Quasi-Monte Carlo Flows

Florian Wenzel[*]
TU Kaiserslautern
Germany
wenzelfl@hu-berlin.de

Alexander Buchholz[*]
ENSAE-CREST, Paris
France
alexander.buchholz@ensae.fr

Stephan Mandt
Univ. of California, Irvine
USA
mandt@uci.edu

Abstract

Normalizing flows provide a general approach to construct flexible variational posteriors. The parameters are learned by stochastic optimization of the variational bound, but inference can be slow due to high variance of the gradient estimator. We propose Quasi-Monte Carlo (QMC) flows which reduce the variance of the gradient estimator by one order of magnitude. First results show that QMC flows lead to faster inference and samples from the variational posterior cover the target space more evenly.

1 Introduction

Variational inference constructs approximations to intractable target distributions by solving an optimization problem. The introduction of the reparametrization gradient and the score function gradient (Kingma and Welling, 2013; Rezende et al., 2014; Ranganath et al., 2014) enabled the application of variational inference to a variety of complex Bayesian model (e.g. variational autoencoders (Kingma and Welling, 2013) and Bayesian deep neural networks (Blundell et al., 2015; Neal, 2012)).

An important problem in variational inference is to design complex variational families which enable better posterior approximations. Rezende and Mohamed (2015) introduced a flexible family of distributions, called normalizing flows. Normalizing flows transform a base distribution through a number of tractable transformations into more complicated distributions. The parameters of the flow are learned by stochastic optimizing using the reparameterization gradient estimator, but the optimization can be slow due to high noise in the gradient estimator.

In this paper, we introduce Quasi-Monte Carlo (QMC) flows which aim on reducing the variance of the gradient estimator in order to achieve faster inference. We build on the idea of Quasi-Monte Carlo variational inference (QMCVI), recently introduced by Buchholz et al. (2018). QMC flows are constructed by replacing the Monte-Carlo samples of the base distribution in the normalizing flow by samples from a Quasi-Monte Carlo sequence. QMC flows reduce the variance of the gradient estimator by one order of magnitude and lead to posterior samples that cover the target space more evenly.

2 Background

Variational Inference Inference in probabilistic models aims to compute the posterior distribution $p(z|x)$ of the hidden variables z given the data x. Since this distribution is often intractable, the idea behind variational inference (Jordan et al., 1999) is to approximate the posterior by a variational distribution $q_\lambda(z)$. The optimal variational parameters λ are learned by maximizing the so-called evidence lower bound (ELBO), $\mathcal{L}(\lambda) = \mathbb{E}_{q_\lambda(z)}[\log p(x, z) - \log q_\lambda(z)]$. For some models, specific properties of the model can be exploited to obtain closed-form updates to optimize the ELBO (e.g.

[*]equal contributions

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

Jähnichen et al., 2018; Wenzel et al., 2018). Here we focus on a general approach by using the reparameterization gradient estimator.

Normalizing flows In order to obtain a good approximation to the posterior it is crucial to use a rich enough variational distribution. A way of increasing the flexibility of the approximation is by using normalizing flows (Rezende and Mohamed, 2015). Normalizing flows are defined as a transformation of a random variable $z \sim q(z)$ through an invertible mapping $f : \mathbb{R}^d \mapsto \mathbb{R}^d$. The density of $z' = f(z)$ is given as

$$q(z') = q(z) \left| \det \frac{\partial f^{-1}}{\partial z'} \right| = q(z) \left| \det \frac{\partial f}{\partial z} \right|^{-1},$$

using the change of variables formula and the inverse function theorem. When stacking several of these transformations one obtains the normalizing flow

$$\begin{aligned} z_K &= f_K \circ \cdots \circ f_2 \circ f_1(z_0), \\ \log q_K(z_K) &= \log q_0(z_0) - \sum_{k=1}^{K} \log \left| \det \frac{\partial f_k}{\partial z_{k-1}} \right|. \end{aligned} \tag{1}$$

Normalizing flows allow for easy computations of expectations with respect to $q(z_K)$ making use of $\mathbb{E}_{q_K}[h(z)] = \mathbb{E}_{q_0}[h(f_K \circ \cdots \circ f_2 \circ f_1(z_0))]$. Hence only samples from $q(z_0)$ are required. A normalizing flow can lead to arbitrary complex variational distributions $q(z_K)$ if more layers are added.

An example of a simple flow is the planar flow: $f(z) = z + uh(w^t z + b)$, where $\lambda = \{w \in \mathbb{R}^d, u \in \mathbb{R}^d, b \in \mathbb{R}\}$ and h is a smooth element wise non-linearity. Other examples of flows are e.g. radial flows (Rezende et al., 2014), inverse autoregressive flows (Kingma et al., 2016), Sylvester flows (Berg et al., 2018) and neural autoregressive flows (Huang et al., 2018).

All these approaches have in common that they are trained by stochastic optimization of a variational bound (ELBO). The gradients of the variational bound are estimated based on samples from the base distribution $q(z_0)$. In this paper, we replace samples from the base distribution by samples from a Quasi-Monte Carlo (QMC) sequence. This leads to a gradient estimator with lower variance and, therefore, to faster optimization.

Quasi-Monte Carlo We recall now the ideas behind QMC. Low discrepancy sequences, also called QMC sequences, are used for integrating a function ψ over the $[0, 1]^d$ hypercube. When using standard i.i.d. samples on $[0, 1]^d$, the error of the approximation is $\mathcal{O}(N^{-1})$. QMC achieves a rate of convergence in terms of the MSE of $\mathcal{O}\left(N^{-2}(\log N)^{2d-2}\right)$ if ψ is sufficiently regular (Leobacher and Pillichshammer, 2014). This is achieved by a deterministic sequence that covers the hypercube more evenly. From a theoretical perspective the performance of QMC deteriorates with the dimension. However, in practice QMC tends to work well even as the dimension increases.

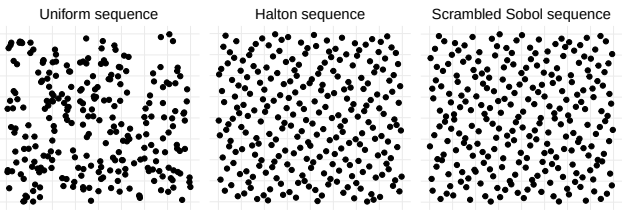

Figure 1: MC (left), QMC (center) and RQMC (right) sequences of length $N = 256$ on $[0, 1]^2$. QMC and RQMC tend to cover the target space more evenly.

On a high level, QMC sequences are constructed such that the number of points that fall in a rectangular volume is proportional to the volume. This idea is closely linked to stratification. Halton sequences e.g. are constructed using coprime numbers (Halton, 1964). Sobol sequences are based on the reflected binary code (Antonov and Saleev, 1979). More details on QMC can be found in Niederreiter (1992); Leobacher and Pillichshammer (2014); Dick et al. (2013).

QMC sequences are deterministic and, therefore, inconvenient for constructing estimators. Carefully reintroducing randomness while preserving the structure of the sequence leads to randomized QMC (RQMC) sequences. RQMC can be used for obtaining unbiased estimators. For illustration purposes, we show different sequences in Figure 1.

QMC or RQMC can be used for integration not only with respect to uniform distributions. To work with other distributions the initial sequence on $[0,1]^d$ is transformed by a function Γ that maps uniform random samples to the desired distribution (e.g. the inverse cumulative distribution function). This function has to be sufficiently smooth. Constructing QMC sequences typically costs $\mathcal{O}(N \log N)$ (Gerber and Chopin, 2015).

3 Quasi-Monte Carlo Flows

We propose Quasi-Monte Carlo (QMC) flows, a flexible class of variational distributions that exhibit low variance for gradient estimation. To construct a QMC flow we take an ordinary normalizing flow (1) and turn it into a QMC flow by replacing the base distribution $q(z_0)$ by a randomized QMC sequence (RQMC).

In normalizing flows, the base distribution $q(z_0)$ is typically a Gaussian distribution (parameterized by a neural network), i.e. $q(z_0) = \mathcal{N}(z_0|\mu(x), \Sigma(x))$ and $\mu(x), \Sigma(x)$ are non-linear functions of the data. Samples from the normalizing flow $z_k \sim q_k(z_k)$ are obtained by first, generating a sample $z_0 \sim q(z_0)$ and then pushing it through the flow (1). The base sample z_0 is generated via the reparameterization trick, $z_0 = r(\epsilon) = \mu + \Sigma^{\frac{1}{2}}\varepsilon$ and $\varepsilon \sim \mathcal{N}(0, I)$.

To obtain a QMC flow we only add a slight change to the generative process. A sequence of samples $\{z_1, \ldots, z_N\}$ from the QMC flow is generated as follows. We first generate a randomized QMC (RQMC) sequence $\{u_1, \ldots, u_N\}$ and apply the flow

$$z_n = f_K \circ \cdots \circ f_2 \circ f_1 \circ r(\tilde{\varepsilon}_n), \tag{2}$$

where $\tilde{\varepsilon}_n = \Phi^{-1}(u_n)$ and Φ^{-1} is the inverse Normal cumulative distribution function. The samples from the QMC flow will approximate the same distribution as samples from the original flow, but have the advantage that the reparameterization gradient estimator exhibits lower variance.

Low variance gradient estimation. The parameters of the normalizing flow are learned by stochastic optimization of the ELBO using the reparameterization gradient (Kingma and Welling, 2013; Rezende et al., 2014),

$$\nabla_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{p(\varepsilon)}[\nabla_\lambda \log p(x, f_\lambda(\varepsilon)) - \nabla_\lambda \log q_K(f_\lambda(\varepsilon)|\lambda)] = \mathbb{E}_{p(\varepsilon)}[g_\lambda(\varepsilon)]. \tag{3}$$

We obtain an estimator of the gradient by approximating this expectation by

$$\nabla_\lambda \mathcal{L}(\lambda \approx \frac{1}{N}\sum_{n=1}^{N} g_\lambda(\tilde{\varepsilon}_n),$$

where the samples $\tilde{\varepsilon}_n$ are based on a RQMC sequence (1). This is in contrast to the standard approach to normalizing flows where the gradient estimator is based on Monte Carlo samples from a Normal distribution, i.e. $\varepsilon_n \sim \mathcal{N}(0, I)$.

Our RQMC based gradient estimator is unbiased and exhibits lower variance than the ordinary MC based gradient estimator. The asymptotic variance of the gradient estimator for standard normalizing flows is $\mathcal{O}(N^{-1})$. Provided sufficient regularity of $g_\lambda(\varepsilon)$, the variance of our gradient estimator is one order of magnitude lower $\mathcal{O}(N^{-2})$. For a proof see Buchholz et al. (2018).

Faster convergence and more diverse samples. Due to reduced gradient noise, QMC flows can be faster learned than standard normalizing flows which we show empirically in the experiments. Buchholz et al. (2018) provide a theoretical analysis and show in the simplified setting of stochastic gradient descent with a constant learning rate, that the RQMC based gradient estimator leads to an asymptotic converge that is one order of magnitude faster than the standard MC based approach.

A second advantage of QMC flows are, that samples pushed through the flow are more diverse, if the transformation via the flow is sufficiently smooth. That means that QMC flow samples cover the target space more evenly (low discrepancy) and estimators based on those samples have lower variance. In the experiments we show that estimating the mean of the target distribution is more accurate using our approach.

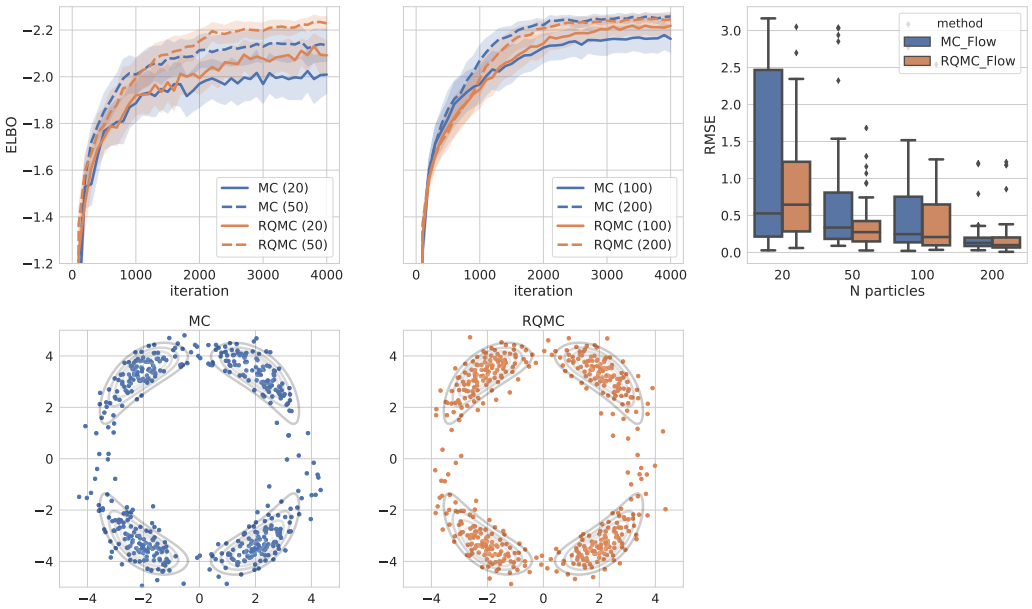

Figure 2: Top row, left and middle: Convergence of the ELBO based on different sample sizes. The experiments have been repeated 50 times to get confidence intervals. The gradient estimators based on RQMC points show a faster convergence. For 200 samples both MC and RQMC behave similarly. Top row, right: RMSE of the mean of the estimators based on MC or RQMC points pushed through the flow. For smaller sample sizes RQMC yields more precise estimators. Bottom row: Samples pushed through the flow. RQMC points cover the target space more evenly.

4 Experiments

We fit a normalizing flow to a 2-dimensional multimodal distribution with an energy function given as

$$U(z) = \frac{1}{2}\left(\frac{||z|| - 2}{0.4}\right) - \log\left(e^{-1/2\left[\frac{z_1-2}{0.6}\right]^2} + e^{-1/2\left[\frac{z_2+2}{0.6}\right]^2}\right).$$

We use a planar flow with 32 layers. The results of the training over 4000 steps using an Adagrad optimizer (Duchi et al., 2011) and different sample sizes for the construction of the stochastic gradient estimator are shown in Figure 2.

For samples sizes of $20, 50, 100$, the RQMC approaches show a faster convergence to a higher value of the ELBO. For a sample size of 200 this effect disappears as the variance of the gradient estimator becomes negligible for both MC and RQMC sampling. A second advantage of QMC for normalizing flows is the fact that points, that get pushed through the flow, allow a more precise estimation of expectations with respect to the transformation.

5 Conclusion

Depending on the setting, QMC may lead to substantial improvements of the optimization procedure required for variational inference with normalizing flows. As the use of QMC comes with almost no computational overhead, our method is an easy way of improving existing implementations.

The concept of QMC flows can also be used in the setting of more involved approaches than the standard normalizing flow, e.g. the inverse autoregressive flow (Kingma et al., 2016) or Sylvester flow (Berg et al., 2018). We will leave the investigation for future studies.

Acknowledgments

We like to thank Marius Kloft for fruitful discussions. This work was partly funded by the German Research Foundation (DFG) award KL 2698/2-1, by the Federal Ministry of Science and Education (BMBF) awards 031B0187B, 01IS18051A and a GENES doctoral research scholarship.

References

Antonov, I. A. and Saleev, V. (1979). An economic method of computing LPτ-sequences. *USSR Computational Mathematics and Mathematical Physics*, 19(1):252–256.

Berg, R. v. d., Hasenclever, L., Tomczak, J. M., and Welling, M. (2018). Sylvester normalizing flows for variational inference. *arXiv preprint arXiv:1803.05649*.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1613–1622.

Buchholz, A., Wenzel, F., and Mandt, S. (2018). Quasi-monte carlo variational inference. In *International Conference on Machine Learning*, pages 667–676.

Dick, J., Kuo, F. Y., and Sloan, I. H. (2013). High-dimensional integration: The quasi-monte carlo way. *Acta Numerica*, 22:133–288.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Gerber, M. and Chopin, N. (2015). Sequential quasi monte carlo. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3):509–579.

Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural autoregressive flows. *arXiv preprint arXiv:1804.00779*.

Jähnichen, P., Wenzel, F., Kloft, M., and Mandt, S. (2018). Scalable generalized dynamic topic models. In *AISTATS*.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4743–4751. Curran Associates, Inc.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Leobacher, G. and Pillichshammer, F. (2014). *Introduction to quasi-Monte Carlo integration and applications*. Springer.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Niederreiter, H. (1992). *Random number generation and quasi-Monte Carlo methods*. SIAM.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pages 814–822.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1278–1286.

Wenzel, F., Galy-Fajou, T., Donner, C., Kloft, M., and Opper, M. (2018). Efficient gaussian process classification using polya-gamma data augmentation. *arXiv*.