Approximate Bayesian Inference

Mohammad Emtiyaz Khan
AIP (RIKEN), Tokyo

http://emtiyaz.github.io

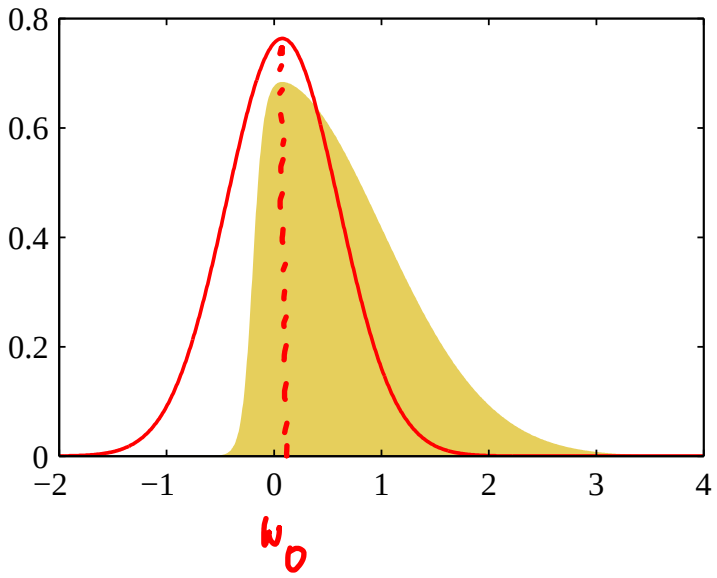
emtiyaz.khan@riken.jp

June 28-29, 2018

Contents

1 Laplace's Method

A straightforward choice is a Gaussian distribution. The main idea behind Laplace's method is to find a Gaussian approximation of the unnormalized posterior distribution $p^*(\mathbf{w}) := p(\mathcal{D}|\mathbf{w})p(\mathbf{w})$ centered at a maximum $\mathbf{w}_0$.



The normalizing constant of a Gaussian is available in closed-form, making the approximation easier to compute than the posterior distribution.

To find the covariance of the approximation, we Taylor- expand:

$$\log p^*(\mathbf{w}) = \log p^*(\mathbf{w}_0) + \frac{c}{2}(\mathbf{w} - \mathbf{w}_0)^\top \mathbf{H}(\mathbf{w}_0)(\mathbf{w} - \mathbf{w}_0) + \ldots,$$

where $\mathbf{H}(\mathbf{w})$ is the Hessian. Therefore, the Hessian can be used as the covariance:

$$q^*(\mathbf{w}) := \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{H}(\mathbf{w}_0)^{-1}).$$

Gaussian approximation.

Question: Can you think of one example where Laplace's method is exact, and one example where Laplace's method may not work?

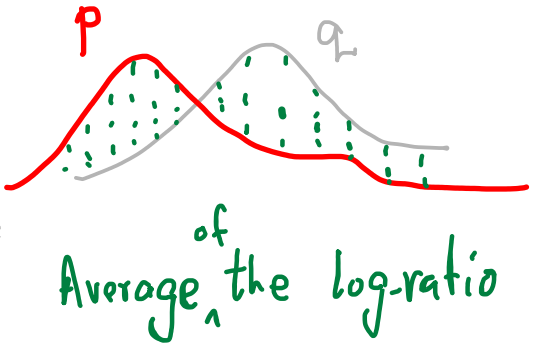
A method called Integrated Nested Laplace Approximation is shown to work well for a class of latent Gaussian models [Rue et al., 2009]. Laplace's method has been applied to neural networks [Barber and Bishop, 1998, MacKay, 2003, Ritter et al., 2018].

2 Variational Inference (VI)

Laplace's method exploits the local information at a MAP estimate, but this might not be accurate. Methods, such as variational inference and expectation propagation, improve accuracy by using a *global average*.

local → global

To measure the "goodness" of an approximation $q(\mathbf{w})$ we need to define distance between distributions.



Average of the log-ratio

Kullback-Leibler Divergence

The KL divergence enables us to measure a distance between two densities p and q,

$$\mathbb{D}_{KL}[q \,\|\, p] := -\int q(\mathbf{w}) \log \frac{p(\mathbf{w})}{q(\mathbf{w})}.$$

It is not a "proper" distance measure because $\mathbb{D}_{KL}[q \,\|\, p] \neq \mathbb{D}_{KL}[p \,\|\, q]$.

Bayesian Inference as Optimization

We can express Bayesian inference as an optimization problem:

$$\max_{q \in \mathcal{P}} \mathcal{L}_{VI}(q) := \mathbb{E}_{q(w)}\left[\log p(\mathcal{D}|\mathbf{w})\right] - \mathbb{D}_{KL}[q(\mathbf{w}) \,\|\, p(\mathbf{w})].$$

This is equivalent to Bayesian inference when $\mathcal{P}$ contains the posterior distribution. We can see this by rewriting the problem as,

$$\mathcal{L}_{VI}(q) = \log p(\mathcal{D}) - \mathbb{D}_{KL}[q(\mathbf{w}) \,\|\, p(\mathbf{w}|\mathcal{D})].$$

Maximizing $\mathcal{L}_{VI}$ is equivalent to minimizing the second term which has a minimum value of 0 at $q^*(w) := p(\mathbf{w}|\mathcal{D})$. Since $\mathcal{D}_{KL} \geq 0$, $\log p(\mathcal{D}) \geq \mathcal{L}_{VI}$ (a lower bound).

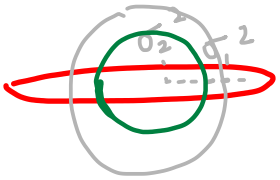
By relaxing the optimization problem, we can compute approximations. This is called variational inference (VI). The objective $\mathcal{L}_{VI}$ is called the evidence lower bound (ELBO) or the variational objective. The approximation q is called the variational distribution. Several other names for VI are variational Bayes, minimum description-length, and ensemble learning.

3 Mean-Field VI

One straightforward way is to restrict the space of densities $\mathcal{P}$ such that the optimization problem becomes easier, e.g., we can use a mean-field approximation:

Mean-field theory

$$p(\mathbf{w}|\mathcal{D}) \approx q(\mathbf{w}) := \prod_{i=1}^{D} q_i(w_i),$$



where w_i is the i'th dimension of $\mathbf{w}$ and q_i is an arbitrary distribution over w_i. The distribution q is the mean-field variational distribution.

$D_{KL}[q\|p(w|D)]$

$= -\mathrm{Tr}\left(\begin{bmatrix}\sigma_1^2 & \\ & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}\sigma_q^2 & \\ & \sigma_q^2\end{bmatrix}\right)$

$-\log\left|\begin{bmatrix}\sigma_1^2 & \\ & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}\sigma_q^2 & \\ & \sigma_q^2\end{bmatrix}\right|$

$= \frac{\sigma_q^2}{\sigma_1^2} + \frac{\sigma_q^2}{\sigma_2^2}$

$-2\log\sigma_q^2 + \text{cnst}$

Question: Consider,

$$p(\mathbf{w}|\mathcal{D}) := \mathcal{N}\left(\begin{bmatrix}w_1\\w_2\end{bmatrix}\middle|\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & 0\\0 & \sigma_2^2\end{bmatrix}\right)$$
$$\approx q(\mathbf{w}) := \mathcal{N}(w_1|0,\sigma_q^2)\mathcal{N}(w_2|0,\sigma_q^2),$$

where the variance σ_q^2 of $q_1(w_1)$ is same as that of $q_2(w_1)$.

What is the value of σ_q^2 that minimizes $\mathbb{D}_{KL}[q(\mathbf{w}) \,\|\, p(\mathbf{w}|\mathcal{D})]$? What about minimizing $\mathbb{D}_{KL}[p(\mathbf{w}|\mathcal{D}) \,\|\, q(\mathbf{w})]$?

$D_{KL}[\mathcal{N}(\mu_0,\Sigma_0) \| \mathcal{N}(\mu_1,\Sigma_1)]$ (Q, P)

$:= \frac{1}{2}\left[\mathrm{Tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - \log\left|\Sigma_1^{-1}\Sigma_0\right| - D\right]$

$\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} = \frac{2}{\sigma_q^2}$

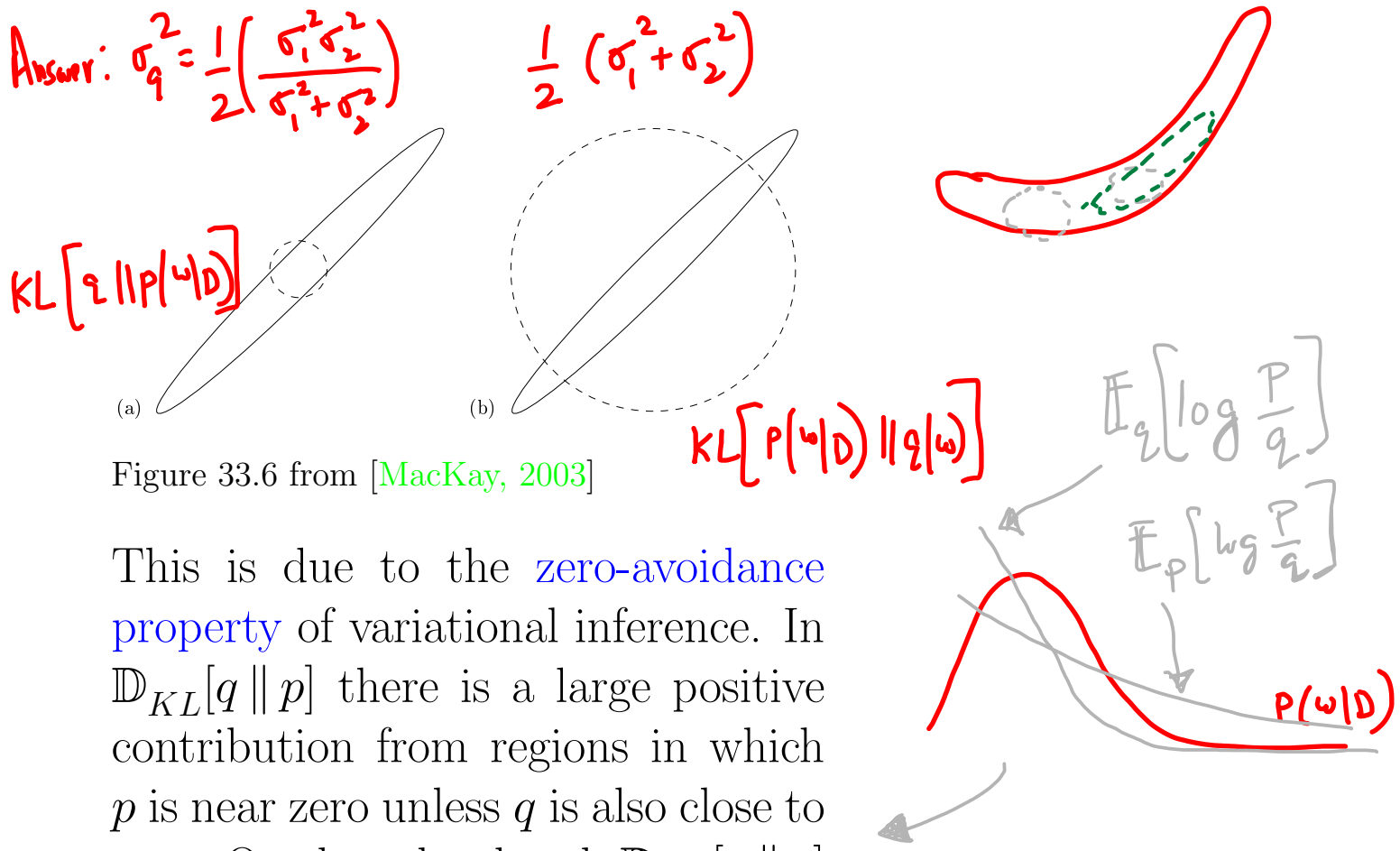
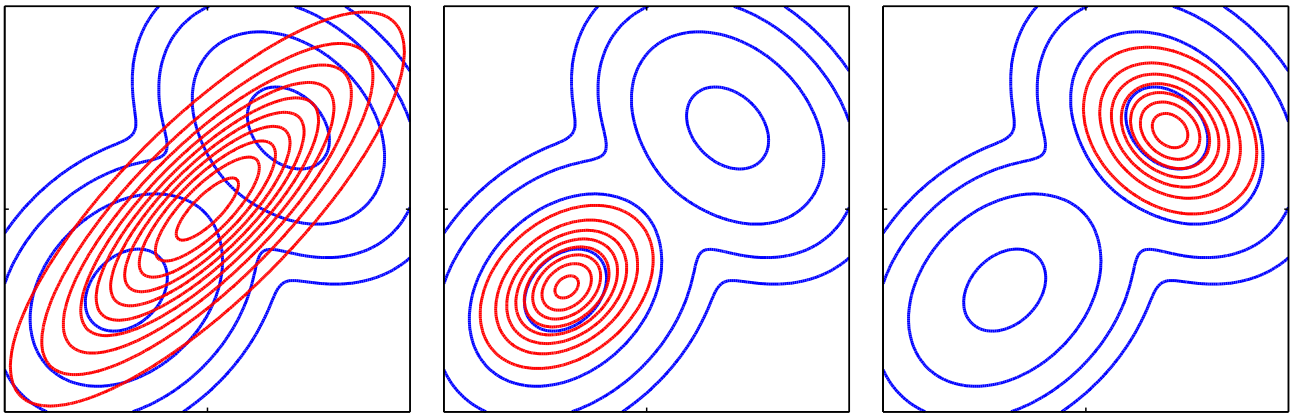


Figure 33.6 from [MacKay, 2003]

This is due to the zero-avoidance property of variational inference. In $\mathbb{D}_{KL}[q \,\|\, p]$ there is a large positive contribution from regions in which p is near zero unless q is also close to zero. On the other hand, $\mathbb{D}_{KL}[p \,\|\, q]$ is minimized by q that covers the *mass* of p.

Minimizing $\mathbb{D}_{KL}[p(\mathbf{w}|\mathcal{D}) \,\|\, q(\mathbf{w})]$ leads to a different type of method, e.g., expectation propagation. The two methods obtain very different approximations.



Taken from [Bishop, 2006] Figure 10.3.

Optimality Condition

The optimal solution for mean-field VI takes the following form:

$$\log q_i^*(w_i) = \frac{\mathbb{E}_{q_{/i}^*(\boldsymbol{w}_{/i})}\left[\log p(\mathcal{D}, \mathbf{w})\right]}{\int \mathbb{E}_{q_{/i}^*(\boldsymbol{w}_{/i})}\left[\log p(\mathcal{D}, \mathbf{w})\right] dw_i},$$

where $q_{/i}^*(\mathbf{w}_{/i}) := \prod_{j \neq i} q_j^*(w_j)$. For a class of *conditionally-conjugate* models, this update is easy to perform using coordinate descent.

Question: Suppose we want to approximate

$$p(\mathbf{w}|\mathcal{D}) := \mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \middle| \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \lambda_1 & \lambda_{12} \\ \lambda_{12} & \lambda_2 \end{bmatrix}^{-1}\right),$$

by a factorized $q_1(w_1)q_2(w_2)$. What is the optimal form of q_1 and q_2?

Coordinate-Descent

When computing $q_i(w_i)$ is easy given $q(\mathbf{w}_{/i})$, we can use a coordinate-descent algorithm to optimize ELBO.

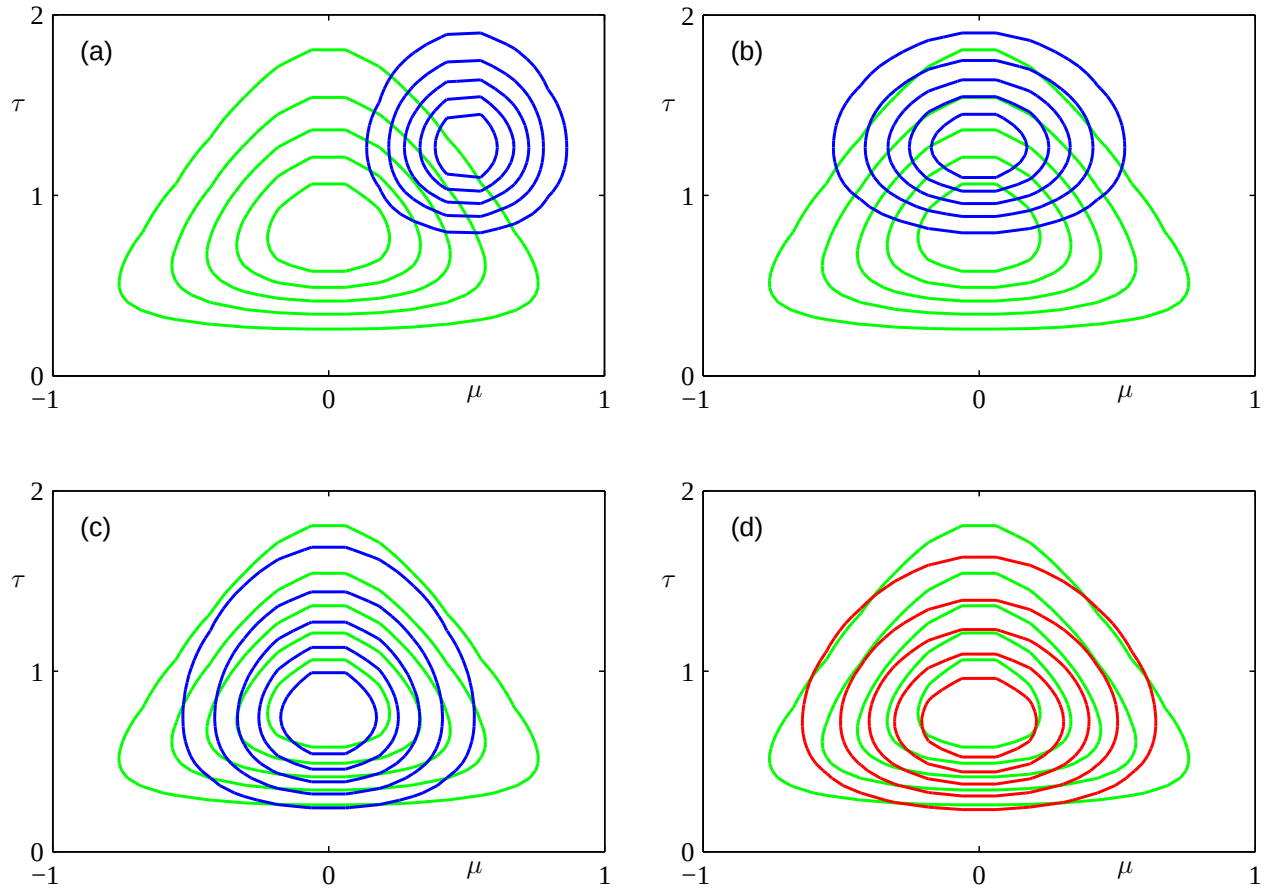


Figure 1: Coordinate descent for mean-field VI in $\mathcal{N}(\mathcal{D}|\mu, \sigma^2)$ with Gaussian prior on μ and Gamma prior on σ ([Bishop, 2006] Fig. 10.4).

Question: Derive the update for mean-field VI in a Bayesian linear-regression model $y_n \approx w_1 x_{n1} + w_2 x_{n2}$, i.e., $p(w_1, w_2|\mathcal{D}) \approx q_1(w_1) q_2(w_2)$

$$(I + X^T X)^{-1} \mu_* = X^T y$$

$$\Sigma_* = \left[\mathrm{diag}\left(I + X^T X\right)\right]^{-1}$$

This kind of update arises due to a conjugacy property of pairs of exponential-family distribution, which we describe next.

Conjugate Exponential-Family Models

An exponential-family prior distribution takes the following form:

$$p_{\eta_0}(\mathbf{w}) := h(\mathbf{w}) \exp\left[\boldsymbol{\phi}(\mathbf{w})^T \boldsymbol{\eta}_0 - A(\boldsymbol{\eta}_0)\right].$$

The prior and likelihood are conjugate distributions when the likelihood/prior can be expressed in the same form with respect to $\mathbf{w}$:

$$\begin{aligned} p(\mathcal{D}|\mathbf{w}) &= \exp\left[\boldsymbol{\phi}_1(\mathcal{D})^T \boldsymbol{\eta}_1(\mathbf{w}) - A_1(\mathbf{w})\right], \\ &= \exp\left[\boldsymbol{\phi}(\mathbf{w})^T \boldsymbol{\eta}_{10}(\mathcal{D}) - f_{10}(\mathcal{D})\right], \end{aligned}$$

for some functions $\boldsymbol{\eta}_{10}$ and f_{10}. The posterior distribution in this case is available in closed-form:

$$p(\mathbf{w}|\mathcal{D}) \propto \exp\left[\boldsymbol{\phi}(\mathbf{w})^T \{\boldsymbol{\eta}_1(\mathcal{D}) + \boldsymbol{\eta}_0\}\right].$$

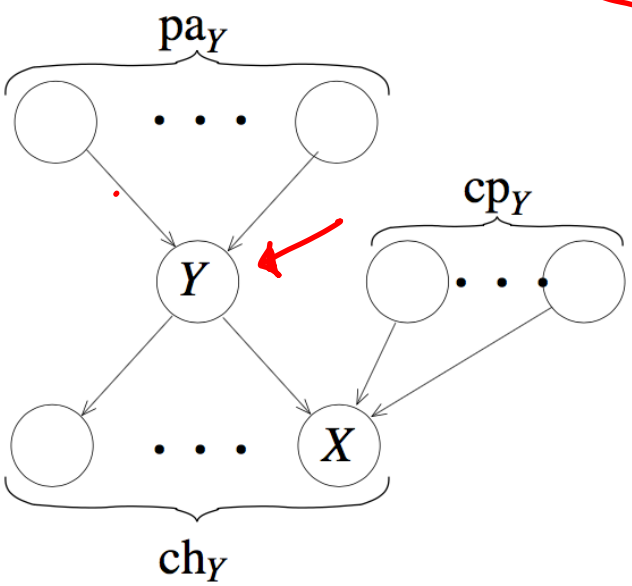
Note that a closed-form expression does not necessarily mean that the computation is easy.

Conditionally-Conjugate Models

Using the conjugacy property, efficient mean-field VI can be performed on conditionally-conjugate graphical models [Beal, 2003].

Given a Bayesian network over $\mathbf{w}$, we denote the set of the parents and children of note w_i by pa_i and ch_i respectively. We also denote the set of co-parent of a child w_j by cp_{ij}. The optimality condition can be written as follows:

$$\log q_i^*(w_i) = \mathbb{E}_{q^*_{/i}(\boldsymbol{w}_{/i})}\left[\log p(w_i|\mathrm{pa}_i)\right] + \sum_{j\in\mathrm{ch}_i} \mathbb{E}_{q^*_{/i}(\boldsymbol{w}_{/i})}\left[\log p(w_j|w_i,\mathrm{cp}_{ij})\right] + \mathrm{cnst.}$$



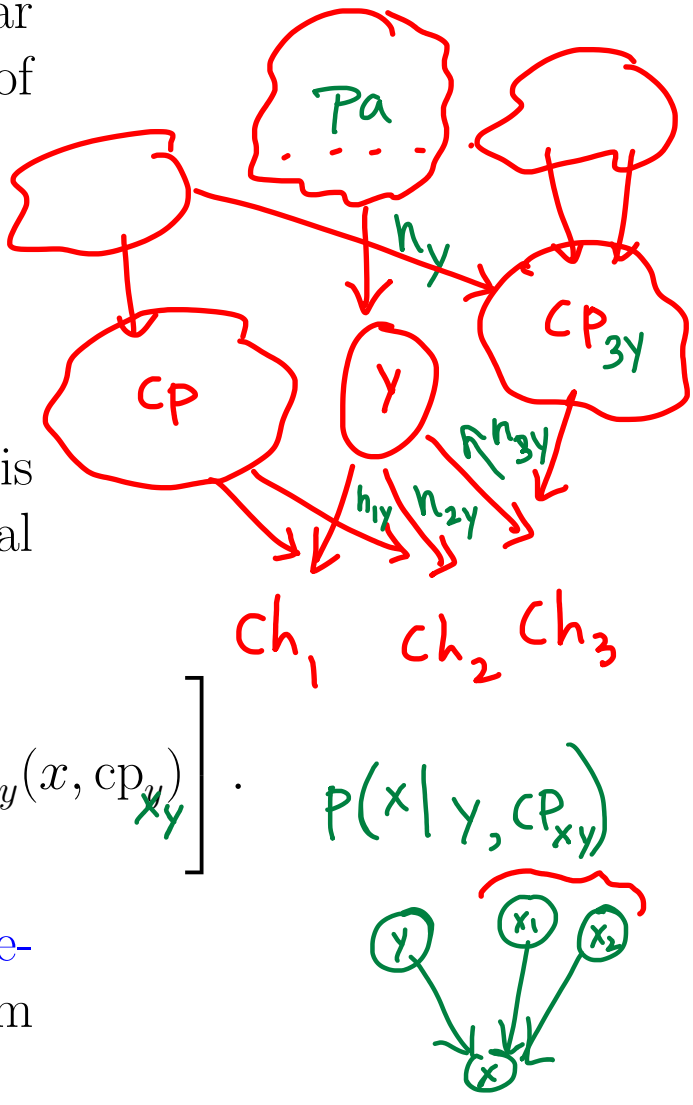
Variational Message Passing

The optimal distribution can be computed locally by simply adding *messages* from neighbors.

Consider the factor $y \to x$:

$$\begin{aligned}\log p(y|\mathrm{pa}_y) &= \boldsymbol{\eta}_y(\mathrm{pa}_y)^\top \boldsymbol{\phi}_y(y) + f_y(y) + g_y(\mathrm{pa}_y), \\ \log p(x|y,\mathrm{cp}_y) &= \boldsymbol{\eta}_x(y,\mathrm{cp}_y)^\top \boldsymbol{\phi}_x(x) + f_x(x) + g_x(y,\mathrm{cp}_x), \\ &= \boldsymbol{\eta}_{xy}(x,\mathrm{cp}_y)^\top \boldsymbol{\phi}_y(y) + f_{xy}(X,\mathrm{cp}_y).\end{aligned}$$

where the last line follows due to the conjugacy property which ensures that log of a factor is a multi-linear function of the sufficient-statistics of all of the variables involved in it.



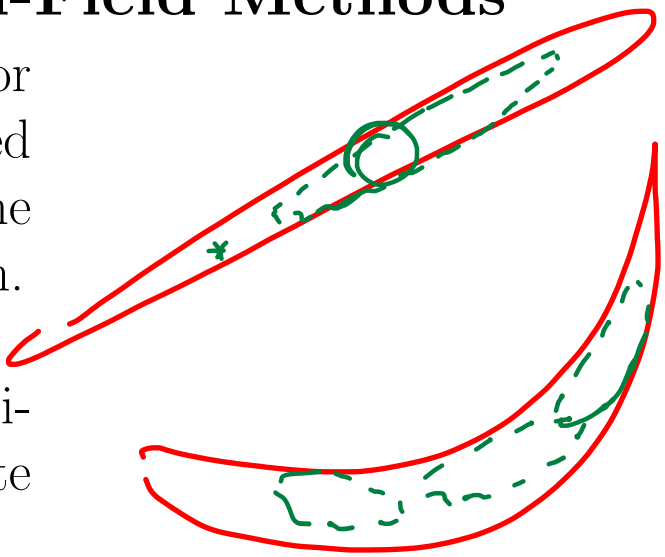
The optimal natural-parameter is obtained by summing the natural parameters of its neighbors:

$$\boldsymbol{\eta}_y^* = \mathbb{E}_{q^*_{/y}(\boldsymbol{w}_{/y})}\left[\boldsymbol{\eta}_y(\mathrm{pa}_y) + \sum_{x \in \mathrm{ch}_y} \boldsymbol{\eta}_{xy}(x,\mathrm{cp}_y)\right].$$

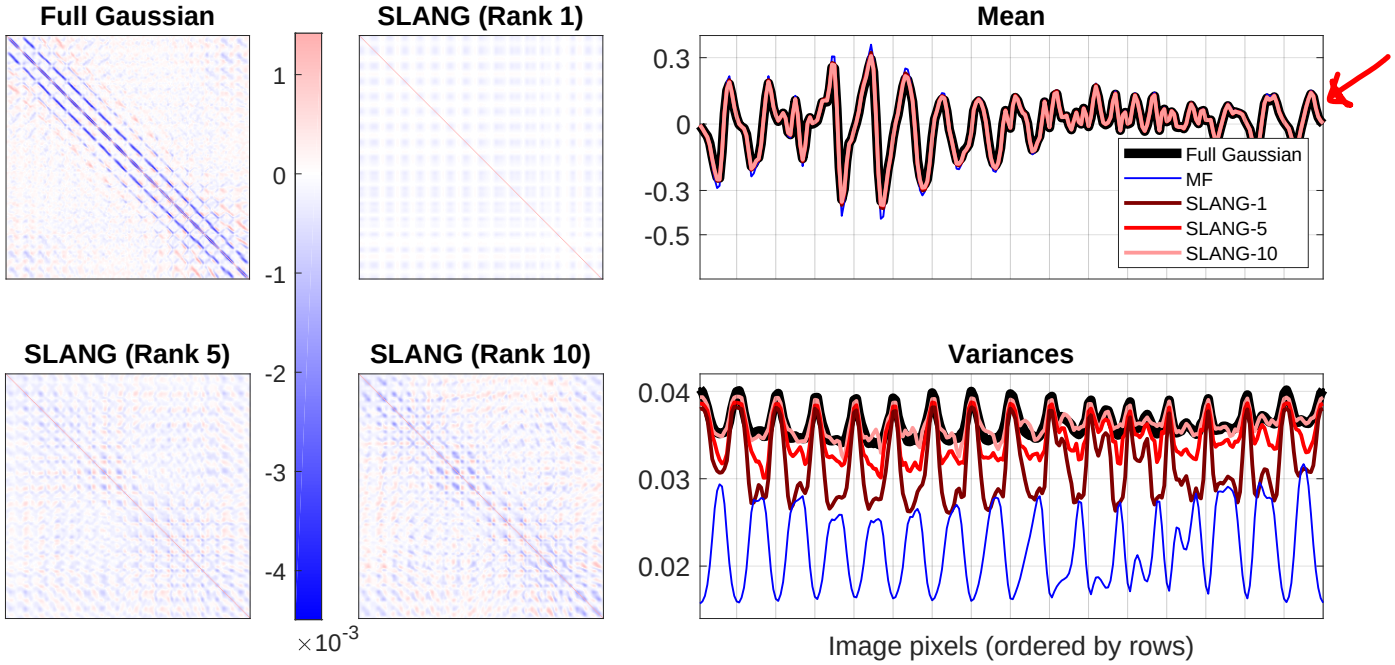
This is the variational message-passing (VMP) algorithm [Winn and Bishop, 2005].

Pros and Cons of Mean-Field Methods

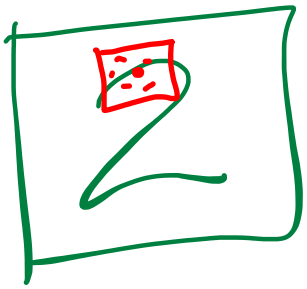
Inference could be very fast for some models. Also, we do not need to make any approximations for the form of the variational distribution.



Mean-field underestimates the variance which could be very inaccurate in many situations.



The methods discussed so far only work for conditionally-conjugate models (i.e., not for logistic regression and DNNs).



4 Gradient-Based VI

An alternative to mean-field is to choose q to be of a specific parametric form, and then optimize the lower bound with respect to the parameters of q. Given parameters $\boldsymbol{\theta}$, we denote the distribution as $q_{\boldsymbol{\theta}}(\mathbf{w})$, for example, $\boldsymbol{\theta} := \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ for a Gaussian approximation with mean $\mathbf{m}$ and covariance $\boldsymbol{\Sigma}$.

$q_\theta(w)$

$\mathcal{P}$ = "a small class of parameteric distribution"

The lower bound can be written as a function of $\boldsymbol{\theta}$: $\mathcal{L}_{VI}(\theta) :=$ (approximate Posterior / Prior)

likelihood

$$= \mathbb{E}_{q_\theta(\boldsymbol{w})}\left[\log p(\mathcal{D}|\mathbf{w})\right] - \mathbb{D}_{KL}[q_\theta(\mathbf{w}) \,\|\, p(\mathbf{w})],$$

Data-fit

Reg

$$= \mathbb{E}_{q_\theta(\boldsymbol{w})}\left[\mathcal{L}_{MAP}(\mathbf{w}) - \log q_\theta(\mathbf{w})\right].$$

$p(D,w)$

$$= \mathbb{E}_q\left[\log \frac{p(D|w)\,p(w)}{q_\theta(w)}\right]$$

This is attractive due to its similarity to the MAP objective. If we can compute *unbiased* stochastic gradients, we can use a stochastic-gradient method, e.g., SGD.

"Hat"

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \rho \widehat{\frac{\partial \mathcal{L}_{VI}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}},$$

where ρ is a learning rate.

$$\mathbb{E}_q\left[\log q_\theta(w)\right] = \log|\Sigma|$$

$$\mathcal{L}_{MAP}(\theta) = \sum_{i=1}^{N} \log p(y_i|f_\theta(x_i)) + \log p(\theta)$$

How to compute stochastic gradients?

Stochastic Gradients I

REINFORCE is an approach to compute stochastic gradients of generic functions that can be written as an expectation of q_θ [Williams, 1992]. It is based on the log-derivative trick:

$$\frac{\partial q_\theta}{\partial \boldsymbol{\theta}} = q_\theta \frac{\partial \log q_\theta}{\partial \boldsymbol{\theta}}.$$

$$\frac{\partial}{\partial \theta} \mathbb{E}_{q_\theta}[f(\omega)] = \int f(\omega) \frac{\partial}{\partial \theta} q_\theta(\omega) d\omega = \int f(\omega) \frac{\partial \log q_\theta(\omega)}{\partial \theta} q_\theta(\omega) d\omega = \mathbb{E}_{q_\theta(\omega)}\left[f(\omega) \frac{\partial \log q_\theta(\omega)}{\partial \theta}\right] \approx f(\omega_*) \left[\frac{\partial \log q_\theta(\omega)}{\partial \theta}\right]_{\omega=\omega_*}$$

$$\frac{\partial \log q_\theta(\omega)}{\partial \theta} = \frac{1}{q_\theta(\omega)} \frac{\partial q_\theta(\omega)}{\partial \theta}$$

The REINFORCE gradient estimator with one sample $\mathbf{w}_* \sim q_\theta(\mathbf{w})$ and a data-minibatch is given by:

$$\frac{\partial}{\partial \theta} \mathbb{E}_{q_\theta}[\log q_\theta(\omega)]$$

$$\frac{\partial \mathcal{L}_{VI}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \approx \frac{\partial \log q_\theta(\mathbf{w})}{\partial \boldsymbol{\theta}} \left\{ \widehat{\mathcal{L}}_{MAP}(\mathbf{w}) - \log q_\theta(\mathbf{w})) - 1 \right\}.$$

$\widehat{\mathcal{L}}_{MAP}(\mathbf{w}) \approx N \log p(y_n | f_\omega(x)) + \log p(\theta)$

This type of approximation is also referred to as a doubly stochastic-gradient method.

REINFORCE is widely applicable, but might suffer from higher variance. Some variance reduction methods are discussed in [Ranganath et al., 2014].

Stochastic Gradients II

When q_θ can be *reparameterized* in terms of a simpler parameter-free distribution, we can use the *reparameterization trick* [Kingma and Welling, 2013], e.g., let $q_\theta := \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2))$ with $\boldsymbol{\theta} := \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$, then a sample from q_θ can be written as,

$$\mathbf{w}(\boldsymbol{\theta}; \boldsymbol{\epsilon}) = \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|0, \mathbf{I}).$$

Here is a stochastic gradient with one Monte-Carlo sample $\mathbf{w}^*$:

$$\frac{\partial \mathcal{L}_{VI}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \approx \frac{\partial \mathbf{w}^*}{\partial \boldsymbol{\theta}} \frac{\partial \mathcal{L}_{MAP}(\mathbf{w}^*)}{\partial \mathbf{w}} - \frac{\partial \mathbf{w}^*}{\partial \boldsymbol{\theta}} \frac{\partial \log q_\theta(\mathbf{w}^*)}{\partial \mathbf{w}} - \frac{\partial \log q_\theta(\mathbf{w}^*)}{\partial \boldsymbol{\theta}}.$$

This approximation makes use of the gradient of the objective and usually has lower variance than REINFORCE estimator, however it is only applicable when the distribution is reparameterizable.

Application to BNNs

We now discuss recent VI methods for Bayesian neural networks (BNNs). The most common approach is to use a Gaussian prior $p(\mathbf{w})$ and a Gaussian approximation $q_\theta(\mathbf{w})$ with $\boldsymbol{\theta} := \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. We can then use a reparameterization trick to compute the gradients.

An alternative is to use the Bonnet's and Price's theorems [Opper and Archambeau, 2009, Rezende et al., 2014] to express the gradients of the expectation of $f(\mathbf{w})$ with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in terms of the gradient and Hessian of $f(\mathbf{w})$,

$$\nabla_\mu \mathbb{E}_q\left[f(\mathbf{w})\right] = \mathbb{E}_q\left[\nabla_w f(\mathbf{w})\right],$$
$$\nabla_\Sigma \mathbb{E}_q\left[f(\mathbf{w})\right] = \tfrac{1}{2}\mathbb{E}_q\left[\nabla^2_{ww} f(\mathbf{w})\right].$$

We can also avoid computing Hessian by using a Gauss-Newton approximation [Graves, 2011]:

$$\nabla_\Sigma \mathbb{E}_q\left[f(\mathbf{w})\right] = \tfrac{1}{2}\mathbb{E}_q\left[\nabla_w f(\mathbf{w}) \nabla_w f(\mathbf{w})^\top\right].$$

We can perform VI just by using backpropagation. An alternative method is Bayes-by-Backprop [Blundell et al., 2015].

5 Natural-Gradient VI

Variational distributions has a Riemannian manifold associated with them. We can exploit it to improve convergence and also to obtain simple updates. Overall, this leads to methods that unify message-passing and gradient-based methods.

Exponential-Family Approximations

We will focus on exponential-family variational distribution,

$$q_\lambda(\mathbf{w}) := h(\mathbf{w}) \exp\left[\boldsymbol{\lambda}^\top \boldsymbol{\phi}(\mathbf{w}) - A(\boldsymbol{\lambda})\right],$$

where $\boldsymbol{\lambda}$ is the natural-parameter. We also need to define expectation parameter and Fisher information matrix (FIM):

$$\boldsymbol{\mu}(\boldsymbol{\lambda}) := \mathbb{E}_{q_\lambda}[\boldsymbol{\phi}(\mathbf{w})],$$
$$\mathbf{F}(\boldsymbol{\lambda}) := \mathbb{E}_{q_\lambda}[\nabla_\lambda \log q_\lambda(\mathbf{w}) \nabla_\lambda \log q_\lambda(\mathbf{w})^\top].$$

We will use the following properties:

$$\boldsymbol{\mu}(\boldsymbol{\lambda}) := \nabla_\lambda A(\boldsymbol{\lambda}), \qquad \mathbf{F}(\boldsymbol{\lambda}) := \nabla^2_\lambda A(\boldsymbol{\lambda}).$$

We will assume a minimal representation which makes sure that mapping between $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ is one to one.

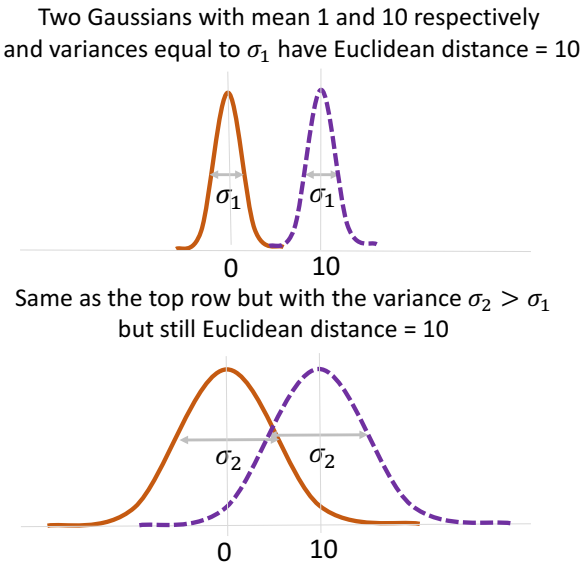
Natural Gradients

Given the FIM, natural gradients in the natural-parameter space are defined as follows:

$$\widetilde{\nabla}_\lambda f(\boldsymbol{\lambda}) := \mathbf{F}(\boldsymbol{\lambda})^{-1}\nabla_\lambda \mathcal{L}(\boldsymbol{\lambda}).$$

The FIM specifies a Riemannian geometry which gives a more natural way of measuring distances between distributions than the Euclidean distance used in SGD.

$$\begin{aligned}\boldsymbol{\lambda}_{t+1} &= \boldsymbol{\lambda}_t + \rho_t \nabla_\lambda \mathcal{L}(\boldsymbol{\lambda}_t),\\ &= \arg\max_{\lambda}\ \boldsymbol{\lambda}^\top \nabla_\lambda \mathcal{L}(\boldsymbol{\lambda}_t) - \frac{1}{2\rho_t}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_t\|^2.\end{aligned}$$



Replacing the Euclidean distance $\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_t\|^2$ by a Riemannian metric, $(\boldsymbol{\lambda}-\boldsymbol{\lambda}_t)^\top \mathbf{F}(\boldsymbol{\lambda}_t)(\boldsymbol{\lambda}-\boldsymbol{\lambda}_t)$, we get the natural-gradient descent:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \alpha_t \widetilde{\nabla}_\lambda \mathcal{L}(\boldsymbol{\lambda}_t).$$

Optimality Condition

Natural-gradient is not only useful to improve convergence, but also naturally appears in the optimality condition:

$$\boldsymbol{\lambda}^* = \mathbf{F}(\boldsymbol{\lambda}^*)^{-1} \nabla_\lambda \mathbb{E}_{q_{\lambda^*}} \left[\mathcal{L}_{MAP}(\mathbf{w}) \right].$$

The optimal natural-parameter is equal to the *natural-gradient of expected MAP objective.*

Natural-Gradient Computation

The FIM might be expensive to compute, but in some cases we can simplify the computation using the following relationship:

$$\mathbf{F}(\boldsymbol{\lambda})^{-1} \nabla_\lambda f(\boldsymbol{\lambda}) = \nabla_\mu f(\boldsymbol{\lambda}).$$

For example, for a Gaussian distribution, ∇_μ is much easier to compute than an explicit computation of the FIM.

Message-Passing using Natural Gradients

Suppose we want to compute a Gaussian approximation for the following lower bound with a Gaussian prior $p(\mathbf{w})$,

$$\sum_{n=1}^{N} \mathbb{E}_{q_\lambda(\boldsymbol{w})} \left[\log p(y_n|f_w(\mathbf{x}_n))\right] - \mathbb{D}_{KL}[q_\lambda(\mathbf{w}) \,\|\, p(\mathbf{w})].$$

The natural-gradient of the second term is equal to $\boldsymbol{\eta} - \boldsymbol{\lambda}$. Using this we can obtain a stochastic natural-gradient descent update:

$$\boldsymbol{\lambda}_{t+1} = (1-\alpha_t)\boldsymbol{\lambda}_t + \alpha_t \left[\boldsymbol{\eta} + N\nabla_\mu \mathbb{E}_{q_\lambda(\boldsymbol{w})} \left[\log p(y_n|f_w(\mathbf{x}_n))\right]\right].$$

In general, the natural-gradient of terms that are conjugate to q_λ is very simple to compute using an update similar to conditionally-conjugate models. This algorithm is proposed in [Khan and Lin, 2017].

For conditionally-conjugate models, this approach reduces to stochastic variational inference [Hoffman et al., 2013].

Natural-Gradient VI for BNNs

Natural-gradient VI for BNN is also simpler than gradient-based VI.

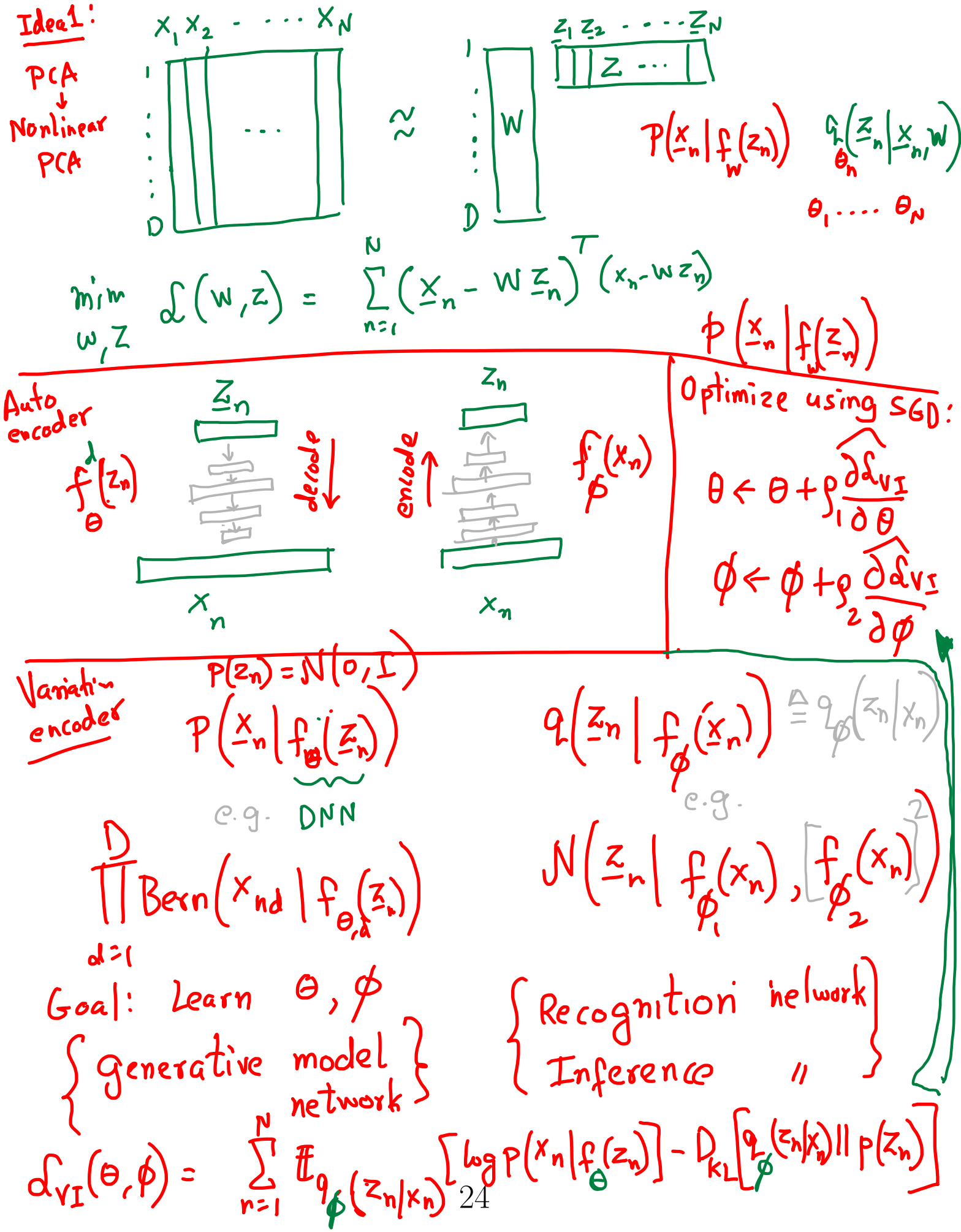
$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t - \beta_t \frac{\widehat{\nabla} \log p(y_n|f_{w_t}(\mathbf{x}_n)) + \tilde{\lambda}\boldsymbol{\mu}_t}{\mathbf{s}_{t+1} + \tilde{\lambda}},$$

$$\mathbf{s}_{t+1} = (1-\beta_t)\mathbf{s}_t + \beta_t \, \widehat{\nabla}^2 \log p(y_n|f_w(\mathbf{x}_n)),$$

where we have used one data example n and one Monte-Carlo (MC) sample $\boldsymbol{\theta}_t \sim \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2)$ with $\boldsymbol{\sigma}_t^2 := 1/[N(\mathbf{s}_t + \tilde{\lambda})]$ and $\tilde{\lambda} := \lambda/N$.

If we replace Hessian by a Gauss-Newton approximation, this is equivalent to a weight-perturbed RMSprop optimizer. A version with the Adam optimizer is derived in [Khan et al., 2018].

6 Variational Auto-Encoders



7 Further reading

- For Laplace's method, read Section 4.4 in [Bishop, 2006] and Chapter 27 in [MacKay, 2003].
- For VI, read Chapter 10 in [Bishop, 2006] and Chapter 33 in [MacKay, 2003].
- For more details on reformulation of Bayesian inference as an optimization problem, see [Zhu et al., 2014] or [Williams, 1980].
- See more on minimum description-length principle in [Hinton and Van Camp, 1993].
- A good reference for exponential-family distributions is [Wainwright and Jordan, 2008], Chapter 3.
- For natural-gradients and information geometry, [Amari, 2016] is an easy to read book.

References

[Amari, 2016] Amari, S. (2016). *Information geometry and its applications*. Springer.

[Barber and Bishop, 1998] Barber, D. and Bishop, C. M. (1998). Ensemble learning in Bayesian neural networks. *Generalization in Neural Networks and Machine Learning*, 168:215–238.

[Beal, 2003] Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of Cambridge.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Blundell et al., 2015] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural networks. In *International Conference on Machine Learning*, pages 1613–1622.

[Graves, 2011] Graves, A. (2011). Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, pages 2348–2356.

[Hinton and Van Camp, 1993] Hinton, G. E. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Annual Conference on Computational Learning Theory*, pages 5–13.

[Hoffman et al., 2013] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

[Khan and Lin, 2017] Khan, M. E. and Lin, W. (2017). Conjugate-computation variational inference: converting variational inference in non-conjugate models to inferences in conjugate models. In *International Conference on Artificial Intelligence and Statistics*, pages 878–887.

[Khan et al., 2018] Khan, M. E., Nielsen, D., Tangkaratt, V., Lin, W., Gal, Y., and Srivastava, A. (2018). Fast and scalable bayesian deep learning by weight-perturbation in adam. In *Proceedings of the 35th International Conference on Machine Learning*.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

[MacKay, 2003] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[Opper and Archambeau, 2009] Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792.

[Ranganath et al., 2014] Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *International conference on Artificial Intelligence and Statistics*, pages 814–822.

[Rezende et al., 2014] Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286.

[Ritter et al., 2018] Ritter, H., Botev, A., and Barber, D. (2018). A scalable laplace approximation for neural networks. In *International Conference on Learning Representations*.

[Rue et al., 2009] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of Royal Statistical Sociecty, Series B*, 71:319–392.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1–2:1–305.

[Williams, 1980] Williams, P. M. (1980). Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, 31(2):131–144.

[Williams, 1992] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

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

[Zhu et al., 2014] Zhu, J., Chen, N., and Xing, E. P. (2014). Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1):1799–1847.

List of concepts

(Notes)

(Notes)