Implicit Reparameterization Gradients

Michael Figurnov, Shakir Mohamed, Andriy Mnih

Presenter: Abdul Fatir

National University of Singapore

Table of contents

Introduction

Introduction

- Backpropagation through a stochastic node is an important problem in machine learning.
- Optimization of $\mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z})]$ requires computation of $\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z})]$.
- Objective of stochastic variational inference[3] includes one such expectation

$$\mathcal{L}(\mathbf{x}, \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

Previous Methods

Score-function Gradient Estimators

These estimators transform the integral into an expectation using the "log-trick".

$$
\begin{aligned}
\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}\left[f(\mathbf{z})\right] &= \nabla_\phi \int f(\mathbf{z}) q_\phi(\mathbf{z}) d\mathbf{z} \\
&= \int f(\mathbf{z}) q_\phi(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z}) d\mathbf{z} \\
&= \mathbb{E}_{q_\phi(\mathbf{z})}\left[f(\mathbf{z}) \nabla_\phi \log q_\phi(\mathbf{z})\right]
\end{aligned}
$$

Benefits

Works even when $f(\mathbf{z})$ is not differentiable.

Issues

This gradient estimator has high variance. Methods have been proposed in the literature to control the variance.

Pathwise Gradient Estimators

Commonly known as the "reparameterization trick", these estimators replace probability distributions with a deterministic and differentiable transformation $g(\phi, \boldsymbol{\varepsilon})$ of a fixed base distribution.

$$\begin{aligned}\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}\left[f(\mathbf{z})\right] &= \nabla_\phi \mathbb{E}_{q_\phi(\varepsilon)}\left[f(g(\phi, \boldsymbol{\varepsilon}))\right] \\ &= \mathbb{E}_{q_\phi(\varepsilon)}\left[\nabla_\mathbf{z} f(g(\phi, \boldsymbol{\varepsilon}))\nabla_\phi g(\phi, \boldsymbol{\varepsilon})\right]\end{aligned}$$

Benefits

This method can easily be applied to the local-scale family, distributions with tractable quantile function, and their derivatives.

Issues

Many standard distributions such as Gamma, Beta, $\mathrm{Dirichlet}$, $\mathrm{Wishart}$, etc. do not meet the requirements of this trick.

Recent Advances

Surrogate Distributions

Reparametrizable surrogate distributions such as $\mathrm{GumbelSoftmax}$ for $\mathrm{Categorical}$[2], $\mathrm{Kumaraswamy}$ for Beta[5], etc. have been proposed to approximate the respective distributions.

Generalized Reparameterizations

Methods such as Generalized Reparameterization Gradients (GRG)[6] and Rejection Sampling Variational Inference (RSVI)[4] have been proposed that build upon score-function gradients and reparameterization.

Implicit Reparameterization Gradients

Explicit Reparameterization

- Requires a standardization function $\mathcal{S}_\phi(\mathbf{z})$ such that $\mathcal{S}_\phi(\mathbf{z}) = \boldsymbol{\varepsilon} \sim p(\boldsymbol{\varepsilon})$.
- Requires $\mathcal{S}_\phi(\mathbf{z})$ to be invertible.
- $\mathbf{z} \sim q_\phi(\mathbf{z}) \Leftrightarrow \mathbf{z} = \mathcal{S}_\phi^{-1}(\boldsymbol{\varepsilon})$ and $\boldsymbol{\varepsilon} \sim p(\boldsymbol{\varepsilon})$

$$\begin{aligned}\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z})] &= \mathbb{E}_{q(\boldsymbol{\varepsilon})}[\nabla_\phi f(\mathcal{S}_\phi^{-1}(\boldsymbol{\varepsilon}))] \\ &= \mathbb{E}_{q(\boldsymbol{\varepsilon})}[\nabla_\mathbf{z} f(\mathcal{S}_\phi^{-1}(\boldsymbol{\varepsilon}))\nabla_\phi \mathcal{S}_\phi^{-1}(\boldsymbol{\varepsilon})]\end{aligned}$$

Implicit Reparameterization[1]

Eliminates the requirement of invertible $\mathcal{S}_\phi(\mathbf{z})$.

Derivation

$$\nabla_\phi \mathbb{E}_{q_\phi(\mathbf{z})}[f(\mathbf{z})] = \mathbb{E}_{q(\boldsymbol{\varepsilon})}[\nabla_\mathbf{z} f(\mathcal{S}_\phi^{-1}(\boldsymbol{\varepsilon}))\nabla_\phi \mathcal{S}_\phi^{-1}(\boldsymbol{\varepsilon})] \quad (1)$$

$$= \mathbb{E}_{q_\phi(\mathbf{z})}[\nabla_\mathbf{z} f(\mathbf{z})\nabla_\phi \mathbf{z}] \quad (2)$$

$$\frac{d\mathcal{S}_\phi(\mathbf{z})}{d\phi} = \frac{d\boldsymbol{\varepsilon}}{d\phi} = 0 \quad (3)$$

$$\frac{\partial \mathcal{S}_\phi(\mathbf{z})}{\partial \mathbf{z}}\frac{d\mathbf{z}}{d\phi} + \frac{\partial \mathcal{S}_\phi(\mathbf{z})}{\partial \phi} = 0 \quad (4)$$

$$\boxed{\nabla_\phi \mathbf{z} = -(\nabla_\mathbf{z}\mathcal{S}_\phi(\mathbf{z}))^{-1}\nabla_\phi \mathcal{S}_\phi(\mathbf{z})} \quad (5)$$

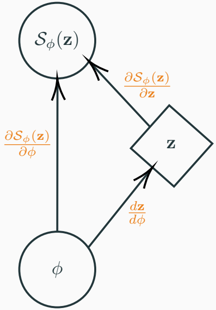


Figure 1

Examples

Normal Distribution

- $\mathcal{S}_\phi(\mathbf{z}) = \frac{\mathbf{z}-\boldsymbol{\mu}}{\boldsymbol{\sigma}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
- Explicit Reparameterization

$$\mathcal{S}_\phi^{-1}(\boldsymbol{\varepsilon}) = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\varepsilon} \Rightarrow \frac{d\mathbf{z}}{d\boldsymbol{\mu}} = \mathbf{1} \text{ and } \frac{d\mathbf{z}}{d\boldsymbol{\sigma}} = \boldsymbol{\varepsilon}$$

- Implicit Reparameterization

$$\frac{d\mathbf{z}}{d\boldsymbol{\mu}} = -\frac{d\mathcal{S}_\phi(\mathbf{z})/d\boldsymbol{\mu}}{d\mathcal{S}_\phi(\mathbf{z})/d\mathbf{z}} = \mathbf{1} \text{ and } \frac{d\mathbf{z}}{d\boldsymbol{\sigma}} = -\frac{d\mathcal{S}_\phi(\mathbf{z})/d\boldsymbol{\sigma}}{d\mathcal{S}_\phi(\mathbf{z})/d\mathbf{z}} = \frac{\mathbf{z}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

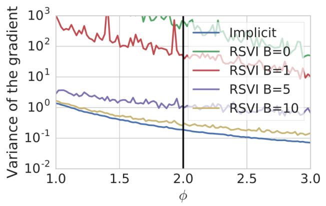
Cumulative Distribution Function

- $\mathcal{S}_\phi(\mathbf{z}) = F_\phi(\mathbf{z}) \sim \mathrm{Uniform}(0, 1)$
- $\nabla_\phi \mathbf{z} = -\frac{\nabla_\phi F_\phi(\mathbf{z})}{q_\phi(\mathbf{z})}$

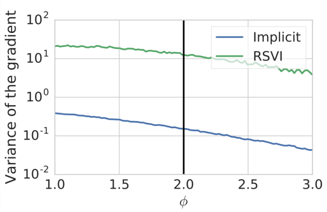
Experiments

Gradient of Cross-entropy

- Gradient of cross-entropy is required for minimization of KL-divergence.
- Variance of the gradient was observed for toy Dirichlet and Von Mises distributions.



(a) Dirichlet



(b) Von Mises

Figure 2: Comparison of RSVI and Implicit Gradients

Latent Dirichlet Allocation

- Variational Inference was performed using a neural network to model the Dirichlet variational posterior over topics.
- Experiments were performed on **20 Newsgroups** and **RCV1** datasets.

Model	Training method	20 Newsgroups	RCV1
LDA [5]	Implicit reparameterization	876 ± 7	896 ± 6
	RSVI $B = 1$	1066 ± 7	1505 ± 33
	RSVI $B = 5$	968 ± 18	1075 ± 15
	RSVI $B = 10$	887 ± 10	953 ± 16
	RSVI $B = 20$	865 ± 11	907 ± 13
	SVI	964 ± 4	1330 ± 4
LN-LDA [41]	Explicit reparameterization	875 ± 6	951 ± 10

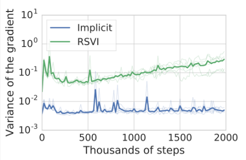
Figure 3: Text Perplexity

- Implicits gradients performed better or as good as earlier approaches and also learn sparse topic weights.

Variational Auto-Encoders

- Non-normal priors and variational posteriors used with VAEs.
- These models performed better than Normal in terms of test negative log-likelihood.
- Implicit gradients outperform RSVI on VAEs with Von Mises posterior.

(a) Von Mises, Uniform Prior



(b) Gradient Variance

Figure 4: VAE with Von Mises Posterior

Conclusion

Conclusion

- An unbiased estimator of gradients with respect to parameters of a probability distribution in a stochastic graph is presented.
- The gradients exhibit low variance and do not require inversion of standardization function.
- Even distributions without analytic expression of CDFs are supported by means of automatic differentiation of an efficient numerical method.
- Solves the problem of gradient estimation for many distributions such as Gamma, Beta, $\mathrm{Dirichlet}$, *etc*.

Questions?

References i

M. Figurnov, S. Mohamed, and A. Mnih.
Implicit reparameterization gradients.
CoRR, abs/1805.08498, 2018.

E. Jang, S. Gu, and B. Poole.
Categorical Reparameterization with Gumbel-Softmax.
ArXiv e-prints, Nov. 2016.

D. P. Kingma and M. Welling.
Auto-Encoding Variational Bayes.
ArXiv e-prints, Dec. 2013.

C. A. Naesseth, F. J. R. Ruiz, S. W. Linderman, and D. M. Blei.
Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms.
ArXiv e-prints, Oct. 2016.

E. Nalisnick and P. Smyth.
Stick-Breaking Variational Autoencoders.
ArXiv e-prints, May 2016.

F. J. R. Ruiz, M. K. Titsias, and D. M. Blei.
The Generalized Reparameterization Gradient.
ArXiv e-prints, Oct. 2016.