Variational Inference

Outline

- Laplace Approximation
- Motivation for variational inference
- Mean field assumption
- Variational Bayes
- Example 1: Univariate Gaussian
- Example 2: Linear Regression
- Conclusion

Laplace Approximation

- Aims to find a Gaussian approximation to a (intractable) continuous probability distribution

Posterior: $p(\boldsymbol{\theta}|\mathcal{D}) = \frac{1}{Z} e^{-E(\boldsymbol{\theta})}$ where $E(\boldsymbol{\theta}) \; = \; -\log p(\boldsymbol{\theta}, \mathcal{D})$

Idea: Taylor series expansion around the mode of E($\boldsymbol{\theta}$)

$$E(\boldsymbol{\theta}) \approx E(\boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{g} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{H}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

$$\mathbf{g} \triangleq \nabla E(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}^*}, \;\; \mathbf{H} \triangleq \frac{\partial^2 E(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \,\big|_{\boldsymbol{\theta}^*}$$

Laplace Approximation

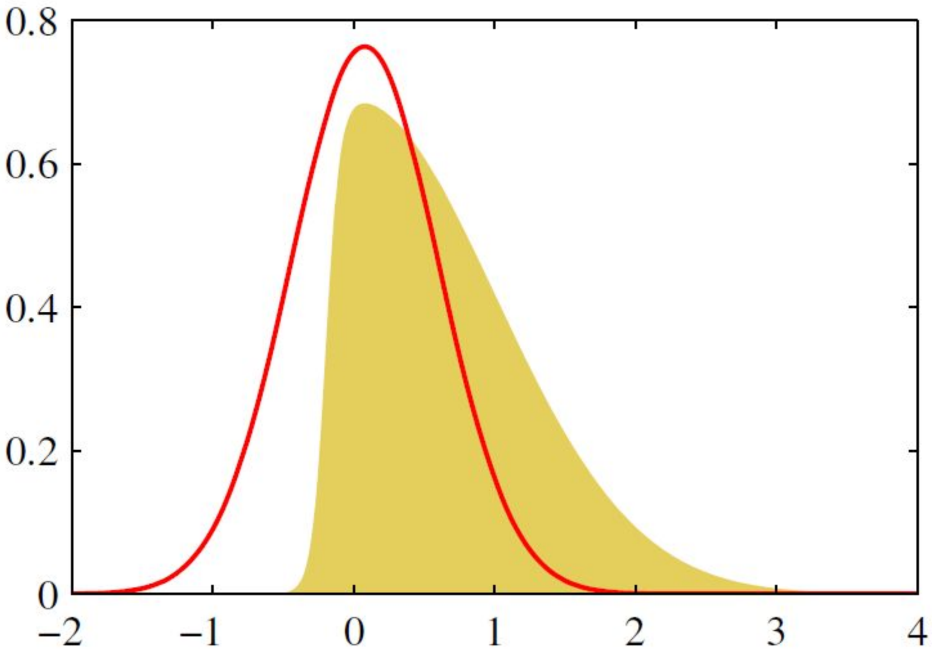
$$\begin{aligned}\hat{p}(\boldsymbol{\theta}|\mathcal{D}) &\approx \frac{1}{Z}e^{-E(\boldsymbol{\theta}^*)}\exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\mathbf{H}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\right] \\ &= \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\theta}^*,\mathbf{H}^{-1})\end{aligned}$$

Posterior is approximated by a Gaussian distribution

$$Z = p(\mathcal{D}) \approx \int \hat{p}(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} = e^{-E(\boldsymbol{\theta}^*)}(2\pi)^{D/2}|\mathbf{H}|^{-\frac{1}{2}}$$

Laplace approximation to marginal likelihood

Laplace Approximation



Motivation

- Important when it is difficult to compute the posterior distribution P(x | D) of the variables x given the data D (e.g: non-conjugacy for continuous variables, exponentially many hidden states for discrete variables)
- **Main Idea:** Approximate the posterior distribution p*(x) by a more tractable distribution q(x) chosen from a family of simple(r) distributions. q(x) is easy to integrate or has an analytic form.
- Choose the q(x) which “best” approximates p*(x) => choose q(x) which maximizes some form of similarity with the true posterior
- Turned the integration problem to an optimization problem !

Variational Inference

- Judge the quality of approximation using (Reverse) KL divergence:

$$\mathbb{KL}(q||p^*) = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p^*(\mathbf{x})}$$

- This is hard to compute since calculating p*(x) requires knowledge of the normalization constant Z => use unnormalized posterior distribution
- p~(x) = p*(x) Z
- Objective: Minimize

$$J(q) \triangleq \mathbb{KL}(q||\tilde{p})$$

Variational Inference

$$L(q) \triangleq -J(q) = -\mathbb{KL}\left(q \| p^*\right) + \log Z \leq \log Z = \log p(\mathcal{D})$$

Objective: Maximize L(q) i.e. the lower bound on the log likelihood of observing the data

$$J(q) = \mathbb{E}_q\left[\log q(\mathbf{x})\right] + \mathbb{E}_q\left[-\log \tilde{p}(\mathbf{x})\right] = -\mathbb{H}\left(q\right) + \mathbb{E}_q\left[E(\mathbf{x})\right]$$

Helmholtz Free Energy

Entropy

Expected Energy

- q(x) needs to be zero when p*(x) is zero => Reverse KL divergence is zero forcing => q(x) will under-estimate the support of p*(x)

Mean Field assumption

- Posterior is fully factorized => $q(\mathbf{x}) = \prod_i q_i(\mathbf{x}_i)$
- Rewriting our objective function with this assumption:

$$
\begin{aligned}
L(q_j) &= \sum_{\mathbf{x}} \prod_i q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right] \\
&= \sum_{\mathbf{x}_j} \sum_{\mathbf{x}_{-j}} q_j(\mathbf{x}_j) \prod_{i \neq j} q_i(\mathbf{x}_i) \left[\log \tilde{p}(\mathbf{x}) - \sum_k \log q_k(\mathbf{x}_k) \right] \\
&= \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log f_j(\mathbf{x}_j) - \sum_{\mathbf{x}_j} q_j(\mathbf{x}_j) \log q_j(\mathbf{x}_j) + \text{const}
\end{aligned}
$$

Mean Field assumption

$$\log f_j(\mathbf{x}_j) \triangleq \sum_{\mathbf{x}_{-j}} \prod_{i \neq j} q_i(\mathbf{x}_i) \log \tilde{p}(\mathbf{x}) = \mathbb{E}_{-q_j} \left[\log \tilde{p}(\mathbf{x})\right]$$

Expectation wrt to all q_i except j

Rewriting $L(q_j)$

$$L(q_j) = -\mathbb{KL}\left(q_j || f_j\right)$$

Update equation for q_j

$$\log q_j(\mathbf{x}_j) = \mathbb{E}_{-q_j}\left[\log \tilde{p}(\mathbf{x})\right]$$

Do coordinate descent wrt each variable using the above update !

Variational Bayes

- Method to infer the parameters of a model. Use mean field assumption on the parameters:

$$p(\boldsymbol{\theta}|\mathcal{D}) \approx \prod_k q(\boldsymbol{\theta}_k)$$

VB for Univariate Gaussian

Likelihood:

$$\mathcal{N}(x|\mu,\sigma^2) \quad = \quad \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

Let $\lambda = 1/\sigma^2$ Using conjugate **prior** to compare against true posterior :

$$p(\mu,\lambda) = \mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})\mathrm{Ga}(\lambda|a_0,b_0)$$

Un-normalized posterior:

$$\log\tilde{p}(\mu,\lambda) \quad = \quad \log p(\mu,\lambda,\mathcal{D}) = \log p(\mathcal{D}|\mu,\lambda) + \log p(\mu|\lambda) + \log p(\lambda)$$

VB for Univariate Gaussian

Mean Field approximation to posterior: $q(\mu, \lambda) = q_\mu(\mu) q_\lambda(\lambda)$

Update equations:

$$\log q_\mu(\mu) \quad = \quad \mathbb{E}_{q_\lambda}\left[\log p(\mathcal{D}|\mu, \lambda) + \log p(\mu|\lambda)\right]$$

$$q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \kappa_N^{-1})$$

$$\mu_N = \frac{\kappa_0 \mu_0 + N\overline{x}}{\kappa_0 + N}, \; \kappa_N = (\kappa_0 + N)\mathbb{E}_{q_\lambda}\left[\lambda\right]$$

VB for Univariate Gaussian

Update equations:

$$\log q_\lambda(\lambda) \quad = \quad \mathbb{E}_{q_\mu}\left[\log p(\mathcal{D}|\mu, \lambda) + \log p(\mu|\lambda) + \log p(\lambda)\right]$$

$$q_\lambda(\lambda) = \mathrm{Ga}(\lambda|a_N, b_N),$$

$$\begin{aligned} a_N &= a_0 + \frac{N+1}{2} \\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_{q_\mu}\left[\kappa_0(\mu - \mu_0)^2 + \sum_{i=1}^N (x_i - \mu)^2\right] \end{aligned}$$

VB for Univariate Gaussian

Computing expectations:

$$\mathbb{E}_{q(\mu)}[\mu] = \mu_N$$

$$\mathbb{E}_{q(\mu)}[\mu^2] = \frac{1}{\kappa_N} + \mu_N^2$$

$$\mathbb{E}_{q(\lambda)}[\lambda] = \frac{a_N}{b_N}$$

Final updates:

$$\mu_N = \frac{\kappa_0\mu_0 + N\overline{x}}{\kappa_0 + N} \qquad a_N = a_0 + \frac{N+1}{2}$$

$$\kappa_N = (\kappa_0 + N)\frac{a_N}{b_N} \qquad b_N = b_0 + \kappa_0(\mathbb{E}[\mu^2] + \mu_0^2 - 2\mathbb{E}[\mu]\mu_0) + \frac{1}{2}\sum_{i=1}^{N}(x_i^2 + \mathbb{E}[\mu^2] - 2\mathbb{E}[\mu]x_i)$$

VB for Univariate Gaussian

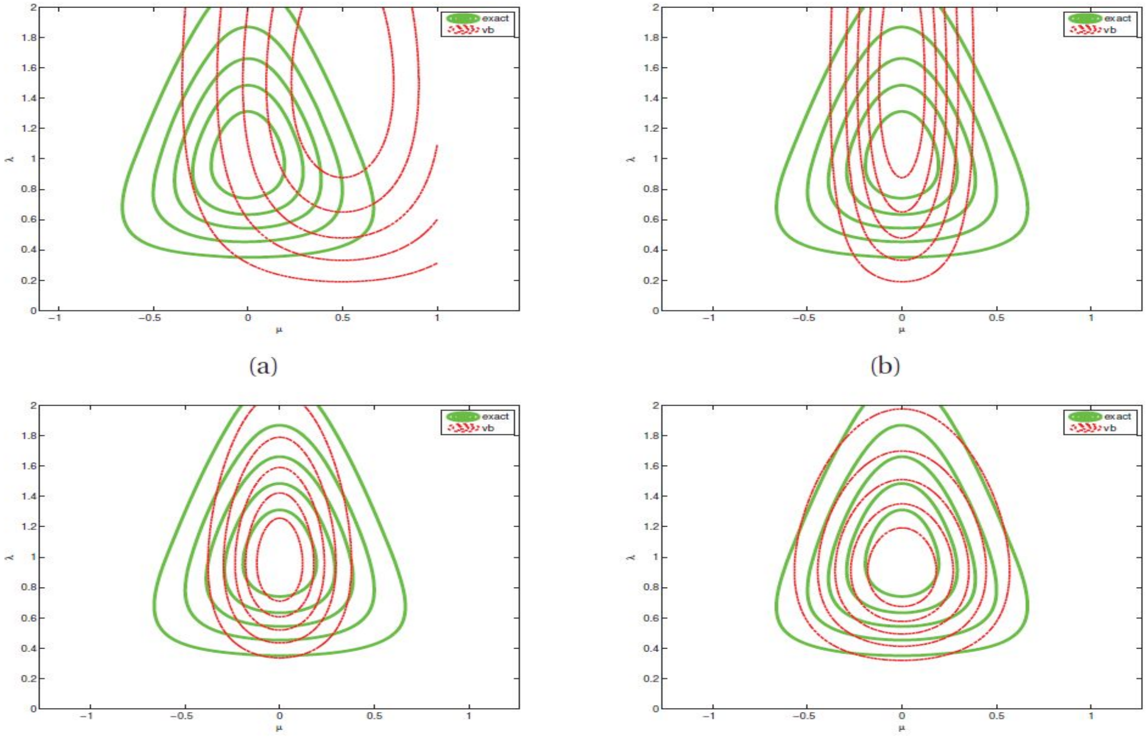
Calculate the objective function:

$$L(q) = \int\int q(\mu,\lambda) \log \frac{p(\mathcal{D},\mu,\lambda)}{q(\mu,\lambda)} d\mu d\lambda$$

$$L(q) = \frac{1}{2}\log\frac{1}{\kappa_N} + \log\Gamma(a_N) - a_N \log b_N + \text{const}$$

Evaluate this function at each iteration. Terminate when its increments become small.

VB for Univariate Gaussian



VB: Linear Regression

Likelihood:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{X}\mathbf{w}, \lambda^{-1})$$

Prior:

$$p(\mathbf{w}, \lambda, \alpha) \quad = \quad \mathcal{N}(\mathbf{w}|\mathbf{0}, (\lambda\alpha)^{-1}\mathbf{I})\mathrm{Ga}(\lambda|a_0^\lambda, b_0^\lambda)\mathrm{Ga}(\alpha|a_0^\alpha, b_0^\alpha)$$

Mean Field assumption:

$$q(\mathbf{w}, \alpha, \lambda) = q(\mathbf{w}, \lambda)q(\alpha)$$

After solving the update equations:

$$q(\mathbf{w}, \alpha, \lambda) \quad = \quad \mathcal{N}(\mathbf{w}|\mathbf{w}_N, \lambda^{-1}\mathbf{V}_N)\mathrm{Ga}(\lambda|a_N^\lambda, b_N^\lambda)\mathrm{Ga}(\alpha|a_N^\alpha, b_N^\alpha)$$

Summary

Advantages:

- Reduces integration to an optimization problem
- Well defined termination criteria. Easy to debug.
- Mean field assumption “automatically” picks the family of distributions
- Arguably a more principled approach than sampling

Disadvantages:

- Not consistent !
- Not as out-of-the-box as sampling