Topics on Natural Gradient, Variational Inference, and Catastrophic Forgetting

Jiaxin Shi

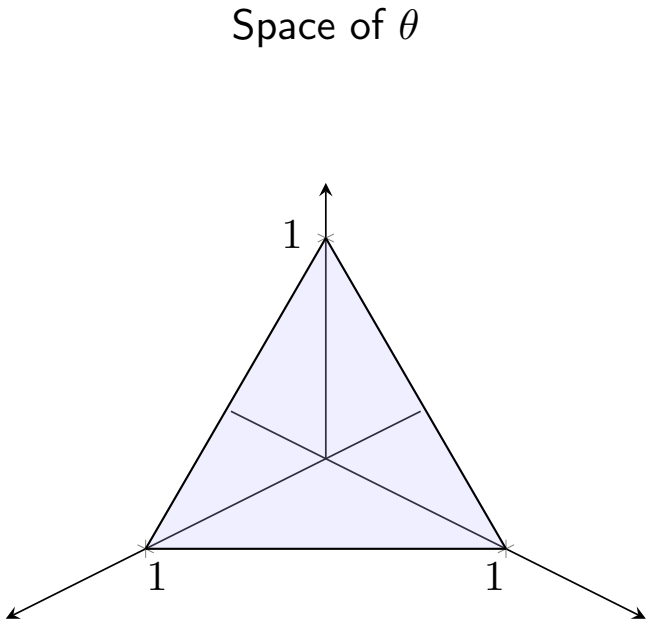
Tsinghua University

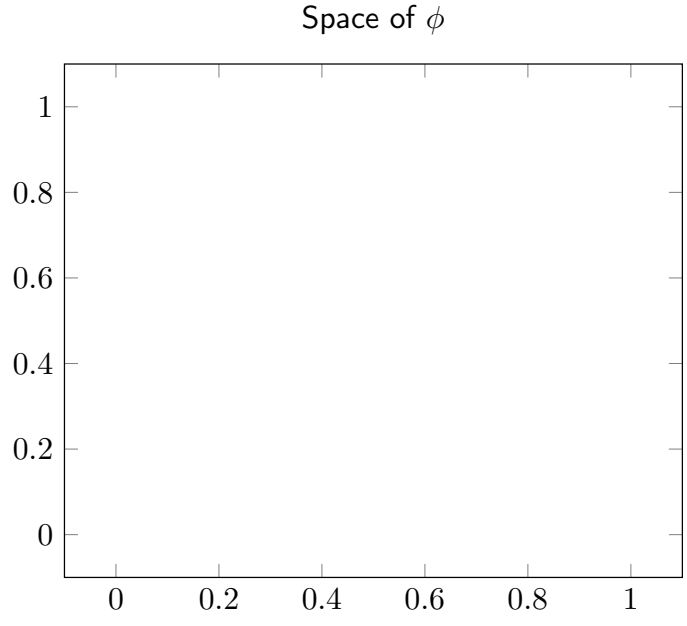
December 22, 2018

Outline

- Information Geometry, Natural Gradient
- Natural Gradient for Point Estimation (NGPE)
- Natural Gradient for Variational Inference (NGVI)
  - Natural Gradient is Mirror Descent
  - NGVI and Variational Message Passing (VMP)
- NGVI with Gaussian Variational Approximations
  - NGVI for Bayesian Neural Networks
  - Variational Adaptive-Newton (VAN)
  - NGVI as Noisy NGPE
- Catastrophic Forgetting
  - Continual Learning, Bayesian Updates, and EWC
  - An Information-Geometry View

Information Geometry





$$p(\mathbf{x};\theta) = p^*(\mathbf{x};\phi(\theta))$$

How about working in space of probability distributions $\{p(\mathbf{x};\theta)\}$?

Natural Gradient

Definition

For any problem with loss function

$$h(\theta) = f(q(\mathbf{x}; \theta)), \tag{1}$$

its natural gradient:

$$\tilde{\nabla} h = \mathbf{F}(\theta)^{-1} \nabla_\theta h, \tag{2}$$

where $\mathbf{F}(\theta)$ is the Fisher information matrix:

$$\begin{aligned} \mathbf{F}(\theta) &= \mathbb{E}_{q(\mathbf{x};\theta)} \left[\nabla_\theta \log q(\mathbf{x}; \theta) \log q(\mathbf{x}; \theta)^\top \right] \\ &= -\mathbb{E}_{q(\mathbf{x};\theta)} [\nabla^2 \log q(\mathbf{x}; \theta)] \end{aligned}$$

[Amari, 1998]

Natural Gradient

Geometric Interpretation (1)

- Gradient descent:

$$\theta_{t+1} = \theta_t - \rho_t \nabla_\theta h(\theta). \tag{3}$$

- Equivalent formulation:

$$\theta_{t+1} = \arg\min_\theta h(\theta_t) + \nabla_\theta h(\theta)^\top (\theta - \theta_t) + \frac{1}{2\rho_t} \|\theta - \theta_t\|^2. \tag{4}$$

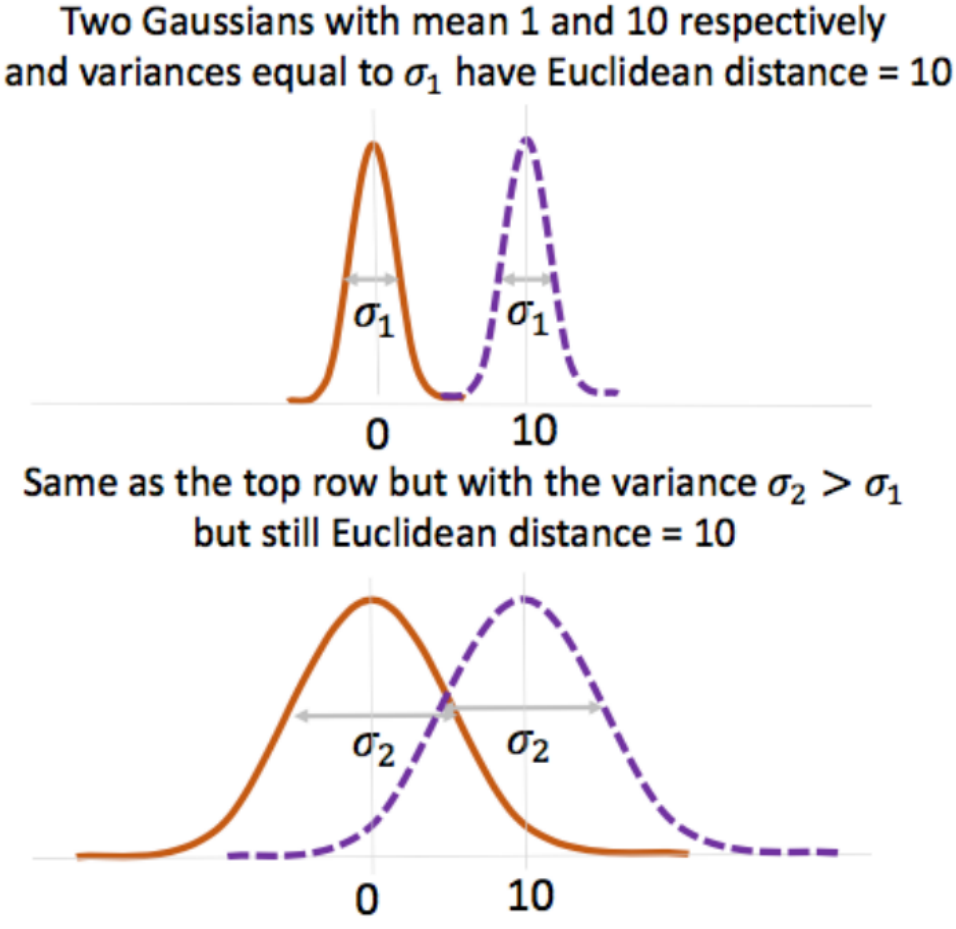
- Considering geometry of q:

$$\theta_{t+1} = \arg\min_\theta h(\theta_t) + \nabla_\theta h(\theta)^\top (\theta - \theta_t) + \frac{1}{\rho_t} \mathrm{KL}[q_{\theta_t} \| q_{\theta_{t+1}}] \tag{5}$$

Natural Gradient

Geometric Interpretation (2)

Euclidean distance is inappropriate:



(Amari 1999, Sato 2001, Honkela et.al. 2010, Hoffman et.al. 2013, Khan and Lin 2017)

Natural Gradient

KL Divergence and Fisher Information

$$\mathbb{E}_{q(\mathbf{x}|\theta)} \nabla_\theta \log q(\mathbf{x}|\theta) = 0$$

$$\log q(\mathbf{x}|\theta) = \log q(\mathbf{x}|\theta_t) + \nabla_\theta \log q(\mathbf{x}|\theta_t)^\top (\theta - \theta_t) + \frac{1}{2}(\theta - \theta_t)^\top \nabla_\theta^2 \log q(\mathbf{x}|\theta_t)(\theta - \theta_t) + O((\theta - \theta_t)^3).$$

$$\begin{aligned} \mathrm{KL}[q_{\theta_t} \| q_\theta] &= \mathbb{E}_{q(\mathbf{x}|\theta_t)}[\log q(\mathbf{x}|\theta_t) - \log q(\mathbf{x}|\theta)] \\ &= -\frac{1}{2}(\theta - \theta_t)^\top \mathbb{E}_{q(\mathbf{x}|\theta_t)}[\nabla_\theta^2 \log q(\mathbf{x}|\theta_t)](\theta - \theta_t) + O((\theta - \theta_t)^3) \\ &\approx \frac{1}{2}(\theta - \theta_t)^\top \mathbf{F}(\theta_t)(\theta - \theta_t). \end{aligned}$$

Natural Gradient

Geometric Interpretation (2)

Remember to consider the geometery of q:

$$\theta_{t+1} = \arg\min_{\theta} h(\theta_t) + \nabla_\theta h(\theta)^\top (\theta - \theta_t) + \frac{1}{\rho_t} \mathrm{KL}[q_{\theta_t} \| q_{\theta_{t+1}}].$$

Plugging in the local quadratic approximation

$$\theta_{t+1} = \arg\min_{\theta} h(\theta_t) + \nabla_\theta h(\theta)^\top (\theta - \theta_t) + \frac{1}{2\rho_t} (\theta_{t+1} - \theta_t)^\top \mathbf{F}(\theta_t)(\theta_{t+1} - \theta_t)$$

We get the natural gradient update:

$$\theta_{t+1} = \theta_t + \rho_t \mathbf{F}(\theta_t)^{-1} \theta_t \tag{6}$$

- **Note** More rigorous derivation can show

$$-\sqrt{2} \frac{\tilde{\nabla} h}{\|\nabla\|_{\mathbf{F}^{-1}}} = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \arg\min_{d: \mathrm{KL}[q_{\theta+d} \| q_\theta] \leq \epsilon^2} h(\theta + d). \tag{7}$$

Natural Gradient for Point Estimation (NGPE)

Natural gradient as Approximate Second-order Methods: Special Case

Consider MLE for a neural network:

$$h(\theta) = -\frac{1}{|S|} \sum_{(\mathbf{x},y)\in S} \log p(y|\mathbf{x};\theta). \tag{8}$$

If the loss $L(y, f(\mathbf{x};\theta))$ is not defined as $-\log p(y|\mathbf{x};\theta)$, we could assume

$$p(y|\mathbf{x};\theta) \propto \exp\{-L(y; f(\mathbf{x};\theta))\}. \tag{9}$$

Fisher information matrix:

$$\mathbf{F} = \mathbb{E}_{p(\mathbf{x},y)}[\nabla \log p(y,\mathbf{x}|\theta) \nabla \log p(y,\mathbf{x}|\theta)^\top] = -\mathbb{E}_{p(\mathbf{x},y)}[\mathbf{H}_{\log p(y|\mathbf{x};\theta)}].$$

Empirical fisher:

$$\begin{aligned}\hat{\mathbf{F}} &= \mathbb{E}_{Q_\mathbf{x}} \mathbb{E}_{p(y|\mathbf{x})}[\nabla \log p(y|\mathbf{x};\theta) \nabla \log p(y|\mathbf{x};\theta)^\top] \\ &= -\frac{1}{|S|} \sum_{\mathbf{x}\in S_\mathbf{x}} \mathbb{E}_{p(y|\mathbf{x})}[\mathbf{H}_{\log p(y|\mathbf{x};\theta)}]\end{aligned}$$

Hessian of h:

$$\mathbf{H}_h = -\frac{1}{|S|} \sum_{\mathbf{x}\in S_\mathbf{x}} \mathbb{E}_{Q(y|\mathbf{x})}[\mathbf{H}_{\log p(y|\mathbf{x};\theta)}]$$

Natural Gradient for Point Estimation (NGPE)

Natural gradient as Approximate Second-order Methods: Justification

Let

$$\mathbf{z} = f(\mathbf{x}; \theta) \tag{10}$$

Gauss-Newton Approximation for $L = \frac{1}{2}\|\mathbf{y} - \mathbf{z}\|^2$:

$$\mathbf{G} = \frac{1}{|S|} \sum_{(\mathbf{x}, y) \in S} \nabla_\theta f(\mathbf{x}; \theta) \nabla_\theta f(\mathbf{x}; \theta)^\top$$

Generalized Gauss-Newton Approximation for arbitrary loss $L(y, \mathbf{z})$:

$$\mathbf{G} = \frac{1}{|S|} \sum_{(\mathbf{x}, y) \in S} \nabla_\theta f(\mathbf{x}; \theta) \mathbf{H}_L \nabla_\theta f(\mathbf{x}; \theta)^\top, \quad \mathbf{H}_L = \nabla^2_{\mathbf{z}} L.$$

When $L = -\log f(\mathbf{x}; \theta)$, $\mathbf{H}_L = \frac{1}{f(\mathbf{x};\theta)^2}.$, then we have

$$\mathbf{G} = \nabla_\theta L \nabla_\theta L^\top. \tag{11}$$

[Martens, 2014, Bottou et al., 2018]

Natural Gradient for Variational Inference (NGVI)

Variational Inference (VI)

Consider a latent-variable model that takes the following form:

$$p(\mathcal{D}, \mathbf{z}) = \left[\prod_{i=1}^{N} \log p(\mathcal{D}_i|\mathbf{z})\right] p(\mathbf{z}). \tag{12}$$

$p(\mathbf{z})$ is an exponential family, denoted by

$$p(\mathbf{z}) = h(\mathbf{z}) \exp\{\boldsymbol{\eta}_0^\top \phi(\mathbf{z}) - A(\boldsymbol{\eta}_0)\}. \tag{13}$$

Variational Inference works by approximating $p(\mathbf{z}|\mathcal{D})$ by another exponential family distribution $q(\mathbf{z}|\boldsymbol{\lambda})$:

$$q_{\boldsymbol{\lambda}}(\mathbf{z}) = h(\mathbf{z}) \exp\{\boldsymbol{\lambda}^\top \phi(\mathbf{z}) - A(\boldsymbol{\lambda})\}. \tag{14}$$

The variational objective is

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum_{i=1}^{N} \mathbb{E}_{q_{\boldsymbol{\lambda}}} \log p(\mathcal{D}_i|\mathbf{z}) + \mathbb{E}_{q_{\boldsymbol{\lambda}}} \left[\frac{p(\mathbf{z})}{q_{\boldsymbol{\lambda}}(\mathbf{z})}\right]. \tag{15}$$

Natural Gradient for Variational Inference (NGVI)

A straightforward approach to maximize $\mathcal{L}$:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho_t \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}). \tag{16}$$

which is equivalent to

$$\boldsymbol{\lambda}_{t+1} = \arg\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_t) - \frac{1}{2\rho_t} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_t\|^2. \tag{17}$$

This simple approach uses Euclidean distances while ignoring the information geometry of $q_{\boldsymbol{\lambda}}(\mathbf{z})$. Natural gradient VI fixes this issue by:

$$\boldsymbol{\lambda}_{t+1} = \arg\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_t) - \frac{1}{\rho_t} \mathrm{KL}[q_{\boldsymbol{\lambda}}(\mathbf{z}) \| q_{\boldsymbol{\lambda}_t}(\mathbf{z})]. \tag{18}$$

Plugging in the local approximation:

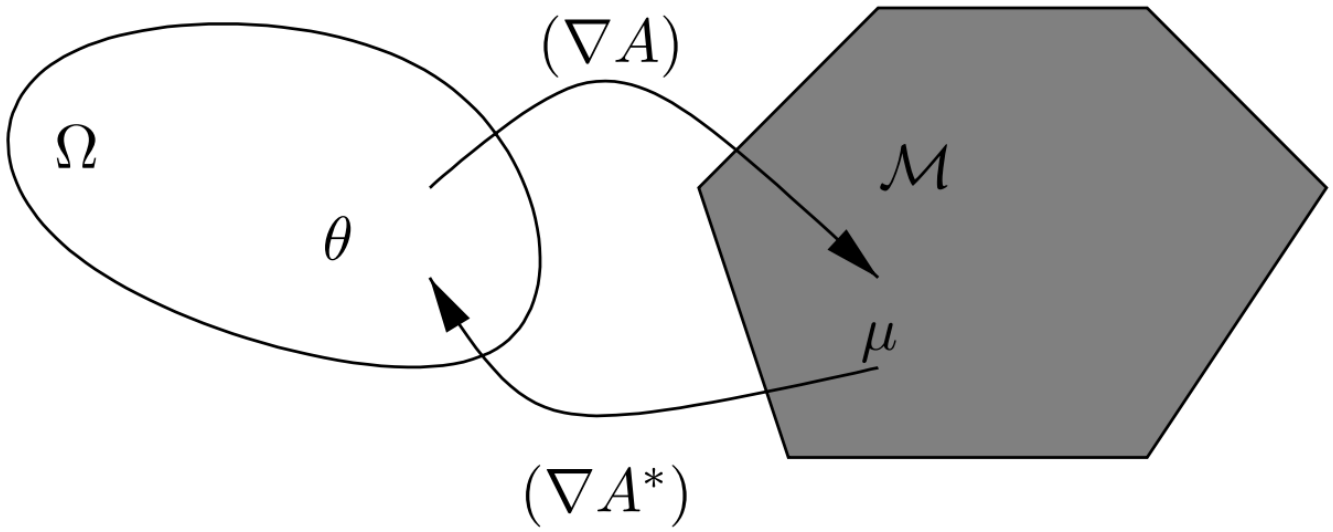
$$\boldsymbol{\lambda}_{t+1} = \arg\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_t) - \frac{1}{2\rho_t} (\boldsymbol{\lambda} - \boldsymbol{\lambda}_t)^\top \mathbf{F}(\boldsymbol{\lambda}_t)(\boldsymbol{\lambda} - \boldsymbol{\lambda}_t). \tag{19}$$

Then the solution is:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho_t \mathbf{F}(\boldsymbol{\lambda})^{-1} \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_t), \tag{20}$$

where $\tilde{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_t) = \mathbf{F}(\boldsymbol{\lambda}_t)^{-1} \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_t)$ is known as the natural gradient.

Exponetial Families



The mean parameters of $q_{\boldsymbol{\lambda}}(\mathbf{z})$:

$$\boldsymbol{\mu} = \mathbb{E}_{q_{\boldsymbol{\lambda}}}[\phi(\mathbf{z})].$$

Duality between the natural parameter and the mean parameter:

$$\nabla A(\boldsymbol{\lambda}) = \boldsymbol{\mu}, \quad \nabla A^*(\boldsymbol{\mu}) = \boldsymbol{\lambda}.$$

[Wainwright et al., 2008]

Natural Gradient for Variational Inference (NGVI)

Natural Gradient is Mirror Descent

Let the form of $\mathcal{L}$ with respect to $\boldsymbol{\mu}$ be $\mathcal{L}^*(\boldsymbol{\mu})$. We have

$$
\begin{aligned}
\tilde{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_t) &= \mathbf{F}(\boldsymbol{\lambda}_t)^{-1} \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}_t) \\
&= \mathbf{F}(\boldsymbol{\lambda}_t)^{-1} \left[\nabla_{\boldsymbol{\lambda}} \boldsymbol{\mu} \big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_t} \right]^\top \nabla_{\boldsymbol{\mu}} \mathcal{L}^*(\boldsymbol{\mu}_t) \\
&= \left[\nabla^2 A(\boldsymbol{\lambda}_t) \right]^{-1} \nabla^2 A(\boldsymbol{\lambda}_t) \cdot \nabla_{\boldsymbol{\mu}} \mathcal{L}^*(\boldsymbol{\mu}_t) \\
&= \nabla_{\boldsymbol{\mu}} \mathcal{L}^*(\boldsymbol{\mu}_t).
\end{aligned}
$$

Mirror Descent

Legendre-Frenchel Convex Conjugate

Convex conjugate:

$$\varphi^*(\mathbf{x}) = \sup_{\mathbf{y}} \left\{ \mathbf{y}^\top \mathbf{x} - \varphi(\mathbf{y}) \right\} \tag{21}$$

- $\varphi^{**} = \varphi$,
- $\varphi(\mathbf{x}) + \varphi^*(\mathbf{y}) \geq \mathbf{x}^\top \mathbf{y}$.

Mirror descent for $f(\mathbf{x})$:

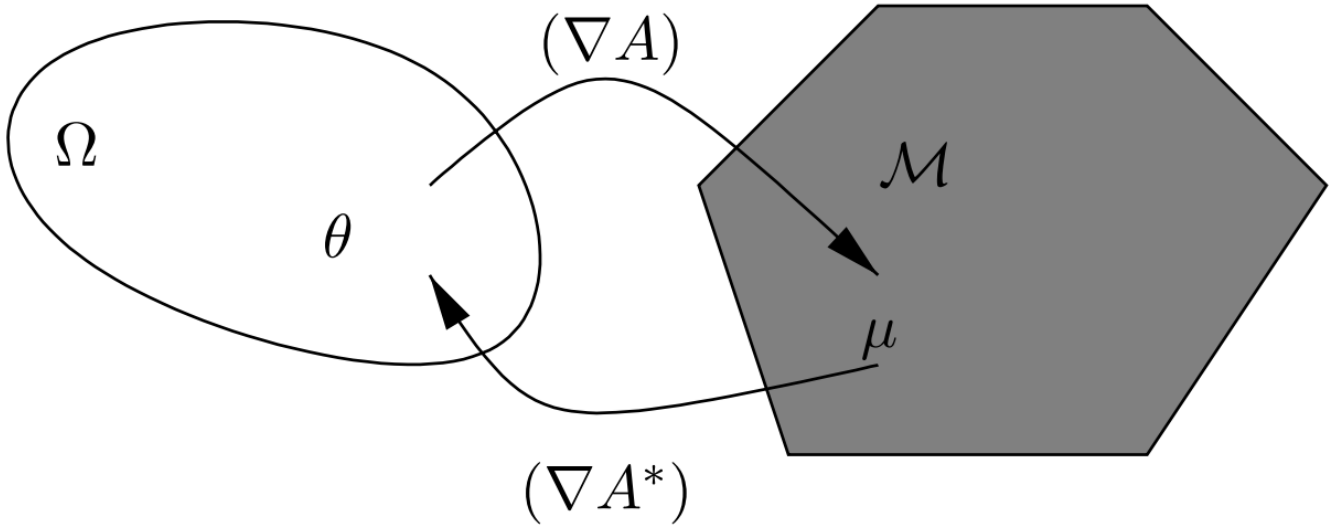
$$\mathbf{x}_{t+1} = \arg\max_{\mathbf{x}} \mathbf{x}^\top \nabla_{\mathbf{x}} f(\mathbf{x}) - \frac{1}{\rho_t} \mathbb{B}_{\varphi^*}(\mathbf{x} \| \mathbf{x}_t). \tag{22}$$

Optimal solution:

$$\mathbf{x}_{t+1} = \nabla \varphi^*(\nabla \varphi(\mathbf{x}) - \rho_t \nabla_{\mathbf{x}} f(\mathbf{x})). \tag{23}$$

Mirror Descent

Fenchel's duality in Exponential Families



Given an exponential family:

$$p(\mathbf{x}) = h(\boldsymbol{\lambda}) \exp\{\theta^\top \phi(\mathbf{x}) - A(\boldsymbol{\lambda})\} \tag{24}$$

Convex conjugate of $A(\boldsymbol{\lambda})$:

$$A^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \boldsymbol{\mu} - A(\boldsymbol{\lambda}). \tag{25}$$

Natural Gradient for Variational Inference (NGVI)

Natural Gradient is Mirror Descent

Mirror descent for $\mathcal{L}^*(\boldsymbol{\mu})$:

$$\boldsymbol{\mu}_{t+1} = \arg\max_{\boldsymbol{\mu}} \boldsymbol{\mu}^\top \nabla_{\boldsymbol{\mu}} \mathcal{L}^*(\boldsymbol{\mu}) - \frac{1}{\rho_t} \mathbb{B}_{A^*}(\boldsymbol{\mu} \| \boldsymbol{\mu}_t). \tag{26}$$

Note for exponential families:

$$\mathbb{B}_{A^*}(\boldsymbol{\mu} \| \boldsymbol{\mu}_t) = \mathrm{KL}[q(\boldsymbol{\mu}) \| q(\boldsymbol{\mu}_t)] \tag{27}$$

Mirror descent update

$$\boldsymbol{\mu}_{t+1} = \nabla A \left(\nabla A^*(\boldsymbol{\mu}_t) - \rho_t \nabla_{\boldsymbol{\mu}} \mathcal{L}^*(\boldsymbol{\mu}_t) \right).$$

We can rewrite it as

$$\boldsymbol{\mu}_{t+1} = \nabla A(\boldsymbol{\lambda} - \rho_t \tilde{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda})). \tag{28}$$

[Raskutti and Mukherjee, 2013]

Natural Gradient for Variational Inference (NGVI)

Given

$$\tilde{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}_t) = \nabla_{\boldsymbol{\mu}}\mathcal{L}^*(\boldsymbol{\mu}_t),$$

now we use this duality to compute the $\tilde{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$.

$$\begin{aligned}\tilde{\nabla}_{\boldsymbol{\lambda}}\mathrm{KL}[q_{\boldsymbol{\lambda}}(\mathbf{z})\|p(\mathbf{z})] &= \nabla_{\boldsymbol{\mu}}\left[(\boldsymbol{\lambda}-\boldsymbol{\eta}_0)^\top\boldsymbol{\mu} + A(\boldsymbol{\eta}_0) - A(\boldsymbol{\lambda})\right] \\ &= \boldsymbol{\lambda}-\boldsymbol{\eta}_0 + \nabla_{\boldsymbol{\mu}}\boldsymbol{\lambda}\cdot\boldsymbol{\mu} - \nabla_{\boldsymbol{\mu}}A(\boldsymbol{\lambda}) \\ &= \boldsymbol{\lambda}-\boldsymbol{\eta}_0.\end{aligned}$$

Thus we have

$$\tilde{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda}) = \boldsymbol{\eta}_0 - \boldsymbol{\lambda} + \sum_{i=1}^{N}\nabla_{\boldsymbol{\mu}}\mathbb{E}_q \log p(\mathcal{D}_i|\mathbf{z})\big|_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda})}. \tag{29}$$

For simplicity, we let $\tilde{\mathbf{g}}_i(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\mu}}\mathbb{E}_q \log p(\mathcal{D}_i|\mathbf{z})\big|_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda})}$:

$$\boldsymbol{\lambda}_{t+1} = (1-\rho_t)\boldsymbol{\lambda}_t + \rho_t\left[\boldsymbol{\eta}_0 + \sum_{i=1}^{N}\tilde{\mathbf{g}}_i(\boldsymbol{\lambda}_t)\right]. \tag{30}$$

Natural Gradient for Variational Inference (NGVI)

NGVI and Variational Message Passing (VMP) (1)

Besides the gradient based optimization, we could also check the optimality condition. By setting $\tilde{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}(\boldsymbol{\lambda})$ to be zero, we have:

$$\boldsymbol{\lambda}^* = \boldsymbol{\eta}_0 + \sum_{i=1}^{N} \tilde{\mathbf{g}}_i(\boldsymbol{\lambda}^*). \tag{31}$$

If the likelihood term is also conjugate, i.e., $p(\mathcal{D}_i|\mathbf{z}) \propto \exp\{-\boldsymbol{\eta}_i\phi(\mathbf{z})\}$:

$$\boldsymbol{\lambda}^* = \boldsymbol{\eta}_0 + \sum_{i=1}^{N} \boldsymbol{\eta}_i \tag{32}$$

This is known to be variational message passing (VMP) in conjugate exponential-family graphical models.

Natural Gradient for Variational Inference (NGVI)

NGVI and Variational Message Passing (VMP) (2)

- VMP, Infer.NET [Winn and Bishop, 2005].

$$\boldsymbol{\lambda}^* = \boldsymbol{\eta}_0 + \sum_{i=1}^{N} \boldsymbol{\eta}_i \tag{33}$$

- Stochastic Variational Inference (SVI) [Hoffman et al., 2013]
  - NGVI with stochastic approximation for the data term.

$$\boldsymbol{\lambda}_{t+1} = (1-\rho_t)\boldsymbol{\lambda}_t + \rho_t(\boldsymbol{\eta}_0 + \frac{N}{|B|}\sum_{i\in B}\boldsymbol{\eta}_i) \tag{34}$$

- Conjugate Computation Variational Inference [Khan and Lin, 2017]
  - Generalization of VMP to nonconjugate models (i.e., NGVI).

$$\boldsymbol{\lambda}_{t+1} = (1-\rho_t)\boldsymbol{\lambda}_t + \rho_t \left[\boldsymbol{\eta}_0 + \sum_{i=1}^{N} \tilde{\mathbf{g}}_i(\boldsymbol{\lambda}_t)\right]. \tag{35}$$

  - Also allows stochastic approximations as in SVI.

Natural Gradient for Variational Inference (NGVI)

NGVI and Variational Message Passing (VMP) (3)

With $\boldsymbol{\lambda}^* = \boldsymbol{\eta}_0 + \sum_{i=1}^{N} \tilde{\mathbf{g}}_i(\boldsymbol{\lambda}^*)$. The optimal $q_{\boldsymbol{\lambda}}$ has the form:

$$
\begin{aligned}
q_{\boldsymbol{\lambda}}(\mathbf{z}) &\propto h(\mathbf{z}) \exp\left\{ \left[\boldsymbol{\eta}_0 + \sum_{i=1}^{N} \tilde{\mathbf{g}}_i(\boldsymbol{\lambda}^*) \right]^\top \phi(\mathbf{z}) \right\} \\
&\propto h(\mathbf{z}) \exp\left\{ \boldsymbol{\eta}_0^\top \phi(\mathbf{z}) \right\} \exp\left\{ \left[\sum_{i=1}^{N} \tilde{\mathbf{g}}_i(\boldsymbol{\lambda}^*) \right]^\top \phi(\mathbf{z}) \right\} \\
&\propto p(\mathbf{z}) \prod_{i=1}^{N} \exp\left\{ \tilde{\mathbf{g}}_i(\boldsymbol{\lambda}^*)^\top \phi(\mathbf{z}) \right\}.
\end{aligned}
$$

$p(\mathcal{D}_i|\mathbf{z})$ is replaced by $\exp\left\{ \tilde{\mathbf{g}}_i(\boldsymbol{\lambda}^*)^\top \phi(\mathbf{z}) \right\}$. For local probabilistic models, where $p(\mathbf{z}) = \prod_{n=1}^{N} p(\mathbf{z}_n)$

- Finding $\boldsymbol{\lambda}^*$ for each $\mathbf{z}_n$ is impractical.
- Idea of amortized inference: Replacing $\tilde{\mathbf{g}}_i(\boldsymbol{\lambda}^*)$ with a neural network $r(\mathcal{D}_i; \mathbf{w})$ [Johnson et al., 2016, Luo et al., 2018].

NGVI with Gaussian Variational Approximations

NGVI for Bayesian Neural Networks (1)

Below we apply NGVI to Bayesian neural networks (BNN). $\mathbf{z}$ is now the weights $\mathbf{w}$, and the joint likelihood is

$$p(\mathbf{y}, \mathbf{w}|\mathbf{X}) = \left[\prod_{i=1}^{N} p(y_i|\mathbf{w}, \mathbf{x}_i)\right] p(\mathbf{w}). \tag{36}$$

The prior is chosen to be:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \tau^{-1}\mathbf{I}). \tag{37}$$

Its natural parameters are $\boldsymbol{\eta}_0 = \{\mathbf{0}, -\tau\mathbf{I}/2\}$. We define the variational posterior over $\mathbf{w}$ as a Gaussian distribution:

$$q_{\boldsymbol{\lambda}}(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{V}), \tag{38}$$

where $\boldsymbol{\lambda} = \{\mathbf{V}^{-1}\mathbf{m}, -\frac{1}{2}\mathbf{V}^{-1}\}$. The mean parameters $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$, where

$$\boldsymbol{\mu}_1 = \mathbf{m}, \quad \boldsymbol{\mu}_2 = \mathbf{m}\mathbf{m}^\top + \mathbf{V}. \tag{39}$$

NGVI with Gaussian Variational Approximations

NGVI for Bayesian Neural Networks (2)

Remember

$$\tilde{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}) = \boldsymbol{\eta}_0 - \boldsymbol{\lambda} + \sum_{i=1}^{N} \nabla_{\boldsymbol{\mu}} \mathbb{E}_q \log p(y_i|\mathbf{x}_i, \mathbf{w})\big|_{\boldsymbol{\mu}=\boldsymbol{\mu}(\boldsymbol{\lambda})}. \tag{40}$$

To compute $\tilde{\mathbf{g}}_i(\boldsymbol{\lambda})$, we use the gradient estimators for Gaussian expectations [Graves, 2011]:

$$\nabla_{\mathbf{m}} \mathbb{E}_{\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{V})} \left[f(\mathbf{w})\right] = \mathbb{E}_{\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{V})} \left[\nabla_{\mathbf{w}} f(\mathbf{w})\right],$$
$$\nabla_{\mathbf{V}} \mathbb{E}_{\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{V})} \left[f(\mathbf{w})\right] = \frac{1}{2} \mathbb{E}_{\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{V})} \left[\nabla^2_{\mathbf{w}} f(\mathbf{w})\right].$$

For simplicity, we denote

$$\mathbf{g}_i(\mathbf{w}) = \nabla_{\mathbf{w}} \log p(y_i|\mathbf{x}_i, \mathbf{w}),\ \mathbf{H}_i(\mathbf{w}) = \nabla^2_{\mathbf{w}} \log p(y_i|\mathbf{x}_i, \mathbf{w}).$$

Thus we have

$$\nabla_{\boldsymbol{\mu}_1} \mathbb{E}_q \left[\log p(y_i|\mathbf{x}_i, \mathbf{w})\right] = \mathbb{E}_q \left[\mathbf{g}_i(\mathbf{w})\right] - \mathbb{E}_q \left[\mathbf{H}_i(\mathbf{w})\right] \mathbf{m}$$
$$\nabla_{\boldsymbol{\mu}_2} \mathbb{E}_q \left[\log p(y_i|\mathbf{x}_i, \mathbf{w})\right] = \frac{1}{2} \mathbb{E}_q \left[\mathbf{H}_i(\mathbf{w})\right]$$

NGVI with Gaussian Variational Approximations

Variational Adaptive-Newton

So the natural gradient update can be written as

$$\mathbf{V}_{t+1}^{-1} = (1-\rho_t)\mathbf{V}_t^{-1} + \rho_t \left[\tau\mathbf{I} - \sum_{i=1}^{N} \mathbb{E}_q\left[\mathbf{H}_i(\mathbf{w})\right]\right],$$

$$\mathbf{m}_{t+1} = \mathbf{m}_t + \rho_t \mathbf{V}_{t+1}\left(\sum_{i=1}^{N} \mathbb{E}_q[\mathbf{g}_i(\mathbf{w})] - \tau\mathbf{m}_t\right).$$

[Khan et al., 2017]

NGVI with Gaussian Variational Approximations

NGVI as noisy NGPE (1)

Generalized Gaussian Newton Approximation

Gauss-Newton

$$\frac{N}{M}\sum_{i\in\mathcal{M}}\nabla^2_{\theta\theta}f_i(\theta) \approx \frac{N}{M}\sum_{i\in\mathcal{M}}[\nabla_\theta f_i(\theta)]^2$$

$$\approx N\left[\frac{1}{M}\sum_{i\in\mathcal{M}}\nabla_\theta f_i(\theta)\right]^2$$

Gradient-Magnitude

NGVI with Gaussian Variational Approximations

NGVI as noisy NGPE (2)

Adam

while not converged **do**
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\mu}$
  Randomly sample a data example $\mathcal{D}_i$
  $\mathbf{g} \leftarrow -\nabla \log p(\mathcal{D}_i|\boldsymbol{\theta})$
  $\mathbf{m} \leftarrow \gamma_1 \mathbf{m} + (1-\gamma_1)\,\mathbf{g}$
  $\mathbf{s} \leftarrow \gamma_2 \mathbf{s} + (1-\gamma_2)\,(\mathbf{g} \circ \mathbf{g})$
  $\hat{\mathbf{m}} \leftarrow \mathbf{m}/(1-\gamma_1^t), \quad \hat{\mathbf{s}} \leftarrow \mathbf{s}/(1-\gamma_2^t)$
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \alpha\, \hat{\mathbf{m}}/(\sqrt{\hat{\mathbf{s}}} + \delta)$
  $t \leftarrow t+1$
end while

Vadam

while not converged **do**
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, $\boldsymbol{\sigma} \leftarrow 1/\sqrt{N\mathbf{s} + \lambda}$
  Randomly sample a data example $\mathcal{D}_i$
  $\mathbf{g} \leftarrow -\nabla \log p(\mathcal{D}_i|\boldsymbol{\theta})$
  $\mathbf{m} \leftarrow \gamma_1 \mathbf{m} + (1-\gamma_1)\,(\mathbf{g} + \lambda\boldsymbol{\mu}/N)$
  $\mathbf{s} \leftarrow \gamma_2 \mathbf{s} + (1-\gamma_2)\,(\mathbf{g} \circ \mathbf{g})$
  $\hat{\mathbf{m}} \leftarrow \mathbf{m}/(1-\gamma_1^t), \quad \hat{\mathbf{s}} \leftarrow \mathbf{s}/(1-\gamma_2^t)$
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \alpha\, \hat{\mathbf{m}}/(\sqrt{\hat{\mathbf{s}}} + \lambda/N)$
  $t \leftarrow t+1$
end while

Concurrent work [Zhang et al., 2017, Khan et al., 2018]

Catastrophic Forgetting

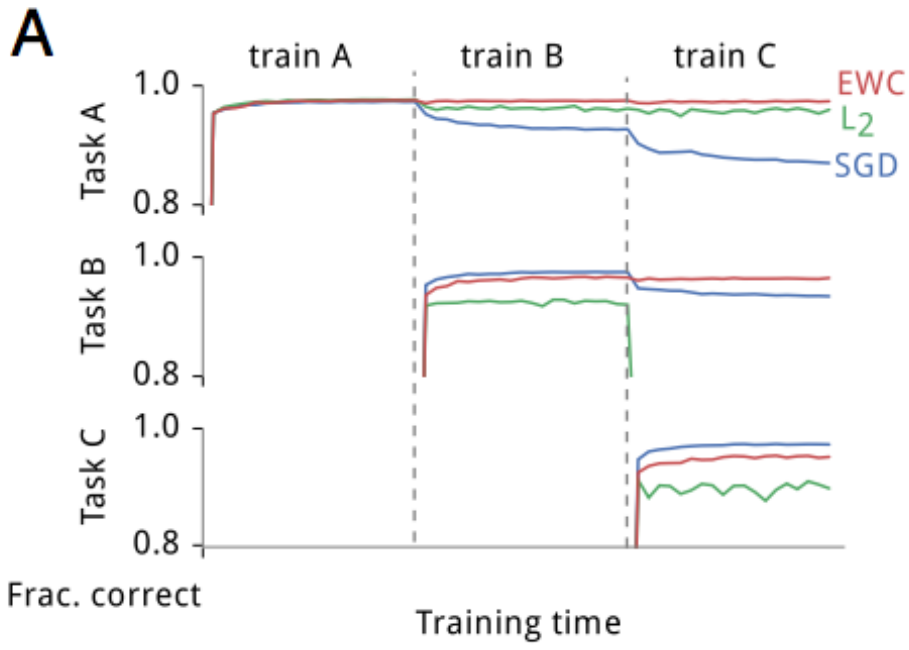


Figure: [Kirkpatrick et al., 2017]

Find a solution $\mathbf{w}_B$ for task B that also performs well on task A.

- **Claim** Bayesian inference is a natural solution to continual learning:

$$p(\mathbf{w}|D_A) \propto p(D_A|\mathbf{w})p(\mathbf{w}),$$
$$p(\mathbf{w}|D_B, D_A) \propto p(D_B|\mathbf{w})p(\mathbf{w}|D_A).$$

Elastic Weight Consolidation (EWC) [Kirkpatrick et al., 2017]

$$\log p(\mathbf{w}|D_A, D_B) = \log p(D_B|\mathbf{w}) + \log p(\mathbf{w}|D_A) + C. \tag{41}$$

Use a diagonal Laplace approximation to $\log p(\mathbf{w}|D_A)$:

$$\log p(\mathbf{w}|D_A, D_B) \approx \log p(D_B|\mathbf{w}) + \frac{1}{2}\sum_i [\mathbf{F}(\mathbf{w}_A)]_{ii}(\mathbf{w} - \mathbf{w}_A)^2 + C. \tag{42}$$

where

$$\begin{aligned}\mathbf{F}(\mathbf{w}_A) &= \mathbb{E}_{p(D|\mathbf{w}_A)}[\nabla_{\mathbf{w}} \log p(D|\mathbf{w}_A) \nabla_{\mathbf{w}} \log p(D|\mathbf{w}_A)^\top] \\ &= -\mathbb{E}_{p(D|\mathbf{w}_A)}[\nabla^2_{\mathbf{w}} \log p(D|\mathbf{w}_A)].\end{aligned}$$

Catastrophic Forgetting

An Information Geometry View (1)

- Many configurations of $\mathbf{w}$ will result in the same performance because the over-parameterization of neural networks.
- It is likely that there is a solution $\mathbf{w}_B$ for task B that is not far from the solution of task A, $\mathbf{w}_A$. This motivates the following optimization problem:

$$\mathbf{w}_B = \arg\min_{\mathbf{w}} -\log p(D_B|\mathbf{w}) + \lambda d(\mathbf{w}, \mathbf{w}_A), \tag{43}$$

- A naive choice: $d(\mathbf{w}, \mathbf{w}') = \|\mathbf{w} - \mathbf{w}'\|^2$.
- The optimal distance measure between the two parameters should be

$$d(\mathbf{w}, \mathbf{w}_A) = \mathrm{KL}\left[p(D|\mathbf{w}) \| p(D|\mathbf{w}_A)\right]. \tag{44}$$

- Problem: we need to store all the A dataset when learning from B.

Catastrophic Forgetting
An Information Geometry View (2)

- We show that EWC is in fact a local quadratic approximation to the KL:

$$\log p(D|\mathbf{w}) = \log p(D|\mathbf{w}_A) + \nabla_{\mathbf{w}} \log p(D|\mathbf{w}_A)^\top (\mathbf{w} - \mathbf{w}_A) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_A)^\top \nabla^2_{\mathbf{w}} \log p(D|\mathbf{w}_A)(\mathbf{w} - \mathbf{w}_A) + O((\mathbf{w} - \mathbf{w}_A)^3).$$

$$\begin{aligned}\mathrm{KL}[p(D|\mathbf{w}_A)\|p(D|\mathbf{w})] &= \mathbb{E}_{p(D|\mathbf{w}_A)}[\log p(D|\mathbf{w}_A) - \log p(D|\mathbf{w})] \\ &= -\frac{1}{2}(\mathbf{w} - \mathbf{w}_A)^\top \mathbb{E}_{p(D|\mathbf{w}_A)}[\nabla^2_{\mathbf{w}} \log p(D|\mathbf{w}_A)](\mathbf{w} - \mathbf{w}_A) + O((\mathbf{w} - \mathbf{w}_A)^3) \\ &\approx \frac{1}{2}(\mathbf{w} - \mathbf{w}_A)^\top \mathbf{F}(\mathbf{w}_A)(\mathbf{w} - \mathbf{w}_A).\end{aligned}$$

- In EWC

$$d(\mathbf{w}, \mathbf{w}_A) = \frac{1}{2}(\mathbf{w}-\mathbf{w}_A)^\top \mathbf{F}(\mathbf{w}_A)(\mathbf{w}-\mathbf{w}_A) \approx \frac{1}{2}\sum_i [\mathbf{F}(\mathbf{w}_A)]_{ii}(\mathbf{w}_i - \mathbf{w}_{A,i})^2,$$

which is clearly using a diagonal approximation to the Fisher matrix.

References I

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011.

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

Matthew Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.

Mohammad Emtiyaz Khan and Wu Lin. Conjugate-computation variational inference: Converting variational inference in non-conjugate models to inferences in conjugate models. *arXiv preprint arXiv:1703.04265*, 2017.

References II

Mohammad Emtiyaz Khan, Wu Lin, Voot Tangkaratt, Zuozhu Liu, and Didrik Nielsen. Variational adaptive-newton method for explorative learning. *arXiv preprint arXiv:1711.05560*, 2017.

Mohammad Emtiyaz Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. *arXiv preprint arXiv:1806.04854*, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, page 201611835, 2017.

Yucen Luo, Tian Tian, Jiaxin Shi, Jun Zhu, Ning Chen, and Bo Zhang. Semi-crowdsourced clustering with deep generative models. In *Advances in neural information processing systems*, 2018.

James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

References III

Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *arXiv preprint arXiv:1310.7780*, 2013.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

John Winn and Christopher M Bishop. Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694, 2005.

Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. *arXiv preprint arXiv:1712.02390*, 2017.