

Nordic probabilistic AI school

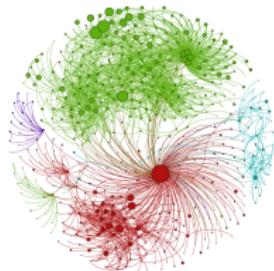
Variational Inference and Optimization

Helge Langseth, Andrés Masegosa, and Thomas Dyhre Nielsen

June 14, 2022

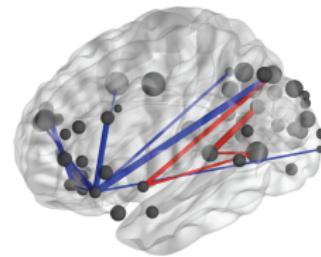
Introduction

Examples



Communities discovered in a 3.7M node network of U.S. Patents

[Gopalan and Blei, PNAS 2013]



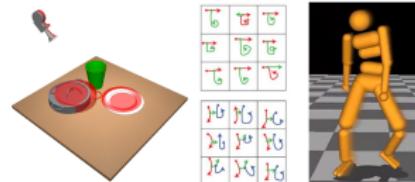
Neuroscience analysis of 220 million fMRI measurements

[Manning et al., PLOS ONE 2014]



Analysis of 1.7M taxi trajectories, in Stan

[Kucukelbir et al., 2016]



Scenes, concepts and control.

[Eslami et al., 2016, Lake et al. 2015]

Images borrowed from David Blei et al.: *Variational Inference: Foundations and Modern Methods* (NeurIPS Tutorial, 2016)

Common challenges in many real-world projects:

- **Modelling:** Efficient representations, incorporate domain expert knowledge, ...
- **Data:** Missing data, erroneous data, low signal-to-noise ratio, ...
- **Scalability:** Large number of variables, large number of observations, ...
- **Robustness:** Statistical variations, concept drift, adversarial attacks, ...
- **Trustworthiness:** Uncertainty awareness, , ...
- **Regulations:** Transparency, bias, ...

Our strategy: Probabilistic Machine Learning

- Build a **probabilistic model**.
- Apply **probabilistic inference algorithms**.

ProbAi – 2022

Bayesian Machine Learning = Probabilistic model + Bayesian inference

- **Likelihood-part:** A probabilistic model typically defined by $p(x | \theta)$.
- **Prior:** $p(\theta)$ reflects our *a priori* belief about the parameters θ .

Bayesian Machine Learning = Probabilistic model + Bayesian inference

- **Likelihood-part:** A probabilistic model typically defined by $p(\mathbf{x} | \boldsymbol{\theta})$.
- **Prior:** $p(\boldsymbol{\theta})$ reflects our *a priori* belief about the parameters $\boldsymbol{\theta}$.

Now we can calculate the posterior over $\boldsymbol{\theta}$ given observations $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$,

$$p(\boldsymbol{\theta} | \mathcal{D}) = \frac{p(\boldsymbol{\theta}) p(\mathcal{D} | \boldsymbol{\theta})}{p(\mathcal{D})},$$

... and, e.g., the predictive distribution of a new observation \mathbf{x}' :

$$p(\mathbf{x}' | \mathcal{D}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}' | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta}.$$

Bayesian Machine Learning = Probabilistic model + Bayesian inference

- **Likelihood-part:** A probabilistic model typically defined by $p(\mathbf{x} | \boldsymbol{\theta})$.
- **Prior:** $p(\boldsymbol{\theta})$ reflects our *a priori* belief about the parameters $\boldsymbol{\theta}$.

Now we can calculate the posterior over $\boldsymbol{\theta}$ given observations $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$,

$$p(\boldsymbol{\theta} | \mathcal{D}) = \frac{p(\boldsymbol{\theta}) p(\mathcal{D} | \boldsymbol{\theta})}{p(\mathcal{D})},$$

... and, e.g., the predictive distribution of a new observation \mathbf{x}' :

$$p(\mathbf{x}' | \mathcal{D}) = \int_{\boldsymbol{\theta}} p(\mathbf{x}' | \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta}.$$

Being Bayesian means maintaining a distribution over $\boldsymbol{\theta}$.

Using a point-estimate for $\boldsymbol{\theta}$ is **probabilistic** (but not **Bayesian**) ML.

Example: Linear regression

A Bayesian linear regression with univariate explanatory variables:

Likelihood – $p(\mathcal{D} | \theta)$: $p(y_i | \mathbf{w}, x_i, \sigma_y^2) = \mathcal{N}(w_0 + w_1 \cdot x_i, \sigma_y^2)$

Note! The observation noise, σ_y^2 is known, so the parameter-set is simply $\theta = \{\mathbf{w}\}$.

Prior – $p(\theta)$: $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \sigma_w^2)$

Bayesian Linear regression – Full model:

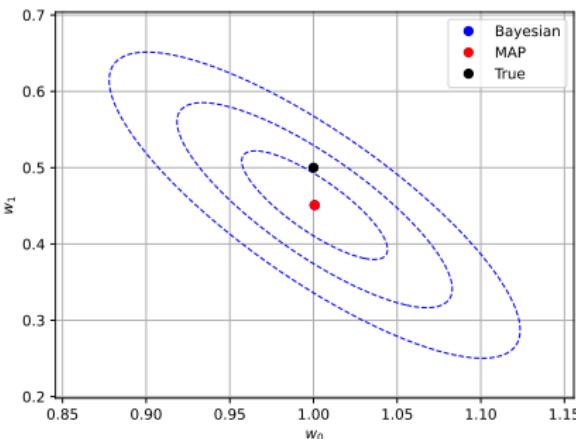
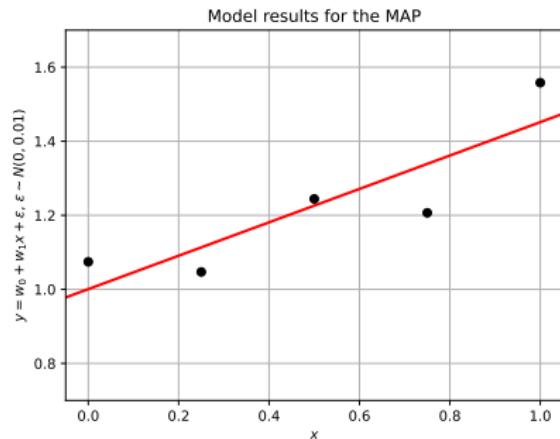
$$p(\mathcal{D}, \theta) = p(\{y_i\}_{i=1}^n, \mathbf{w} | \{\mathbf{x}_i\}_{i=1}^n, \sigma_y^2, \sigma_w^2) = \overbrace{p(\mathbf{w} | \sigma_w^2)}^{p(\theta)} \overbrace{\prod_{i=1}^n p(y_i | \mathbf{w}, \mathbf{x}_i, \sigma_y^2)}^{p(\mathcal{D} | \theta)}$$

Example: Linear regression – MAP vs (fully) Bayesian

Bayes linear regression w/ some fake data:

- We have generated $N = 5$ examples from $y_i = 1.0 + 0.5 \cdot x_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$.
- Weights unknown a priori, so here we use the vague priors $w_j \sim \mathcal{N}(0, 100^2)$.

Results for the fully Bayesian model and the MAP:



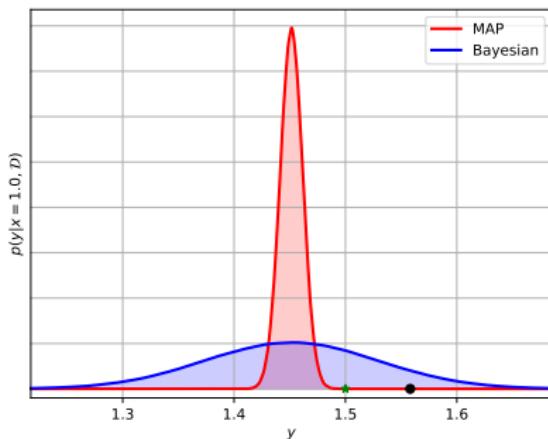
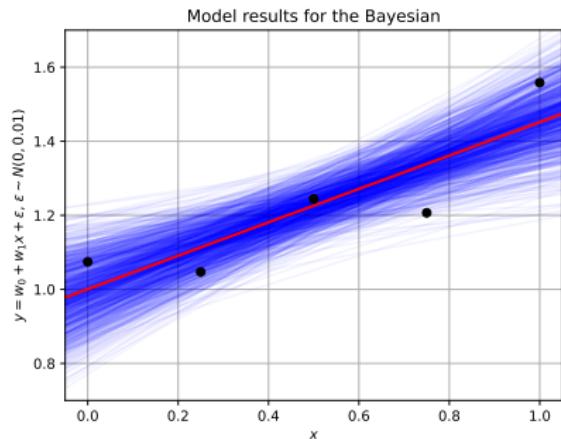
- **MAP:** Reasonable point estimate; No model uncertainty;
- **Bayes:** Model uncertainty around same MAP estimate;

Example: Linear regression – MAP vs (fully) Bayesian

Bayes linear regression w/ some fake data:

- We have generated $N = 5$ examples from $y_i = 1.0 + 0.5 \cdot x_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 0.1^2)$.
- Weights unknown a priori, so here we use the vague priors $w_j \sim \mathcal{N}(0, 100^2)$.

Results for the fully Bayesian model and the MAP:



- **MAP:** Reasonable point estimate; No model uncertainty; Predictive uncertainty degenerated to observation noise: poor fit wrt. true value and observation.
- **Bayes:** Model uncertainty around same MAP estimate; Captures model uncertainty well; Predictive distribution reasonable.

Bayesian inference is in principle easy using Bayes' rule:

$$p(\boldsymbol{\theta} | \mathcal{D}) = \frac{p(\boldsymbol{\theta}) p(\mathcal{D} | \boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\boldsymbol{\theta}) p(\mathcal{D} | \boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) p(\mathcal{D} | \boldsymbol{\theta}) d\boldsymbol{\theta}}$$

Note! This can only be solved analytically for **some simple models** (e.g., linear regression), but typically not for any of the really interesting models.

The big plan today: Use **optimization** to approximate $p(\boldsymbol{\theta} | \mathcal{D})$

What we want:

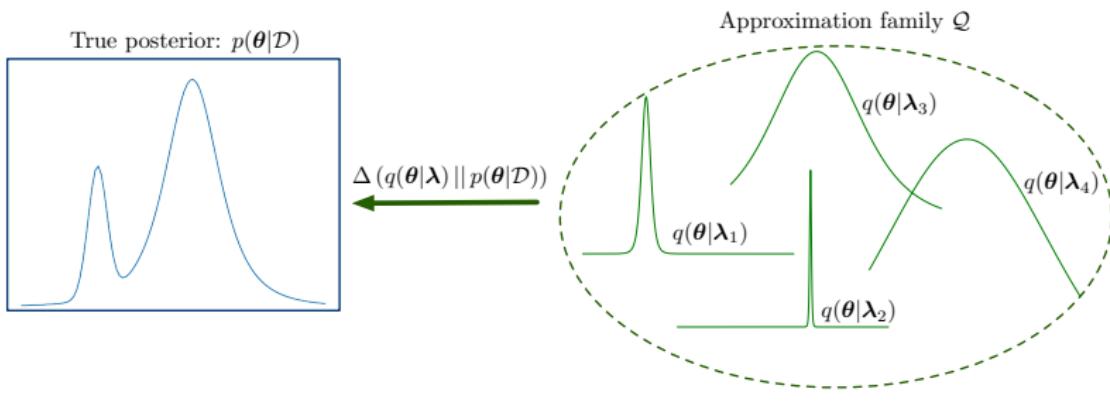
- Computationally efficient;
- Well-behaved objective;
- Easy integration with other frameworks.

What we don't want:

- No purely sampled-based techniques (like Gibbs sampling);
- No degenerate solutions (point estimators like MAP).

Variational Bayes: Approximate inference by optimization

Approximate inference through optimization – Main idea



Approximate inference through optimization

- **General goal:** Somehow approximate $p(\theta | \mathcal{D})$ with a $q(\theta | \mathcal{D})$.
- **Note!** We use $q(\theta)$ as a short-hand for $q(\theta | \mathcal{D})$.

Formalization of approximate inference through optimization:

Given a family of tractable distributions \mathcal{Q} and a distance measure between distributions Δ , choose

$$\hat{q}(\theta) = \arg \min_{q \in \mathcal{Q}} \Delta(q(\theta) || p(\theta | \mathcal{D})).$$

Decisions to be made:

- ➊ How to define $\Delta(\cdot || \cdot)$ so that we end up with a high-quality solution?
 - How to work with $\Delta(q(\theta) || p(\theta | \mathcal{D}))$ when we don't know what $p(\theta | \mathcal{D})$ is?
- ➋ How to define a family of distributions \mathcal{Q} that is both flexible enough to generate good approximations and restrictive enough to support efficient calculations?

Desiderata

To use Δ to measure the distance from an object f to an object g it would be relevant to require that Δ has the following properties:

Positivity: $\Delta(f \parallel g) \geq 0$ and $\Delta(f \parallel g) = 0$ if and only if $f = g$.

Symmetry: $\Delta(f \parallel g) = \Delta(g \parallel f)$

Triangle: For objects f , g , and h we have that $\Delta(f \parallel g) \leq \Delta(f \parallel h) + \Delta(h \parallel g)$.

Desiderata

To use Δ to measure the distance from an object f to an object g it would be relevant to require that Δ has the following properties:

Positivity: $\Delta(f \parallel g) \geq 0$ and $\Delta(f \parallel g) = 0$ if and only if $f = g$.

Symmetry: $\Delta(f \parallel g) = \Delta(g \parallel f)$

Triangle: For objects f , g , and h we have that $\Delta(f \parallel g) \leq \Delta(f \parallel h) + \Delta(h \parallel g)$.

Standard choice when working with probability distributions

The **Kullback-Leibler divergence** is the standard distance measure:

$$\text{KL}(f \parallel g) = \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \log \left(\frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta} \sim f} \left[\log \left(\frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})} \right) \right].$$

Notice that while $\text{KL}(f \parallel g)$ obeys the positivity criterion, it satisfies neither symmetry nor the triangle inequality. It is thus **not a proper distance measure**.

Information-projection

- Minimizes $\text{KL}(q||p) = -\mathbb{E}_{\theta \sim q}[\log p(\theta)] - \mathcal{H}_q$.
- Preference given to q that has:
 - High q -probability allocated to p -probable regions.
 - Small q in any region where p is small.
“ $p(\theta) = 0 \implies q(\theta) = 0$ ”.
 - High entropy (\sim variance)

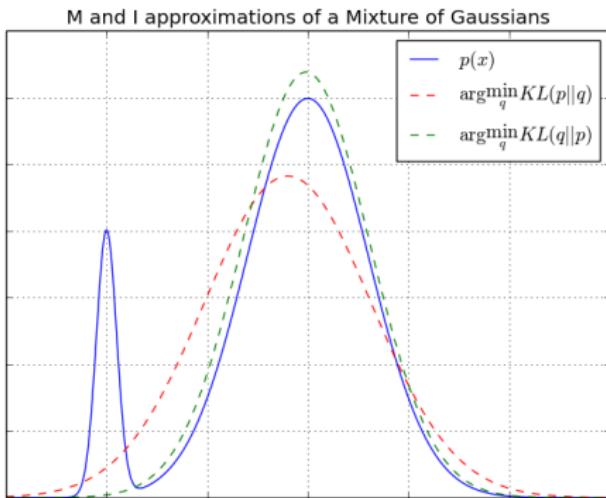
Moment-projection

- Minimizes $\text{KL}(p||q) = -\mathbb{E}_{\theta \sim p}[\log q(\theta)] - \mathcal{H}_p$.
- Preference given to q that has:
 - High q -probability allocated to p -probable regions.
 - $q(\theta) > 0$ in any region where p is non-negligible.
“ $p(\theta) > 0 \implies q(\theta) > 0$ ”
 - No explicit focus of entropy

Cheat-sheet:

- Entropy:** $\mathcal{H}_f = -\int_{\theta} f(\theta) \log(f(\theta)) d\theta = -\mathbb{E}_f [\log(f(\theta))]$.
- KL-divergence:** $\text{KL}(f||g) = \mathbb{E}_f \left[\log \left(\frac{f(\theta)}{g(\theta)} \right) \right] = -\mathbb{E}_f [\log(g(\theta))] - \mathcal{H}_f$.
- Intuition:** Cheat a bit (measure-zero, limit-zero-rates, etc.) and think
 $\text{If } g(\theta_0) \approx 0, \text{ then } -\mathbb{E}_{\theta \sim f}[\log g(\theta)] \text{ becomes 'huge' unless } f(\theta_0) \approx 0$
 because $\lim_{x \rightarrow 0^+} \log(x)$ diverges, while $\lim_{x \rightarrow 0^+} x \cdot \log(x) = 0$.

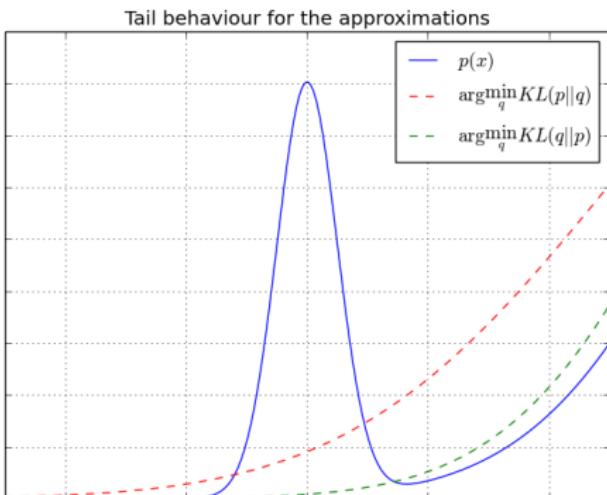
Moment and Information projection – main difference



Example: Approximating a Mix-of-Gaussians by a single Gaussian

- Moment projection – optimizing $KL(p||q)$ – has slightly larger variance.
- Similar mean values, but Information projection – optimizing $KL(q||p)$ – focuses mainly on the most prominent mode.

Moment and Information projection – main difference



Example: Approximating a Mix-of-Gaussians by a single Gaussian

- Moment projection – optimizing $KL(p||q)$ – has slightly larger variance.
- Similar mean values, but Information projection – optimizing $KL(q||p)$ – focuses mainly on the most prominent mode.
- M-projection is **zero-avoiding**, while I-projection is **zero-forcing**.

VB uses information projections:

Variational Bayes relies on **information projections**, i.e., approximates $p(\boldsymbol{\theta} \mid \mathcal{D})$ by

$$\hat{q}(\boldsymbol{\theta}) = \arg \min_{q \in \mathcal{Q}} \text{KL}(q(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \mathcal{D}))$$

- **Positives:**

- Clever interpretation when used for Bayesian machine learning.
 - We will end up with an objective that lower-bounds the marginal log likelihood, $\log p(\mathcal{D})$.
- Very efficient when combined with cleverly chosen \mathcal{Q} .

- **Negatives:**

- May result in *zero-forcing* behaviour.
 - Typical choice of \mathcal{Q} can make this issue even more prominent.

ELBO: Evidence Lower-BOund

Notice how we can rearrange the KL divergence as follows:

$$\begin{aligned} \text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})) &= \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} \right] = \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta}) \cdot p(\mathcal{D})}{p(\boldsymbol{\theta}|\mathcal{D}) \cdot p(\mathcal{D})} \right] \\ &= \log p(\mathcal{D}) - \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathcal{D})} \right] = \log p(\mathcal{D}) - \mathcal{L}(q) \end{aligned}$$

Evidence Lower Bound (ELBO): $\mathcal{L}(q) = -\mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathcal{D})} \right] = \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right].$

ELBO: Evidence Lower-Bound

Notice how we can rearrange the KL divergence as follows:

$$\begin{aligned} \text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})) &= \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} \right] = \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta}) \cdot p(\mathcal{D})}{p(\boldsymbol{\theta}|\mathcal{D}) \cdot p(\mathcal{D})} \right] \\ &= \log p(\mathcal{D}) - \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathcal{D})} \right] = \log p(\mathcal{D}) - \mathcal{L}(q) \end{aligned}$$

Evidence Lower Bound (ELBO): $\mathcal{L}(q) = -\mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathcal{D})} \right] = \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right].$

VB focuses on ELBO:

$$\log p(\mathcal{D}) = \mathcal{L}(q) + \text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D}))$$

Since $\log p(\mathcal{D})$ is constant wrt. the distribution q it follows:

- We can minimize $\text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D}))$ by maximizing $\mathcal{L}(q)$
- This is **computationally simpler** because it uses $p(\boldsymbol{\theta}, \mathcal{D})$ and not $p(\boldsymbol{\theta}|\mathcal{D})$.
- $\mathcal{L}(q)$ is a **lower bound** of $\log p(\mathcal{D})$ because $\text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})) \geq 0$.

~~ Look for $\hat{q}(\boldsymbol{\theta}) = \arg \max_{q \in \mathcal{Q}} \mathcal{L}(q)$.

ELBO: Evidence Lower-Bound

Notice how we can rearrange the KL divergence as follows:

$$\begin{aligned} \text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D})) &= \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} \right] = \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta}) \cdot p(\mathcal{D})}{p(\boldsymbol{\theta}|\mathcal{D}) \cdot p(\mathcal{D})} \right] \\ &= \log p(\mathcal{D}) - \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathcal{D})} \right] = \log p(\mathcal{D}) - \mathcal{L}(q) \end{aligned}$$

Evidence Lower Bound (ELBO): $\mathcal{L}(q) = -\mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathcal{D})} \right] = \mathbb{E}_{\boldsymbol{\theta} \sim q} \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right].$

Summary:

- We started out looking for $\arg \min_{q \in \mathcal{Q}} \text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D}))$.
- Didn't know how to calculate $\text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D}))$ because $p(\boldsymbol{\theta}|\mathcal{D})$ is unknown.
- Still, we can find the optimal approximation by maximizing $\mathcal{L}(q)$:

$$\arg \max_{q \in \mathcal{Q}} \mathcal{L}(q) = \arg \min_{q \in \mathcal{Q}} \text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathcal{D}))$$

Variational Bayes w/ Mean Field

The mean field assumption

What we have ...

We now have the first building-block of the approximation:

$$\Delta(q \parallel p) = \text{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathcal{D})) ,$$

and avoided the issue with $p(\boldsymbol{\theta} \mid \mathcal{D})$ by focusing on $\mathcal{L}(q)$.

We still need the set \mathcal{Q} :

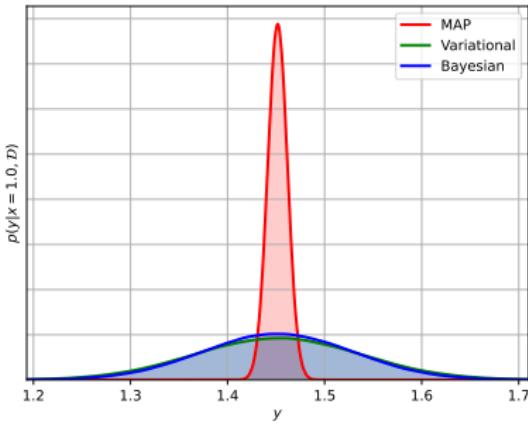
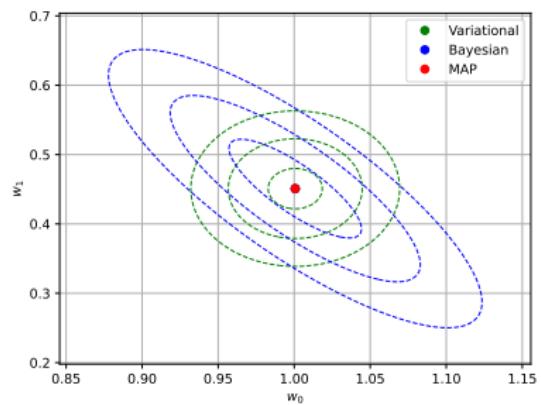
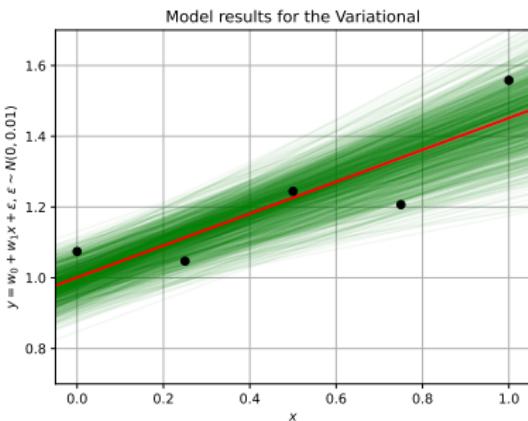
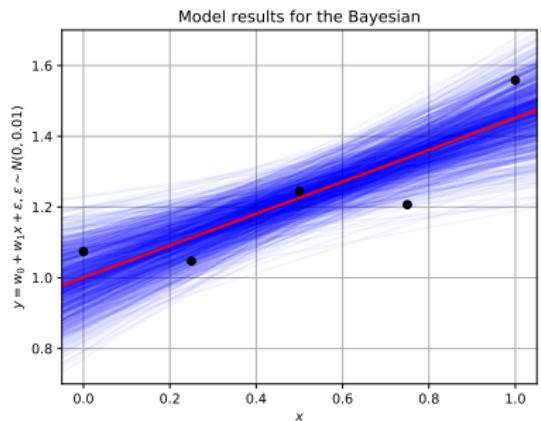
Very often you will see the **mean field assumption**, which states that \mathcal{Q} consists of distributions that **factorize** according to the equation

$$q(\boldsymbol{\theta}) = \prod_i q_i(\theta_i) .$$

This may seem like a very restricted set, but you'll get to see some tricks later ...

VB-MF example – “sanity check”

Bayes linear regression with likelihood $y_i \mid \{w_0, w_1, x_i, \sigma_y^2\} = \mathcal{N}(w_0 + w_1 x_i, \sigma_y^2)$.



Setup:

- We have observed \mathcal{D} , and can calculate the full joint $p(\boldsymbol{\theta}, \mathcal{D})$.
- We use the ELBO as our objective, and assume $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\cdot | \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.

Algorithm:

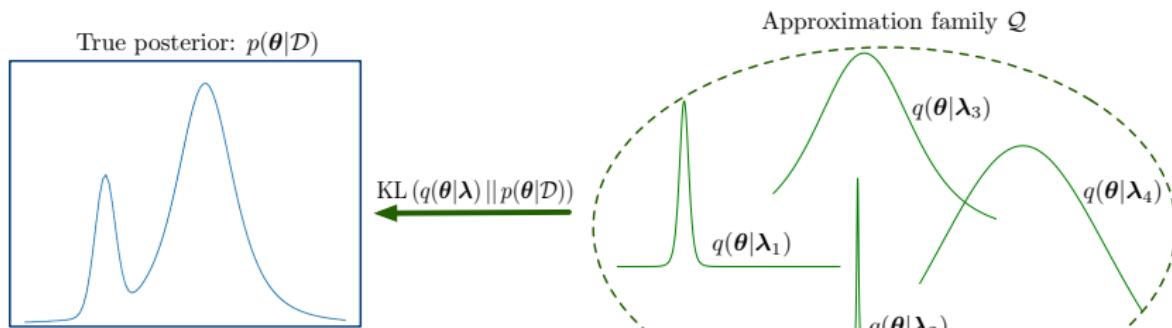
Repeat until negligible improvement in terms of $\mathcal{L}(q)$:

- ① For each j :
 - Somehow choose $\boldsymbol{\lambda}_j$ to maximize $\mathcal{L}(q)$, based on \mathcal{D} and $\{\boldsymbol{\lambda}_i\}_{i \neq j}$.
- ② Calculate the new $\mathcal{L}(q)$.

Solving the VB optimization

Recap: What is variational inference?

VI: Approximate the true posterior distribution $p(\theta | \mathcal{D})$ with a **variational distribution** from a tractable family of distributions \mathcal{Q} . The family is indexed by the parameters λ .



Our computational challenge:

Fit the variational parameters $\hat{\lambda}$ so that the “distance” $\text{KL}(q(\theta | \lambda) || p(\theta | \mathcal{D}))$ is minimized:

$$q(\theta | \hat{\lambda}) = \arg \min_{q \in \mathcal{Q}} \text{KL}(q(\theta) || p(\theta | \mathcal{D})) = \arg \max_{\lambda} \mathcal{L}(q(\theta | \lambda))$$

Solving the VB equation one θ_j at the time

We will maximize $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right]$ under the assumption that $q(\cdot)$ factorizes.

Let us pick one j , utilize that $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ under MF, and keep $q_{\neg j}(\cdot)$ fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})]\end{aligned}$$

Solving the VB equation one θ_j at the time

We will maximize $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right]$ under the assumption that $q(\cdot)$ factorizes.

Let us pick one j , utilize that $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ under MF, and keep $q_{\neg j}(\cdot)$ fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})]\end{aligned}$$

Notation-trick:

For the term $\mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})]$ we simply define $\tilde{f}_j(\theta_j)$ so that

$$\log \tilde{f}_j(\theta_j) := \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})].$$

Solving the VB equation one θ_j at the time

We will maximize $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right]$ under the assumption that $q(\cdot)$ factorizes.

Let us pick one j , utilize that $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ under MF, and keep $q_{\neg j}(\cdot)$ fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})]\end{aligned}$$

Notation-trick:

For the term $\mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})]$ we simply define $\tilde{f}_j(\theta_j)$ so that

$$\log \tilde{f}_j(\theta_j) := \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})].$$

We next define the *normalized version* by $f_j(\theta_j) := \frac{\tilde{f}_j(\theta_j)}{\int_{\boldsymbol{\theta}} \tilde{f}_j(\theta_j) d\boldsymbol{\theta}}$.

In all, this means that

$$\mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})] = \log f_j(\theta_j) + c_1$$

Solving the VB equation one θ_j at the time

We will maximize $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right]$ under the assumption that $q(\cdot)$ factorizes.

Let us pick one j , utilize that $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ under MF, and keep $q_{\neg j}(\cdot)$ fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \boxed{\mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})]} - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \boxed{\log f_j(\theta_j)} - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})] + c_1\end{aligned}$$

Solving the VB equation one θ_j at the time

We will maximize $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right]$ under the assumption that $q(\cdot)$ factorizes.

Let us pick one j , utilize that $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ under MF, and keep $q_{\neg j}(\cdot)$ fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \log f_j(\theta_j) - \boxed{\mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})]} + c_1\end{aligned}$$

Simplification:

Notice that $\log q(\boldsymbol{\theta}) = \log q_j(\theta_j) + \log q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ (under MF). Therefore

$$\begin{aligned}\boxed{\mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})]} &= \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q_j(\theta_j) + \log q_{\neg j}(\boldsymbol{\theta}_{\neg j})] \\ &= \mathbb{E}_{q_j} [\log q_j(\theta_j)] + \mathbb{E}_{q_{\neg j}} [\log q_{\neg j}(\boldsymbol{\theta}_{\neg j})] \\ &= \boxed{\mathbb{E}_{q_j} [\log q_j(\theta_j)]} + c_2,\end{aligned}$$

because $\mathbb{E}_{q_{\neg j}} [\log q_{\neg j}(\boldsymbol{\theta}_{\neg j})]$ is constant wrt. $q_j(\cdot)$.

Solving the VB equation one θ_j at the time

We will maximize $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right]$ under the assumption that $q(\cdot)$ factorizes.

Let us pick one j , utilize that $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ under MF, and keep $q_{\neg j}(\cdot)$ fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \log f_j(\theta_j) - \mathbb{E}_{q_j} [\log q_j(\theta_j)] + c\end{aligned}$$

Solving the VB equation one θ_j at the time

We will maximize $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right]$ under the assumption that $q(\cdot)$ factorizes.

Let us pick one j , utilize that $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ under MF, and keep $q_{\neg j}(\cdot)$ fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \log f_j(\theta_j) - \mathbb{E}_{q_j} [\log q_j(\theta_j)] + c\end{aligned}$$

Almost there:

Recall that $f_j(\theta_j)$ integrates to 1, and is per definition non-negative.

We can therefore regard it as a density function for θ_j , and get

$$\begin{aligned}\mathbb{E}_{q_j} \log f_j(\theta_j) - \mathbb{E}_{q_j} [\log q_j(\theta_j)] &= -\mathbb{E}_{q_j} [\log q_j(\theta_j) - \log f_j(\theta_j)] \\ &= -\text{KL}(q_j(\theta_j) || f_j(\theta_j))\end{aligned}$$

Solving the VB equation one θ_j at the time

We will maximize $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right]$ under the assumption that $q(\cdot)$ factorizes.

Let us pick one j , utilize that $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ under MF, and keep $q_{\neg j}(\cdot)$ fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \log f_j(\theta_j) - \mathbb{E}_{q_j} [\log q_j(\theta_j)] + c \\ &= -\text{KL}(q_j(\theta_j) || f_j(\theta_j)) + c\end{aligned}$$

Solving the VB equation one θ_j at the time

We will maximize $\mathcal{L}(q) = \mathbb{E}_q \left[\log \frac{p(\boldsymbol{\theta}, \mathcal{D})}{q(\boldsymbol{\theta})} \right]$ under the assumption that $q(\cdot)$ factorizes.

Let us pick one j , utilize that $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$ under MF, and keep $q_{\neg j}(\cdot)$ fixed.

$$\begin{aligned}\mathcal{L}(q) &= \mathbb{E}_q [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_q [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})] - \mathbb{E}_{q_j} \mathbb{E}_{q_{\neg j}} [\log q(\boldsymbol{\theta})] \\ &= \mathbb{E}_{q_j} \log f_j(\theta_j) - \mathbb{E}_{q_j} [\log q_j(\theta_j)] + c \\ &= -\text{KL}(q_j(\theta_j) || f_j(\theta_j)) + c\end{aligned}$$

We get the following result:

The ELBO is maximized wrt. q_j by choosing it equal to $f_j(\theta_j)$:

$$q_j(\theta_j) = \frac{1}{Z} \exp (\mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})])$$

... and to get there we had to make the following assumptions:

- Mean field: $q(\boldsymbol{\theta}) = \prod_i q_i(\theta_i)$, and specifically $q(\boldsymbol{\theta}) = q_j(\theta_j) \cdot q_{\neg j}(\boldsymbol{\theta}_{\neg j})$.
- We optimize wrt. $q_j(\cdot)$, while keeping $q_{\neg j}(\cdot)$ fixed – i.e., we do coordinate ascent in probability distribution space.

Setup

- We have observed \mathcal{D} , and can calculate the full joint $p(\boldsymbol{\theta}, \mathcal{D})$.
- We use the ELBO as our objective, and assume $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\theta_j | \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to **optimize the parameterization $\boldsymbol{\lambda}_j$** .

The CAVI algorithm

Repeat until negligible improvement in terms of $\mathcal{L}(q)$:

- For each j :
 - Somehow choose $\boldsymbol{\lambda}_j$ to maximize $\mathcal{L}(q)$, based on \mathcal{D} and $\{\boldsymbol{\lambda}_i\}_{i \neq j}$.
- Calculate the new $\mathcal{L}(q)$.

Setup

- We have observed \mathcal{D} , and can calculate the full joint $p(\boldsymbol{\theta}, \mathcal{D})$.
- We use the ELBO as our objective, and assume $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\theta_j | \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.

The CAVI algorithm

Repeat until negligible improvement in terms of $\mathcal{L}(q)$:

- For each j :
 - Calculate $\mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})]$ using current estimates for $q_i(\cdot | \boldsymbol{\lambda}_i)$, $i \neq j$.
 - Choose $\boldsymbol{\lambda}_j$ so that $q_j(\theta_j | \boldsymbol{\lambda}_j) \propto \exp (\mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})])$.
- Calculate the new $\mathcal{L}(q)$.

Setup

- We have observed \mathcal{D} , and can calculate the full joint $p(\boldsymbol{\theta}, \mathcal{D})$.
- We use the ELBO as our objective, and assume $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\theta_j | \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.

The CAVI algorithm

Repeat until negligible improvement in terms of $\mathcal{L}(q)$:

- For each j :
 - Calculate $\mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})]$ using current estimates for $q_i(\cdot | \boldsymbol{\lambda}_i)$, $i \neq j$.
 - Choose $\boldsymbol{\lambda}_j$ so that $q_j(\theta_j | \boldsymbol{\lambda}_j) \propto \exp(\mathbb{E}_{q_{\neg j}} [\log p(\boldsymbol{\theta}, \mathcal{D})])$.
- Calculate the new $\mathcal{L}(q)$.

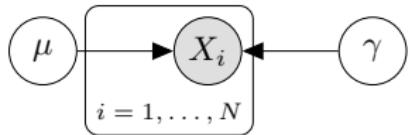
The procedure gives us the $q(\boldsymbol{\theta} | \boldsymbol{\lambda}) \in \mathcal{Q}$ that is closest to $p(\boldsymbol{\theta} | \mathcal{D})$, even though we **do not know** what $p(\boldsymbol{\theta} | \mathcal{D})$ is. Quite remarkable!

Setup

- We have observed \mathcal{D} , and can calculate the full joint $p(\boldsymbol{\theta}, \mathcal{D})$.
- We use the ELBO as our objective, and assume $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\theta_j | \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.

A simple Gaussian model

A Gaussian model with unknown mean and precision

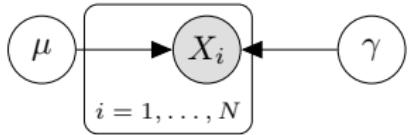


- $X_i | \{\mu, \gamma\} \sim \mathcal{N}(\mu, 1/\gamma)$
- $\mu \sim \mathcal{N}(0, \tau^{-1})$
- $\gamma \sim \text{Gamma}(\alpha, \beta)$

The probability model

$$p(\mathcal{D}, \overbrace{\mu, \gamma}^{\theta} | \tau, \alpha, \beta) = \prod_{i=1}^N p(x_i | \mu, \gamma^{-1}) p(\mu | 0, \tau^{-1}) p(\gamma | \alpha, \beta)$$

A Gaussian model with unknown mean and precision



- $X_i | \{\mu, \gamma\} \sim \mathcal{N}(\mu, 1/\gamma)$
- $\mu \sim \mathcal{N}(0, \tau^{-1})$
- $\gamma \sim \text{Gamma}(\alpha, \beta)$

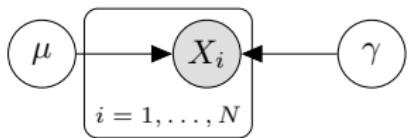
The probability model

$$p(\mathcal{D}, \overbrace{\mu, \gamma}^{\theta} | \tau, \alpha, \beta) = \prod_{i=1}^N p(x_i | \mu, \gamma^{-1}) p(\mu | 0, \tau^{-1}) p(\gamma | \alpha, \beta)$$

...after taking the log

$$\log p(\mathcal{D}, \mu, \gamma | \tau, \alpha, \beta) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu | 0, \tau^{-1}) + \log p(\gamma | \alpha, \beta)$$

A Gaussian model with unknown mean and precision



- $X_i | \{\mu, \gamma\} \sim \mathcal{N}(\mu, 1/\gamma)$
- $\mu \sim \mathcal{N}(0, \tau^{-1})$
- $\gamma \sim \text{Gamma}(\alpha, \beta)$

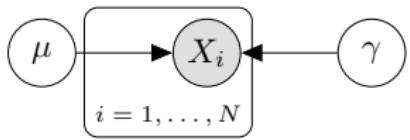
The probability model

$$p(\mathcal{D}, \overbrace{\mu, \gamma}^{\theta} | \tau, \alpha, \beta) = \prod_{i=1}^N p(x_i | \mu, \gamma^{-1}) p(\mu | 0, \tau^{-1}) p(\gamma | \alpha, \beta)$$

The variational model (full mean field)

$$q(\mu, \gamma) = q(\mu)q(\gamma), \quad \min_q \text{KL}(q(\mu)q(\gamma) || p(\mu, \gamma | \mathcal{D}))$$

A Gaussian model with unknown mean and precision



- $X_i | \{\mu, \gamma\} \sim \mathcal{N}(\mu, 1/\gamma)$
- $\mu \sim \mathcal{N}(0, \tau^{-1})$
- $\gamma \sim \text{Gamma}(\alpha, \beta)$

The probability model

$$p(\mathcal{D}, \overbrace{\mu, \gamma}^{\theta} | \tau, \alpha, \beta) = \prod_{i=1}^N p(x_i | \mu, \gamma^{-1}) p(\mu | 0, \tau^{-1}) p(\gamma | \alpha, \beta)$$

The variational model (full mean field)

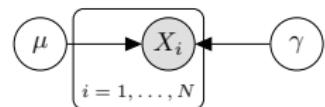
$$q(\mu, \gamma) = q(\mu)q(\gamma), \quad \min_q \text{KL}(q(\mu)q(\gamma) || p(\mu, \gamma | \mathcal{D}))$$

where

- $q(\mu) = \mathcal{N}(\nu_q, \tau_q^{-1})$
- $q(\gamma) = \text{Gamma}(\alpha_q, \beta_q)$

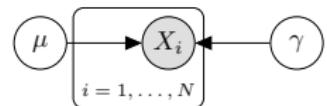
We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c =$$



We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c =$$

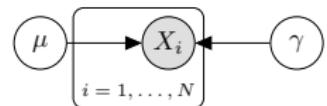


Recall

$$\log p(\mathbf{x}, \mu, \gamma) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu) + \log p(\gamma)$$

We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c =$$

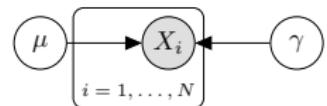


Recall

$$\log p(\mathbf{x}, \mu, \gamma) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu) + \log p(\gamma)$$

We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c$$

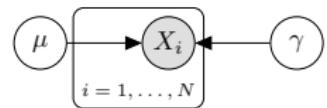


Recall

$$\log p(\mathbf{x}, \mu, \gamma) = \sum_{i=1}^N \log p(x_i | \mu, \gamma^{-1}) + \log p(\mu) + \log p(\gamma)$$

We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c$$



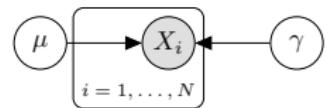
Recall

$$\log p(x_i | \mu, \gamma^{-1}) = \mathcal{N}(\mu, \gamma^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\gamma) - \frac{\gamma}{2} (x_i - \mu)^2$$

$$\log p(\mu) = \mathcal{N}(0, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (\mu)^2$$

We choose the variational distribution so that

$$\log q(\mu) = \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c$$



Recall

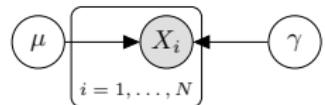
$$\log p(x_i | \mu, \gamma^{-1}) = \mathcal{N}(\mu, \gamma^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\gamma) - \frac{\gamma}{2} (x_i - \mu)^2$$

$$\log p(\mu) = \mathcal{N}(0, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (\mu)^2$$

VB for simple Gaussian model: updating $q(\mu)$

We choose the variational distribution so that

$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c \\ &= \sum_{i=1}^N \mathbb{E}_{q_\gamma} \left[-\frac{\gamma}{2} (x_i - \mu)^2 \right] - \frac{\tau}{2} (\mu)^2 + c\end{aligned}$$

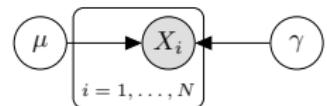


Recall

$$\log p(x_i | \mu, \gamma^{-1}) = \mathcal{N}(\mu, \gamma^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\gamma) - \frac{\gamma}{2} (x_i - \mu)^2$$

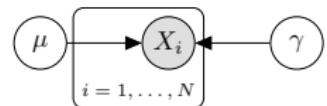
$$\log p(\mu) = \mathcal{N}(0, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (\mu)^2$$

We choose the variational distribution so that



$$\begin{aligned}
 \log q(\mu) &= \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c \\
 &= \sum_{i=1}^N \mathbb{E}_{q_\gamma} \left[-\frac{\gamma}{2} (x_i - \mu)^2 \right] - \frac{\tau}{2} (\mu)^2 + c \\
 &= -\frac{1}{2} \mathbb{E}_{q_\gamma} [\gamma] \left(\sum_{i=1}^N x_i^2 + N \cdot \mu^2 - 2\mu \sum_{i=1}^N x_i \right) - \frac{\tau}{2} (\mu)^2 + c
 \end{aligned}$$

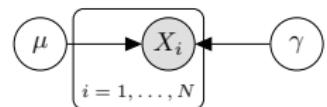
We choose the variational distribution so that



$$\begin{aligned}
 \log q(\mu) &= \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c \\
 &= \sum_{i=1}^N \mathbb{E}_{q_\gamma} \left[-\frac{\gamma}{2} (x_i - \mu)^2 \right] - \frac{\tau}{2} (\mu)^2 + c \\
 &= -\frac{1}{2} \mathbb{E}_{q_\gamma} [\gamma] \left(\sum_{i=1}^N x_i^2 + N \cdot \mu^2 - 2\mu \sum_{i=1}^N x_i \right) - \frac{\tau}{2} (\mu)^2 + c \\
 &= -\frac{1}{2} (\mathbb{E}_{q_\gamma} [\gamma] \cdot N + \tau) \mu^2 + \left(\mathbb{E}_{q_\gamma} [\gamma] \sum_{i=1}^N x_i \right) \mu + c
 \end{aligned}$$

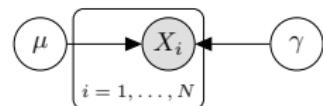
We choose the variational distribution so that

$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c \\ &= -\frac{1}{2} (\mathbb{E}_{q_\gamma} [\gamma] \cdot N + \tau) \mu^2 + \left(\mathbb{E}_{q_\gamma} [\gamma] \sum_{i=1}^N x_i \right) \mu + c\end{aligned}$$



We choose the variational distribution so that

$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c \\ &= -\frac{1}{2} (\mathbb{E}_{q_\gamma} [\gamma] \cdot N + \tau) \mu^2 + \left(\mathbb{E}_{q_\gamma} [\gamma] \sum_{i=1}^N x_i \right) \mu + c\end{aligned}$$

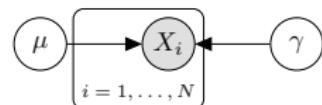


Recall the normal distribution

$$\begin{aligned}\log q(\mu | \nu_q, \tau_q^{-1}) &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau_q) - \frac{\tau_q}{2} (\mu - \nu_q)^2 \\ &= -\frac{1}{2} \tau_q \mu^2 + \tau_q \nu_q \mu + c\end{aligned}$$

We choose the variational distribution so that

$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c \\ &= -\frac{1}{2} (\mathbb{E}_{q_\gamma} [\gamma] \cdot N + \tau) \mu^2 + \left(\mathbb{E}_{q_\gamma} [\gamma] \sum_{i=1}^N x_i \right) \mu + c\end{aligned}$$

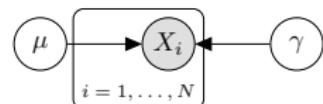


Recall the normal distribution

$$\begin{aligned}\log q(\mu | \nu_q, \tau_q^{-1}) &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau_q) - \frac{\tau_q}{2} (\mu - \nu_q)^2 \\ &= -\frac{1}{2} \tau_q \mu^2 + \tau_q \nu_q \mu + c\end{aligned}$$

We choose the variational distribution so that

$$\begin{aligned}\log q(\mu) &= \mathbb{E}_{q_\gamma} [\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^N \mathbb{E}_{q_\gamma} [\log p(x_i | \mu, \gamma^{-1})] + \log p(\mu) + c \\ &= -\frac{1}{2} (\mathbb{E}_{q_\gamma} [\gamma] \cdot N + \tau) \mu^2 + \left(\mathbb{E}_{q_\gamma} [\gamma] \sum_{i=1}^N x_i \right) \mu + c\end{aligned}$$



Thus, we see that $q(\mu)$ is normally distributed with

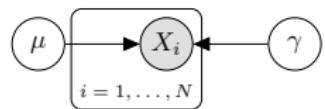
- precision $\tau_q \leftarrow \mathbb{E}_{q_\gamma} [\gamma] \cdot N + \tau$
- mean $\nu_q \leftarrow \tau_q^{-1} \left(\mathbb{E}_{q_\gamma} [\gamma] \sum_{i=1}^N x_i \right)$

Recall the normal distribution

$$\begin{aligned}\log q(\mu | \nu_q, \tau_q^{-1}) &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau_q) - \frac{\tau_q}{2} (\mu - \nu_q)^2 \\ &= -\frac{1}{2} \tau_q \mu^2 + \tau_q \nu_q \mu + c\end{aligned}$$

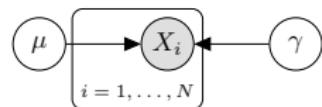
We choose the variational distribution so that

$$\log q(\gamma) = \mathbb{E}_{q_\mu} [\log p(\mathbf{x}, \mu, \gamma)] + c$$



We choose the variational distribution so that

$$\log q(\gamma) = \mathbb{E}_{q_\mu} [\log p(\mathbf{x}, \mu, \gamma)] + c$$



Thus, we see that $q(\gamma)$ is Gamma distributed with

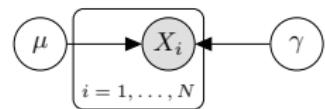
- $\alpha_q \leftarrow \frac{N}{2} + \alpha$
- $\beta_q \leftarrow \beta + \frac{1}{2} \sum_{i=1}^N \mathbb{E}_{q_\mu} [(x_i - \mu)^2]$

Note that:

- $\mathbb{E}_{q_\mu} [(x_i - \mu)^2] = x_i^2 + \mathbb{E}_{q_\mu} [\mu^2] - 2 \cdot x_i \cdot \mathbb{E}_{q_\mu} [\mu]$
- $\mathbb{E}_{q_\mu} [\mu^2] = \text{Var}(\mu) + \mathbb{E}_{q_\mu} [\mu]^2$

We choose the variational distribution so that

$$\log q(\gamma) = \mathbb{E}_{q_\mu} [\log p(\mathbf{x}, \mu, \gamma)] + c$$



Thus, we see that $q(\gamma)$ is Gamma distributed with

- $\alpha_q \leftarrow \frac{N}{2} + \alpha$
- $\beta_q \leftarrow \beta + \frac{1}{2} \sum_{i=1}^N \mathbb{E}_{q_\mu} [(x_i - \mu)^2]$

Note that:

- $\mathbb{E}_{q_\mu} [(x_i - \mu)^2] = x_i^2 + \mathbb{E}_{q_\mu} [\mu^2] - 2 \cdot x_i \cdot \mathbb{E}_{q_\mu} [\mu]$
- $\mathbb{E}_{q_\mu} [\mu^2] = \text{Var}(\mu) + \mathbb{E}_{q_\mu} [\mu]^2$

Recall

$$\log q(\gamma | \alpha_q, \beta_q) = \alpha_q \cdot \log(\beta_q) + (\alpha_q - 1) \log(\gamma) - \beta_q \cdot \gamma - \log(\Gamma(\alpha_q))$$

Monitoring the ELBO

The variational updating rules are guaranteed to never decrease the ELBO $\mathcal{L}(q)$:

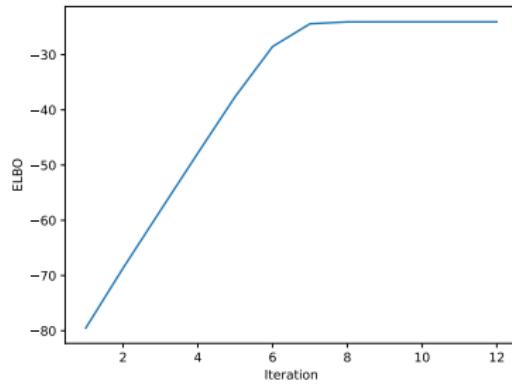
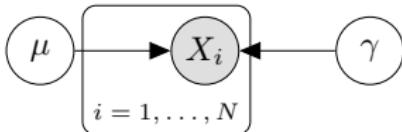
$$\mathcal{L}(q) = \mathbb{E}_q \log p(\mathbf{x}, \mu, \gamma | \tau, \alpha, \beta) - \mathbb{E}_q \log q(\mu, \gamma)$$

$$= \sum_{i=1}^N \mathbb{E}_q \log p(x_i | \mu, \gamma) + \mathbb{E}_q \log p(\mu | 0, \tau) + \mathbb{E}_q \log p(\gamma | \alpha, \beta) - \mathbb{E}_q \log q(\mu) - \mathbb{E}_q \log q(\gamma)$$

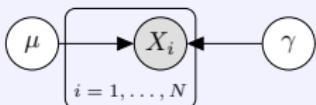
at any updating step. With some pencil pushing we arrive at a somewhat complicated but closed form expression (not shown here).

Monitoring the ELBO can be useful for

- Assessing convergence
- Doing debugging
- ...



Code Task: VB for a simple Gaussian model

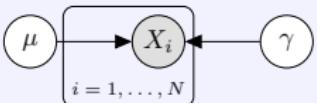


- $X_i | \{\mu, \gamma\} \sim \mathcal{N}(\mu, 1/\gamma)$
- $\mu \sim \mathcal{N}(0, \tau)$
- $\gamma \sim \text{Gamma}(\alpha, \beta)$

In this task you need to use mean-field, and look for $q(\mu, \gamma) = q(\mu) \cdot q(\gamma)$ that best approximates $p(\mu, \gamma | \mathcal{D})$ wrt. the VB measure $\text{KL}(q || p)$.

- Go though the notebook
student_simple_model.ipynb
and try to link the code to the derivations in the slides.
- Implement the update rules for $q(\mu)$ and $q(\gamma)$ (from the slides) in the notebook.
- Experiment with the model and the data set; try changing the prior and the data generating process.

Code Task: VB for a simple Gaussian model



- $X_i | \{\mu, \gamma\} \sim \mathcal{N}(\mu, 1/\gamma)$
- $\mu \sim \mathcal{N}(0, \tau)$
- $\gamma \sim \text{Gamma}(\alpha, \beta)$

Variational Updating Equation for $q(\mu) = \mathcal{N}(\nu_q, \tau_q^{-1})$

- precision $\tau_q \leftarrow \mathbb{E}_{q_\gamma}[\gamma] \cdot N + \tau$
- mean $\nu_q \leftarrow \tau_q^{-1} \left(\mathbb{E}_{q_\gamma}[\gamma] \sum_{i=1}^N x_i \right)$

Variational Updating Equation for $q(\gamma) = \text{Gamma}(\alpha_q, \beta_q)$

- $\alpha_q \leftarrow \frac{N}{2} + \alpha$
- $\beta_q \leftarrow \beta + \frac{1}{2} \sum_{i=1}^N \mathbb{E}_{q_\mu}[(x_i - \mu)^2]$

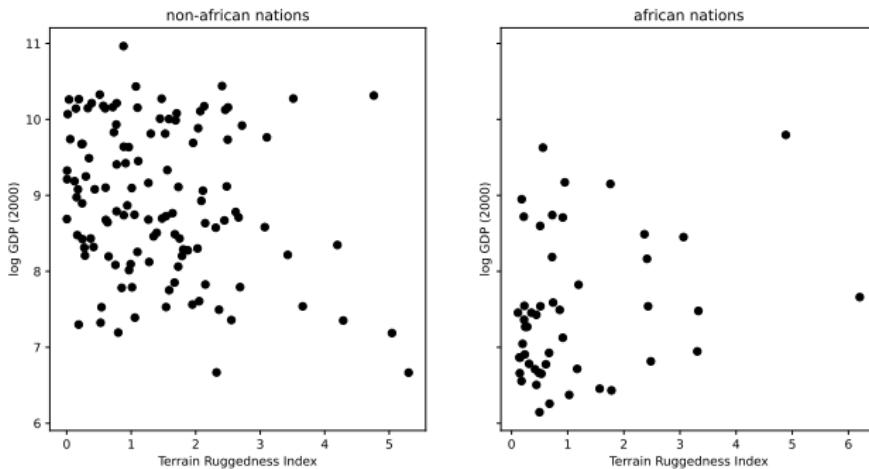
Note that:

- $\mathbb{E}_{q_\mu}[(x_i - \mu)^2] = x_i^2 + \mathbb{E}_{q_\mu}[\mu^2] - 2 \cdot x_i \cdot \mathbb{E}_{q_\mu}[\mu]$
- $\mathbb{E}_{q_\mu}[\mu^2] = \text{Var}(\mu) + \mathbb{E}_{q_\mu}[\mu]^2$
- $\mathbb{E}_{q_\gamma}[\gamma] = \frac{\alpha_q}{\beta_q}$

Bayesian linear regression

Real Data Example

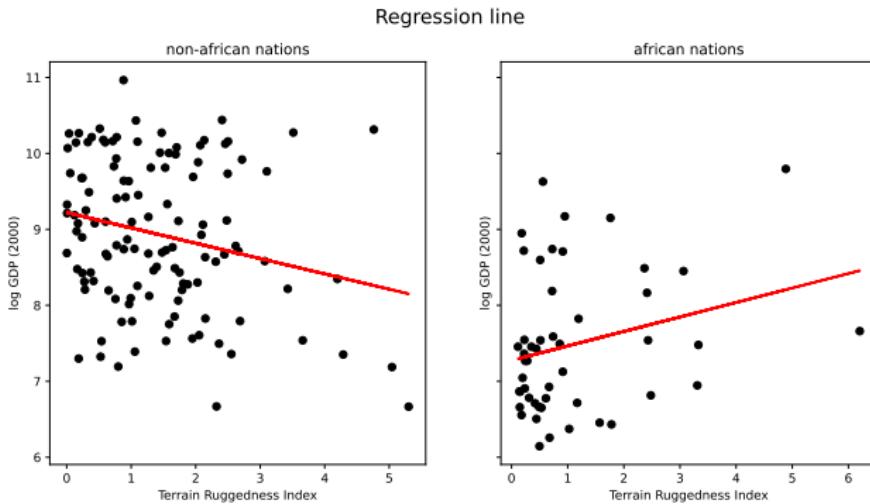
Scatter plot of data



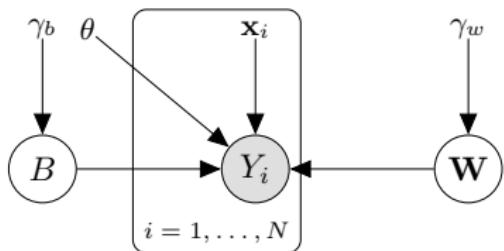
Relationship between topographic heterogeneity and GDP per capita

- Terrain ruggedness or bad geography is related to poorer economic performance outside of Africa.

Real Data Example

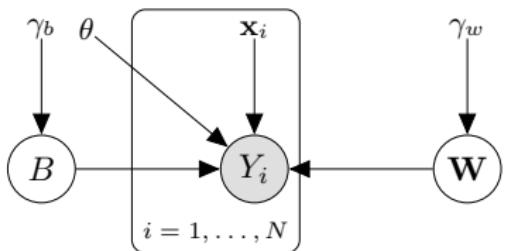


The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i | \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The Bayesian linear regression model

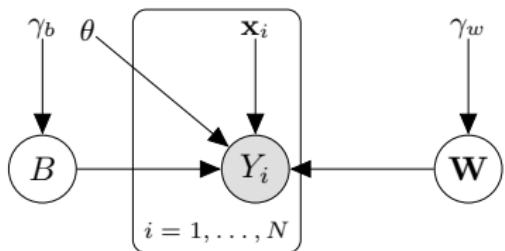


- Num. of data dim: M
- Num. of data inst: N
- $Y_i | \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The probability model

$$p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^N p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) p(\mathbf{w} | \gamma_w) p(b | \gamma_b)$$

The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i | \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

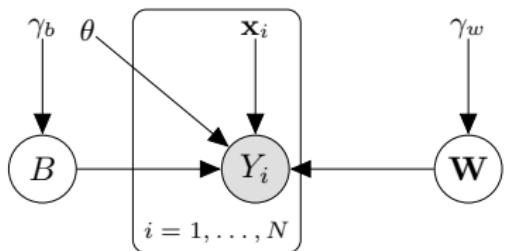
The probability model

$$p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^N p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) p(\mathbf{w} | \gamma_w) p(b | \gamma_b)$$

...after taking the log

$$\log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) + \log p(\mathbf{w} | \gamma_w) + \log p(b | \gamma_b)$$

The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i | \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

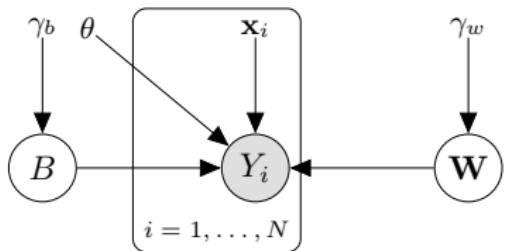
The probability model

$$p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^N p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) p(\mathbf{w} | \gamma_w) p(b | \gamma_b)$$

...after taking the log

$$\begin{aligned} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) &= \sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) + \log p(\mathbf{w} | \gamma_w) + \log p(b | \gamma_b) \\ &= \sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) + \sum_{j=1}^M \log p(w_j | \gamma_w) + \log p(b | \gamma_b) \end{aligned}$$

The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i | \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

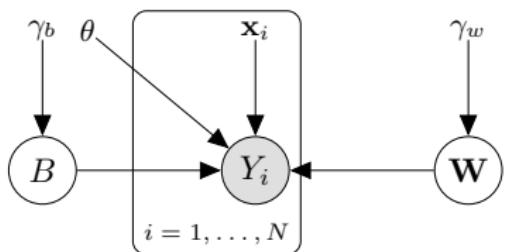
The probability model

$$p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^N p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) p(\mathbf{w} | \gamma_w) p(b | \gamma_b)$$

The variational model (full mean field)

$$q(\cdot) = q(b | \cdot) \prod_{i=1}^M q(w_i | \cdot)$$

The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i | \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The probability model

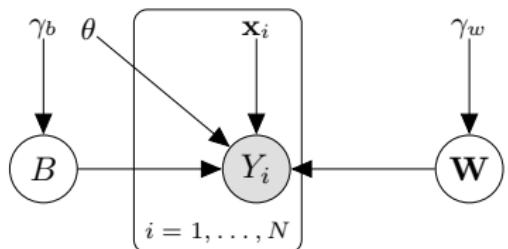
$$p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^N p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) p(\mathbf{w} | \gamma_w) p(b | \gamma_b)$$

The variational updating rules (full mean field) - with some pencil pushing

$q(w_j)$ is normally distributed with

- precision $\tau_j \leftarrow (\gamma_w + \theta \sum_{i=1}^N (x_{ij}^2))$
- mean $\mu_j \leftarrow \tau_j^{-1} \theta \sum_{i=1}^N x_{ij} (y_i - (\sum_{k \neq j} x_{ik} \mathbb{E}(W_k) + \mathbb{E}(B)))$

The Bayesian linear regression model



- Num. of data dim: M
- Num. of data inst: N
- $Y_i | \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The probability model

$$p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^N p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) p(\mathbf{w} | \gamma_w) p(b | \gamma_b)$$

The variational updating rules (full mean field) - with some pencil pushing

$q(b)$ is normally distributed with

- precision $\tau \leftarrow (\gamma_b + \theta N)$
- mean $\mu \leftarrow \tau^{-1} \theta \sum_{i=1}^N (y_i - \mathbb{E}(\mathbf{W}^\top) \mathbf{x}_i)$

Supplementary

We choose the variational distribution so that

$$\log q(w_j) = \mathbb{E}_{q \neg w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c$$

We choose the variational distribution so that

$$\log q(w_j) = \mathbb{E}_{q \neg w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c$$

Recall

$$\log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) + \sum_{j=1}^M \log p(w_j | \gamma_w) + \log p(b | \gamma_b)$$

We choose the variational distribution so that

$$\log q(w_j) = \mathbb{E}_{q \neg w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c$$

Recall

$$\log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) + \sum_{j=1}^M \log p(w_j | \gamma_w) + \log p(b | \gamma_b)$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\log q(w_j) = \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c$$

Recall

$$\log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) = \sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) + \sum_{j=1}^M \log p(w_j | \gamma_w) + \log p(b | \gamma_b)$$

We choose the variational distribution so that

$$\log q(w_j) = \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c$$

The normal distribution

$$\log p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\theta) - \frac{\theta}{2} (y_i - (\mathbf{w}^\top \mathbf{x}_i + b))^2$$

$$\log p(w_j | \gamma_w) = \log \mathcal{N}(w_j | 0, \gamma_w^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\gamma_w) - \frac{\gamma_w}{2} w_j^2$$

We choose the variational distribution so that

$$\log q(w_j) = \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c$$

The normal distribution

$$\log p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\theta) - \frac{\theta}{2} (y_i - (\mathbf{w}^\top \mathbf{x}_i + b))^2$$

$$\log p(w_j | \gamma_w) = \log \mathcal{N}(w_j | 0, \gamma_w^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\gamma_w) - \frac{\gamma_w}{2} w_j^2$$

We choose the variational distribution so that

$$\begin{aligned}\log q(w_j) &= \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c \\ &= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^N \mathbb{E}((y_i - (\mathbf{W}^\top \mathbf{x}_i + B))^2) + c\end{aligned}$$

The normal distribution

$$\log p(y_i | \mathbf{x}_i, \mathbf{w}, b, \theta) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\theta) - \frac{\theta}{2} (y_i - (\mathbf{w}^\top \mathbf{x}_i + b))^2$$

$$\log p(w_j | \gamma_w) = \log \mathcal{N}(w_j | 0, \gamma_w^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\gamma_w) - \frac{\gamma_w}{2} w_j^2$$

We choose the variational distribution so that

$$\begin{aligned}\log q(w_j) &= \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c \\ &= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^N \mathbb{E}((y_i - (\mathbf{W}^\top \mathbf{x}_i + B))^2) + c\end{aligned}$$

Expanding the square

$$(y - (\mathbf{w}^\top \mathbf{x} + b))^2 = y^2 + \mathbf{x}^\top \mathbf{w} \mathbf{w}^\top \mathbf{x} + b^2 + 2\mathbf{w}^\top \mathbf{x}b - 2y\mathbf{w}^\top \mathbf{x} - 2yb$$

$$\mathbf{x}^\top \mathbf{w} \mathbf{w}^\top \mathbf{x} = x_j^2 w_j^2 + \sum_{h, k \neq j} x_k x_h w_k w_h + 2x_j w_j \sum_{k \neq j} x_k w_k$$

We choose the variational distribution so that

$$\begin{aligned}\log q(w_j) &= \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c \\ &= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^N \mathbb{E}((y_i - (\mathbf{W}^\top \mathbf{x}_i + B))^2) + c\end{aligned}$$

Expanding the square

$$(y - (\mathbf{w}^\top \mathbf{x} + b))^2 = y^2 + \mathbf{x}^\top \mathbf{w} \mathbf{w}^\top \mathbf{x} + b^2 + 2\mathbf{w}^\top \mathbf{x} b - 2y\mathbf{w}^\top \mathbf{x} - 2yb$$

$$\mathbf{x}^\top \mathbf{w} \mathbf{w}^\top \mathbf{x} = \mathbf{x}_j^\top \mathbf{w}_j^2 + \sum_{h, k \neq j} x_k x_h w_k w_h + 2x_j w_j \sum_{k \neq j} x_k w_k$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\begin{aligned}\log q(w_j) &= \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c \\ &= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^N \mathbb{E}((y_i - (\mathbf{W}^\top \mathbf{x}_i + B))^2) + c \\ &= -\frac{\gamma_w}{2} w_j^2 - \theta \sum_{i=1}^N \left(\frac{1}{2} x_{ij}^2 w_j^2 + w_j \left(\sum_{k \neq j} x_{ij} x_{ik} \mathbb{E}(W_k) + x_{ij} \mathbb{E}(B) - y x_{ij} \right) \right) + c\end{aligned}$$

Expanding the square

$$(y - (\mathbf{w}^\top \mathbf{x} + b))^2 = y^2 + \mathbf{x}^\top \mathbf{w} \mathbf{w}^\top \mathbf{x} + b^2 + 2\mathbf{w}^\top \mathbf{x} b - 2y\mathbf{w}^\top \mathbf{x} - 2yb$$

$$\mathbf{x}^\top \mathbf{w} \mathbf{w}^\top \mathbf{x} = x_j^2 w_j^2 + \sum_{h, k \neq j} x_k x_h w_k w_h + 2x_j w_j \sum_{k \neq j} x_k w_k$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\begin{aligned}\log q(w_j) &= \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c \\ &= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^N \mathbb{E}((y_i - (\mathbf{W}^\top \mathbf{x}_i + B))^2) + c \\ &= -\frac{\gamma_w}{2} w_j^2 - \theta \sum_{i=1}^N \left(\frac{1}{2} x_{ij}^2 w_j^2 + w_j \left(\sum_{k \neq j} x_{ik} x_{ik} \mathbb{E}(W_k) + x_{ij} \mathbb{E}(B) - y x_{ij} \right) \right) + c \\ &= -\frac{1}{2} (\gamma_w + \theta \sum_{i=1}^N (x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^N x_{ij} (y_i - (\sum_{k \neq j} x_{ik} \mathbb{E}(W_k) + \mathbb{E}(B)))) + c\end{aligned}$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\begin{aligned}\log q(w_j) &= \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c \\ &= -\frac{1}{2} (\gamma_w + \theta \sum_{i=1}^N (x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^N x_{ij} (y_i - (\sum_{k \neq j} x_{ik} \mathbb{E}(W_k) + \mathbb{E}(B)))) + c\end{aligned}$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\begin{aligned}\log q(w_j) &= \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c \\ &= -\frac{1}{2} (\gamma_w + \theta \sum_{i=1}^N (x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^N x_{ij} (y_i - (\sum_{k \neq j} x_{ik} \mathbb{E}(W_k) + \mathbb{E}(B)))) + c\end{aligned}$$

Recall the normal distribution

$$\begin{aligned}\log p(w | \mu, \tau) &= \log \mathcal{N}(w | \mu, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (w - \mu)^2 \\ &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{1}{2} \tau w^2 - \frac{\tau}{2} \mu^2 + w \tau \mu\end{aligned}$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\begin{aligned}\log q(w_j) &= \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c \\ &= -\frac{1}{2} (\gamma_w + \theta \sum_{i=1}^N (x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^N x_{ij} (y_i - (\sum_{k \neq j} x_{ik} \mathbb{E}(W_k) + \mathbb{E}(B)))) + c\end{aligned}$$

Recall the normal distribution

$$\begin{aligned}\log p(w | \mu, \tau) &= \log \mathcal{N}(w | \mu, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (w - \mu)^2 \\ &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{1}{2} \textcolor{red}{\tau} w^2 - \frac{\tau}{2} \mu^2 + w \textcolor{green}{\tau \mu}\end{aligned}$$

VB for Bayesian linear regression: updating $q(w_j)$

We choose the variational distribution so that

$$\begin{aligned}\log q(w_j) &= \mathbb{E}_{q \sim w_j} \log p(\cdot | \mathbf{x}, \theta, \gamma_w, \gamma_b) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i | \mathbf{x}_i, \mathbf{W}, B, \theta) + \log p(w_j | \gamma_w) + c \\ &= -\frac{1}{2} (\gamma_w + \theta \sum_{i=1}^N (x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^N x_{ij} (y_i - (\sum_{k \neq j} x_{ik} \mathbb{E}(W_k) + \mathbb{E}(B)))) + c\end{aligned}$$

Thus, we see that $q(w_j)$ is normally distributed with

- precision $\tau \leftarrow (\gamma_w + \theta \sum_{i=1}^N (x_{ij}^2))$
- mean $\mu \leftarrow \tau^{-1} \theta \sum_{i=1}^N x_{ij} (y_i - (\sum_{k \neq j} x_{ik} \mathbb{E}(W_k) + \mathbb{E}(B)))$

Recall the normal distribution

$$\begin{aligned}\log p(w | \mu, \tau) &= \log \mathcal{N}(w | \mu, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (w - \mu)^2 \\ &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{1}{2} \textcolor{red}{\tau} w^2 - \frac{\tau}{2} \mu^2 + w \textcolor{green}{\tau \mu}\end{aligned}$$

VB for Bayesian linear regression: updating $q(b)$

We choose the variational distribution so that

$$\begin{aligned}\log q(b) &= \mathbb{E}_{q \sim w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(b \mid \gamma_b) + c \\ &= \dots \\ &= -\frac{1}{2}(\gamma_b + \theta N)b^2 + b \left(\theta \sum_{i=1}^N (y_i - \mathbb{E}(\mathbf{W}^\top) \mathbf{x}_i) \right) + c\end{aligned}$$

VB for Bayesian linear regression: updating $q(b)$

We choose the variational distribution so that

$$\begin{aligned}\log q(b) &= \underset{q \sim w_j}{\mathbb{E}} \log p(\cdot \mid \mathbf{W}, B, \theta, \gamma) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(b \mid \gamma_b) + c \\ &= \dots \\ &= -\frac{1}{2} (\color{red}{\gamma_b + \theta N}) b^2 + b \left(\color{green}{\theta} \sum_{i=1}^N (y_i - \mathbb{E}(\mathbf{W}^\top) \mathbf{x}_i) \right) + c\end{aligned}$$

Recall the normal distribution

$$\begin{aligned}\log p(b \mid \mu, \tau) &= \log \mathcal{N}(b \mid \mu, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (b - \mu)^2 \\ &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{1}{2} \color{red}{\tau} b^2 - \frac{\tau}{2} \mu^2 + b \color{green}{\tau \mu}\end{aligned}$$

VB for Bayesian linear regression: updating $q(b)$

We choose the variational distribution so that

$$\begin{aligned}\log q(b) &= \underset{q \sim w_j}{\mathbb{E}} \log p(\cdot \mid \mathbf{W}, B, \theta, \gamma) + c = \sum_{i=1}^N \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(b \mid \gamma_b) + c \\ &= \dots \\ &= -\frac{1}{2} (\color{red}{\gamma_b + \theta N}) b^2 + b \left(\color{green}{\theta \sum_{i=1}^N (y_i - \mathbb{E}(\mathbf{W}^\top) \mathbf{x}_i)} \right) + c\end{aligned}$$

Thus, we get that $q(b)$ is normally distributed with

- precision $\tau \leftarrow (\gamma_b + \theta N)$
- mean $\mu \leftarrow \tau^{-1} \theta \sum_{i=1}^N (y_i - \mathbb{E}(\mathbf{W}^\top) \mathbf{x}_i)$

Recall the normal distribution

$$\begin{aligned}\log p(b \mid \mu, \tau) &= \log \mathcal{N}(b \mid \mu, \tau^{-1}) = -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{\tau}{2} (b - \mu)^2 \\ &= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{1}{2} \color{red}{\tau} b^2 - \frac{\tau}{2} \mu^2 + b \color{green}{\tau \mu}\end{aligned}$$