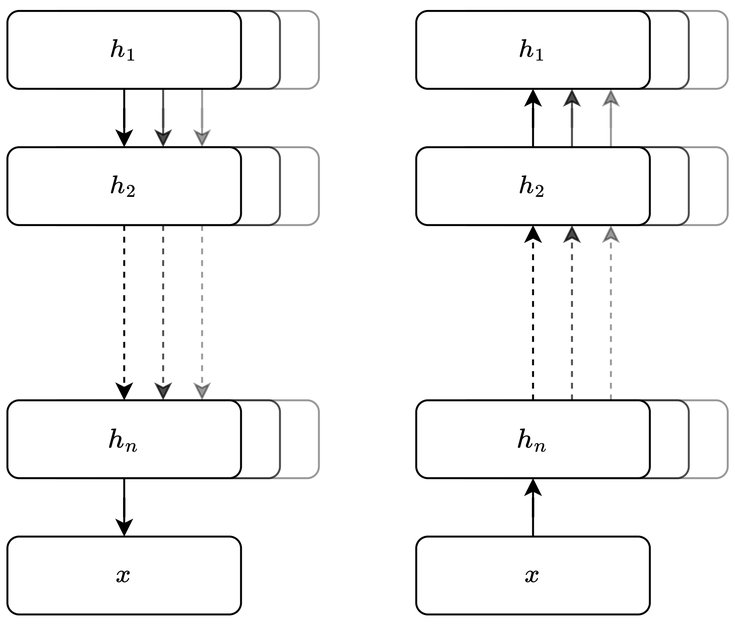
Reweighted wake-sleep (RWS) can perform Bayesian inference in a very general class of models. However, the number of samples required for effective importance weighting is exponential in the number of latents; getting this many importance samples is intractable in all but the smallest models. Here, we develop *massively parallel RWS*, which circumvents this issue by drawing K samples of all n latent variables and reasoning about all K^n possible combinations of samples.

Reasoning about K^n combinations might seem intractable, but the required computations can be performed in polynomial time by exploiting conditional independencies. Our algorithm and implementation of MPRWS show considerable improvements over standard RWS, which draws K samples from the full joint.

Background

We seek the the parameters, ϕ, of an approximate posterior $\mathrm{Q}_\phi(\mathbf{h}|\mathbf{x})$ with latents $\mathbf{h}$ and data $\mathbf{x}$.

RWS [2] draws K samples from an underlying approximate posterior, then uses importance weighting to provide a better estimate of the true posterior P_θ. RWS then updates its approximate posterior towards the importance-weighted estimate of the true posterior.



Standard reweighted wake-sleep. Draws K samples from the joint latent space.

However, naively importance weighting, as we do in methods like RWS, does not provide enough samples for accurate estimation of the posterior [3], making practical applications of this method difficult.

In past work [1] we circumvent this problem with a massively parallel scheme drawing K samples for each latent variable, then *effectively* obtaining K^n samples by considering all combinations of K samples for each of the n latents. Let each individual sample for each separate latent variable be h_i^k, where k indexes the sample and i indexes the latent variable.

To sample all K copies of the full joint latent space, TMC [1] uses an IID distribution over the K samples, $z_i^1, \ldots, z_i^K$,

$$\mathrm{Q}_{\mathrm{TMC}}(h|x) = \prod_{i=1}^{n} \prod_{k \in \mathcal{K}} \mathrm{Q}_{\mathrm{TMC}}\left(h_i^k \middle| h_j \text{ for all } j \in \mathrm{qa}(i)\right). \quad (1)$$

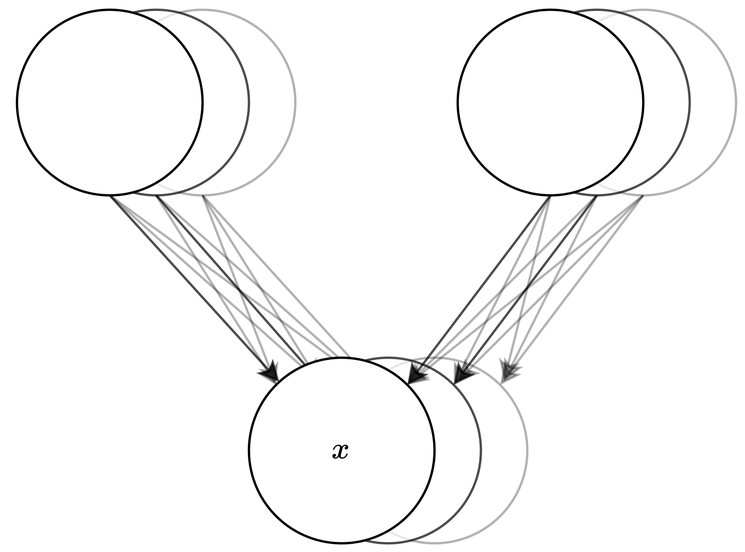
$\mathrm{qa}(i) =$ indices of parents of the ith latent variable under the approximate posterior.

The present work generalises this, allowing dependencies (in Q) between the K samples for a single latent, and deriving parallel update rules for an RWS which averages over combinations of samples.

$$\mathrm{Q}_{\mathrm{MP}}(h|x) = \prod_{i=1}^{n} \mathrm{Q}_{\mathrm{MP}}(h_i|h_j \text{ for all } j \in \mathrm{qa}(i)). \quad (2)$$

There are no formal constraints on these dependencies. However, there are practical constraints, namely that we need to be able to efficiently compute the single-particle marginals

Methods



Here, $K = 3$ and $n = 2$, so we reason over $K^n = 9$ combinations of samples.

Algorithm 1: Massively Parallel RWS

Require: Data x, Prior P_θ, Proposal Q_{MP}, $K \geq 1$

for $i \leftarrow 1$ to n **do**

Sample $z_i \sim \mathrm{Q}_{\mathrm{MP}}(z_i|z_j \text{ for all } j \in \mathrm{qa}(i))$

$z \leftarrow \{z_1, \ldots, z_{i-1}\} \cup z_i$

$f^i_{k_i, \mathbf{k}_{\mathrm{pa}(i)}}(z) \leftarrow \frac{\mathrm{P}_\theta\left(z_i^{k_i} \middle| z_j^{k_j} \text{ for all } j \in \mathrm{pa}(i)\right)}{\mathrm{Q}_{\mathrm{MP}}\left(z_i^{k_i} | x, z_j \text{ for all } j \in \mathrm{qa}(i)\right)}$

end for

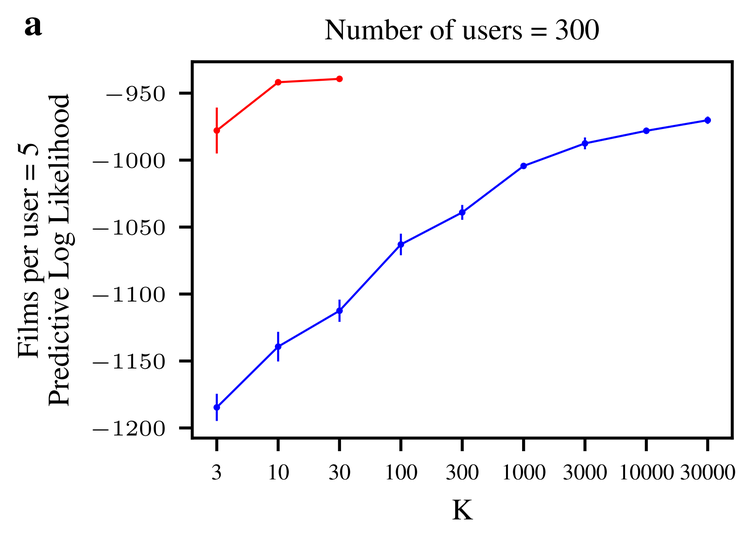
$f^x_{\mathbf{k}_{\mathrm{pa}(x)}}(z) \leftarrow \mathrm{P}_\theta\left(x \middle| z_j^{k_j} \text{ for all } j \in \mathrm{pa}(x)\right)$

$\mathcal{P}_{\mathrm{MP}}(z) \leftarrow \frac{1}{K^n} \sum_{\mathbf{k}^n} f^x_{\mathbf{k}_{\mathrm{pa}(x)}}(z) \prod_i f^i_{k_i, \mathbf{k}_{\mathrm{pa}(i)}}(z)$

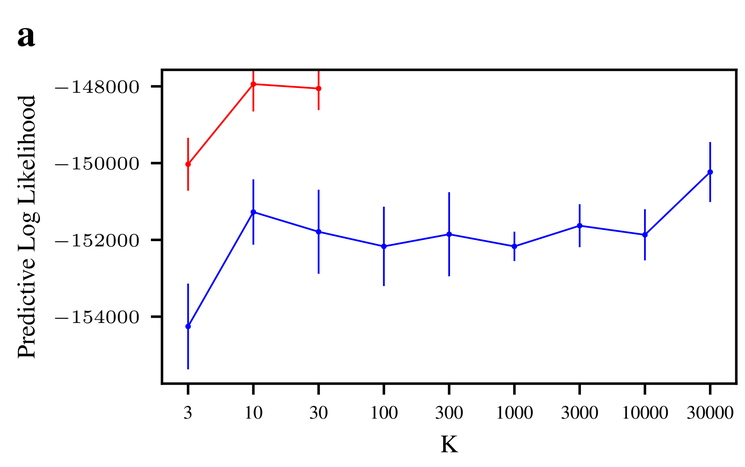
$\Delta\theta_{\mathrm{MP}} \leftarrow \nabla_\theta \log \mathcal{P}_{\mathrm{MP}}(z)$

$\Delta\phi_{\mathrm{MP}} \leftarrow \nabla_\phi(-\log \mathcal{P}_{\mathrm{MP}}(z))$

Results



MPRWS vs global RWS on the `movielens` dataset.



MPRWS vs global RWS on the NYC bus breakdown dataset.

References

[1] Laurence Aitchison. "Tensor Monte Carlo: particle methods for the GPU era". In: *Advances in Neural Information Processing Systems* 32 (2019).

[2] Jörg Bornschein and Yoshua Bengio. "Reweighted wake-sleep". In: *arXiv preprint arXiv:1406.2751* (2014).

[3] Sourav Chatterjee and Persi Diaconis. "The sample size required in importance sampling". In: *The Annals of Applied Probability* 28.2 (2018), pp. 1099–1135.