Appendix

We analyze the geometric structure of VPNG to show its insight. Then we provide more details for the experiments.

A. Analysis on the geometric structure of VPNG

As discussed in Hoffman et al. (2013), the traditional natural gradient points to the steepest ascent direction of the ELBO in the symmetric KL divergence space of the variational distribution q. Mathematically, for the ELBO function as in Equation 1, the traditional natural gradient points to the direction of the solution to the following optimization problem, as $\epsilon \to 0$:

$$\begin{aligned} \arg\max_{\Delta\boldsymbol{\lambda}} \quad & \mathcal{L}(\boldsymbol{\lambda}+\Delta\boldsymbol{\lambda}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \mathrm{KL}_{\text{sym}}(q(\boldsymbol{\lambda})||q(\boldsymbol{\lambda}+\Delta\boldsymbol{\lambda})) \le \epsilon. \end{aligned}$$

In fact, denote $\boldsymbol{\eta} = \begin{pmatrix}\boldsymbol{\lambda}\\ \boldsymbol{\theta}\end{pmatrix}$, the reparameterization for $\mathbf{z} = g(\mathbf{x}, \boldsymbol{\varepsilon}; \boldsymbol{\lambda})$ and $p_{\mathbf{x}'}(\boldsymbol{\eta}) = p(\mathbf{x}' \,|\, \mathbf{z} = g(\mathbf{x}, \boldsymbol{\varepsilon}; \boldsymbol{\lambda}); \boldsymbol{\theta})$ to be the reparameterized predictive distribution, our VPNG (as defined in Equation 14) shares similar geometric structures and points to the direction of the solution to the following optimization problem, as $\epsilon \to 0$:

$$\begin{aligned} \arg\max_{\Delta\boldsymbol{\eta}} \quad & \mathcal{L}(\boldsymbol{\eta}+\Delta\boldsymbol{\eta}) \\ \text{s.t.} \quad & \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathrm{KL}_{\text{sym}}(p_{\mathbf{x}'}(\boldsymbol{\eta})||p_{\mathbf{x}'}(\boldsymbol{\eta}+\Delta\boldsymbol{\eta}))\right] \le \epsilon. \end{aligned} \tag{18}$$

Here the expectation on $\boldsymbol{\varepsilon}$ takes with respect to the parameter-free distribution $s(\boldsymbol{\varepsilon})$ in the reparameterization.

Proof. The proof for the above fact is similar with the proof for the traditional natural gradient as in Hoffman et al. (2013). Ideally, we want to find a (possibly approximate) Riemannian metric $G(\boldsymbol{\eta})$ to capture the geometric structure of the expected symmetric KL divergence $\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathrm{KL}_{\text{sym}}(p_{\mathbf{x}'}(\boldsymbol{\eta})||p_{\mathbf{x}'}(\boldsymbol{\eta}+\Delta\boldsymbol{\eta}))\right]$:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathrm{KL}_{\text{sym}}(p_{\mathbf{x}'}(\boldsymbol{\eta})||p_{\mathbf{x}'}(\boldsymbol{\eta}+\Delta\boldsymbol{\eta}))\right] \approx \Delta\boldsymbol{\eta}^\top G(\boldsymbol{\eta})\Delta\boldsymbol{\eta} + o(\|\Delta\boldsymbol{\eta}\|^2).$$

By making first-order Taylor approximation on $p_{\mathbf{x}'}(\boldsymbol{\eta}+\Delta\boldsymbol{\eta})$ and $\log p_{\mathbf{x}'}(\boldsymbol{\eta}+\Delta\boldsymbol{\eta})$, we get

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathrm{KL}_{\text{sym}}(p_{\mathbf{x}'}(\boldsymbol{\eta})||p_{\mathbf{x}'}(\boldsymbol{\eta}+\Delta\boldsymbol{\eta}))\right] \\ =&\mathbb{E}_{\boldsymbol{\varepsilon}}[\int (p_{\mathbf{x}'}(\boldsymbol{\eta}+\Delta\boldsymbol{\eta}) - p_{\mathbf{x}'}(\boldsymbol{\eta})) \\ &\qquad \cdot (\log p_{\mathbf{x}'}(\boldsymbol{\eta}+\Delta\boldsymbol{\eta}) - \log p_{\mathbf{x}'}(\boldsymbol{\eta}))d\mathbf{x}'] \\ =&\mathbb{E}_{\boldsymbol{\varepsilon}}[\int (\nabla_{\boldsymbol{\eta}} p_{\mathbf{x}'}(\boldsymbol{\eta})^\top \Delta\boldsymbol{\eta}) \cdot (\nabla_{\boldsymbol{\eta}} \log p_{\mathbf{x}'}(\boldsymbol{\eta})^\top \Delta\boldsymbol{\eta}) d\mathbf{x}' \\ &\qquad + O(\|\Delta\boldsymbol{\eta}\|^3)] \\ =&\mathbb{E}_{\boldsymbol{\varepsilon}}[\int p_{\mathbf{x}'}(\boldsymbol{\eta}) \cdot (\nabla_{\boldsymbol{\eta}} \log p_{\mathbf{x}'}(\boldsymbol{\eta})^\top \Delta\boldsymbol{\eta}) \\ &\qquad \cdot (\nabla_{\boldsymbol{\eta}} \log p_{\mathbf{x}'}(\boldsymbol{\eta})^\top \Delta\boldsymbol{\eta}) d\mathbf{x}' + O(\|\Delta\boldsymbol{\eta}\|^3)] \\ =&\Delta\boldsymbol{\eta}^\top F_r \Delta\boldsymbol{\eta} + O(\|\Delta\boldsymbol{\eta}\|^3). \end{aligned}$$

The term $O(\|\Delta\boldsymbol{\eta}\|^3)$ is negligible compared to the first term when $\epsilon \to 0$. Hence, we could take $G(\boldsymbol{\eta})$ to be just F_r, the variational predictive Fisher information as defined in Section 3.3. By Amari (1998)'s analysis on natural gradients, we know that the solution to Equation 18 points to the direction of $G(\boldsymbol{\eta})^{-1} \cdot \nabla_{\boldsymbol{\lambda},\boldsymbol{\theta}}\mathcal{L} = \nabla^{\text{VPNG}}_{\boldsymbol{\lambda},\boldsymbol{\theta}}\mathcal{L}$, when $\epsilon \to 0$. □

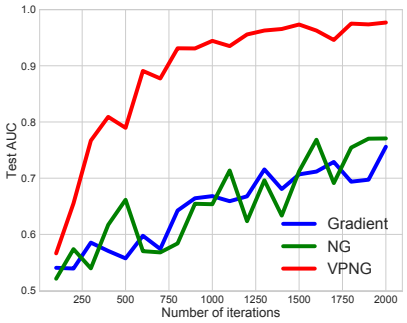

Figure 4. Bayesian Logistic regression test AUC-iteration learning curve.

B. More details for the experiments

For the Bayesian Logistic regression experiment, we show the test AUC-iteration curve as in Figure 4. It can be seen that the VPNG behaves more stable compared to the baseline methods.

For the VAE and the VMF experiments, we chose hyperparameters for all methods based on the training ELBO at the end of the time budget.

For the traditional natural gradient and VPNG, we applied the dampening factor μ. More precisely, we take VPNG updates as $\hat{\nabla}^{\text{VPNG}}_{\boldsymbol{\theta},\boldsymbol{\lambda}}\mathcal{L} = (\hat{F}_r + \mu I)^{-1} \cdot \nabla_{\boldsymbol{\theta},\boldsymbol{\lambda}}\mathcal{L}$ and traditional natural gradient updates as $\hat{\nabla}^{\text{NG}}_{\boldsymbol{\lambda}}\mathcal{L} = (\hat{F}_q + \mu I)^{-1} \cdot \nabla_{\boldsymbol{\lambda}}\mathcal{L}$). This is also applied in the Bayesian Logistic regression experiment in Section 5.1.

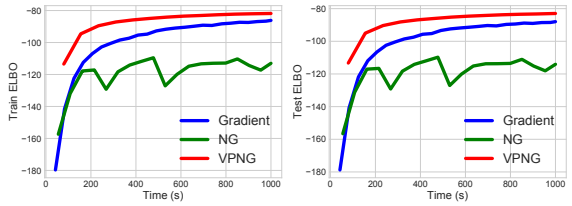

Figure 5. VAE learning curves on binarized MNIST, without exponential moving averages

For the VAE and the VMF experiments, we applied the K-FAC approximation (Martens & Grosse, 2015) to efficiently approximate the Fisher information matrices, in the NG and VPNG computations. For the VPNGs, we view the VAE model as a 6-layer neural network and the VMF model as a 4-layer neural network. For the traditional NGs, we view both the VAE model and the VMF model as 3-layer neural networks. We apply K-FAC on these models to efficiently approximate the Fisher information matrices with respect to the model distributions, given the samples from the variational distributions.

To apply the K-FAC approximation, we will need to compute matrix multiplications and matrix inversions for some non-diagonal large square matrices (i.e. the $\bar{A}_{0,0}$ matrices in the K-FAC paper (Martens & Grosse, 2015) and some other matrices that are computed during the K-FAC approximation process). In order to make the algorithms faster, we applied low-rank approximations for some large matrices of these forms by sparse eigenvalue decompositions. All of these large matrices are positive semi-definite. For each such large matrix M, we keep only the $K \cdot \ln(\dim(M))$ dimensions of it with the largest eigenvalues and K is a hyperparameter that can be tuned.

For NG and VPNG, we applied the exponential moving averages for all matrices $\bar{A}_{i,i}$ and $\bar{G}_{i+1,i+1}$ (again, we use the notations in the K-FAC paper (Martens & Grosse, 2015)) to make the learning process more stable. We found that, by adding the exponential moving average technique, our VPNG performs similarly to the case without this technique, while the traditional natural gradient is much more stable and efficient. If we do not apply the exponential moving average technique, the traditional natural gradients will not perform well. As an example, we show the performances of all methods without the exponential moving average technique in the VAE experiment in Figure 5.

We found that NG and VPNG performed similarly with respect to the dampening factor μ, the exponential moving average decay parameter and the low-rank approximation function parameter K. However, different step sizes are needed to get the best performance from these two methods. We grid searched the step sizes and report the best one.