Approximate Inference

Bayesian Inference

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

Rev'd Thomas Bayes (1702–1761)

- Bayes rule tells us how to do inference about hypotheses from data.
- Learning and prediction can be seen as forms of inference.

Bayesian Inference

$$P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}\|\theta, m)$	likelihood of parameters θ in model m
$P(\theta\|m)$	prior probability of θ
$P(\theta\|\mathcal{D}, m)$	posterior of θ given data $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D}, m) = \int P(x|\theta, \mathcal{D}, m)P(\theta|\mathcal{D}, m)d\theta$$

Model Comparison:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta, m)P(\theta|m)\, d\theta$$

Motivation

- One of central tasks: Bayesian Inference

  can be latent variables or model parameters in general

  - Evaluation of posterior distribution p(Z | X), and expectation of certain function w.r.t. this posterior. E.g. in EM, expected log likelihood w.r.t. p(Z | X)

$$\mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X})}[f(\boldsymbol{Z})]$$

- Unfortunately, the expectation is typically intractable.

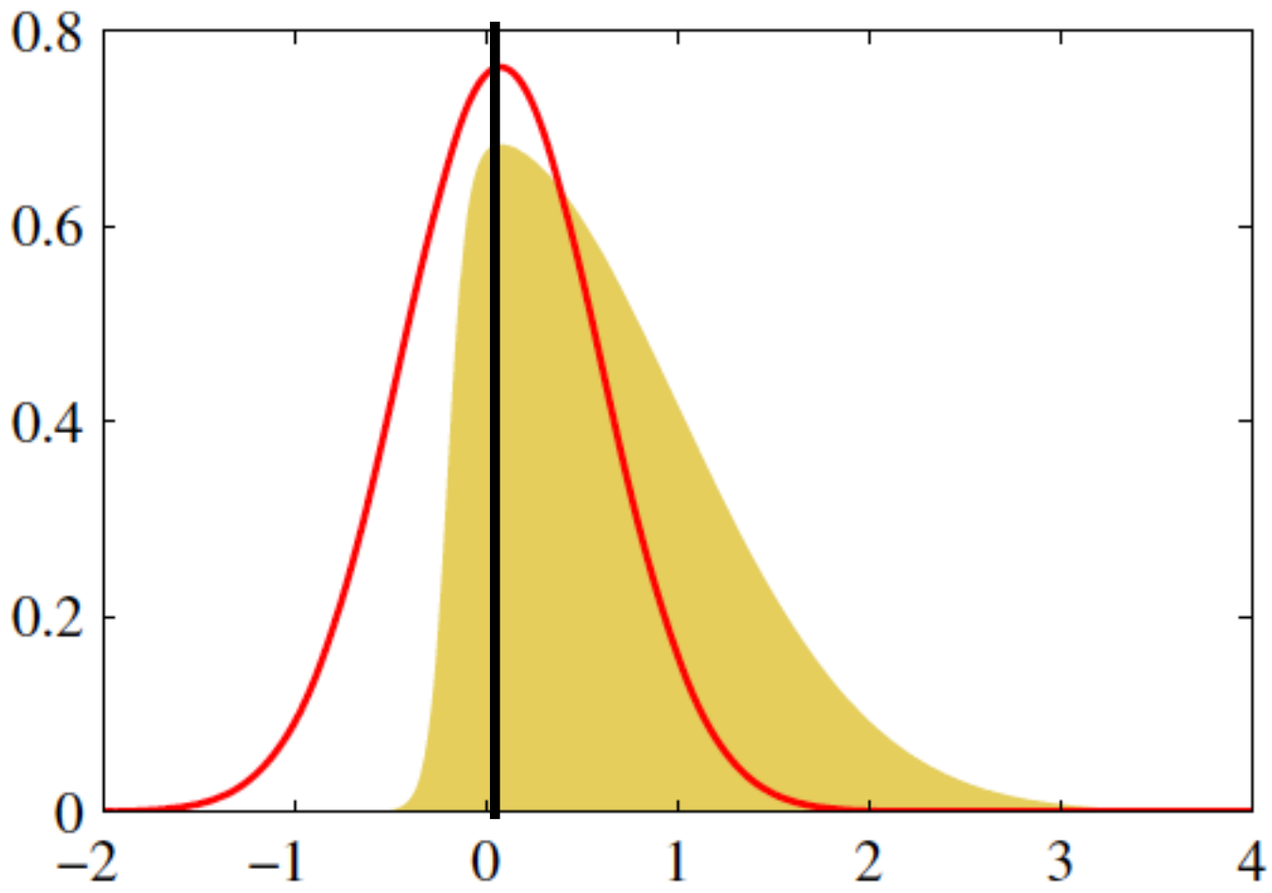
- Approximation schemes are required

  - Deterministic techniques: **Laplace approximation**, **variational inference**

  - Stochastic techniques: **Markov Chain Monte Carlo** (MCMC)

Laplace Approximation

- The idea
  - using Gaussian distribution to approximate the target posterior distribution around the mode



Single Variable Case

Target distribution: $$p(z) = \frac{1}{Z} f(z)$$

Find the mode: $$\left. \frac{df(z)}{dz} \right|_{z=z_0} = 0$$

2nd order Taylor expansion: $$\ln f(z) \simeq \ln f(z_0) - \frac{1}{2} A (z - z_0)^2$$

$$A = - \left. \frac{d^2}{dz^2} \ln f(z) \right|_{z=z_0}$$

$$f(z) \simeq f(z_0) \exp \left\{ -\frac{A}{2} (z - z_0)^2 \right\}$$

The Gaussian approximation: $$q(z) = \left(\frac{A}{2\pi} \right)^{1/2} \exp \left\{ -\frac{A}{2} (z - z_0)^2 \right\}$$

Multi-variable Case

$$\ln f(\mathbf{z}) \simeq \ln f(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)$$

Hessian matrix: $\mathbf{A} = -\nabla\nabla \ln f(\mathbf{z})|_{\mathbf{z}=\mathbf{z}_0}$

$$f(\mathbf{z}) \simeq f(\mathbf{z}_0)\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\}$$

$$q(\mathbf{z}) = \frac{|\mathbf{A}|^{1/2}}{(2\pi)^{M/2}}\exp\left\{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^{\mathrm{T}}\mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right\} = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

Some Remarks on Lap. Approx.

- When will the Lap. Approx. be better?
  - Single-mode target distribution
  - Expected to be better with increasingly number of observed data points (Why? Hint: central limit theorem)
- Finding the mode
  - Typically require numerical optimization
- Though simple and intuitive, the computational complexity is indeed a problem particularly for high-dimensional space. (Why?)

Variational Inference

- Calculus of variations: 18th century by Euler and Lagrange
- The idea
  - Using a family of parametric distribution to approximate the posterior distribution

$$q(\mathbf{Z}|\boldsymbol{\omega}) \sim p(\mathbf{Z}|\mathbf{X})$$

- Turn an inference problem to an optimization problem

Leonhard Euler
1707–1783

Euler was a Swiss mathematician and physicist who worked in St. Petersburg and Berlin and who is widely considered to be one of the greatest mathematicians of all time. He is certainly the most prolific, and his collected works fill 75 volumes. Amongst his many contributions, he formulated the modern theory of the function, he developed (together with Lagrange) the calculus of variations, and he discovered the formula $e^{i\pi} = -1$, which relates four of the most important numbers in mathematics. During the last 17 years of his life, he was almost totally blind, and yet he produced nearly half of his results during this period.

- Decompose the log marginal probability:

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

Evidence lower bound (ELBO):

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- The ELBO can be derived either by the non-negativity of KL distance or the Jensen's inequality.

- Choose a restricted family of variational distribution

  - Then optimize w.r.t. (the parameters of) variational distribution.

Factorized Distributions as Variational Distributions

- Factorized distributions (mean field theory)

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$$

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}
\end{aligned}$$

$$\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const} \qquad \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i$$

Evaluate the derivative of ELBO w.r.t. to q(Z), and let it be zero.

$$\ln q_j^\star(\mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const.}$$

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\mathbf{Z})]\right)\, \mathrm{d}\mathbf{Z}_j}$$

Maximizing the ELBO via coordinate ascent,
i.e. update each q(Zj) alternatively with others fixed.

Guaranteed to converge to local minima.

Exponential Family Conditionals

- Suppose each conditional is in an **exponential family** (e.g. Gaussian, Bernoulli, etc.).

$$p(z_j \mid z_{-j}, x) = h(z_j) \exp\{\eta(z_{-j}, x)^\top t(z_j) - a(\eta(z_{-j}, x))\}$$

Natural parameters **Sufficient statistics**

Mean field variational inference is straightforward

– Compute the log of the conditional

$$\log p(z_j \mid z_{-j}, x) = \log h(z_j) + \eta(z_{-j}, x)^\top t(z_j) - a(\eta(z_{-j}, x))$$

– Compute the expectation with respect to $q(z_{-j})$

$$\mathrm{E}[\log p(z_j \mid z_{-j}, x)] = \log h(z_j) + \mathrm{E}[\eta(z_{-j}, x)]^\top t(z_j) - \mathrm{E}[a(\eta(z_{-j}, x))]$$

– Noting that the last term does not depend on q_j, this means that

$$q^*(z_j) \propto h(z_j) \exp\{\mathrm{E}[\eta(z_{-j}, x)]^\top t(z_j)\}$$

and the normalizing constant is $a(\mathrm{E}[\eta(z_{-j}, x)])$.

So, the optimal $q(z_j)$ is in the same exponential family as the conditional.

Coordinate ascent algorithm

- Give each hidden variable a variational parameter ν_j, and put each one in the same exponential family as its model conditional,

$$q(z_{1:m} \mid \nu) = \prod_{j=1}^{m} q(z_j \mid \nu_j)$$

- The coordinate ascent algorithm iteratively sets each natural variational parameter ν_j equal to the expectation of the natural conditional parameter for variable z_i given all the other variables and the observations,

$$\nu_j^* = \mathrm{E}[\eta(z_{-j}, x)].$$

Variational Mixture of Gaussians

- **Bayesian Mixture of Gaussians**
  - Put prior distribution over the hidden variables z and Gaussian parameters
  - Evaluate the posterior distribution over these hidden variables/parameters, i.e. not just a point estimate, we obtain the entire distribution over hidden variables.
  - Intractable posterior distribution.

The Generative Process:

1. Draw $\mu_k \sim \mathcal{N}(0, \tau^2)$ for $k = 1 \ldots K$.
2. For $i = 1 \ldots n$:
   (a) Draw $z_i \sim \text{Mult}(\pi)$;
   (b) Draw $x_i \sim \mathcal{N}(\mu_{z_i}, \sigma^2)$

- The intractable posterior

$$p(\mu_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}$$

Tractable

Intractable

Mean field family

$$q(\mu_{1:K}, z_{1:n}) = \prod_{k=1}^{K} q(\mu_k \mid \tilde{\mu}_k, \tilde{\sigma}_k^2) \prod_{i=1}^{n} q(z_i \mid \phi_i)$$

- Update variational distribution $q(z_i)$

$$q^*(z_i) \propto \exp\{\mathrm{E}_{-i}[\log p(\mu_{1:K}, z_i, z_{-i}, x_{1:n})]\}$$

$$\log p(\mu_{1:K}, z_i, z_{-i}, x_{1:n}) =$$
$$\log p(\mu_{1:k}) + \left(\sum_{j \neq i} \log p(z_j) + \log p(x_j \mid z_j)\right) + \log p(z_i) + \log p(x_i \mid z_i)$$

$$q^*(z_i) \propto \exp\{\log \pi_{z_i} + \mathrm{E}[\log p(x_i \mid \mu_{z_i})]\}$$

$$\mathrm{E}[\log p(x_i \mid \mu_i)]\} = -(1/2)\log 2\pi - x_i^2/2 + x_i \mathrm{E}[\mu_{z_i}] - \mathrm{E}[\mu_{z_i}^2]/2$$

Thus the coordinate update for $q(z_i)$ is

$$q^*(z_i = k) \propto \exp\{\log \pi_k + x_i \mathrm{E}[\mu_k] - \mathrm{E}[\mu_k^2]/2\}$$

(A)

- Update variational distribution $q(\mu_k)$

– What is the conditional distribution of μ_k given $x_{1:n}$ and $z_{1:n}$?

– Intuitively, this is the posterior Gaussian mean with the data being the observations that were assigned (in $z_{1:n}$) to the kth cluster.

– Let's put the prior and posterior, which are Gaussians, in their canonical form. The parameters are

$$\begin{aligned}\hat{\lambda}_1 &= \lambda_1 + \textstyle\sum_{i=1}^n z_i^k x_i \\ \hat{\lambda}_2 &= \lambda_2 + \textstyle\sum_{i=1}^n z_i^k).\end{aligned}$$

– So, the optimal variational family is going to be a Gaussian with natural parameters

$$\begin{aligned}\tilde{\lambda}_1 &= \lambda_1 + \textstyle\sum_{i=1}^n \mathrm{E}[z_i^k] x_i \\ \tilde{\lambda}_2 &= \lambda_2 + \textstyle\sum_{i=1}^n \mathrm{E}[z_i^k]\end{aligned}$$

– Finally, because z_i^k is an indicator, its expectation is its probability, i.e., $q(z_i = k)$.

It's convenient to specify the Gaussian prior in its mean parameterization, and we need the expectations of the variational posterior for the updates on z_i.

– The mapping from natural parameters to mean parameters is

$$\begin{aligned} \mathrm{E}[X] &= \eta_1/\eta_2 \\ \mathrm{Var}(X) &= 1/\eta_2 \end{aligned}$$

(Note: this is an alternative parameterization of the Gaussian, appropriate for the conjugate prior of the unit-variance likelihood.

– So, the variational posterior mean and variance of the cluster component k is

$$\begin{aligned} \mathrm{E}[\mu_k] &= \frac{\lambda_1 + \sum_{i=1}^n \mathrm{E}[z_i^k] x_i}{\lambda_2 + \sum_{i=1}^n \mathrm{E}[z_i^k]} \\ \mathrm{Var}(\mu_k) &= 1/(\lambda_2 + \sum_{i=1}^n \mathrm{E}[z_i^k]) \end{aligned}$$

We'd rather specify a prior mean and variance.

– For the Gaussian conjugate prior, we map

$$\eta = \langle \mu/\sigma^2, 1/\sigma^2 \rangle.$$

– This gives the variational update in mean parameter form,

$$\begin{aligned} \mathrm{E}[\mu_k] &= \frac{\mu_0/\sigma_0^2 + \sum_{i=1}^n \mathrm{E}[z_i^k] x_i}{1/\sigma_0^2 + \sum_{i=1}^n \mathrm{E}[z_i^k]} \quad \textbf{(B)} \\ \mathrm{Var}(\mu_k) &= 1/(1/\sigma_0^2 + \textstyle\sum_{i=1}^n \mathrm{E}[z_i^k]). \quad \textbf{(C)} \end{aligned}$$

These are the usual Bayesian updates with the data weighted by its variational probability of being assigned to cluster k.

- The ELBO is the sum of two terms,

$$\left(\sum_{k=1}^K \mathrm{E}[\log p(\mu_k)] + \mathrm{H}(q(\mu_k)) \right) + \left(\sum_{i=1}^n \mathrm{E}[\log p(z_i)] + \mathrm{E}[\log p(x_i \mid z_i, \mu_{1:K})] + \mathrm{H}(q(z_i)) \right)$$

- The expectations in these terms are the following.
  - The expected log prior over mixture locations is

$$\mathrm{E}[\log p(\mu_k)] = -(1/2)\log 2\pi\sigma_0^2 - \mathrm{E}[\mu_k^2]/2\sigma_0^2 + \mathrm{E}[\mu_k]\mu_0/\sigma_0^2 - \mu_0^2/2\sigma_0^2,$$

    where $\mathrm{E}[\mu_k] = \tilde{\mu}_k$ and $\mathrm{E}[\mu_k^2] = \tilde{\sigma}_k^2 + \tilde{\mu}_k^2$.
  - The expected log prior over mixture assignments is not random,

$$\mathrm{E}[\log p(z_i)] = \log(1/K)$$

  - The entropy of each variational location posterior is

$$\mathrm{H}(q(\mu_k)) = (1/2)\log 2\pi\tilde{\sigma}_k^2 + 1/2.$$

    If you haven't seen this, work it out at home by computing $-\mathrm{E}[\log q(\mu_k)]$.
  - The entropy of each variational assignment posterior is

$$\mathrm{H}(q(z_i)) = -\sum_{k=1}^{K} \phi_{ij}\log\phi_{ij}$$

- Now we can describe the coordinate ascent algorithm.
  - We are given data $x_{1:n}$, hyperparameters μ_0 and σ_0^2, and a number of groups K.
  - The variational distributions are
    - n variational multinomials $q(z_i)$
    - K variational Gaussians $q(\mu_k \mid \tilde{\mu}_k, \tilde{\sigma}_k^2)$.
  - Repeat until the ELBO converges:

1. For each data point x_i
   - Update the variational multinomial $q(z_i)$ from Equation (A)
2. For each cluster $k = 1 \dots K$
   - Update the mean and variance from Equation (B) and Equation (C)

- We can obtain a posterior decomposition of the data.
  - Points are assigned to $\arg\max_k \phi_{i,k}$.
  - Cluster means are estimated as $\tilde{\mu}_k$.

- We can approximate the predictive distribution with a mixture of Gaussians, each at the expected cluster mean. This is

$$p(x_{\text{new}} \mid x_{1:n}) \approx \frac{1}{K} \sum_{k=1}^{K} p(x_{\text{new}} \mid \tilde{\mu}_k),$$

where $p(x \mid \tilde{\mu}_k)$ is a Gaussian with mean $\tilde{\mu}_k$ and unit variance.