Adaptive Importance Sampling

$$\int P(x) = \int \frac{P(x)}{q(x)} q(x) \approx \frac{1}{N} \sum_{x_i=1}^{N} \frac{P(x_i)}{q(x_i)} \equiv \mu^N \quad where \quad x_i \sim q(x)$$

uncertainty (squared):

$$Var(\mu^N) = \frac{1}{N} \left[\int \frac{P(x)}{q(x)} P(x)\, dx - \left(\int P(x)\, dx \right)^2 \right]$$

- minimize the uncertainty with respect to q
  - it suffices to minimize: $\log\left(\int \frac{P(x)}{q(x)} P(x)\, dx \right)$
  - by Jensen's inequality: $\geq \int \left(\log \frac{P(x)}{q(x)} \right) P(x)\, dx$
  - which is the Kullback-Leibler divergence $KL(P \| q)$

Adaptive Importance Sampling

Note:

- $(0 \leq)\ Var(\mu^N)\ = 0$ if and only if $P = q$
- $(0 \leq) KL(P \| q) = 0$ if and only if $P = q$
- $(0 \leq) KL(q \| P) = 0$ if and only if $P = q$

Although not guaranteed, there is a good chance to decrease $Var(\mu^N)$ while minimizing $KL(P \| q)$ or $KL(q \| P)$ with respect to q.

Adaptive Importance Sampling

- Conventional adaptive Importance Sampling approach:
  - restrict q to be a Gaussian or Student T mixture with a fixed number of components K:

$$q(x)=\sum_{i=1}^{K}\alpha_i q_i(x|\mu_i,\Sigma_i,\nu_i) \quad \textit{where} \quad q_i\in\{\mathcal{N},\mathcal{T}\}$$

  - minimize $KL(P\|q) \rightarrow$ EM-like parameter updates
  - no a priori information about q (i.e. flat priors for the parameters)

Adaptive Importance Sampling

- Variational Bayes adaptive Importance Sampling approach:
  - restrict q to be a Gaussian or Student T mixture with a fixed number of components K:

$$q(x)=\sum_{i=1}^{K}\alpha_i q_i(x|\mu_i,\Sigma_i,\nu_i) \quad \textit{where} \quad q_i\in\{\mathcal{N},\mathcal{T}\}$$

  - minimize $KL(q\|P)\rightarrow$ EM-like hyperparameter updates
  - include prior information about $\alpha_i, \mu_i, \Sigma_i, \nu_i$
- Very detailed in “Pattern Recognition and Machine learning” (Christopher M. Bishop)

Variational Bayes

Definition: Hyperparameter

Consider a probabilistic model with parameters $\boldsymbol{\theta}$. Then the parameters of the prior distribution are called hyperparameters **h**:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta} | \boldsymbol{h})$$

Variational Bayes

Example: Hyperparameter

Be **X** data known to be Gaussian distributed, but with unknown mean and covariance:

$$P(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{X} \sim g(\boldsymbol{x}) \qquad \boldsymbol{\mu}, \boldsymbol{\Sigma} \text{ unknown}$$

Suppose the prior can be written as follows:

$$P(\underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}}_{\text{parameters}} \mid \underbrace{\boldsymbol{m}, \beta, \boldsymbol{W}, \nu}_{\text{hyperparameters}}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}, (\beta \boldsymbol{\Lambda})^{-1})\, \mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}, \nu)$$

Then m, β, W and ν are hyperparameters.

Variational Bayes

- We end up with an EM-like algorithm for the hyperparameters.
- Can do Importance Sampling by taking the mode of $p(\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\nu}|\boldsymbol{X})$ as parameters for q.
- Can include knowledge from previous sampling runs.
- Can even correctly include Markov Chain prerun on multimodal target.

Variational Bayes (details)

Given data **X**, we can write for the evidence of any arbitrary model:

$$\ln p(\boldsymbol{X}) = \mathcal{L}(u) + KL(u \| p)$$

where we define:

$$\mathcal{L}(u) \equiv \int u(\boldsymbol{Z}, \boldsymbol{\theta}) \ln \left\{ \frac{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})}{u(\boldsymbol{Z}, \boldsymbol{\theta})} \right\} d\boldsymbol{Z}\, d\boldsymbol{\theta}$$

$$KL(u \| p) \equiv -\int u(\boldsymbol{Z}, \boldsymbol{\theta}) \ln \left\{ \frac{p(\boldsymbol{Z}, \boldsymbol{\theta} | \boldsymbol{X})}{u(\boldsymbol{Z}, \boldsymbol{\theta})} \right\} d\boldsymbol{Z}\, d\boldsymbol{\theta}$$

$\boldsymbol{\theta}$: model parameters ($\alpha_i, \mu_i, \Sigma_i, \nu_i$)

Z: latent variables (next slide)

u: an arbitrary proper probability distribution

Variational Bayes (details)

- We want: $p(\boldsymbol{Z}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu} | \boldsymbol{X})$
- We need: analytically tractable approximation
  - assume $u(\boldsymbol{Z}, \boldsymbol{\theta}) = u(\boldsymbol{Z}) u(\boldsymbol{\theta})$
  - maximize $\mathcal{L}(u)$ $\left(\Leftrightarrow\right.$ minimize $\left. KL(u \| p)\right)$

Variational Bayes (details)

Definition: Latent (hidden) variables

Be **X** = {x_1, … ,x_N} the (visible) data, then **Z** = {z_1, … ,z_N} is called latent if **Z** contains information which cannot uniquely be determined from **X** but inferred in a probabilistic model.

Example (Gaussian mixtures):

$$g(x) = \sum_i \alpha_i \mathcal{N}_i(x | \mu_i, \Sigma_i)$$

$$\boldsymbol{X} \sim g(x)$$

Then the index i is a latent variable