Natural Gradients via the Variational Predictive Distribution

Da Tang
Columbia University
datang@cs.columbia.edu

Rajesh Ranganath
New York University
rajeshr@cims.nyu.edu

Abstract

Variational inference transforms posterior inference into parametric optimization thereby enabling the use of latent variable models where it would otherwise be impractical. However, variational inference can be finicky when different variational parameters control variables that are strongly correlated under the model. Traditional natural gradients that use the variational approximation fail to correct for correlations as they are based on factorizations that break correlations. We develop a new natural gradient using the variational predictive distribution. The variational predictive distribution captures how changing the variational parameters changes the predictive distribution for new data. A toy example shows the theoretical insight. We demonstrate the empirical value of our method on a deep generative model of images.

1 Introduction

Variational inference [3] transforms posterior inference in latent variable models into optimization. It posits a parametric approximating family and tries to find the distribution in this family that minimizes the Kullback-Leibler (KL) divergence to the posterior. Variational inference makes posterior computation practical where it would not be otherwise.

Unfortunately, the objective function for variational inference can be tricky to optimize. Consider the following example: we have n data points drawn from a bivariate Gaussian. The Gaussian has an unknown mean and a fixed covariance matrix with variance 1 and covariance $1 - \varepsilon$ for a small constant $\varepsilon > 0$. Then the two parameters for the variational approximation control two variables that have strong correlations. For this example, the gradient can have a direction almost orthogonal to the optimal update direction as in Figure 1. Further, natural gradients for variational inference [1], that adjust for the non-Euclidean nature of probability distributions, fail to change the gradient direction when the variational approximation factorizes.

To deal with the pathological curvature of the variational inference objective, we define a new type of natural gradient. This natural gradient rescales the gradient with the inverse of the Fisher information of the variational predictive distribution. Figure 1 shows that the natural gradient based on the variational predictive distribution points almost directly to the optimum. We show this approach outperforms vanilla stochastic gradient approaches on variational autoencoders [4, 9] on images.

2 Background

Latent variable models and variational inference Consider a model of n data points $\mathbf{x}_{1:n}$ that includes an unobserved variable $\mathbf{z}_{1:n}$ for each data point and a latent variable shared across data $\boldsymbol{\beta}$. The joint distribution for this model is

$$p(\boldsymbol{\beta}, \mathbf{z}, \mathbf{x}) = p(\boldsymbol{\beta}) \prod_{i=1}^{n} p(\mathbf{z}_i \,|\, \boldsymbol{\beta}) p(\mathbf{x}_i \,|\, \boldsymbol{\beta}, \mathbf{z}_i).$$

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

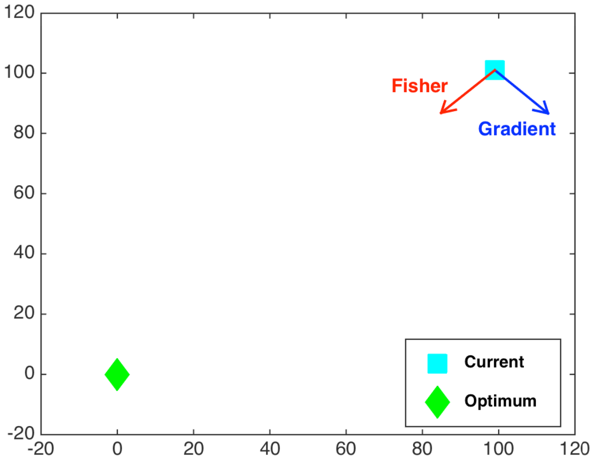

Figure 1: The normal gradient direction (Gradient) and our new Fisher-information based natural gradient direction (Fisher) for the toy example. In this example, the gradient direction and the normal natural gradient [1] direction are the same (Gradient).

Given $\boldsymbol{\beta}$, the data and data-specific (local) variables are independent. This family of models has been studied in many ways [1, 2, 8].

The central task in working with a latent variable model is computing the posterior distribution $p(\boldsymbol{\beta}, \mathbf{z} \,|\, \mathbf{x})$. For most models, computing the posterior distribution requires approximations. One approximation is variational inference [3]. Variational inference posits a distribution $q(\mathbf{z}, \boldsymbol{\beta}; \boldsymbol{\lambda})$ over the latent variables indexed by parameter $\boldsymbol{\lambda}$. The algorithm consists of maximizing the evidence lower bound (ELBO):

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_q[\log p(\mathbf{x} \,|\, \boldsymbol{\beta}, \mathbf{z})] - \mathrm{KL}(q(\boldsymbol{\beta}, \mathbf{z}; \boldsymbol{\lambda}))||p(\boldsymbol{\beta}, \mathbf{z})) \tag{1}$$

Maximizing the ELBO minimizes the KL divergence to the posterior. The family q is chosen to be simple. One example is the mean-field family, where $q(\boldsymbol{\beta}, \mathbf{z}; \boldsymbol{\lambda}) = q(\boldsymbol{\beta}; \boldsymbol{\lambda}_{\boldsymbol{\beta}}) \prod_{i=1}^{n} q(\mathbf{z}_i; \boldsymbol{\lambda}_{z_i})$.

q-Fisher Information Stochastic variational inference (SVI) [1] uses stochastic optimization to maximize Equation 1. To improve efficiency, SVI uses the *natural gradient* of the ELBO. The natural gradient is defined as $\nabla^{\text{natural}}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}) = F_q^{-1} \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda})$. Here F_q is the Fisher information matrix of the variational distribution:

$$F_q = \mathbb{E}_{q(\boldsymbol{\beta}, \mathbf{z}|\mathbf{x}; \boldsymbol{\lambda})} \left[\nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\beta}, \mathbf{z}|\mathbf{x}; \boldsymbol{\lambda}) \cdot \nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\beta}, \mathbf{z}|\mathbf{x}; \boldsymbol{\lambda})^\top \right].$$

We call this matrix the q-Fisher information matrix. Using this matrix can improve the convergence rate of SVI by taking advantage of the non-Euclidean geometry of probability spaces.

3 The variational predictive distribution

When is the q-Fisher information insufficient? While the q-Fisher information helps optimize the ELBO, sometimes it is insufficient. Consider the following example with bivariate Gaussian likelihood that has unknown mean with an isotropic Gaussian prior:

$$p(\mathbf{x}_{1:n}, \boldsymbol{\mu}) = p(\boldsymbol{\mu}; 0, I_2) \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 1-\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix}\right) \tag{2}$$

We choose a mean-field approximation where $q(\boldsymbol{\mu}; \boldsymbol{\lambda}) = \mathcal{N}(\lambda_1, \sigma^2)\mathcal{N}(\lambda_2, \sigma^2)$, with σ fixed.

The posterior distribution for this problem is analytic: $p(\boldsymbol{\mu}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}', \Sigma')$ with the covariance $\Sigma' = (n\Sigma^{-1} + I_2)^{-1}$ and the mean $\boldsymbol{\mu}' = (n \cdot I_2 + \Sigma)^{-1} \cdot \sum_{i=1}^{n} \mathbf{x}_i$. Hence the optimal solution for the variational parameter $\boldsymbol{\lambda}$ should be $\boldsymbol{\mu}'$. The gradient of the objective function $\mathcal{L}(\boldsymbol{\lambda})$ is

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}) = -\boldsymbol{\lambda} + \frac{1}{n}\Sigma^{-1} \cdot \left(-\boldsymbol{\lambda} + \sum_{i=1}^{n} \mathbf{x}_i \right).$$

The precision matrix Σ^{-1} is pathological. It has an eigenvector $\mathbf{v}_1 = \frac{1}{\sqrt{2}}(1,1)^\top$ with eigenvalue $\frac{1}{2-\varepsilon}$ and an eigenvector $\mathbf{v}_2 = \frac{1}{\sqrt{2}}(1,-1)^\top$ with eigenvalue $\frac{1}{\varepsilon}$. As a result, gradient methods will almost always go along the direction of the eigenvector $\mathbf{v}_2$, as shown in Figure 1. Moreover, natural gradients fail to resolve this. The q-Fisher information matrix of this problem is diagonal. It cannot help resolve the extreme curvature between the parameters λ_1 and λ_2.

The variational predictive Fisher information The pathology of the ELBO for the model in Equation 2 comes from the ill-conditioned covariance matrix Σ. The covariance matrix of the posterior can correct for this pathology since its covariance matrix is $\Sigma' \approx \frac{1}{n}\Sigma$. The variational approximation should capture this curvature. The disconnect lies in that variational inference with the mean-field assumption cannot capture the underlying correlation. This issue can arise in any approximating family that does not contain the posterior.

The problem is that the q-Fisher information matrix measures how parameter perturbations alter the distribution of the latent variables regardless of the model. We bring the model back into the picture by considering how perturbations of the variational parameters alter the predictive distribution of new data. This can be done via the variational predictive distribution

$$r(\mathbf{x}' \mid \mathbf{x}_i; \boldsymbol{\lambda}) = \int p(\mathbf{x}' \mid \mathbf{z}_i, \boldsymbol{\beta}) q(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\beta}; \boldsymbol{\lambda}) q(\boldsymbol{\beta}; \boldsymbol{\lambda}) d\mathbf{z}_i d\boldsymbol{\beta}.$$

The Fisher information—the **variational predictive Fisher information** matrix—of this distribution is

$$F_r = \mathbb{E}_{Q_{\mathbf{x}_i}, r(\mathbf{x}' \mid \mathbf{x}_i; \boldsymbol{\lambda})} \left[\nabla_\lambda \log r(\mathbf{x}' \mid \mathbf{x}_i; \boldsymbol{\lambda}) \cdot \nabla_{\boldsymbol{\lambda}} \log r(\mathbf{x}' \mid \mathbf{x}_i; \boldsymbol{\lambda})^\top \right], \tag{3}$$

is an expectation over some distribution on the new data points $\mathbf{x}_i'$ (we marginalize over the empirical distribution $Q_{\mathbf{x}_i}$ for $\mathbf{x}_i$). This matrix can capture curvature. The following theorem (proved in the Appendix) makes this precise:

Theorem 1. *For the example in Equation 2, its variational predictive Fisher information is* $F_r = (\Sigma + \sigma^2 \cdot I_2)^{-1}$.

Therefore, when σ is small, our variational predictive Fisher information contains the curvature we want to correct. Hence, we apply an update $\delta\boldsymbol{\lambda} = F_r^{-1} \cdot \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda})$ (we call this the **variational predictive natural gradient**) at the point $\boldsymbol{\lambda}$. The algorithm can then optimize towards the true mean rather than get stuck on the line $\lambda_1 - \lambda_2 = 0$, as shown in Figure 1.

4 Variational inference with approximate curvature

To use the variational predictive Fisher information (Equation 3), we need to compute the variational predictive distribution, differentiate it, and take an expectation with respect to it. These steps will be tractable for specific choices of model and variational approximation. We address how to approximate it in a broader setting here.

First, note the term inside of Equation 3 is a derivative with respect to the integral that defines r. In the general case, we can differentiate and use score function-style estimators used in black box variational inference [8]. However, in the case where q is reparameterizable [4, 9], i.e., a draw from q can be written as a deterministic transformation g of parameter-free noise, the computation simplifies. We detail the steps below.

1. We use a Monte Carlo estimate of r by sampling from $r'(\mathbf{x}' \mid \mathbf{x}_i; \boldsymbol{\lambda}) = p(\mathbf{x}' | \boldsymbol{\beta}', \mathbf{z}_i')$ with latent variables $\mathbf{z}$ drawn from q. Because of reparameterization, we can differentiate r' with respect to $\boldsymbol{\lambda}$.
2. Even with r', the outer expectation for the variational predictive Fisher information may be intractable. We can use Monte Carlo again to approximate this expectation. Let us denote the estimation as $F'_{r'}$.
3. This approximation $F'_{r'}$ might be non-invertible. We add a small dampening parameter μ to ensure invertibility. This parameter could be fixed or dynamically adjusted.

We set the learning rates using RMSProp [10], and set the regularization parameter μ to be a constant in our experiment.

5 Experiments

We study our variational posterior Fisher information natural gradient method on MNIST [5]. MNIST contains 70,000 images (60,000 for training and 10,000 for testing) of handwritten digits, each of size 28 × 28. We train a variational autoencoder (VAE) [4], which is a type of deep latent Gaussian models [9], on the binarized version of this dataset.

For each input image $\mathbf{x}_i$, there is a 100-dimensional latent representation $\mathbf{z}_i$ with a standard normal prior. Our variational distribution $q(\mathbf{z} \mid \mathbf{x}; \boldsymbol{\lambda})$ is a pointwise Gaussian distribution for $\mathbf{z}$ whose mean and standard deviation values are computed through a 3-layer feedforward neural network with input $\mathbf{x}$. The likelihood is a pointwise Bernoulli distribution whose logit values are computed through another 3-layer feedforward neural network.

Notice that this model is different with the original model in Equation 1. The model has no global latent variables, only global parameters. We optimize the parameters in both the variational distribution and the model distribution together. Our approach still applies in this setting. To compute variational posterior Fisher information matrices efficiently, we can view the entire VAE structure as a 6-layer neural network with a stochastic layer between the third and fourth layer. We then apply the tridiagonal block-wise Kronecker product approximation that Martens and Grosse [6] proposed. This enables us to compute Fisher information matrices faster in feed-forward neural networks. To further improve efficiency, we use low-rank approximations of large matrices. Finally, we use exponential moving averages of all quantities related to the variational predictive distribution.

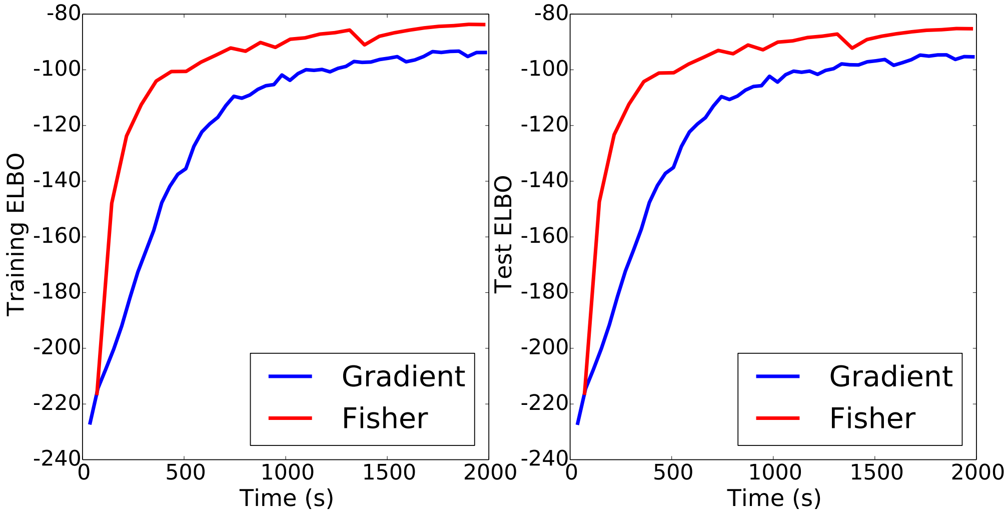

Figure 2: The learning curves for the VAE experiments on the binarized MNIST dataset. We show the training and test ELBO values for both the variational predictive natural gradient method (Fisher) and vanilla stochastic gradient optimization (Gradient). The test ELBO values are computed over the whole test set and the training ELBO values are computed over a fixed set of 10000 randomly-chosen (out of the whole 60000) images.

We compare the variational predictive natural gradient method with vanilla stochastic gradient optimization. We allow both methods to run for 2000 seconds (we found similar results at longer runtimes) and select the best step sizes by grid search. We use RMSProp [10] with decay 0.99. Figure 2 shows the results. Our method outperforms vanilla stochastic gradient optimization on both the training and test sets. The intuitive reason for the performance gain stems from that the VAE parameters control pixels that are highly correlated across images. The variational predictive natural gradient corrects for this correlation.

6 Conclusion

We introduced natural gradients based on the variational predictive distribution. They capture the parameter dependences in variational inference. We demonstrated the value on a toy model and a deep generative model of images.

Acknowledgments

The authors want to thank Jaan Altosaar for useful discussions.

References

[1] Matthew Hoffman, David Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[2] Matthew Johnson, David Duvenaud, Alex Wiltschko, Ryan Adams, and Sandeep Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems*, 2016.

[3] Michael Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[4] Diederik. Kingma and Max. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[5] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[6] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pages 2408–2417, 2015.

[7] Kevin Murphy. Conjugate bayesian analysis of the gaussian distribution. 2007.

[8] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

[9] Danilo Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

[10] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2): 26–31, 2012.

Appendix

We prove Theorem 1 of the example in Equation 2.

Proof. To compute the variational posterior Fisher information matrix F_r, we first use a method similar to what Murphy [7] proposed to compute the variational predictive distribution. Then we compute its Fisher information matrix.

Recall the definition of the variational predictive distribution r:

$$r(\mathbf{x}' \,|\, \mathbf{x}_i; \boldsymbol{\lambda}) = \int p(\mathbf{x}' \,|\, \mathbf{z}_i, \boldsymbol{\beta}) q(\mathbf{z}_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}; \boldsymbol{\lambda}) q(\boldsymbol{\beta}; \boldsymbol{\lambda}) d\mathbf{z}_i d\boldsymbol{\beta}.$$

In the example, we do not have local latent variables. So this integration equals $\int p(\mathbf{x}' \mid \boldsymbol{\mu}) q(\boldsymbol{\mu}; \boldsymbol{\lambda}) d\boldsymbol{\mu}$. Notice that both $q(\boldsymbol{\mu}; \boldsymbol{\lambda})$ and $p(\mathbf{x}' \mid \boldsymbol{\mu})$ are Gaussian distributions. Hence $r(\mathbf{x}' \,|\, \mathbf{x}_i; \boldsymbol{\lambda})$ should also be a Gaussian distribution.

Using a similar idea as in [7], we write the random variable $\mathbf{x}'$ of the distribution $r(\mathbf{x}' \,|\, \mathbf{x}_i; \boldsymbol{\lambda})$ as

$$\mathbf{x}' = (\mathbf{x}' - \boldsymbol{\mu}) + \boldsymbol{\mu}. \tag{4}$$

The random variables $\mathbf{x}' - \boldsymbol{\mu}$ and $\boldsymbol{\mu}$ are independent of each other, so we know the mean and covariance matrices of $\mathbf{x}'$ should be

$$\begin{cases} \text{mean}(\mathbf{x}') = \text{mean}(\mathbf{x}' - \boldsymbol{\mu}) + \text{mean}(\boldsymbol{\mu}) = \boldsymbol{\lambda}. \\ \text{Cov}(\mathbf{x}') = \text{Cov}(\mathbf{x}' - \boldsymbol{\mu}) + \text{Cov}(\boldsymbol{\mu}) = \Sigma + \sigma^2 \cdot I_2. \end{cases} \tag{5}$$

Therefore, the Fisher information of the distribution $r(\mathbf{x}' \,|\, \mathbf{x}_i; \boldsymbol{\lambda})$ is

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\lambda}}^{\top} \cdot \text{Cov}^{-1}(\mathbf{x}') \cdot \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\lambda}} = (\Sigma + \sigma^2 \cdot I_2)^{-1}.$$

Our variational posterior Fisher information is defined as the expectation of the above Fisher information over the empirical distribution of new data $\mathbf{x}_i$. But since this matrix is unrelated to the data, our overall variational posterior Fisher information matrix should still be $(\Sigma + \sigma^2 \cdot I_2)^{-1}$.

□