Variational Energy Conserving Subsampling

Ola Rønning [1], Thomas Hamelryck[1]

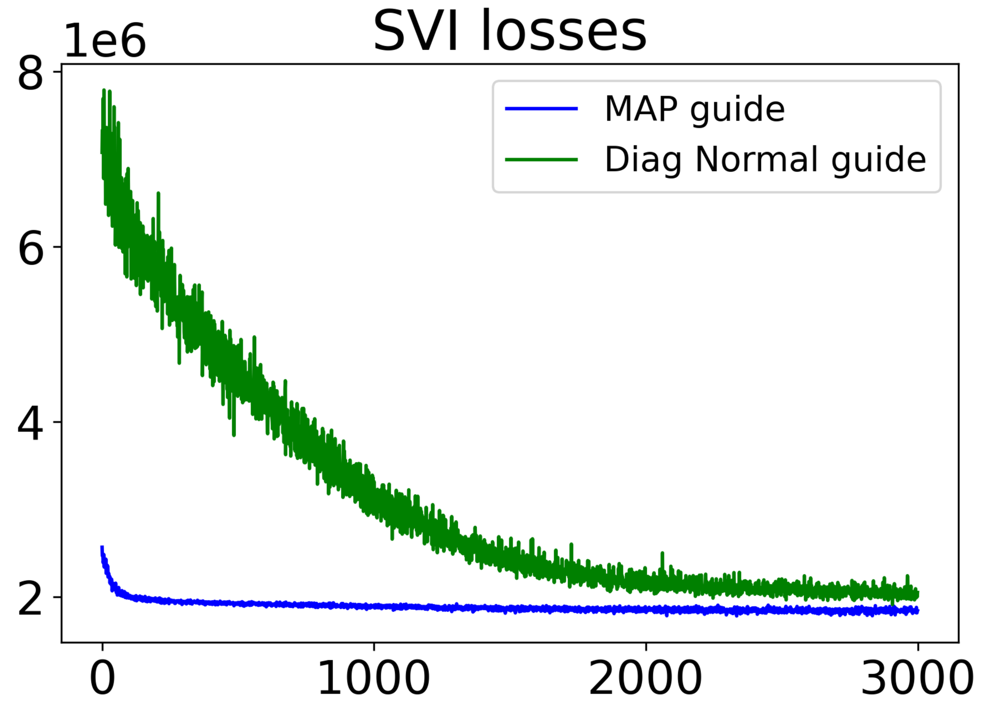
[1] University of Copenhagen, Copenhagen, Denmark.

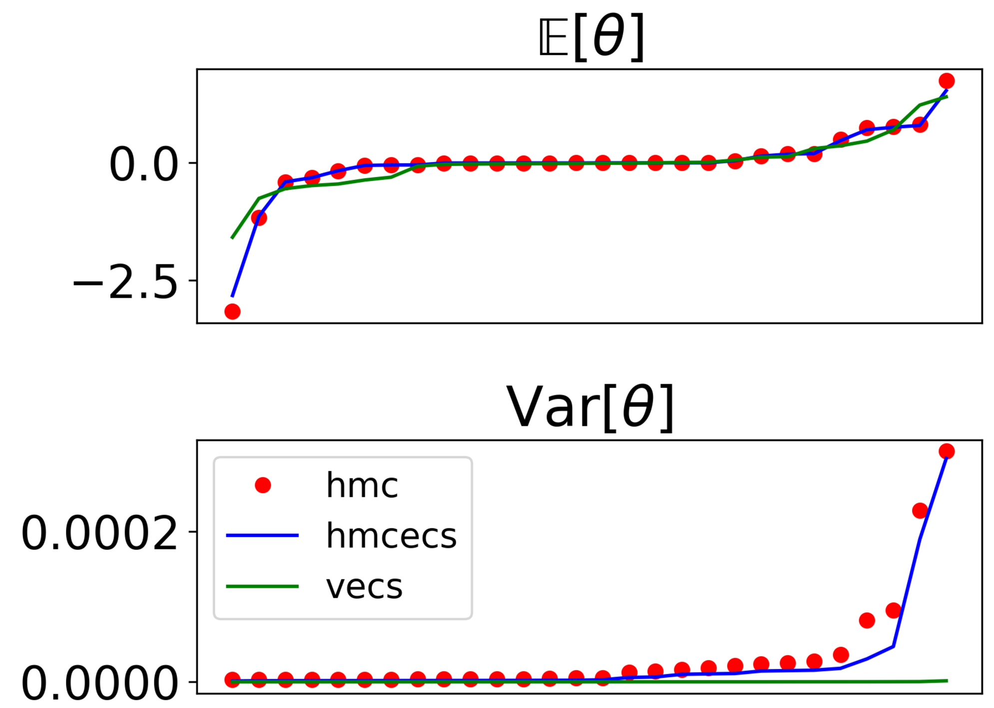
Introduction

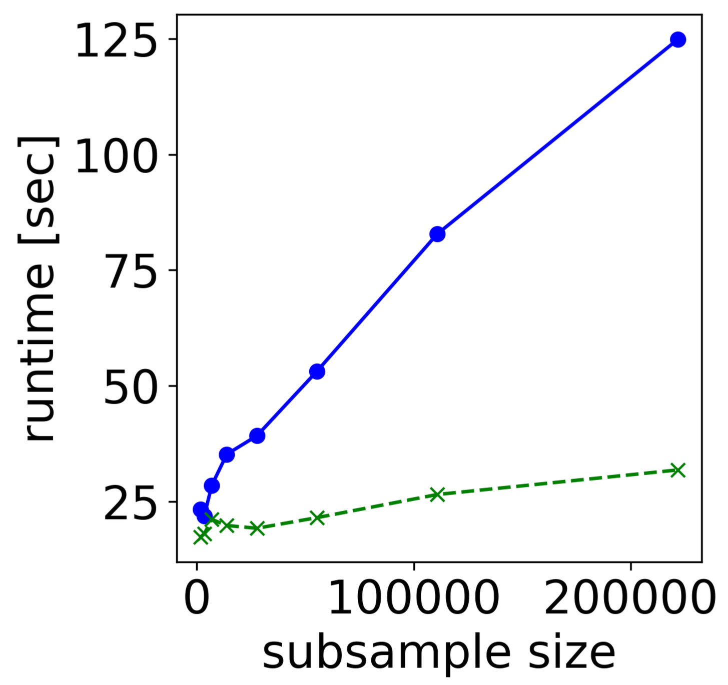
Efficient and accurate Bayesian inference using Markov Chain Monte Carlo methods in the big data case is an open challenge. Recently, [1] proposed Energy Conserving Subsampling (ECS) for Bayesian inference with Hamiltonian Monte Carlo (HMC) on subsets of data. However, their method for subsampling requires computing a memory intensive control variate based on a second order Taylor expansion. To avoid this, we propose a novel control variate we call Variational ECS (VECS). VECS uses an approximate posterior for likelihood estimation, thus preventing the quadratic expansion of feature space. *However, VECS currently fails to effectively incorporate the state of the Markov chain.*

Experiments

For experimentation we use logistic regression on a subset (1.5M examples) of the Higgs dataset.







Perturbed method

For observations $\{x_i\}_n$ and an additive log likelihood ($\log p(x|\theta) = \sum_{i=1}^n \log p(x_i|\theta)$) [2] showed we can produce an unbiased estimate of the log likelihood by

$$\log \hat{p}_s(x|\theta) = \sum_{i=1}^{n} q(x_i, \theta) + d_s(\theta),$$

where $s \sim U$ is a simple random subsample of data indices and

$$d_s(\theta) = \frac{1}{|s|} \sum_{i=1}^{|s|} p_{s_i}(x_{s_i}|\theta) - q(x_{s_i}, \theta).$$

[1] uses a taylor expansion as a proxy

$$q(x_i, \theta) = \log p(x_i|\theta^*) + (\nabla_\theta \log p)(x_i|\theta^*)(\theta - \theta^*) + (\theta - \theta^*)^T (\nabla_\theta^2 \log p)(x_i|\theta^*)(\theta - \theta^*),$$

where θ^* is MAP estimate. Note that if we can compute $\sum_i q_{x_i}(\theta)$ in time proportional to $d^s(\theta)$, we obtain an log likelihood estimator with time complexity proportional to the subsample size.
The estimator $\log \hat{p}_s(x|\theta)$ is a biased estimator of the likelihood. [3] first proposed improving this estimator by

$$\hat{L}(\theta) = \exp\left(\hat{l}(\theta) - \frac{1}{2}\hat{\sigma}^2(\theta)\right), \tag{1}$$

which assumes $\log \hat{p}(\theta|x) \sim \mathcal{N}(\log \hat{p}(\theta|x), \sigma^2)$, where σ^2 is the population variance. [2] showed that with

$$\log \hat{\sigma}^2(\theta) = \left(\frac{n}{m}\right)^2 \sum_i \left(d_{s_i}(\theta) - \bar{d}_{s_i}(\theta)\right)^2,$$

and the proxy expanded around the posterior mode (θ^*) the error of the inferred posterior is $O(\frac{1}{nm^2})$.

Variational Energy Conserving Subsampling

Let $p(x, \theta)$ be a latent model with a prior $\pi(\theta)$ and $\mathcal{Q}$ be a family of parametric variational distributions over $\theta \subseteq \mathbb{R}^d$, then we can infer $q^*_\psi(\theta|x) \in \mathcal{Q} \approx p(\theta|x)$ using VI by optimizing ψ. So we have

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \approx q^*_\psi(\theta|x) \implies p(x|\theta) \approx \frac{p(x)q^*_\psi(\theta|x)}{\pi(\theta)} \implies \log p(x|\theta) \approx \log p(x) + \log q^*_\psi(\theta|x) - \log \pi(\theta) \tag{2}$$

From the evidence lower bound (ELBO) and assuming an additive log-likelihood we have

$$\log p(x) \geq \mathbb{E}_{\phi \sim q^*_\psi}[\overbrace{\sum_i \log p(x_i|\phi)}^{\log p(x|\phi)} + \log \pi(\phi) - \log q^*_\psi(\phi|x)]. \tag{3}$$

Substituting Equation (3) into Equation (2) and considering the contribution of $x_i \in x$ we obtain

$$\log p(x_i|\theta) \geq \mathbb{E}_{\phi \sim q^*_\psi}[\log p(x_i|\phi)] + \mathbb{E}_{\phi \sim q^*_\psi}[\log \pi(\phi)] - \log \pi(\theta) + \log q^*_\psi(\theta|x) - \mathbb{E}_{\phi \sim q^*_{\psi_x}}[q^*_\psi(\phi|x)] \equiv q(x_i, \theta) \tag{4}$$

which is our variational likelihood proxy.
We can precompute $(\forall x_i) q^*_{\psi_x}[\log p(x_i|\phi)]$ to make the complexity of $\sum_i q(x_i, \theta)$ proportional to $d_s(\theta)$.

References

[1] Khue-Dung Dang, Matias Quiroz, Robert Kohn, Tran Minh-Ngoc, and Mattias Villani. Hamiltonian monte carlo with energy conserving subsampling. *Journal of machine learning research*, 20, 2019.

[2] Matias Quiroz, Robert Kohn, Mattias Villani, and Minh-Ngoc Tran. Speeding up mcmc by efficient data subsampling. *Journal of the American Statistical Association*, 2018.

[3] DM Ceperley and Mark Dewing. The penalty method for random walks with uncertain energies. *The Journal of chemical physics*, 110(20):9812–9820, 1999.

Contact:
ola@di.ku.dk

UNIVERSITY OF COPENHAGEN