Variance Reduction Techniques for Sequential Structured Variational Inference

Pierre Thodoroff PIERRE.THODOROFF@MAIL.MCGILL.CA
Harsh Satija HARSH.SATIJA@MAIL.MCGILL.CA

Abstract

There are a lot of underlying similarities between variational Inference and Reinforcement Learning methods. Both the domains have a similar optimization problem and use the same underlying optimization procedure, stochastic optimization by estimating the gradient from the Monte-Carlo samples. However, both domains have their own set of techniques to counter the high-variance of the Monte-Carlo estimates. The goal of the project is to study the parallelisms in both the domains and analyze them to discover if the techniques of one domain can be carried over to the other or vice-versa.

1. Background

1.1. Variational Inference

Variational inference (VI) is a technique used to approximate posteriors in complex latent variables. (Jordan et al., 1999), (Ghahramani & Beal, 2001), i.e. given a model $p(z)p(x|z)$ with latent variables z and observed variables x, we aim to find the posterior $p(z|x)$. VI methods frame a posterior estimation problem as an optimization problem, where the parameters to be optimized aim to find a member of a family of simple probability distributions that is similar to the true posterior distribution(by minimizing the KL divergence between the both). In VI, we aim to maximize the following function:

$$ELBO(q) = \int q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)} dz$$
$$ELBO(\lambda) = \mathbb{E}_{q(z;\lambda)}[\log p(x,z)] - \mathbb{E}_{q(z;\lambda)}[\log q(z;\lambda)]$$

where maximizing Evidence Lower BOund (ELBO) is equivalent to minimizing the KL between the approximate posterior $q(z|x)$ and the true posterior $p(z|x)$. Here λ represents the parameters of the approximate posterior (for example, it can be weights for a neural net).

1.2. Reinforcement Learning

In Reinforcement Learning (RL), an agent interacts with an environment in a sequential manner, where at each time step it takes an action a, and receives a feedback from the environment in form of a reward r. The goal of the agent is to maximize the expected sum of these rewards over time. The activity of an agent can be represented as a trajectory of s, a and r, as $\tau = (s_1, a_1, r_1, \ldots, s_T, a_T)$. An agent's policy determines the distribution of actions over a state, and is denoted by π_θ, where θ represents the parameters of the policy (again can be considered as weights of a neural net). The objective function then becomes:

$$\begin{aligned}\mathbb{J}(\theta) &= \int \pi_\theta(\tau) R(\tau) d\tau \\ &= \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=1}^{T} r(s_t, a_t) \right]\end{aligned}$$

where, $R(\tau) = \sum_{t=1}^{T} r_t$ is return over the course of entire tarjectory.

2. Optimization Procedure

In this section we'll study the general optimization procedure for the above problems.

2.1. Black-Box variational inference

In the VI literature, stochastic optimization is used to solve the optimization problem with respect to general class of reusable variational families and models (Ranganath et al., 2014). The procedure is based on the stochastic gradient estimates of the ELBO (score function):

$$\begin{aligned}\nabla_\lambda \,\text{ELBO}(\lambda) &= \nabla_\lambda \, \mathbb{E}_{q(\mathbf{z}\,;\,\lambda)} \big[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}\,;\,\lambda) \big] \\ &= \mathbb{E}_{q(\mathbf{z}\,;\,\lambda)} \big[\nabla_\lambda \log q(\mathbf{z}\,;\,\lambda) \big(\log p(\mathbf{x}, \mathbf{z}) \\ &\qquad - \log q(\mathbf{z}\,;\,\lambda) \big) \big]\end{aligned}$$

We compute the noisy unbiased gradients of the ELBO

Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

Algorithm 1 Black Box VI

```
Input: model $\log p(x, z)$, variational approximation $q(z; \lambda)$
Output: Variational parameters: $\lambda$
repeat
    $z[s] \sim q$ // Draw samples from q
    $\rho$ = $t$th value of Robins Monroe Sequence
    $\lambda += \rho * \nabla_\lambda$ ELBO($\lambda$)
    $t += 1$
until not converged
```

with Monte Carlo samples from the variational distribution

$$\nabla_\lambda \text{ ELBO}(\lambda) \approx \frac{1}{S}\sum_{s=1}^{S}\big[\big(\log p(\mathbf{x}, \mathbf{z}_s) - \log q(\mathbf{z}_s\ ;\ \lambda)\big)\nabla_\lambda \log q(\mathbf{z}_s\ ;\ \lambda)\big]$$

The reason it is called Black-Box Variational Inference is because the score function and sampling algorithms depend only on the variational distribution and not on the underlying model. Therefore we can easily experiment with different variational approximations and reuse them for a variety of models. The optimization algorithm is presented in 1

2.2. Policy Gradient

In the RL literature, a similar stochastic optimization procedure of the objective function is carried out (also known as REINFORCE (Williams, 1992), (Sutton et al., 1999)). Monte Carlo estimate of the gradient of expected total reward J, is given by:

$$\begin{aligned}\nabla_\theta &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\tau) R(\tau)\right]\\ &= \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \log \pi_\theta(\tau^{(i)})R(\tau^{(i)})\\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})R(\tau^{(i)})\end{aligned}$$

2.3. High level mapping between RL and VI

There has been quite a significant amount of work which studies control as inference (Dayan & Hinton, 1997), (Furmston & Barber, 2010). More recent works by (Theophane Weber, 2015), (Mnih & Gregor, 2014) studies VI as RL and draws a high-level connection between the two domains, summarized in Table 1 . The underlying principle is that in both the problems, the objective function and their gradients take the same form, $\max \int p_\theta(y) f(y) dy$, with respect to parameters θ of distribution $p_\theta(y)$

2.4. Variance Control techniques in VI and RL

Although Monte Carlo estimates of gradients provide an unbiased estimate, but at the same time they are known to suffer from high variance, which may even in some cases render the learning process unable to learn. For this, both RL and VI domains, have come up with their own set of variance reduction techniques. In the subsequent sections, we'll analyze the most common techniques in both the domains.

3. Control Variate

Control variate is a popular method used to reduce the variance of Monte Carlo estimates. The main idea is to modify our estimates such that the expectation stays the same but the variance decrease (**?**). Let the parameter to estimate μ have a statistics m such that $E[m] = \mu$. Furthermore we define a control variate t such that its expected value $E[t] = \tau$ can be calculated. Let's define $m^* = m + c(t - \tau)$ then :

$$\begin{aligned}&E[m] = E[m^*]\\ &Var(m^*) = Var(m) + c^2 Var(t) + 2^* c^* Cov(m, t)\end{aligned}$$

Solving this system for an optimal c yields :

$$\begin{aligned}c^* &= -\frac{Cov(m, t)}{Var(t)}\\ Var(m^*) &= (1 - Corr(m, t)^2) Var(m)\end{aligned}$$

By choosing a control variate that is strongly correlated with our parameter of interest we can reduce the variance of the estimate. The optimal value for c^* can be approximated using samples.
For variational inference we need to stochastically approximate the gradient of the ELBO.

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial \theta} log_{q_\theta}(z^i|x) log(\frac{p(x|z^i)p(z^i)}{q_\theta(z^i|x)})$$

By carefully picking the control variate, it is possible to significantly reduce the variance of this gradient. Several control variates have been designed both in Reinforcement Learning and Variational Inference. However in a sequential setting the samples obtained are highly correlated. The estimates of a^* might be off because of the dependence between sample. We will see several techniques that attempts to overcome this problem in reinforcement learning. Some of those techniques could be used for structured variational inference in sequential settings.

3.1. Control variate in RL

Control variates is an essential tool of reinforcement learning used to stabilize the gradient of REINFORCE. However

Table 1. RL and VI connection

Generic expectation		RL		VI	
Optimization variable	θ	Policy param	θ	Variational pramam	λ
Integration variable	y	Trajectory	τ	Latent trace	z
Distribution	p_θ(y)	Policy	$\pi_\theta(\tau)$	Posterior distribution	$q_\lambda(z\|x)$
Integrand	$f(y)$	Total return	$R(\tau)$	Negative Free Energy	$\log\left(\frac{p(x,z)}{q_\lambda(z\|x)}\right)$

different name are used for different type of control variate:

- **Baselines** Compare the sampled returns to baseline/reference.

$$\begin{aligned}\nabla_\theta \mathbb{E}[R] &\approx \sum_t \nabla_\theta \log q_\theta(z_t) R_t \\ &\approx \sum_t \nabla_\theta \log q_\theta(z_t)(R_t - b)\end{aligned}$$

  . Note that, adding the baseline still gives us an unbiased estimate of the gradient (as $\nabla_\theta \mathbb{E}[b] = 0$), but helps in reducing the variance. Indeed there exists a correlation between the gradient of the expected reward and the gradient of the policy.

- **Advantage Functions** Instead of using a trivial baseline (like average reward), another practice is to learn the approximation of (b), denoted by $V_\phi(z_t)$. This approximation is based on the state at time-step t and can be learned by using regression on returns.

$$b = V_\phi(z_{t-1}) = \mathbb{E}[R_t|z_{t-1}]$$

- **Bootstrapping/TD-learning** Instead of estimating the value based on single reward per time-step, we can take advantage of sequential nature of the problem and learn it over sequence of returns. This is also referred as temporal difference learning in RL literature. The intuition is that single step estimates can be quite noisy, so we try to estimate the rewards over the sequence of n-step returns.

In RL one can interpret the baseline as a way to isolate the impact of taking an action at that specific time step. It helps with the problem of credit assignment. One can see a connection when in sequential variational inference latent variables are samples from q.

All the above different techniques (baseline, advantage, TD, (Sutton & Barto, 1998)) attempt to learn a control variate correlated with the gradient from un-correlated samples. One of the reason why advantage functions in Atari perform well is because they are trained on a bank of experience (experience replay) breaking the correlation existing between samples allowing for a good estimation of a^*. (Mnih et al., 2015)

Comparing and estimating the variance of those different techniques can be a quite a challenging and interesting problem. By defining the estimation of the gradient as a controlled system and analyzing the mixing properties and sample size (Greensmith et al., 2004) provides an exhaustive comparison of those control variates in a partially observable Markov decision process. This report may contain valuable insights into how to do structured variational inference in a sequential setting.

3.2. Control Variates in VI

Control variate are a common tool used in VI especially in the non sequential setting as an accurate estimation of a^* can be obtained (Ranganath et al., 2014). Control variates has proven to be very effective in the situation where iid samples are available. Most of the work in variational inference makes the mean field assumption on q. However this assumption doesn't hold true in any sequential setting. An interesting line of work would be too study variational inference in a sequential setting with a markovian assumptions on q. For example the work around structured variational inference by (Hoffman, 2014) attempts to relax the mean field assumption on q. This line of work may benefit from control variates techniques from reinforcement learning.

4. Rao-Blackwellization

Rao-Blackwellization (Casella & Robert, 1996) reduces the variance of estimates of random variable by replacing it with its conditional expectation with respect to a subset of the variables. For example, Consider two random variables X and Y a function $J(X, Y)$. Assume, we are interested in $E[(X, Y)]$. Let's define another function $\hat{J}(X)$ as $E[J(X, Y)|X]$. Then we have,

$$\begin{aligned}&E[\hat{J}(X)] = E[J(X, Y)] \\ &Var(\hat{J}(X)) = Var(J(X, Y)) - E[(J(X, Y) - \hat{J}(X))^2] \\ \implies &Var(\hat{J}(X)) < Var(J(X, Y))\end{aligned}$$

The intuition behind this technique is that we want to observe the gradient of sub-components of the model (where we look at the conditional probability on just the markov blanket and not the entire graph), therefore limiting the noise in estimates due to effect of non-contributing components. In (Ranganath et al., 2014), the authors show it can be used in context of VI, where essentially the estimate of gradient becomes:

$$\begin{aligned}\nabla_{\lambda_i} \approx \frac{1}{S}\sum_{s=1}^{S}\big[\big(&\log p_i(\mathbf{x},\mathbf{z}_s) \\ &- \log q_i(\mathbf{z}_s\,;\,\lambda_i)\big)\,\nabla_{\lambda_i}\log q_i(\mathbf{z}_s\,;\,\lambda_i)\big]\end{aligned}$$

where q_i is the distribution of variables in the model that depend on the ith variable (λ_i) and the Markov blanket of it denoted by $p_i(x, z_s)$, i.e. the terms in the joint that depend on those variables.

In RL, we can take advantage of sequential nature of the problem and use pre-conditioning on past states, in which case the estimated gradients of the form:

$$\begin{aligned}\mathbb{E}[\nabla_\theta \log q_\theta(z_t)R_t] &= \sum_{j=1}^{t}\left(\sum_{i=j}^{t}(R_i))\right)\nabla_\theta \log q_\theta(z_t|z_{<t}) \\ &= \mathbb{E}_{z_t}[\nabla_\theta \log q_\theta(z_t|z_{t-1})\mathbb{E}_{z>t}[R_t]]\end{aligned}$$

5. Experiments

Implementation of structured variational inference is a complicated task. We decided to start with a simpler one by implementing variance reduction technique for black box variational inference (Ranganath et al., 2014). We then compared the performance of BBVI with and without control variate using a Gaussian mixture model. The model was developed using the framework Edward (Edw). Edward is a python library for probabilistic modeling and inference. It is build on top of Tensorflow (Abadi et al., 2015) enabling automatic differentiation and computational graphs. Control variates were implemented by modifying the computation graph and calculating an estimation of a^* using samples. The dataset was synthetically created from a mixture of 2 Gaussian. The model consists of a Dirichlet prior and a mixture of 2 Gaussian. The parameters were optimized by minimizing the KL divergence using the score function trick between the approximation q and the true probability p.

We can distinctively observe a performance improvements when using control variates.

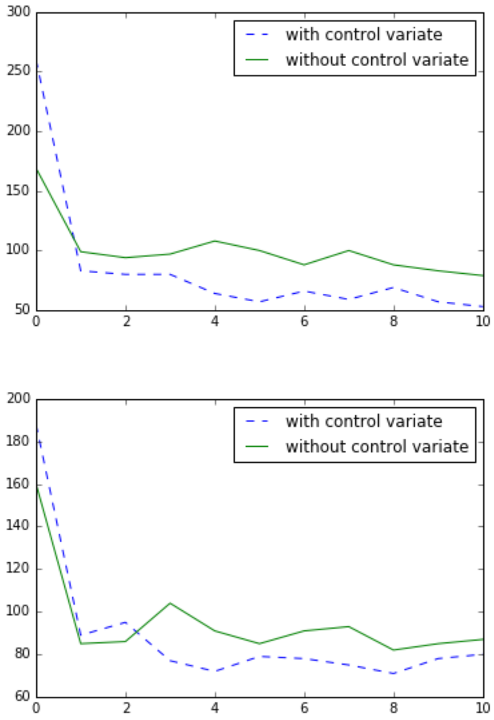

Figure 1. ELBO with and without control variate for mixture Gaussian model

6. Discussion

Recently a lot of work has been done around mean field variational inference and stochastic estimation of the gradient using various variance reduction techniques. They have proven to be quite effective. However, the mean field approximation is a big limitation as there often exists strong correlation between the latent variables, especially in a sequential settings. The problem becomes much more complex when the mean field approximation is relaxed. One of the main problem is the strong correlation that exists between the samples sequentially collected. Reinforcement learning is very familiar with this problem and has designed various techniques to overcome this issue. We believe that tools developed in reinforcement learning can be useful for structured variational inference. In particular control variates developed in reinforcement learning could be used to improve sequential structured variational inference. Finally tools from controlled systems (Markov chains mixing times...) could help analyze the theoretical properties of the various techniques. There is also the scope of using some of the RL techniques such as Actor-Critic (Sutton & Barto, 1998), to use in the context of VI, where instead of synthetically designing a control variate, we can try to learn an approximate function approximation of the same.

6.1. Citations and References

References

Edward. URL http://edwardlib.org/.

Abadi, Martín, Agarwal, Ashish, Barham, Paul, Brevdo, Eugene, Chen, Zhifeng, Citro, Craig, Corrado, Greg S., Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Goodfellow, Ian, Harp, Andrew, Irving, Geoffrey, Isard, Michael, Jia, Yangqing, Jozefowicz, Rafal, Kaiser, Lukasz, Kudlur, Manjunath, Levenberg, Josh, Mané, Dan, Monga, Rajat, Moore, Sherry, Murray, Derek, Olah, Chris, Schuster, Mike, Shlens, Jonathon, Steiner, Benoit, Sutskever, Ilya, Talwar, Kunal, Tucker, Paul, Vanhoucke, Vincent, Vasudevan, Vijay, Viégas, Fernanda, Vinyals, Oriol, Warden, Pete, Wattenberg, Martin, Wicke, Martin, Yu, Yuan, and Zheng, Xiaoqiang. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL http://tensorflow.org/. Software available from tensorflow.org.

Casella, George and Robert, Cshristian. Rao-blackwellisation of sampling schemes. *Biometrika*, 83(1):81, 1996. doi: 10.1093/biomet/83.1.81. URL +http://dx.doi.org/10.1093/biomet/83.1.81.

Dayan, Peter and Hinton, Geoffrey E. Using expectation-maximization for reinforcement learning. *Neural Comput.*, 9(2):271–278, February 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.2.271. URL http://dx.doi.org/10.1162/neco.1997.9.2.271.

Furmston, Thomas and Barber, David. Variational methods for reinforcement learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, pp. 241–248, 2010. URL http://www.jmlr.org/proceedings/papers/v9/furmston10a.html.

Ghahramani, Zoubin and Beal, Matthew J. Propagation algorithms for variational bayesian learning. In Leen, T. K., Dietterich, T. G., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems 13*, pp. 507–513. MIT Press, 2001. URL http://papers.nips.cc/paper/1907-propagation-algorithms-for-variational-bayesian-learning.pdf.

Greensmith, Evan, Bartlett, Peter L., and Baxter, Jonathan. Variance reduction techniques for gradient estimates in reinforcement learning. *J. Mach. Learn. Res.*, 5:1471–1530, December 2004. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1005332.1044710.

Hoffman, Matthew D. Stochastic structured mean-field variational inference. *CoRR*, abs/1404.4114, 2014. URL http://arxiv.org/abs/1404.4114.

Jordan, Michael I., Ghahramani, Zoubin, Jaakkola, Tommi S., and Saul, Lawrence K. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999. ISSN 0885-6125. doi: 10.1023/A:1007665907178. URL http://dx.doi.org/10.1023/A:1007665907178.

Mnih, Andriy and Gregor, Karol. Neural variational inference and learning in belief networks. *CoRR*, abs/1402.0030, 2014. URL http://arxiv.org/abs/1402.0030.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A., Veness, Joel, Bellemare, Marc G., Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K., Ostrovski, Georg, Petersen, Stig, Beattie, Charles, Sadik, Amir, Antonoglou, Ioannis, King, Helen, Kumaran, Dharshan, Wierstra, Daan, Legg, Shane, and Hassabis, Demis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 02 2015. URL http://dx.doi.org/10.1038/nature14236.

Ranganath, Rajesh, Gerrish, Sean, and Blei, David M. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pp. 814–822, 2014. URL http://jmlr.org/proceedings/papers/v33/ranganath14.html.

Sutton, Richard S. and Barto, Andrew G. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.

Sutton, Richard S., Mcallester, David A., Singh, Satinder P., and Mansour, Yishay. Policy gradient methods for reinforcement learning with function approximation. In Solla, Sara A., Leen, Todd K., and Müller, Klaus R. (eds.), *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pp. 1057–1063. The MIT Press, 1999.

Theophane Weber, Nicolas Heess, S. M. Ali Eslami John Schulman David Wingate David Silver. Reinforced variational inference. 2015. URL http://approximateinference.org/2015/accepted/WeberEtAl2015.pdf.

Williams, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning*, pp. 229–256, 1992.