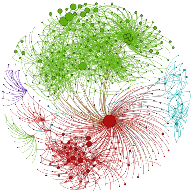
Nordic probabilistic AI school

Variational Inference and Optimization

Helge Langseth, Andrés Masegosa, and Thomas Dyhre Nielsen

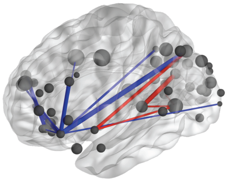
June 17, 2025

Introduction



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

[Gopalan and Blei, PNAS 2013]

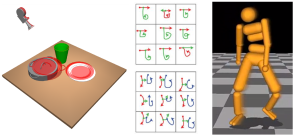


Neuroscience analysis of 220 million fMRI measurements

[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)

Examples

14]

2015]

Image 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**.

Bayesian Machine Learning = Probabilistic model + Bayesian inference

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

Bayesian Machine Learning = Probabilistic model + Bayesian inference

- **Likelihood-part:** A probabilistic model typically defined by $p(\mathbf{x} \mid \boldsymbol{\theta})$.
- **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 $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$,

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

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

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

Bayesian Machine Learning = Probabilistic model + Bayesian inference

- **Likelihood-part:** A probabilistic model typically defined by $p(\mathbf{x} \mid \boldsymbol{\theta})$.
- **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 $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$,

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

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

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

Being Bayesian means maintaining a distribution over θ.
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 – $p(\mathcal{D}\,|\,\boldsymbol{\theta})$**:** $p(y_i\,|\,x_i,\mathbf{w},\sigma_y^2) = \mathcal{N}\left(w_0 + w_1\cdot x_i, \sigma_y^2\right)$

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

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

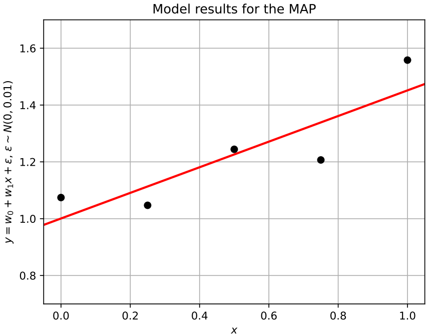
Bayesian Linear regression – Full model:

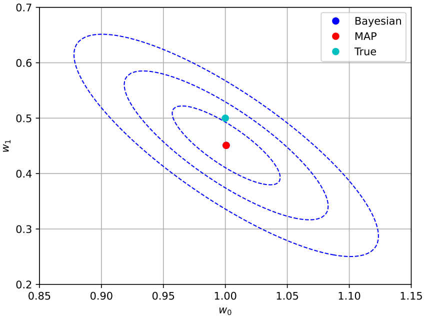
$$p(\mathcal{D},\boldsymbol{\theta}) = p\left(\{y_i\}_{i=1}^n, \mathbf{w}\,|\,\{\mathbf{x}_i\}_{i=1}^n, \sigma_y^2, \sigma_w^2\right) = \overbrace{p(\mathbf{w}\,|\,\sigma_w^2)}^{p(\boldsymbol{\theta})}\overbrace{\prod_{i=1}^n p(y_i\,|\,\mathbf{w},\mathbf{x}_i,\sigma_y^2)}^{p(\mathcal{D}\,|\,\boldsymbol{\theta})}$$

Bayes linear regression with some fake data:

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

Results for the fully Bayesian model and the MAP:



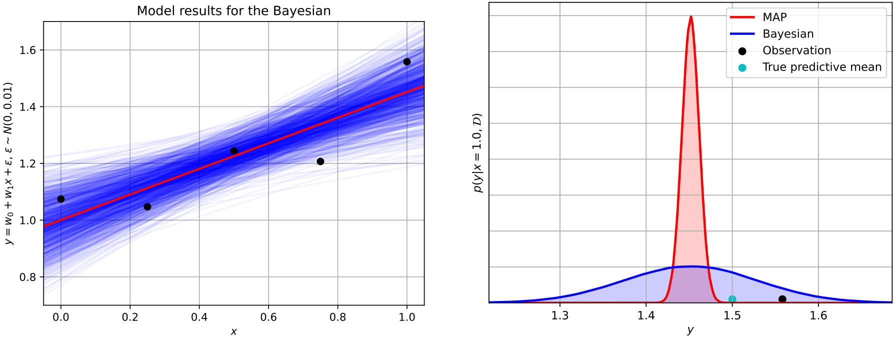


- **MAP:** Reasonable point estimate; No model uncertainty;
- **Bayes:** Model uncertainty around same MAP estimate;

Bayes linear regression with some fake data:

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

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 – Summary

Bayesian inference is in principle easy using Bayes' rule:

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\boldsymbol{\theta})\, p(\mathcal{D} \mid \boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\boldsymbol{\theta})\, p(\mathcal{D} \mid \boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})\, p(\mathcal{D} \mid \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}}$$

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

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

What we want:

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

What we don't want:

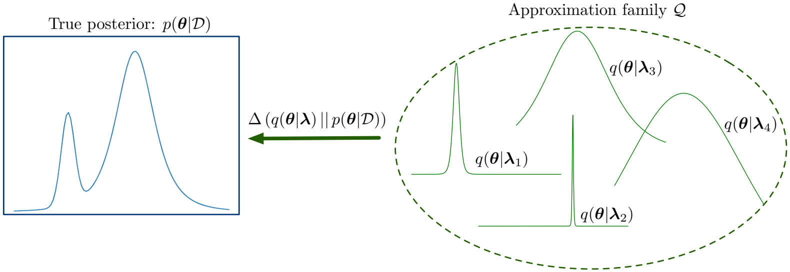
- Purely sampling-based techniques (like Gibbs sampling);
- Degenerate solutions (point estimators like MAP).

FUNDAMENTAL assumption:
It will always be *computationally efficient* to evaluate $p(\mathcal{D}, \boldsymbol{\theta})$ at any given point $\{\mathcal{D}, \boldsymbol{\theta}\}$, e.g., using the simple factorization $p(\mathcal{D}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \cdot p(\mathcal{D} \mid \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_i p(\mathbf{x}_i \mid \boldsymbol{\theta})$.

Variational Bayes: Approximate inference by optimization

Approximate inference through optimization – Main idea

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



Approximate inference through optimization

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

Formalization of approximate inference through optimization:

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

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

Decisions to be made:

1. How to define $\Delta(\cdot||\cdot)$ so that we end up with a high-quality solution?
   - How to work with $\Delta\big(q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta} \mid \mathcal{D})\big)$ when we don't know what $p(\boldsymbol{\theta} \mid \mathcal{D})$ is?
2. 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:** $\Delta(f \,||\, g) \geq 0$ and $\Delta(f \,||\, g) = 0$ if and only if $f = g$.
- **Symmetry:** $\Delta(f \,||\, g) = \Delta(g \,||\, f)$
- **Triangle:** For objects f, g, and h we have that $\Delta(f \,||\, g) \leq \Delta(f \,||\, h) + \Delta(h \,||\, g)$.

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: $\Delta(f\,||\,g) \geq 0$ and $\Delta(f\,||\,g) = 0$ if and only if $f = g$.

Symmetry: $\Delta(f\,||\,g) = \Delta(g\,||\,f)$

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

Standard choice when working with probability distributions

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

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

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

Two alternative KL definitions: $\mathrm{KL}(q||p)$ or $\mathrm{KL}(p||q)$?

Information-projection

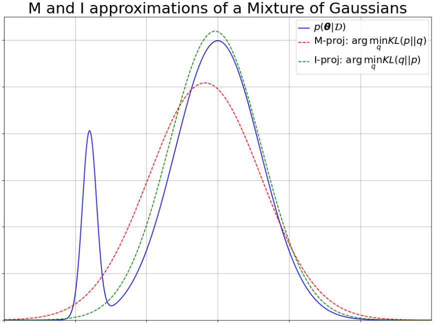
- Minimizes $\mathrm{KL}(q||p) = -\mathbb{E}_{\boldsymbol{\theta}\sim q}[\log p(\boldsymbol{\theta}\,|\,\mathcal{D})] - \mathcal{H}_q$.
- Preference given to q that has:
  1. High q-probability allocated to p-probable regions.
  2. Small q in any region where p is small. "$p(\boldsymbol{\theta}\,|\,\mathcal{D}) \approx 0 \implies q(\boldsymbol{\theta}) \approx 0$".
  3. High entropy ($\sim$ variance)

Moment-projection

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

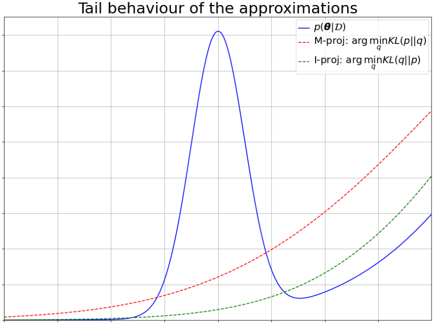
Cheat-sheet:

- **KL-divergence:** $\mathrm{KL}(f||g) = \mathbb{E}_f\left[\log\left(\frac{f(\boldsymbol{\theta})}{g(\boldsymbol{\theta})}\right)\right] = -\mathbb{E}_f\left[\log\left(g(\boldsymbol{\theta})\right)\right] - \mathcal{H}_f$.
- **Entropy:** $\mathcal{H}_f = -\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\,\log\left(f(\boldsymbol{\theta})\right)\mathrm{d}\boldsymbol{\theta} = -\mathbb{E}_f\left[\log\left(f(\boldsymbol{\theta})\right)\right]$.
- **Intuition:** Cheat a bit (measure-zero, limit-zero-rates, etc.) and think *"If $g(\boldsymbol{\theta}_0) \approx 0$, then $-\mathbb{E}_{\boldsymbol{\theta}\sim f}[\log g(\boldsymbol{\theta})]$ becomes 'huge' unless $f(\boldsymbol{\theta}_0) \approx 0$"* because $\lim_{x\to 0^+}\log(x)$ diverges, while $\lim_{x\to 0^+} x\cdot\log(x) = 0$.



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

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



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

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

Variational Bayes setup

VB uses information projections:

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

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

- **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:

$$\boxed{\mathrm{KL}\left(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \mathcal{D})\right)} = \mathbb{E}_{\boldsymbol{\theta} \sim q}\left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \mathcal{D})}\right]$$

ELBO: Evidence Lower-BOund

Notice how we can rearrange the KL divergence as follows:

$$\mathrm{KL}\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}\,|\,\mathcal{D})\right) = \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[\log\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}\,|\,\mathcal{D})}\right] = \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[\log\frac{q(\boldsymbol{\theta})\cdot p(\mathcal{D})}{p(\boldsymbol{\theta}\,|\,\mathcal{D})\cdot p(\mathcal{D})}\right]$$

ELBO: Evidence Lower-BOund

Notice how we can rearrange the KL divergence as follows:

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

ELBO: Evidence Lower-BOund

Notice how we can rearrange the KL divergence as follows:

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

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

ELBO: Evidence Lower-BOund

Notice how we can rearrange the KL divergence as follows:

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

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:

$$\log p(\mathcal{D}) = \mathcal{L}(q) + \mathrm{KL}\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}\,|\,\mathcal{D})\right)$$

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

- We can minimize $\mathrm{KL}\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}\,|\,\mathcal{D})\right)$ 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 $\mathrm{KL}\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}\,|\,\mathcal{D})\right) \geq 0$.

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

Notice how we can rearrange the KL divergence as follows:

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

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

Summary:

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

$$\arg\max_{q\in\mathcal{Q}}\mathcal{L}(q) = \arg\min_{q\in\mathcal{Q}}\mathrm{KL}\left(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}\,|\,\mathcal{D})\right).$$

- It all makes sense: We aim to maximize $\mathcal{L}(q)$, which is a lower-bound of $\log p(\mathcal{D})$.

Variational Bayes w/ Mean Field

The mean field assumption

What we have ...

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

$$\Delta(q \,||\, p) = \mathrm{KL}\left(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \mathcal{D})\right),$$

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 it often works well anyway ...

Wrapping it all up: The VB algorithm under MF

Setup:

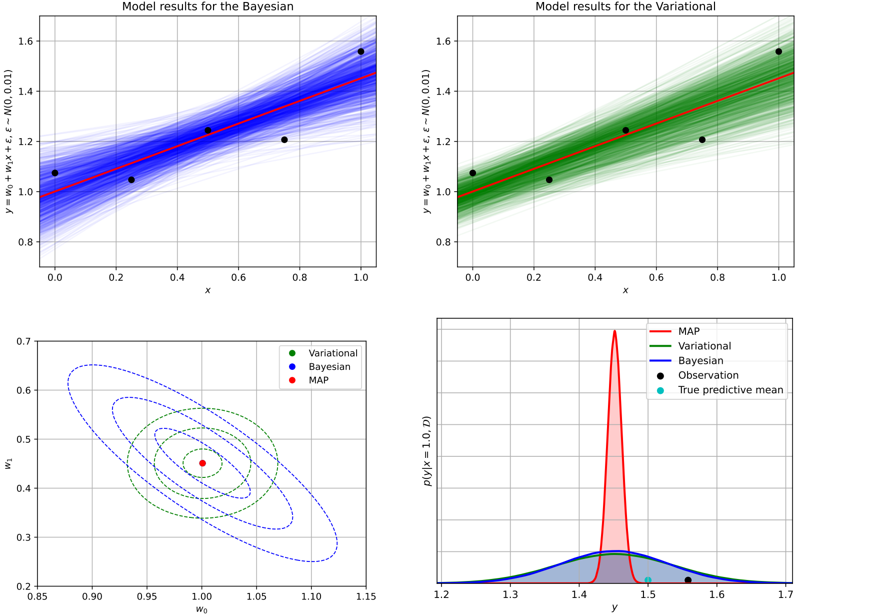
- We have observed $\mathcal{D}$, and can calculate the full joint $p(\boldsymbol{\theta}, \mathcal{D}) = p(\boldsymbol{\theta}) \cdot p(\mathcal{D} \mid \boldsymbol{\theta})$.
- We use the ELBO as our objective, and assume $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\cdot \mid \boldsymbol{\lambda}_j)$, 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 $\mathcal{L}(q)$.

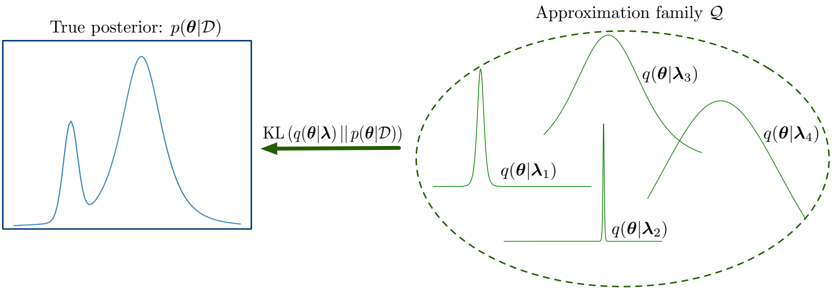
Bayes linear regression with likelihood $y_i \mid \{w_0, w_1, x_i, \sigma_y^2\} = \mathcal{N}(w_0 + w_1\, x_i, \sigma_y^2)$.



Solving the VB optimization

Recap: What is variational inference?

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



Our computational challenge:

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

$$q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\lambda}}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \mathcal{D})\right) = \arg\max_{\boldsymbol{\lambda}} \mathcal{L}\left(q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})\right)$$

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.

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

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

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

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.

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

Notation-trick:

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

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

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

In all, this means that

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

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.

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

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.

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

Simplification:

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

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

because $\mathbb{E}_{q_{\neg j}}\left[\log q_{\neg j}(\boldsymbol{\theta}_{\neg j})\right]$ 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.

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

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.

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

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

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

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.

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

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.

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

We get the following result:

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

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

. . . 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.

VB w/ MF: The CAVI (coordinate ascent variational inference) algorithm

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 that $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\theta_j \mid \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.

VB w/ MF: The CAVI (coordinate ascent variational inference) algorithm

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 that $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\theta_j \mid \boldsymbol{\lambda}_j)$, i.e., we choose the distributional form, while wanting to optimize the parameterization $\boldsymbol{\lambda}_j$.

How this relates to the linear regression example:

- We consider the model $Y|x \sim \mathcal{N}\left(w_0 + w_1 \cdot x, s^2\right)$, $s = 0.1$ known.
- Parameters $\boldsymbol{\theta} = \{w_0, w_1\}$, with vague priors $w_j \sim \mathcal{N}\left(0, 100^2\right)$.
- We want to approximate the posterior $p(\boldsymbol{\theta}|\mathcal{D})$ by $q(\mathbf{w}) = q_0(w_0) \cdot q_1(w_1)$.
- We choose $q_j(w_j|\boldsymbol{\lambda}_j) = \mathcal{N}\left(w_j|\mu_j, \sigma_j^2\right)$ so $\boldsymbol{\lambda}_j = \left\{\mu_j, \sigma_j^2\right\}$.

VB w/ MF: The CAVI (coordinate ascent variational inference) algorithm

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 that $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\theta_j \mid \boldsymbol{\lambda}_j)$, 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 $\mathcal{L}(q)$.

VB w/ MF: The CAVI (coordinate ascent variational inference) algorithm

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 that $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\theta_j \mid \boldsymbol{\lambda}_j)$, 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 $\mathcal{L}(q)$:

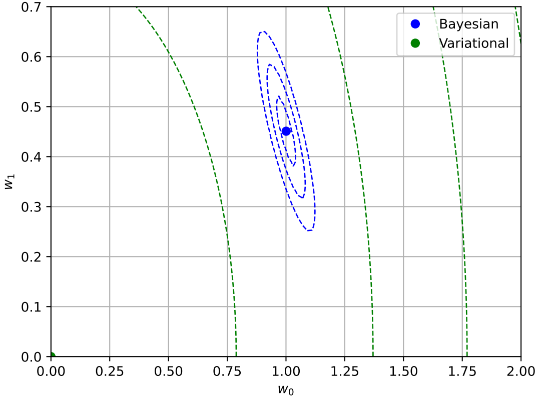
- For each j:
  - Calculate $\mathbb{E}_{q_{\neg j}}[\log p(\boldsymbol{\theta}, \mathcal{D})]$ using current estimates for $q_i(\cdot \mid \boldsymbol{\lambda}_i)$, $i \neq j$.
  - Choose $\boldsymbol{\lambda}_j$ so that $q_j(\theta_j \mid \boldsymbol{\lambda}_j) \propto \exp\left(\mathbb{E}_{q_{\neg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

The procedure is **guaranteed to converge**. If the model is in the conjugate exponential family, the fix-point is guaranteed to be the $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \in \mathcal{Q}$ that is **closest** to $p(\boldsymbol{\theta} \mid \mathcal{D})$, even though **we do not know** what $p(\boldsymbol{\theta} \mid \mathcal{D})$ is. Quite remarkable!

VB w/ MF: The CAVI (coordinate ascent variational inference) algorithm

Setup

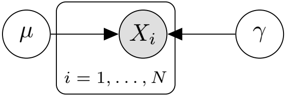
- We have observed $\mathcal{D}$, and can calculate the full joint $p(\boldsymbol{\theta}, \mathcal{D})$.
- We use the ELBO as our objective, and assume that $q(\boldsymbol{\theta})$ factorizes.
- We posit a *variational family* of distributions $q_j(\theta_j \mid \boldsymbol{\lambda}_j)$, i.e., we choose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



A simple Gaussian model

VB for a simple Gaussian model

A Gaussian model with unknown mean and precision

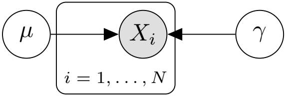


- $X_i \mid \{\mu, \gamma\} \sim \mathcal{N}(\mu, 1/\gamma)$
- $\mu \sim \mathcal{N}(0, \tau^{-1})$
- $\gamma \sim \text{Gamma}(\alpha, \beta)$

The probability model

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

A Gaussian model with unknown mean and precision



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

The probability model

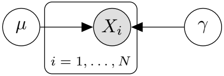
$$p(\mathcal{D}, \overbrace{\mu, \gamma}^{\boldsymbol{\theta}} | \tau, \alpha, \beta) = \prod_{i=1}^{N} p(x_i \mid \mu, \gamma^{-1}) p(\mu \mid 0, \tau^{-1}) p(\gamma \mid \alpha, \beta)$$

The variational model (full mean field)

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

where

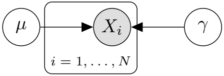
- $q(\mu) = \mathcal{N}(\nu_q, \tau_q^{-1})$
- $q(\gamma) = \text{Gamma}(\alpha_q, \beta_q)$



We choose the variational distribution so that

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

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

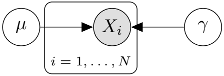


We choose the variational distribution so that

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

Recall

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



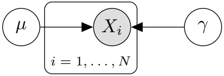
We choose the variational distribution so that

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

Recall

$$\log p(\mathcal{D}, \mu, \gamma) = \sum_{i=1}^{N} \log p(x_i \mid \mu, \gamma^{-1}) + \log p(\mu) + \log p(\gamma)$$

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



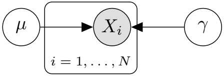
We choose the variational distribution so that

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

Recall

$$\log p(\mathcal{D}, \mu, \gamma) = \sum_{i=1}^{N} \log p(x_i \mid \mu, \gamma^{-1}) + \log p(\mu) + \log p(\gamma)$$

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



We choose the variational distribution so that

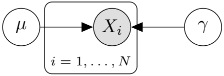
$$\log q(\mu) = \mathbb{E}_{q_\gamma}[\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^{N} \mathbb{E}_{q_\gamma}[\log p(x_i \mid \mu, \gamma^{-1})] + \log p(\mu) + c$$

Recall

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

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

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



We choose the variational distribution so that

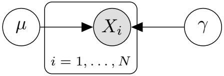
$$\log q(\mu) = \mathbb{E}_{q_\gamma}[\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^{N} \mathbb{E}_{q_\gamma}[\log p(x_i \mid \mu, \gamma^{-1})] + \log p(\mu) + c$$

Recall

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

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

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



We choose the variational distribution so that

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

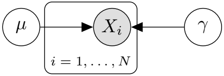
$$= \sum_{i=1}^{N} \mathbb{E}_{q_\gamma}\left[-\frac{\gamma}{2}(x_i - \mu)^2\right] - \frac{\tau}{2}(\mu)^2 + c$$

Recall

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

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

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



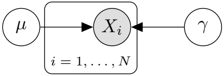
We choose the variational distribution so that

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

$$= \sum_{i=1}^{N} \mathbb{E}_{q_\gamma}\left[-\frac{\gamma}{2}(x_i - \mu)^2\right] - \frac{\tau}{2}(\mu)^2 + c$$

$$= -\frac{1}{2}\mathbb{E}_{q_\gamma}[\gamma]\left(\sum_{i=1}^{N} x_i^2 + N \cdot \mu^2 - 2\mu \sum_{i=1}^{N} x_i\right) - \frac{\tau}{2}(\mu)^2 + c$$

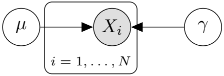
VB for simple Gaussian model: updating $q(\mu)$



We choose the variational distribution so that

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

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

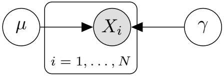


We choose the variational distribution so that

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

$$= -\frac{1}{2}\left(\mathbb{E}_{q_\gamma}[\gamma] \cdot N + \tau\right)\mu^2 + \left(\mathbb{E}_{q_\gamma}[\gamma] \sum_{i=1}^{N} x_i\right)\mu + c$$

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



We choose the variational distribution so that

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

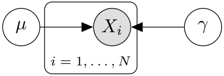
$$= -\frac{1}{2}\left(\mathbb{E}_{q_\gamma}[\gamma] \cdot N + \tau\right)\mu^2 + \left(\mathbb{E}_{q_\gamma}[\gamma] \sum_{i=1}^{N} x_i\right)\mu + c$$

Recall the normal distribution

$$\log q(\mu \mid \nu_q, \tau_q^{-1}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau_q) - \frac{\tau_q}{2}(\mu - \nu_q)^2$$

$$= -\frac{1}{2}\tau_q \mu^2 + \tau_q \nu_q \mu + c$$

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



We choose the variational distribution so that

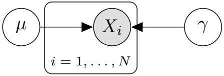
$$\log q(\mu) = \mathbb{E}_{q_\gamma}[\log p(\mathcal{D}, \mu, \gamma)] + c = \sum_{i=1}^{N} \mathbb{E}_{q_\gamma}[\log p(x_i \mid \mu, \gamma^{-1})] + \log p(\mu) + c$$

$$= -\frac{1}{2}\left(\mathbb{E}_{q_\gamma}[\gamma] \cdot N + \tau\right)\mu^2 + \left(\mathbb{E}_{q_\gamma}[\gamma] \sum_{i=1}^{N} x_i\right)\mu + c$$

Recall the normal distribution

$$\log q(\mu \mid \nu_q, \tau_q^{-1}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau_q) - \frac{\tau_q}{2}(\mu - \nu_q)^2$$

$$= -\frac{1}{2}\tau_q \mu^2 + \tau_q \nu_q \mu + c$$



We choose the variational distribution so that

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

$$= -\frac{1}{2}\left(\mathbb{E}_{q_\gamma}[\gamma] \cdot N + \tau\right)\mu^2 + \left(\mathbb{E}_{q_\gamma}[\gamma]\sum_{i=1}^{N} x_i\right)\mu + c$$

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

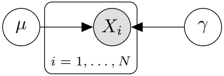
- precision $\tau_q \leftarrow \mathbb{E}_{q_\gamma}[\gamma] \cdot N + \tau$
- mean $\nu_q \leftarrow \tau_q^{-1}\left(\mathbb{E}_{q_\gamma}[\gamma]\sum_{i=1}^{N} x_i\right)$

Recall the normal distribution

$$\log q(\mu \mid \nu_q, \tau_q^{-1}) = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau_q) - \frac{\tau_q}{2}(\mu - \nu_q)^2$$

$$= -\frac{1}{2}\tau_q\mu^2 + \tau_q\nu_q\mu + c$$

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



We choose the variational distribution so that

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

After some pencil-pushing 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:

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

Monitoring the ELBO

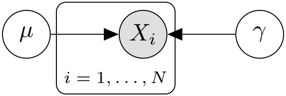
The variational updating rules are guaranteed to never decrease the ELBO $\mathcal{L}(q)$:

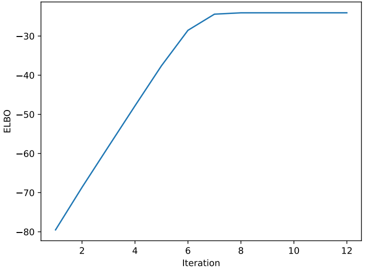
$$\begin{aligned}\mathcal{L}(q) &= \mathop{\mathbb{E}}_q \log p(\mathcal{D}, \mu, \gamma \mid \tau, \alpha, \beta) - \mathop{\mathbb{E}}_q \log q(\mu, \gamma) \\ &= \sum_{i=1}^{N} \mathop{\mathbb{E}}_q \log p(x_i \mid \mu, \gamma) + \mathop{\mathbb{E}}_q \log p(\mu \mid 0, \tau) + \mathop{\mathbb{E}}_q \log p(\gamma \mid \alpha, \beta) - \mathop{\mathbb{E}}_q \log q(\mu) - \mathop{\mathbb{E}}_q \log q(\gamma)\end{aligned}$$

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

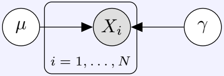
Monitoring the ELBO can be useful for

- Assessing convergence
- Doing debugging
- ...





Code Task: VB for a simple Gaussian model



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

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

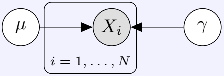
- Go though the notebook

  `students_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

Code Task: VB for a simple Gaussian model



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

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 $q(\gamma) = \text{Gamma}(\alpha_q, \beta_q)$

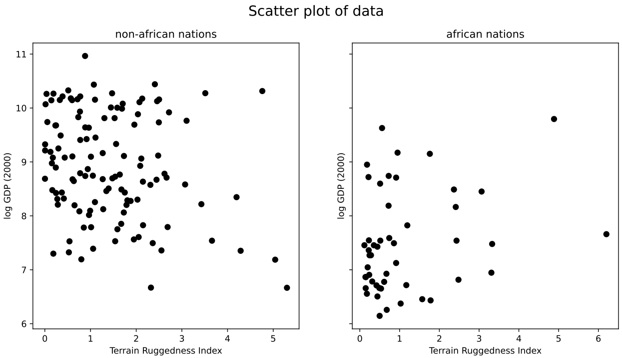
- $\alpha_q \leftarrow \frac{N}{2} + \alpha$
- $\beta_q \leftarrow \beta + \frac{1}{2} \sum_{i=1}^N \mathbb{E}_{q_\mu}[(x_i - \mu)^2]$

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

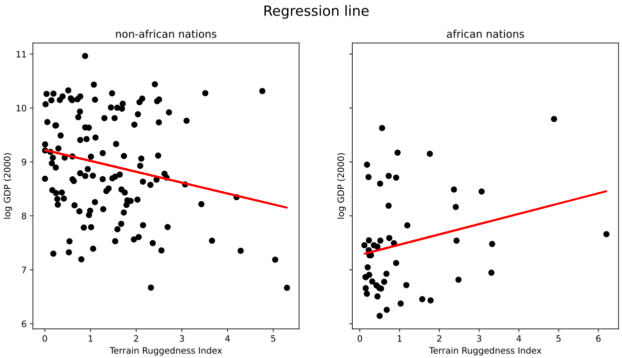
Real Data Example



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

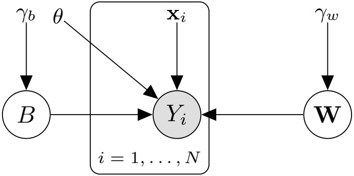


Linear Regression Model

- Negative slope for Non African Nations.
- Positive slope for African Nations.

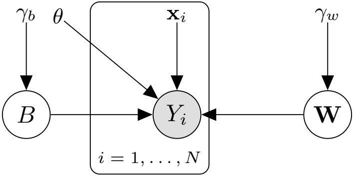
Are these relationships really supported by the data?

The Bayesian linear regression model



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

The Bayesian linear regression model

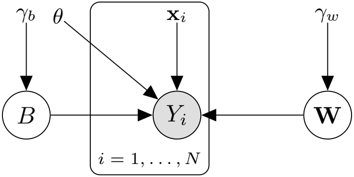


- Num. of data dim: M
- Num. of data inst: N
- $Y_i \mid \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\intercal \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The probability model

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

The Bayesian linear regression model



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

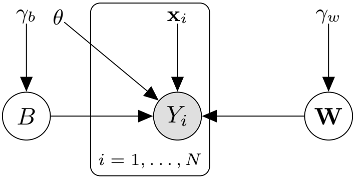
The probability model

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

...after taking the log

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

The Bayesian linear regression model



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

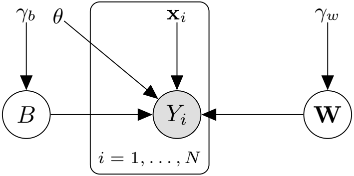
The probability model

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

... after taking the log

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

The Bayesian linear regression model



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

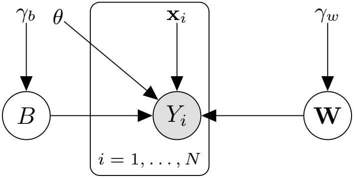
The probability model

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

The variational model (full mean field)

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

The Bayesian linear regression model



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

The probability model

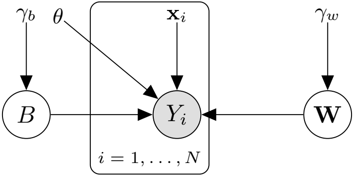
$$p(\cdot \mid \mathbf{x}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w}, b, \theta) p(\mathbf{w} \mid \gamma_w) p(b \mid \gamma_b)$$

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
- $Y_i \mid \{\mathbf{w}, \mathbf{x}_i, b, \theta\} \sim \mathcal{N}(\mathbf{w}^\intercal \mathbf{x}_i + b, 1/\theta)$
- $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \gamma_w^{-1} \mathbf{I}_{M \times M})$
- $B \sim \mathcal{N}(0, \gamma_b^{-1})$

The probability model

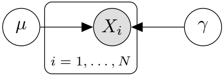
$$p(\cdot \mid \mathbf{x}, \theta, \gamma_w, \gamma_b) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \mathbf{w}, b, \theta) p(\mathbf{w} \mid \gamma_w) p(b \mid \gamma_b)$$

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

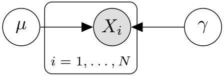
Supplementary



We choose the variational distribution so that

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

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

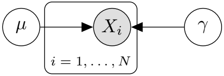


We choose the variational distribution so that

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

Recall

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



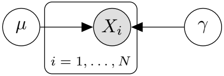
We choose the variational distribution so that

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

Recall

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

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



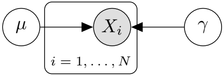
We choose the variational distribution so that

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

Recall

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

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



We choose the variational distribution so that

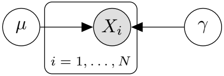
$$\log q(\gamma) = \mathbb{E}_{q_\mu}[\log p(\mathbf{x}, \mu, \gamma)] + c = \sum_{i=1}^{N} \mathbb{E}_{q_\mu}[\log p(x_i \mid \mu, \gamma^{-1})] + \log p(\gamma) + c$$

Recall

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

$$\log p(\gamma \mid \alpha, \beta) = \text{Gamma}(\alpha, \beta) = \alpha \cdot \log(\beta) + (\alpha - 1)\log(\gamma) - \beta \cdot \gamma - \log(\Gamma(\alpha))$$

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



We choose the variational distribution so that

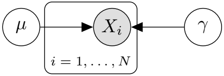
$$\log q(\gamma) = \mathbb{E}_{q_\mu}[\log p(\mathbf{x}, \mu, \gamma)] + c = \sum_{i=1}^{N} \mathbb{E}_{q_\mu}[\log p(x_i \,|\, \mu, \gamma^{-1})] + \log p(\gamma) + c$$

Recall

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

$$\log p(\gamma \,|\, \alpha, \beta) = \text{Gamma}(\alpha, \beta) = \alpha \cdot \log(\beta) + (\alpha - 1)\log(\gamma) - \beta \cdot \gamma - \log(\Gamma(\alpha))$$

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



We choose the variational distribution so that

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

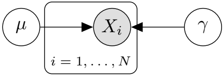
$$= \frac{N}{2}\log(\gamma) - \frac{\gamma}{2}\sum_{i=1}^{N} \mathbb{E}_{q_\mu}[(x_i - \mu)^2] + (\alpha - 1)\log(\gamma) - \beta \cdot \gamma + c$$

Recall

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

$$\log p(\gamma \mid \alpha, \beta) = \text{Gamma}(\alpha, \beta) = \alpha \cdot \log(\beta) + (\alpha - 1)\log(\gamma) - \beta \cdot \gamma - \log(\Gamma(\alpha))$$

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



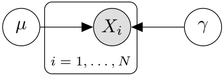
We choose the variational distribution so that

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

$$= \frac{N}{2} \log(\gamma) - \frac{\gamma}{2} \sum_{i=1}^{N} \mathbb{E}_{q_\mu}[(x_i - \mu)^2] + (\alpha - 1) \log(\gamma) - \beta \cdot \gamma + c$$

$$= \left(\frac{N}{2} + \alpha - 1\right) \log(\gamma) - \left(\frac{1}{2} \sum_{i=1}^{N} \mathbb{E}_{q_\mu}[(x_i - \mu)^2] + \beta\right) \cdot \gamma + c$$

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

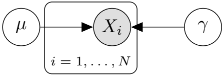


We choose the variational distribution so that

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

$$= \left(\frac{N}{2} + \alpha - 1\right) \log(\gamma) - \left(\frac{1}{2} \sum_{i=1}^{N} \mathbb{E}_{q_\mu}[(x_i - \mu)^2] + \beta\right) \cdot \gamma + c$$

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



We choose the variational distribution so that

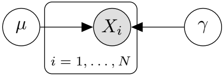
$$\log q(\gamma) = \mathbb{E}_{q_\mu}[\log p(\mathbf{x}, \mu, \gamma)] + c = \sum_{i=1}^{N} \mathbb{E}_{q_\mu}[\log p(x_i \mid \mu, \gamma^{-1})] + \log p(\gamma) + c$$

$$= \left(\frac{N}{2} + \alpha - 1\right) \log(\gamma) - \left(\frac{1}{2} \sum_{i=1}^{N} \mathbb{E}_{q_\mu}[(x_i - \mu)^2] + \beta\right) \cdot \gamma + c$$

Recall

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

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



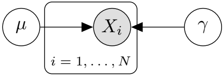
We choose the variational distribution so that

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

$$= \left(\frac{N}{2} + \alpha - 1\right) \log(\gamma) - \left(\frac{1}{2} \sum_{i=1}^{N} \mathbb{E}_{q_\mu}[(x_i - \mu)^2] + \beta\right) \cdot \gamma + c$$

Recall

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



We choose the variational distribution so that

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

$$= \left(\frac{N}{2} + \alpha - 1\right) \log(\gamma) - \left(\frac{1}{2}\sum_{i=1}^{N} \mathbb{E}_{q_\mu}[(x_i - \mu)^2] + \beta\right) \cdot \gamma + c$$

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:

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

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

We choose the variational distribution so that

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

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

We choose the variational distribution so that

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

Recall

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

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

We choose the variational distribution so that

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

Recall

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

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

We choose the variational distribution so that

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

Recall

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

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

We choose the variational distribution so that

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

The normal distribution

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

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

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

We choose the variational distribution so that

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

The normal distribution

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

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

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

We choose the variational distribution so that

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

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

The normal distribution

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

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

We choose the variational distribution so that

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

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

Expanding the square

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

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

We choose the variational distribution so that

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

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

Expanding the square

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

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

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

We choose the variational distribution so that

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

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

$$= -\frac{\gamma_w}{2} w_j^2 - \theta \sum_{i=1}^{N} (\frac{1}{2} x_{ij}^2 w_j^2 + w_j (\sum_{k \neq j} x_{ij} x_{ik} \, \mathbb{E}(W_k) + x_{ij} \, \mathbb{E}(B) - y x_{ij}) + c$$

Expanding the square

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

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

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

We choose the variational distribution so that

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

$$= -\frac{\gamma_w}{2} w_j^2 - \frac{\theta}{2} \sum_{i=1}^{N} \mathbb{E}((y_i - (\mathbf{W}^\intercal \mathbf{x}_i + B))^2) + c$$

$$= -\frac{\gamma_w}{2} w_j^2 - \theta \sum_{i=1}^{N} (\frac{1}{2} x_{ij}^2 w_j^2 + w_j (\sum_{k \neq j} x_{ij} x_{ik} \, \mathbb{E}(W_k) + x_{ij} \, \mathbb{E}(B) - y x_{ij}) + c$$

$$= -\frac{1}{2} (\gamma_w + \theta \sum_{i=1}^{N} (x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^{N} x_{ij} (y_i - (\sum_{k \neq j} x_{ik} \, \mathbb{E}(W_k) + \mathbb{E}(B))) + c$$

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

We choose the variational distribution so that

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

$$= -\frac{1}{2}(\gamma_w + \theta \sum_{i=1}^{N}(x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^{N} x_{ij}(y_i - (\sum_{k \neq j} x_{ik} \, \mathbb{E}(W_k) + \mathbb{E}(B))) + c$$

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

We choose the variational distribution so that

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

$$= -\frac{1}{2}(\gamma_w + \theta \sum_{i=1}^{N}(x_{ij}^2) w_j^2 + w_j \theta \sum_{i=1}^{N} x_{ij}(y_i - (\sum_{k \neq j} x_{ik}\, \mathbb{E}(W_k) + \mathbb{E}(B))) + c$$

Recall the normal distribution

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

$$= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{1}{2}\tau w^2 - \frac{\tau}{2}\mu^2 + w\tau\mu$$

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

We choose the variational distribution so that

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

$$= -\frac{1}{2}\left(\gamma_w + \theta \sum_{i=1}^{N} (x_{ij}^2)\right) w_j^2 + w_j \theta \sum_{i=1}^{N} x_{ij}\left(y_i - \left(\sum_{k \neq j} x_{ik}\, \mathbb{E}(W_k) + \mathbb{E}(B)\right)\right) + c$$

Recall the normal distribution

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

$$= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{1}{2}\tau w^2 - \frac{\tau}{2}\mu^2 + w\tau\mu$$

We choose the variational distribution so that

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

$$= -\frac{1}{2}\left(\gamma_w + \theta \sum_{i=1}^{N} (x_{ij}^2)\right) w_j^2 + w_j \theta \sum_{i=1}^{N} x_{ij}\left(y_i - \left(\sum_{k \neq j} x_{ik}\, \mathbb{E}(W_k) + \mathbb{E}(B)\right)\right) + c$$

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

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

$$= -\frac{1}{2}\log(2\pi) + \frac{1}{2}\log(\tau) - \frac{1}{2}\tau w^2 - \frac{\tau}{2}\mu^2 + w\tau\mu$$

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

We choose the variational distribution so that

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

We choose the variational distribution so that

$$\log q(b) = \underset{q_{\neg w_j}}{\mathbb{E}} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(b \mid \gamma_b) + c$$

$$= \dots$$

$$= -\frac{1}{2}(\gamma_b + \theta N) b^2 + b\left(\theta \sum_{i=1}^{N} (y_i - \mathbb{E}(\mathbf{W})^\intercal \mathbf{x}_i)\right) + c$$

Recall the normal distribution

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

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

We choose the variational distribution so that

$$\log q(b) = \mathop{\mathbb{E}}_{q \neg w_j} \log p(\cdot \mid \mathbf{W}, B, \theta, \boldsymbol{\gamma}) + c = \sum_{i=1}^{N} \mathbb{E} \log p(y_i \mid \mathbf{x}_i, \mathbf{W}, \theta) + \log p(b \mid \gamma_b) + c$$

$$= \dots$$

$$= -\frac{1}{2}(\gamma_b + \theta N) b^2 + b \left(\theta \sum_{i=1}^{N} (y_i - \mathbb{E}(\mathbf{W})^\intercal \mathbf{x}_i) \right) + c$$

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

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

Recall the normal distribution

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

$$= -\frac{1}{2} \log(2\pi) + \frac{1}{2} \log(\tau) - \frac{1}{2} \tau b^2 - \frac{\tau}{2} \mu^2 + b \tau \mu$$