Natural Gradients via the Variational Predictive Distribution

Da Tang[1] and Rajesh Ranganath[2]

Columbia University[1]; New York University.[2]

Variational inference

- Latent variable models: $p(\beta, z, x) = p(\beta)\prod_{i=1}^{n} p(z_i|\beta)p(x_i|\beta, z_i)$.
- Variational inference approximates the posterior through maximizing the evidence lower bound (ELBO):

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

- q-Fisher Information (Hoffman et al., 2013):

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

- Natural gradients with the q-Fisher information adjust for the non-Euclidean nature of probability distributions

Pathological curvatures

- The curvature of the ELBO becomes pathological when different variational parameters control variables that are strongly correlated under the model
- Natural gradients fail to change the gradient direction when the variational approximation factorizes

A toy example - Settings

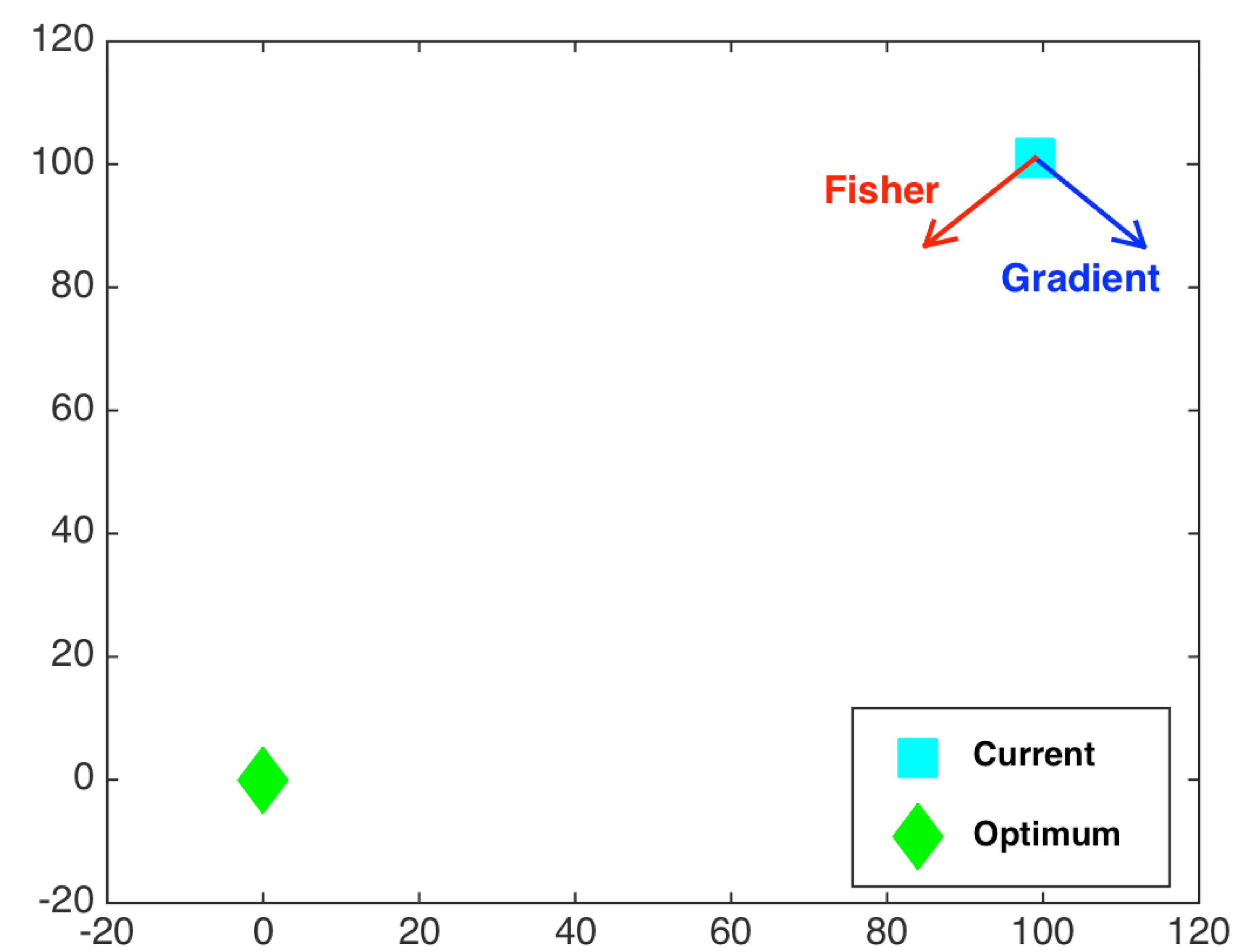
- Model: bivariate Gaussian likelihood with unknown mean and a Gaussian prior:

$$p(x_{1:n}, \mu) = p(\mu; 0, I_2) \cdot \prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \Sigma)$$

where $\Sigma = \begin{pmatrix} 1 & 1-\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix}$, $0 < \varepsilon \ll 1$

- Variational distribution: $q(\mu; \lambda) = \mathcal{N}(\lambda_1, \sigma^2)\mathcal{N}(\lambda_2, \sigma^2)$ with the standard deviation σ fixed

A toy example - Gradient directions



- The gradient and the natural gradient point to the same direction ("Gradient"), which is almost orthogonal to the optimal direction

The variational predictive Fisher information

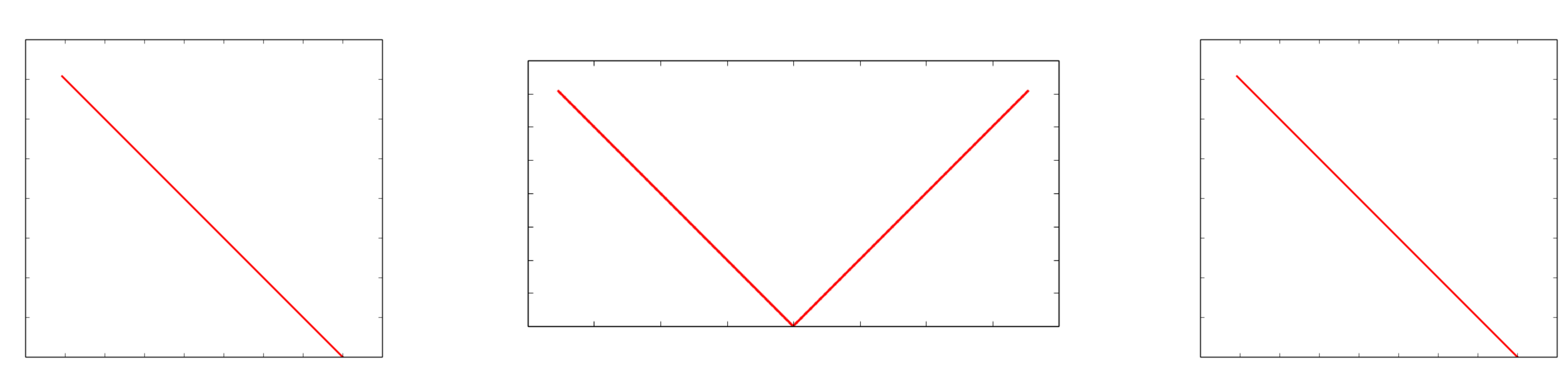
- The variational predictive distribution:

$$r(x'|x_i; \lambda) = \int p(x'|z_i, \beta)q(z_i|x_i, \beta; \lambda)q(\beta; \lambda)dz_i d\beta$$

- The variational predictive Fisher information:

$$F_r = \mathbb{E}_{Q_{x_i}, r(x'|x_i;\lambda)}[\nabla_\lambda \log r(x'|x_i; \lambda) \cdot \nabla_\lambda \log r(x'|x_i; \lambda)^\top]$$

- Eigenspace comparison (the toy example):



(a) Precision matrix Σ^{-1} (b) q-Fisher information F_q (c) our Fisher information F_r

- Natural gradient updates: $\delta\lambda = F_r^{-1} \cdot \nabla_\lambda \mathcal{L}(\lambda)$
- Resolves the curvature issue in the toy example. The algorithm can then optimize towards the optimum ("Fisher")

Variational inference with approximate curvatures

- Use reparameterization tricks
- Apply the following approximation steps:
  - Use a Monte Carlo estimate of the distribution r by sampling from $r'(x'|x_i; \lambda) = p(x'|\beta', z_i')$ with latent variables z drawn from the distribution q
  - Do Monte Carlo again to approximate the expectation of the variational predictive Fisher information
  - Add a small dampening parameter to ensure invertibility

Experiment results

- Train on the **MNIST** dataset (Lecun et al., 1998)
- Pixels are highly correlated:

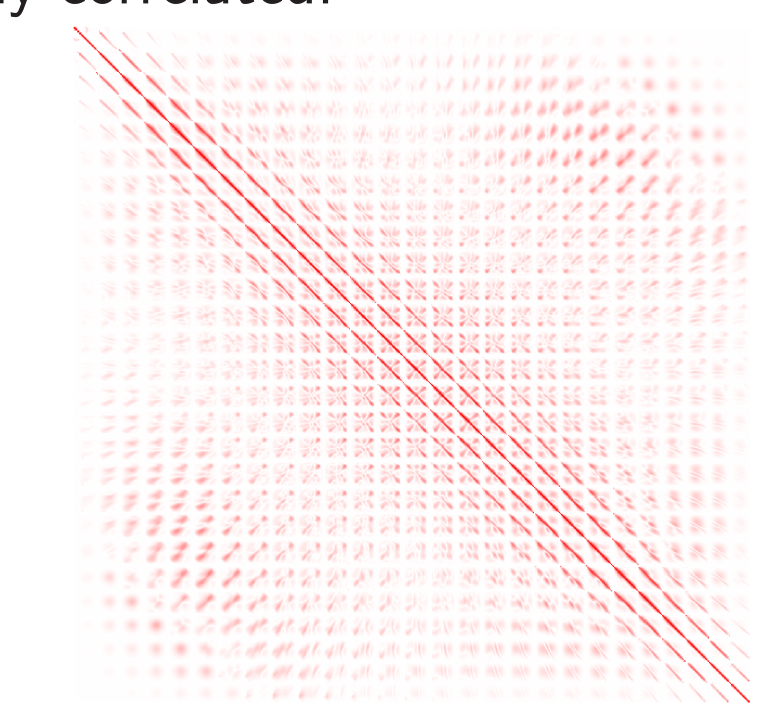
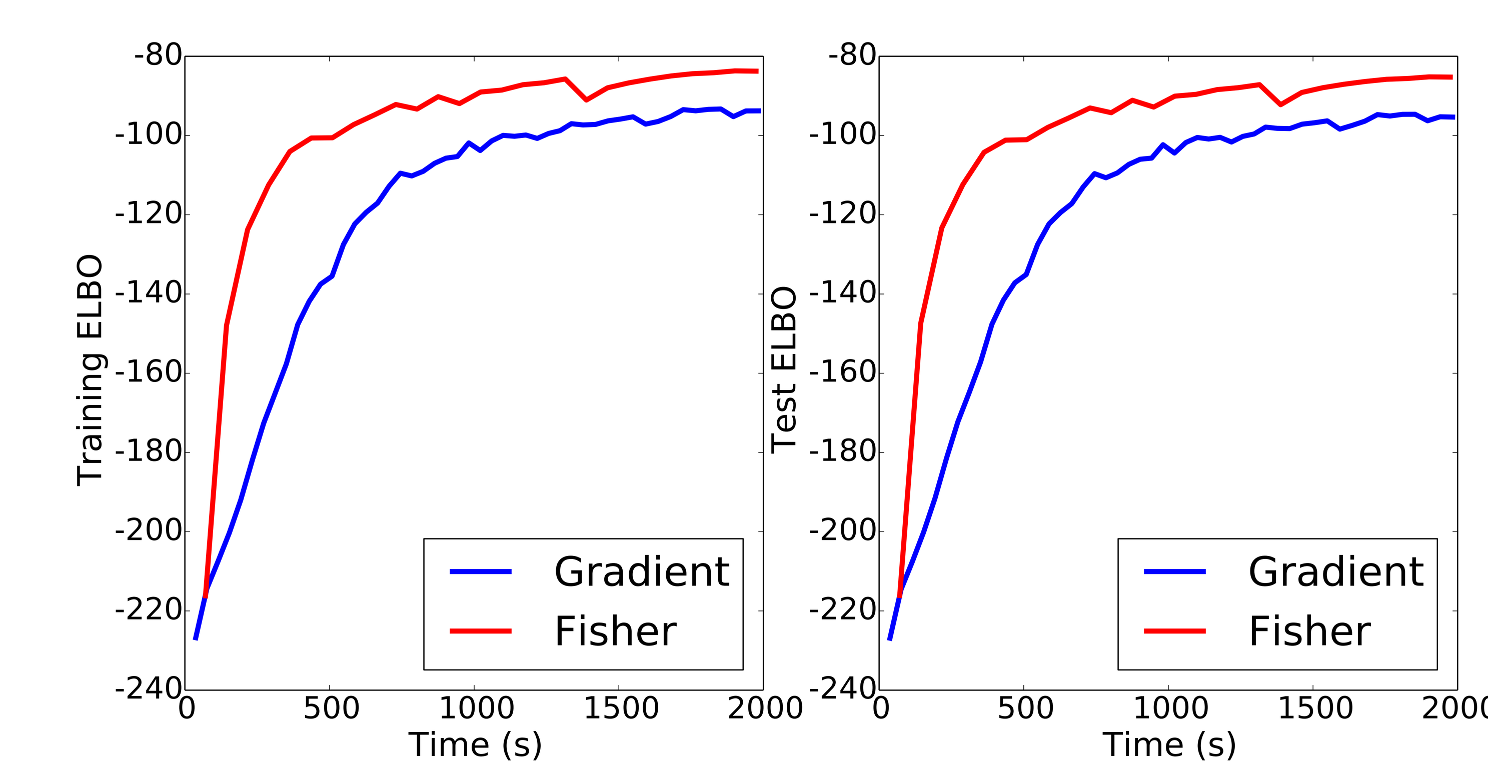


Figure: Pearson correlation coefficients of the training data

- Model: a variational autoencoder (VAE), both the inference and the generative networks are 3-layer feedforward neural networks
- Learning curves:



http://datascience.columbia.edu/

datang@cs.columbia.edu rajeshr@cims.nyu.edu