VarGrad: Gradient Estimator of the Log-Variance Loss Divergence

Grace Yin

March 25, 2022

Abstract

This project aims to explore the variance of the $VarGrad$, which is the gradient estimator of the log-variance loss. To accomplish this mission, this project includes background knowledge of the log-variance, and simulation tests of the $VarGrad$ on logistic models and finite Gaussian mixture models. Finally, this report gives future directions to the studies on the $VarGrad$.

1 Introduction

Variational Inference (VI) plays an important role in Probabilistic Model learning [Zha+19]. It approximates a posterior distribution $p(z|x)$ with a variational distribution $q_\phi(z)$ by minimizing the Kullback-Leibler(KL) divergence between $q_\phi(z)$ to $p(z|x)$, i.e., $\mathrm{KL}(q_\phi(z)||p(z|x))$. Sometimes solving KL divergence is not intractable. Solving the ELBO can be replaced with, because minimizing $\mathrm{KL}(q_\phi(z)||p(z|x))$ is equivalent to maximize the $\mathrm{ELBO}(\phi)$ where

$$\mathrm{ELBO}(\phi) = \mathbb{E}_{q_\phi}\left[\frac{p(x,z)}{q_\phi(z)}\right].$$

Therefore, in a VI problem, it is significant to explore the estimator for the gradient of KL $\nabla_\phi \mathrm{KL}(\phi)$ or the gradient of ELBO $\nabla_\phi \mathrm{ELBO}(\phi)$.

This project is based on the improvements of the 2020 NeurIPS Conference Paper "VarGrad: A Low-Variance Gradient Estimator for Variational Inference" by Lorenz. et eal [Ric+20]. In this paper, the authors brought up a new gradient estimator $VarGrad$, which is the gradient estimator of the Log-Varaince-Loss $\nabla\mathcal{L}_r(\cdot)$

Definition 1. *[Ric+20] For a given distribution $r(z)$, the log-variance loss $\mathcal{L}(\cdot)$ is given by*

$$\mathcal{L}_r(q_\phi(z)||p(z|x)) = \frac{1}{2}Var_r\left(\log\left(\frac{q_\phi(z)}{p(z|x)}\right)\right)$$

$r(z)$ is denoted to be the reference distribution, and there is no limitation of the reference distribution. When the support of $r(z)$ is also the support of $q_\phi(z)$ and $p(z|x)$, log-variance

loss is a divergence [Ric+20]. It was also proved that the gradient of the log-variance loss is equivalent to the gradient of KL when setting $r(z) = q_\phi(z)$ after taking the gradient with respect to ϕ. The proof can be found in the Appendix.

Theorem 1. *[Ric+20] The gradient with respect to ϕ of the log-variance loss, evaluated at $r(z) = q_\phi(z)$, equals the gradient of the KL divergence,*

$$\nabla_\phi \mathcal{L}_r(q_\phi(z)||p(z|x))|_{r=q_\phi} = \nabla_\phi KL(q_\phi(z)||p(z|x))$$

The main contribution of this paper is to build a connection between the new gradient estimator with the leave-one-out estimator which is a state-of-art gradient estimator. Recall that the score function estimator (reinforce estimator) is

$$\hat{g}_{\text{Reinforce}}(\phi) = \frac{1}{S}\sum_{s=1}^{S} \log\left(\frac{q_\phi(z^{(s)})}{p(x, z^{(s)})}\right) \nabla_\phi \log q_\phi(z^{(s)}),\ z^{(s)} \overset{i.i.d}{\sim} q_\phi(z)$$

Since the variance of reinforce estimator is high, control variates are considered to be added to reduce the variance. For instance, the leave-one-out estimator can be obtained by using $S-1$ samples to compute the control variate coefficient and averaging the resulting estimators:

$$\hat{g_{LOO}}(\phi) = \frac{1}{S-1}\left(\sum_{s=1}^{S} f_\phi(z^{(s)})\nabla_\phi \log q_\phi(z^{(s)}) - \bar{f}_\phi \sum_{s=1}^{S} \nabla_\phi \log q_\phi(z^{(s)})\right),\ z^{(s)} \overset{i.i.d}{\sim} q_\phi(z)$$

where

$$f_\phi(z) = \log \frac{q_\phi(z)}{p(x,z)},\ \text{and}\ \bar{f}_\phi = \frac{1}{S}\sum_{s=1}^{S} f_\phi(z^{(s)})$$

On the other hand, computing the empirical variance by doing Monte Carlo samples, the log-variance loss can be approximated by

$$\mathcal{L}_r(q_\phi(z)||p(z|x)) \approx \frac{1}{2(S-1)}\sum_{s=1}^{S}\left(f_\phi(z^{(s)}) - \bar{f}_\phi\right)^2,\ z^{(s)} \overset{i.i.d}{\sim} r(z)$$

Taking the gradient of $\mathcal{L}_r(q_\phi(z)||p(z|x))$ and applying Theorem 1,

$$\hat{g}_{\text{VarGrad}}(\phi) = \nabla_\phi \mathcal{L}_r = \frac{1}{S-1}\left(\sum_{s=1}^{S} f_\phi(z^{(s)})\nabla_\phi \log q_\phi(z^{(s)}) - \bar{f}_\phi \sum_{s=1}^{S} \nabla_\phi \log q_\phi(z^{(s)})\right) = \hat{g}_{LOO}$$

where $z^{(s)} \overset{i.i.d}{\sim} q_\phi(z)$.

Due to this connection, this low-variance gradient estimator can be implemented easily for automatic differentiation engines. We only need to sample $z^{(s)}$ from the variational distribution q_ϕ, apply the stop gradient operator to exclude the gradients of q_ϕ, compute the log-ratio f_ϕ for each sample and the empirical log-variance loss, then take the gradient with ϕ through the log-variance loss. The paper stated that the new gradient estimator $\hat{g}_{VarGrad}$ attains a lower variance than the reinforce estimator in both empirical way and theoretical way.

However, there are several limitations or weaknesses of this paper. First, the analysis of the log-variance loss was not thorough and comprehensive enough. Namely, the authors proved that the gradient of the log-variance loss is equivalent to the gradient estimator of the KL divergence by setting $r(z) = q_\phi$ after taking the gradient, yet they did not provide deep analysis on why do we do this. It seemed that this theorem was not intuition-based, as the conclusion was likely to be derived backward through the result. Had the authors deduced the result forwardly, they might not figure out such a logical connection between the log-variance loss and KL divergence. There was no convincing reason to explain why $\nabla_\phi r(z)$ is not included in the immediate steps in the paper. The authors did not detailedly analyze the reference distribution $r(z)$ in the log-variance loss, as well. In addition, the experiments in this paper may be insufficient, as the experiment portion only covers logistic regression models and a discrete variational autoencoder. The results of the performance of $VarGrad$ may not be convincing by doing classic simple models. They did not provide the plots of the ELBO which are necessary for Variational Inference.

To amend the above flaws, this project intended to include the following steps

- Provide more details of the log-variance loss through its original definition.
- Re-test the VarGrad estimator for logistic regression models. Compare its variance with that of the BBVI (Black Box Variational Inference) gradient estimator.
- Test the VarGrad estimator on a finite Gaussian mixture model. Compare its variance with that of the CAVI (Coordinate Ascent Variational Inference) gradient estimator.
- Derive the gradient estimator of log-variance loss including $\nabla_\phi r(z)$. Compare its variance performance with the re-parametrization gradient estimator.

Due to time constraints, the performance of the convergence of the ELBO will be reserved for future works. Yet they are not included in this project, taking the image of ELBO into consideration is essential in variational inference problems.

2 Background of the Log-variance Loss

The log-variance loss was first introduced by Nusken and Richter [NR20]. The motivation is from stochastic differential equations (SDEs):

$$dX_s^u = (b(X_s^u, s) + \sigma(X_s^u, s)u(S_s^u, s))ds + \sigma(X_s^u, s)dW_s$$

where b is the drift coefficient $\mathbb{R}^d \times [t, T] \to \mathbb{R}^d$ and σ is the diffusion coefficient $\mathbb{R}^d \times [t, T] \to \mathbb{R}^{d\times d}$. The linear growth property of b is defined to be $|b(x, s)| \leq C(1 + |x|)$, and μ is a control term $\mathbb{R}^d \times [t, T] \to \mathbb{R}^d$. $\mathcal{U}$ is denoted to be a set that

$$\mathcal{U} = \{\mu \in C^1(\mathbb{R} \times [t, T]; \mathbb{R}^d) : \mu \text{grows at most linearly in } x\}$$

Denote $\mathbb{P}$ to be the probability measure on $\mathcal{C}$ and $\mathbb{Q}$ to be the path measure on $\mathcal{C}$, then the log-variance loss is defined in Definition 2.

Definition 2. *[NR20]*

$$\mathcal{L}^{\log}_{Var_v}(\mu) = Var_{\mathbb{P}^v}\left(\log \frac{d\mathbb{Q}}{d\mathbb{P}^\mu}\right), \mu \in \mathcal{U}$$

where $v \in \mathcal{U}$.

Note that the choice of v can influence the result of $\mathcal{L}^{\log}_{Var_v}(\mu)$ but $\mathcal{L}^{\log}_{Var_v}(\mu) = 0$ if and only if $\mu = \mu^*$ holds for all the choice of v. The intuition of the connection of the log-variance loss and the KL-divergence comes from the property of the moment loss, which can be defined as

$$\mathcal{L}_{\text{moment}_v}(\mu, y_0) = \mathbb{E}\left[(Y_T^{\mu,v}(y_0) - g(X_T^v))^2\right].$$

where $g(X_T^v) = Y_T^{\mu^*,v}$

Proposition 1. *[NR20] Both* $\mathcal{L}_{moment_v}$ *and* $\mathcal{L}^{\log}_{Var_v}$ *are Gateaux-differentiable at* μ*, and*

$$\left(\frac{\delta}{\delta\mu}\mathcal{L}_{moment_v}(\mu, y_0; \phi)\right)\Big|_{v=\mu} = \left(\frac{\delta}{\delta\mu}\mathcal{L}^{\log}_{Var_v}(\mu; \phi)\right)$$

for all $\phi \in C_b^1(\mathbb{R}^d \times [0, T]; \mathbb{R}^d)$.

Following the Proposition 1,the forward control $v \in \mathcal{U}$ is chosen with the current approximation μ, this is the reason that why the gradient of v is detached from computational graph when differentiating through the loss with respect to ϕ.

3 Experiments

In this section, the gradient estimator of log-variance loss $VarGrad$ was applied to a logistic regression model, and a finite Gaussian mixture model with 3 components. A new gradient estimator of the log-variance loss$\tilde{\nabla}_\phi\mathcal{L}$ was computed by including $\nabla_\phi r(z)$ in the intermediate steps, and it was tested on a simple target distribution with beta distribution. It has been proved the Theorem 2 which states that $VarGrad$ attains lower variance than reinforce estimator theoretically.

Theorem 2. *[Ric+20] Denote* a *to be the control variate coefficient and* δ^{CV} *to be the a control variate correction term, then if*

$$-\frac{\delta_i^{CV}}{\mathbb{E}_{q_\phi}[a^{VarGrad}]} < \frac{1}{2}$$

there exists Monte Carlo samples S_0 *such that* $\forall S \geq S_0$,

$$Var(\hat{g}_{VarGrad,i}(\phi)) \leq Var(\hat{g}_{Reinforce,i}(\phi)).$$

3.1 Logistic Regression

Figure 1 shows the variance of $VarGrad$ is less than BBVI gradient estimator and it is more stable. This result is consistent with that in the paper. Theoretically, in a logistic regression model, $\left|\delta_i^{CV}/\mathbb{E}_{q_\phi}[a^{VarGrad}]\right|$is uniformly bounded over epochs. Therefore, that the condition of Theorem 2 is satisfied leads to a lower variance result. The derivation of BBVI [RGB14] for logistic models and details with the experiment are in Appendix.

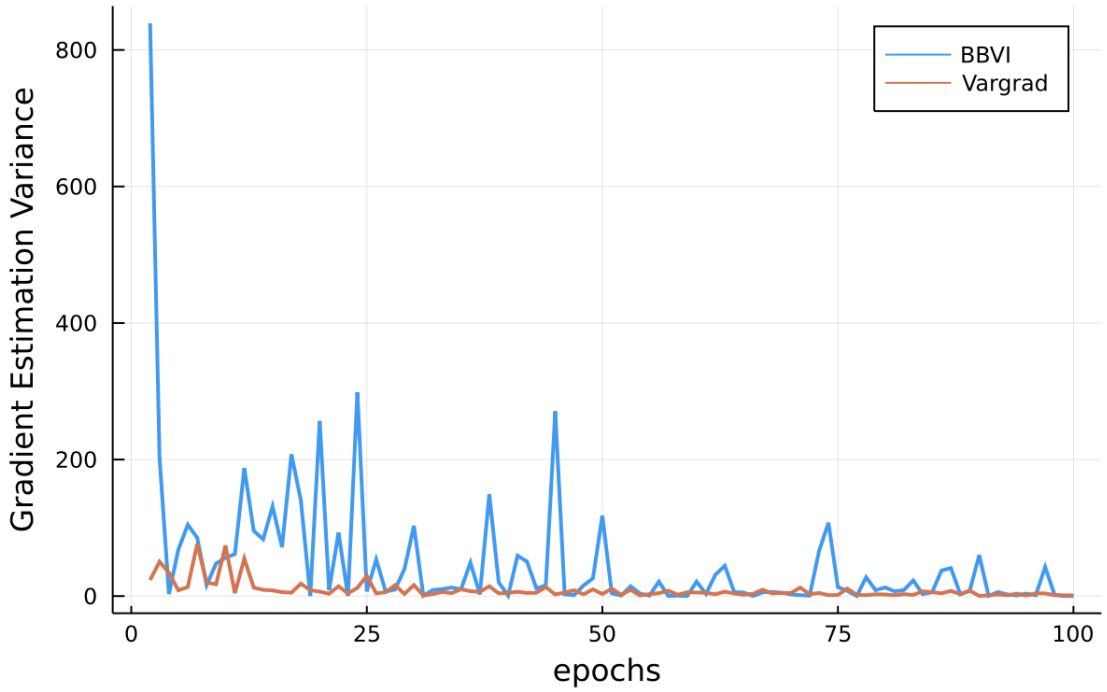

Figure 1: Variance of the BBVI Gradient Estimator and VarGrad for logistic regression models

3.2 Finite Gaussian Mixture Model

Figure 2 shows the variance of the CAVI Gradient Estimator [BKM16] and VarGrad for a spherical finite mixture Gaussian model with two distinct samples. The plot on the left demonstrates that the VarGrad estimator has a lower variance than the CAVI gradient estimator, whereas that on the right indicates that VarGrad may not perform well for a different sample in a spherical mixture Gaussian model. Since $\text{Var}(\hat{g}_{VarGrad,i}(\phi)) \leq \text{Var}(\hat{g}_{Reinforce,i}(\phi))$ has the condition that is for all samples S, $S > S_0$, distinct samples may result in the case that the reinforce estimator gives a lower variance. In addition, it is expected the $VarGrad$ performs better with the increasing of the dimension as the relative error $\delta_i^{CV}/\mathbb{E}[a^{VarGrad}]$ decreases with the increasing of dimensionality. Further experiments can be done on high-dimension Gaussian Mixture Models.

3.3 Gradient Estimator with Log-Variance Loss including $\nabla_\phi r(z)$

Since the log-variance loss is an alternative divergence when when the support of $r(z)$ is also the support of $q(z)$ and $p(z|x)$, the gradient estimator of $\nabla_\phi \mathcal{L}$ including with the $\nabla_\phi r(z)$ term is computed by setting $r(z) = q_\phi(z)$. Thus, a new gradient estimator $\tilde{\nabla}_\phi \mathcal{L}$ can be obtained by

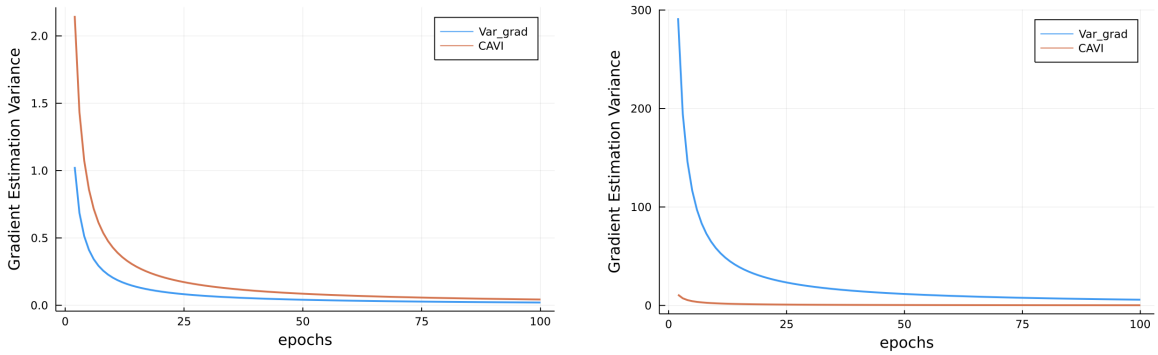

Figure 2: Variance of the CAVI gradient estimator and VarGrad for finite mixture Gaussian models with two different samples.

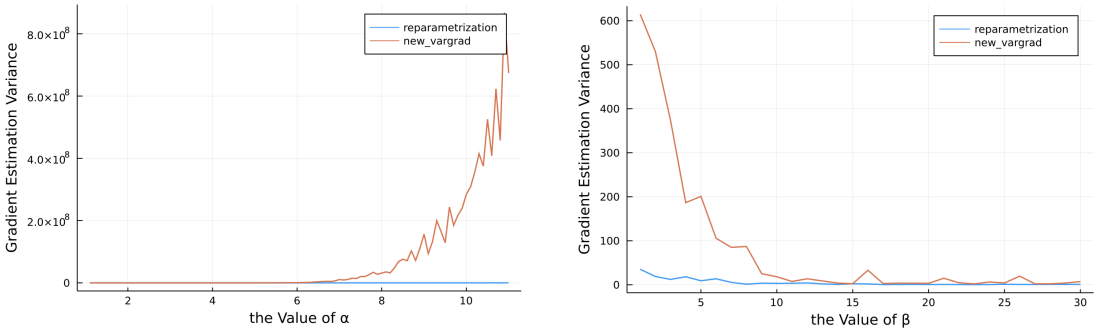

Figure 3: Variance of the new gradient estimator of $\tilde{\nabla}_\phi \mathcal{L}$ and re-parametrization gradient estimator for a beta distribution

$$\tilde{\nabla}_\phi \mathcal{L}_r(q_\phi(z)||p(z|x)) = \int \log\left(\frac{q_\phi(z)}{p(z|x)}\right) \nabla q_\phi(z) + \log^2\left(\frac{q_\phi(z)}{p(z|x)}\right) \nabla q_\phi(z) dz$$
$$- \left(\int \log\left(\frac{q_\phi(z)}{p(z|x)}\right) q_\phi(z) dz\right) \left(\int \nabla q_\phi(z) \left(1 + \log\left(\frac{q_\phi(z)}{p(z|x)}\right)\right) dz\right)$$

The results of its variance performance are in Figure 3. Figure 3 demonstrates that the variance of $\tilde{\nabla}_\phi \mathcal{L}$ is much larger than the re-parametrization gradient estimator for a target distribution, which is beta in this case. $\tilde{\nabla}_\phi \mathcal{L}$ is also unstable with the changes of the values of α and β. This occurs probably because including $\nabla_\phi r(z)$ cannot render the control variate converge to the optimal one. Besides, the increasing gradient terms may increase the computational cost too.

4 Discussion and Future Work

Overall, the results of the experiments on $VarGrad$ suggest that $VarGrad$ has lower variance than reinforcement estimators in a logistic regression model. In a finite Gaussian mixture model, there are some certain occasions where $VarGrad$ has a higher variance since the

results may vary from distinct samples. The gradient estimator $\tilde{\nabla}_\phi \mathcal{L}$ which includes $\nabla_\phi r(z)$ does not perform well compared with re-parametrization gradient estimator for a beta distribution.

There are, however, some limitations in the experiments. First, due to the difficulty in figuring out the autodiff engine for high dimension data, the tests for the logistic regression model and finite GMM are based on a 1-D synthetic dataset. As aforementioned, the advantages of the $VarGrad$ to variance are more obvious with high-dimension data rather than 1-D data. It is particularly worthwhile to test the $VarGrad$ on finite mixture Gaussian models with full structure due to the lack of re-parametrization gradient method for the finite mixture of Gaussian model with full covariance structure. Thereby, it is worthwhile to test whether $VarGrad$ works well on it. Also, it is necessary to inspect the plots of the convergence of the ELBO in Variatioanl Inference.

There are several directions for future works to follow, thus exploring the gradient estimators of the log-variance loss divergence. The first direction, as mentioned in the introduction section, is to investigate the availability of applying $VarGrad$ to complicated models, such as deep neural networks. Similar tests can also be done with $\tilde{\nabla}_\phi \mathcal{L}$. In addition, since the log-variance loss depends on the choice of the reference distribution $r(z)$, how to optimize the log-variance loss with $r(z)$ directly is an open problem. Finally, a thorough examination of the log-variance from the perspective of metric and measure is reserved for future work.

References

[BKM16] D. Blei, A. Kucukelbir, and J. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112 (2016), pp. 859–877.

[NR20] N. Nüsken and L. Richter. *Solving high-dimensional Hamilton-Jacobi-Bellman PDEs using neural networks: perspectives from the theory of controlled diffusions and measures on path space*. 2020. arXiv: 2005.05409 [math.OC].

[RGB14] R. Ranganath, S. Gerrish, and D. Blei. "Black Box Variational Inference". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. Ed. by S. Kaski and J. Corander. Vol. 33. Proceedings of Machine Learning Research. Reykjavik, Iceland: PMLR, 22–25 Apr 2014, pp. 814–822. URL: http://proceedings.mlr.press/v33/ranganath14.html.

[Ric+20] L. Richter et al. "VarGrad: A Low-Variance Gradient Estimator for Variational Inference". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 13481–13492. URL: https://proceedings.neurips.cc/paper/2020/file/9c22c0b51b3202246463e986c7e205df
Paper.pdf.

[Zha+19] C. Zhang et al. "Advances in Variational Inference". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.8 (2019), pp. 2008–2026. DOI: 10.1109/TPAMI.2018.2889774.

5 Appendix

5.1 Proof of Theorem 1

Proof.

$$\nabla_\phi \mathrm{KL}(q_\phi(z)||p(z|x)) = \int \nabla_\phi q_\phi dz + \int \log\left(\frac{q_\phi(z)}{p(z|x)}\right)\nabla_\phi q_\phi(z)dz = \int \log\left(\frac{q_\phi(z)}{p(z|x)}\right)\nabla_\phi q_\phi(z)dz$$

since

$$\int \nabla_\phi q_\phi(z)dz = \nabla_\phi \int q_\phi(z)dz = 0$$

Taking the gradient the log-variance loss,

$$\begin{aligned}
&\nabla_\phi \mathcal{L}_r(q_\phi(z)||p(z|x)) = \frac{1}{2}\nabla_\phi Var_r\left(\log\left(\frac{q_\phi(z)}{p(z|x)}\right)\right) \\
&= \frac{1}{2}\nabla_\phi \int \log^2\left(\frac{q_\phi(z)}{p(z|x)}\right) r(z)dz - \frac{1}{2}\nabla_\phi\left(\int \log\left(\frac{q_\phi(z)}{p(z|x)}\right) r(z)dz\right)^2 \\
&= \int \log\left(\frac{q_\phi(z)}{p(z|x)}\right)\frac{\nabla_\phi q_\phi(z)}{q_\phi(z)} r(z)dz - \left(\int \log\left(\frac{q_\phi(z)}{p(z|x)}\right) r(z)dz\right)\left(\int \log\left(\frac{q_\phi(z)}{q_\phi(z)}\right) r(z)dz\right)
\end{aligned}$$

Setting $r(z) = q_\phi(z)$,

$$\nabla_\phi \mathcal{L}_r(q_\phi(z)||p(z|x)) = \int \log\left(\frac{q_\phi(z)}{p(z|x)}\right)\nabla_\phi q_\phi(z)dz = \nabla_\phi \mathrm{KL}(q_\phi(z)||p(z|x))$$

□

5.2 BBVI for Logistic Regression

The synthetic data $X \in \mathbb{R}^{N\times P}$ and binary variables $y \in \mathbb{R}^N$, the model is $p(y|x,z) \sim \mathrm{Bern}(g(z^T x))$ where $g(\cdot)$ is the inverse-logit function and $z \sim \mathcal{N}(0,\mathcal{I})$. With mean-field assumption, the variational distribution is $q(z|\lambda) = \prod_{i=1}^P \mathcal{N}(z_i|\mu_j,\sigma_j^2)$. Then

$$\log p(y,X,z) = \sum_{i=1}^N [y_i \log(\sigma(z^T x_i) + (1-y_i)\log(1-g(z^T x_i))] + \sum_{j=1}^P p(z_j|\mu=0,\sigma=1)$$

and

$$\log q(z|\lambda) = \sum_{j=1}^P \log p(z_j|\mu_j,\sigma_j^2)$$

where z can be sampled from $q(z|\lambda) \sim \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$. The gradient estimators for μ and σ are

$$\nabla_{\mu_j} \log q(z|\lambda) = \frac{z_j - \mu_j}{\sigma_j^2},$$

$$\nabla_\sigma \log q(z|\lambda) = \left(-\frac{1}{2\sigma_j^2} + \frac{(z_j-\mu_j)^2}{2(\sigma_j^2)^2}\right)\sigma_j^2$$

5.3 CAVI for finite Gaussian Mixture

The full hierarchical model of the mixture of Gaussians is

$$\mu_k \sim \mathcal{N}(0, \sigma^2),\ c_i \sim \text{Categorical}(1/K, \cdots, 1/K),\ x_i|c_i, \mu \sim \mathcal{N}(c_i^T \mu, 1)$$

With mean-field assumption, the variational family has the form

$$q(\mu, c) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i)$$

The updates for m_k^* and s_k^2 are:

$$m_k^* = \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

and

$$s_k^2 = \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

5.4 Gradient Estimator $\tilde{\nabla}_\phi \mathcal{L}$of the Log-Variance Loss

$$\begin{aligned}
&\tilde{\nabla}_\phi \mathcal{L}_r(q_\phi(z)||p(z|x)) = \frac{1}{2}\nabla_\phi Var_r\left(\log\left(\frac{q_\phi(z)}{p(z|x)}\right)\right) \\
&= \frac{1}{2}\nabla_\phi \int \log^2\left(\frac{q_\phi(z)}{p(z|x)}\right) q_\phi(z)dz - \frac{1}{2}\nabla_\phi \left(\int \log\left(\frac{q_\phi(z)}{p(z|x)}\right) q_\phi(z)dz\right)^2 \\
&= \int \log\left(\frac{q_\phi(z)}{p(z|x)}\right) \nabla q_\phi(z) + \log^2\left(\frac{q_\phi(z)}{p(z|x)}\right) \nabla q_\phi(z)dz \\
&= -\left(\int \log\left(\frac{q_\phi(z)}{p(z|x)}\right) q_\phi(z)dz\right)\left(\int \nabla q_\phi(z)\left(1 + \log\left(\frac{q_\phi(z)}{p(z|x)}\right)\right)dz\right)
\end{aligned}$$

5.5 Julia Code

Python codes and Jupyter Notebook can be seen on my Github (user name: graceyin06).