[Paper review 44]

Variational Inference ; A Review for Statisticians

(Blei, et.al , 2018)

[Contents]

1. Abstract
2. Introduction
3. Variational Inference
   1. Problem of Approximate Inference
   2. ELBO
   3. MFVI
   4. CAVI (Coordinate ascent MFVI)
   5. Practicalities
4. A complete example : Bayesian Mixture of Gaussians
   1. (step 1) The variational density of the "mixture assignments"
   2. (step 2) The variational density of the "mixture-component means"
   3. CAVI for the mixture of Gaussians
5. VI with exponential families
   1. Complete conditionals in the exponential family
   2. Conditional conjugacy and Bayesian models
   3. SVI (Stochastic Variational Inference)
6. Discussion

1. Abstract

core problem : **difficult-to-compute pdf**

This paper : review VARIATIONAL INFERENCE

- approximates via optimization
- step 1) posit a family of densities

  step 2) find the member of that family, closest to the target

2. Introduction

VI :faster & easier to scale to LARGE data

$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{z}) p(\mathbf{x} \mid \mathbf{z}).$

- latent variables : $\mathbf{z} = z_{1:m}$

  (help govern the distn of data)

- observation : $x = x_{1:n}$

Draw latent variable from prior $p(z)$

Relate them to observation, via likelihood $p(\mathbf{z} \mid \mathbf{x})$

MCMC

- step 1) construct ergodic Markov chain on z

  (whose stationary distn is posterior $p(\mathbf{z} \mid \mathbf{x})$)

- step 2) sample from the chain (= collect samples from stationary distn)

- step 3) approximate posterior with collected samples

VI is needed, when we need faster speed than MCMC

- when data sets are **large**
- when models are **complex**

Rather than sampling, use optimization!

Key point

- **KEY POINT 1** choose approximating distn to be **flexible**
- **KEY POINT 2** simple enough for efficient optimization

VI vs MCMC

MCMC :

- computationally intensive, but (asymptotically) exact samples
- suited to smaller dataset

VI :

- faster than MCMC (can use stochastic optimization), but do not guarantee exact samples
- suited to largerdataset
- when we want to quickly explore many models

Not only data size, but also **geometry of the posterior**

- multi-mode : VI > MCMC

Modern research in VI

- problem which involve massive data
- using improved optimization method
- easy to apply to a wide class of models
- increase the accuracy of VI (by stretching the boundaries of approx distn)

This paper...

- [Section 2] MVFI, CAVI
- [Section 3] Bayesian Mixture of Gaussians
- [Section 4-1&2] When joint density of z and x are exponential family

  [Section 4-3] SVI

3. Variational Inference

Goal of VI : approximate conditional density of latent variables, given observed variables (= $P(Z \mid X)$)

key : solve using **optimization**

(use family of densities over latent variables, parameterized by free **variational parameters**)

3-1. Problem of Approximate Inference

$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{x})}$

- evidence : $p(\mathrm{x}) = \int p(\mathrm{z}, \mathrm{x}) d\mathrm{z}$ (intractable)

Bayesian Mixture of Gaussians

- K mixture components
- mean params : $\mu = \{\mu_1, \ldots, \mu_K\}$
  - drawn from $p\left(\mu_k\right) = \mathcal{N}\left(0, \sigma^2\right)$
- how to generate x_i ?
  - step 1) choose cluster assignment c_i
  - step 2) draw x_i from $\mathcal{N}\left(c_i^{\top} \mu, 1\right)$
- Full hierarchical model :

$$\begin{aligned} \mu_k &\sim \mathcal{N}\left(0, \sigma^2\right), & k &= 1, \ldots, K, \\ c_i &\sim \text{ Categorical }(1/K, \ldots, 1/K), & i &= 1, \ldots, n . \\ x_i \mid c_i, \mu &\sim \mathcal{N}\left(c_i^{\top} \mu, 1\right) & i &= 1, \ldots, n \end{aligned}$$

- Joint pdf : (latent var : $\mathrm{z} = \{\mu, \mathrm{c}\}$)

  $p(\mu, \mathbf{c}, \mathbf{x}) = p(\mu) \prod_{i=1}^n p\left(c_i\right) p\left(x_i \mid c_i, \mu\right).$
- Evidence :

$$p(\mathrm{x}) = \int p(\mu) \prod_{i=1}^{n} \sum_{c_t} p(c_i)\, p(x_i \mid c_i, \mu)\, \mathrm{d}\mu$$
$$= \sum_{\mathrm{c}} p(\mathrm{c}) \int p(\mu) \prod_{i=1}^{n} p(x_i \mid c_i, \mu)\, \mathrm{d}\mu$$

- time complexity of K-dim : $O(K^n)$

3-2. ELBO

Optimization problem ($q^*(\mathbf{z}) = \underset{q(\mathbf{z}) \in \mathscr{Q}}{\arg\min}\, \mathrm{kL}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}))$)

Minimize KL Divergence :

$$\begin{aligned} \mathrm{KL}(q(\mathrm{z}) \| p(\mathrm{z} \mid \mathrm{x})) &= \mathbb{E}[\log q(\mathrm{z})] - \mathbb{E}[\log p(\mathrm{z} \mid \mathrm{x})] \\ &= \mathbb{E}[\log q(\mathrm{z})] - \mathbb{E}[\log p(\mathrm{z}, \mathrm{x})] + \log p(\mathrm{x}) \end{aligned}$$

Maximize ELBO:

$$\mathrm{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})]$$

Interpretation of ELBO :

$$\begin{aligned} \mathrm{ELBO}(q) &= \mathbb{E}[\log p(\mathbf{z})] + \mathbb{E}[\log p(\mathbf{x} \mid \mathbf{z})] - \mathbb{E}[\log q(\mathbf{z})] \\ &= \mathbb{E}[\log p(\mathbf{x} \mid \mathbf{z})] - \mathrm{kL}(q(\mathbf{z}) \| p(\mathbf{z})) \end{aligned}.$$

- first term) fit well
- second term) regularize well

Relationship between ELBO & $\log p(x)$

- used for **model selection** criterion

EM vs VI

- EM assumes, expectation under $p(z \mid x)$ is computable
- VI does not estimate fixed-model parameters (classical params are treated as latent variables)

3-3. MFVI

Assumption : independency between latent variables

$\rightarrow q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$

Each latent variable z_j is governed by its own variational factor, $q_j(z_j)$

(these variational factors are chosen to MAXIMIZE ELBO)

- ex) choose as Gaussian factor, or categorical factor

Researchers have also studied more complex families

- 1) Structured VI
- 2) Mixture of variational densities

$\rightarrow$ both improve the fidelity of the approximation

(trade-off : dificult to solve variational optimization problem)

Bayesian Mixture of Gaussians (cont)

MFVI : $q(\mu, \mathbf{c}) = \prod_{k=1}^{K} q\left(\mu_k; m_k, s_k^2\right) \prod_{i=1}^{n} q\left(c_i; \varphi_i\right)$.

- $q\left(\mu_k; m_k, s_k^2\right)$: Gaussian distn
- $q\left(c_i; \varphi_i\right)$: its assignment probabilities are a K -vector φ_i

Summary : ELBO is defined by..

- model definition : $p(\mu, \mathbf{c}, \mathbf{x}) = p(\mu) \prod_{i=1}^{n} p\left(c_i\right) p\left(x_i \mid c_i, \mu\right)$
- MFVI : $q(\mu, \mathbf{c}) = \prod_{k=1}^{K} q\left(\mu_k; m_k, s_k^2\right) \prod_{i=1}^{n} q\left(c_i; \varphi_i\right)$

3-4. CAVI (Coordinate ascent MFVI)

CAVI : **iteratively** optimizes each factor of the MF variational density

optimal solution

- conditional : $q_j^*\left(z_j\right) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p\left(z_j \mid \mathbf{z}_{-j}, \mathbf{x}\right)\right]\right\}$.
- joint : $q_j^*\left(z_j\right) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p\left(z_j, \mathbf{z}_{-j}, \mathbf{x}\right)\right]\right\}$.

  (expectation on RHS do not involve j^{th} variational factor $\rightarrow$ valid coordinate update)

Algorithm 1: Coordinate ascent variational inference (CAVI)

Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$

Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$

Initialize: Variational factors $q_j(z_j)$

while *the* ELBO *has not converged* **do**

    for $j \in \{1, \ldots, m\}$ **do**

        Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$

    end

    Compute ELBO$(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})]$

end

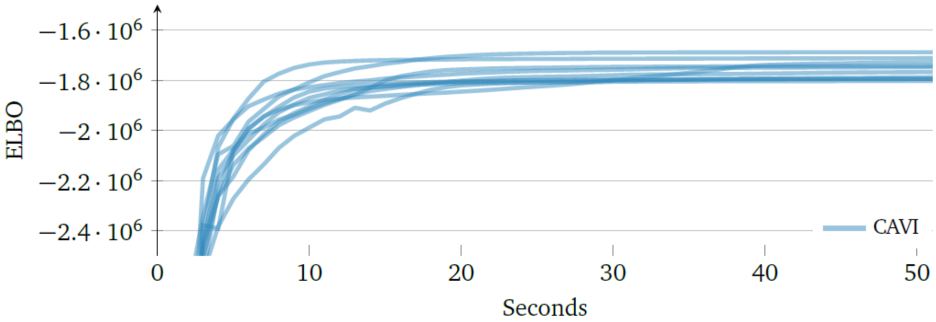
return $q(\mathbf{z})$

(can see that it is closely related to Gibbs sampling)

3-5. Practicalities

Initialization

- ELBO : (usually) non-convex objectiv function
- CAVI only guarantees local optimum (sensitive to initialization)



/

- 10 random initialization, reaching different values!

  (many local optima in ELBO)

- Not always bad!
  - ex) Mixture of Gaussian : many posterior modes
  - exploring latent clusters, predicting new observation

Assessing convergence

- computing the ELBO of full dataset may be undesirable
- proxy! average log predictive of a small held-out dataset

Numerical stability

- probabilities should be between 0~1
- use **log-sum-exp** trick

  $\log[\sum_i \exp(x_i)] = \alpha + \log[\sum_i \exp(x_i - \alpha)].$

4. A complete example : Bayesian Mixture of Gaussians

notation

- K real valued mean params : $\mu = \mu_{1:K}$
- n latent class assignments : $\mathbf{c} = c_{1:n}$

ELBO for mixture of Gaussians

- variational parameters : $\mathbf{m}, \mathbf{s}^2, \varphi$

$$\begin{aligned}\mathrm{ELBO}\left(\mathbf{m}, \mathbf{s}^2, \varphi\right) = & \sum_{k=1}^{K} \mathbb{E}\left[\log p\left(\mu_k\right) ; m_k, s_k^2\right] \\ & + \sum_{i=1}^{n}\left(\mathbb{E}\left[\log p\left(c_i\right) ; \varphi_i\right] + \mathbb{E}\left[\log p\left(x_i \mid c_i, \mu\right) ; \varphi_i, \mathbf{m}, \mathbf{s}^2\right]\right) \\ & - \sum_{i=1}^{n} \mathbb{E}\left[\log q\left(c_i ; \varphi_i\right)\right] - \sum_{k=1}^{K} \mathbb{E}\left[\log q\left(\mu_k ; m_k, s_k^2\right)\right]\end{aligned}$$

CAVI updates each variational parameters in turn

4-1. (step 1)The variational density of the "mixture assignments"

(review) optimal solution : $q_j^*\left(z_j\right) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p\left(z_j, \mathbf{z}_{-j}, \mathbf{x}\right)\right]\right\}$

(1) derive variational update for c_i (cluster assignment)

- $q^*\left(c_i ; \varphi_i\right) \propto \exp\left\{\log p\left(c_i\right) + \mathbb{E}\left[\log p\left(x_i \mid c_i, \mu\right) ; \mathbf{m}, \mathbf{s}^2\right]\right\}.$
    - 1st term) log prior of c_i : $\log p\left(c_i\right) = -\log K$
    - 2nd term) expected log of the c_ith Gaussian density
        - $p\left(x_i \mid c_i, \mu\right) = \prod_{k=1}^{K} p(x_i \mid \mu_k)^{c_{ik}}$
        - Thus,

$$\begin{aligned}\mathbb{E}\left[\log p\left(x_i \mid c_i, \mu\right)\right] & = \sum_k c_{ik} \mathbb{E}\left[\log p\left(x_i \mid \mu_k\right) ; m_k, s_k^2\right] \\ & = \sum_k c_{ik} \mathbb{E}\left[-\left(x_i - \mu_k\right)^2 / 2 ; m_k, s_k^2\right] + \text{ const.} \\ & = \sum_k c_{ik}\left(\mathbb{E}\left[\mu_k ; m_k, s_k^2\right] x_i - \mathbb{E}\left[\mu_k^2 ; m_k, s_k^2\right] / 2\right) + \text{ const.}\end{aligned}$$

- result : $\varphi_{ik} \propto \exp\left\{\mathbb{E}\left[\mu_k ; m_k, s_k^2\right] x_i - \mathbb{E}\left[\mu_k^2 ; m_k, s_k^2\right] / 2\right\}$

4-2. (step 2)The variational density of the "mixture-component means"

(2) variational density $q\left(\mu_k ; m_k, s_k^2\right)$ of k^{th} mixture components

$q\left(\mu_k\right) \propto \exp\left\{\log p\left(\mu_k\right) + \sum_{i=1}^{n} \mathbb{E}\left[\log p\left(x_i \mid c_i, \mu\right) ; \varphi_i, \mathbf{m}_{-k}, \mathbf{s}_{-k}^2\right]\right\}.$

Unnormallized log of $q\left(\mu_k\right)$:

$$
\begin{aligned}
\log q(\mu_k) &= \log p(\mu_k) + \sum_i \mathbb{E}\left[\log p(x_i \mid c_i, \mu); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}_{-k}^2\right] + \text{const.} \\
&= \log p(\mu_k) + \sum_i \mathbb{E}\left[c_{ik} \log p(x_i \mid \mu_k); \varphi_i\right] + \text{ const.} \\
&= -\mu_k^2/2\sigma^2 + \sum_i \mathbb{E}\left[c_{ik}; \varphi_i\right] \log p(x_i \mid \mu_k) + \text{ const.} \\
&= -\mu_k^2/2\sigma^2 + \sum_i \varphi_{ik}\left(-(x_i - \mu_k)^2/2\right) + \text{const.} \\
&= -\mu_k^2/2\sigma^2 + \sum_i \varphi_{ik} x_i \mu_k - \varphi_{ik}\mu_k^2/2 + \text{const.} \\
&= \left(\sum_i \varphi_{ik} x_i\right)\mu_k - \left(1/2\sigma^2 + \sum_i \varphi_{ik}/2\right)\mu_k^2 + \text{const.}
\end{aligned}
$$

where $\varphi_{ik} = \mathbb{E}\left[c_{ik}; \varphi_i\right]$.

Thus, $m_k = \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}, \quad s_k^2 = \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}$

4-3. CAVI for the mixture of Gaussians

Algorithm 2: CAVI for a Gaussian mixture model

Input: Data $x_{1:n}$, number of components K, prior variance of component means σ^2

Output: Variational densities $q(\mu_k; m_k, s_k^2)$ (Gaussian) and $q(c_i; \varphi_i)$ (K-categorical)

Initialize: Variational parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s}^2 = s_{1:K}^2$, and $\boldsymbol{\varphi} = \varphi_{1:n}$

while *the* ELBO *has not converged* **do**

  for $i \in \{1, \ldots, n\}$ **do**

    Set $\varphi_{ik} \propto \exp\{\mathbb{E}[\mu_k; m_k, s_k^2] x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2\}$

  end

  for $k \in \{1, \ldots, K\}$ **do**

$$\text{Set } m_k \longleftarrow \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

$$\text{Set } s_k^2 \longleftarrow \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

  end

  Compute ELBO$(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$

end

return $q(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$

Approximate predictive : (mixture of Gaussians)

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

5. VI with exponential families

(Until now....)

- **MFVI**
- **CAVI** (coordinate ascent algorithm for optimizing ELBO)

- demonstration using **simple mixture of Gaussians**

  (available in closed-form)

Now, will work with **exponential family**

- working with this simplifies VI
- easier to derive CAVI
- section 5-1) general case

  section 5-2) conditionally conjugate models

  section 5-3) SVI

5-1. Complete conditionals in the exponential family

suppose **complete conditional** is "exponential family" :

$$p\left(z_{j} \mid \mathbf{z}_{-j}, \mathbf{x}\right)=h\left(z_{j}\right) \exp \left\{\eta_{j}\left(\mathbf{z}_{-j}, \mathbf{x}\right)^{\top} z_{j}-a\left(\eta_{j}\left(\mathbf{z}_{-j}, \mathbf{x}\right)\right)\right\}$$

CAVI with MFVI

- coordinate update of $q_{j}^{*}\left(z_{j}\right) \propto \exp \left\{\mathbb{E}_{-j}\left[\log p\left(z_{j} \mid \mathbf{z}_{-j}, \mathbf{x}\right)\right]\right\}$ =

  $$\begin{aligned} q\left(z_{j}\right) \propto & \exp \left\{\mathbb{E}\left[\log p\left(z_{j} \mid \mathbf{z}_{-j}, \mathbf{x}\right)\right]\right\} \\ & =\exp \left\{\mathbb{E}\left[\log h\left(z_{j}\right) \exp \left\{\eta_{j}\left(\mathbf{z}_{-j}, \mathbf{x}\right)^{\top} z_{j}-a\left(\eta_{j}\left(\mathbf{z}_{-j}, \mathbf{x}\right)\right)\right\}\right]\right\} \\ & =\exp \left\{\log h\left(z_{j}\right)+\mathbb{E}\left[\eta_{j}\left(\mathbf{z}_{-j}, \mathbf{x}\right)\right]^{\top} z_{j}-\mathbb{E}\left[a\left(\eta_{j}\left(\mathbf{z}_{-j}, \mathbf{x}\right)\right)\right]\right\} \\ & \propto h\left(z_{j}\right) \exp \left\{\mathbb{E}\left[\eta_{j}\left(\mathbf{z}_{-j}, \mathbf{x}\right)\right]^{\top} z_{j}\right\} \end{aligned} .$$

- set $v_{j}=\mathbb{E}\left[\eta_{j}\left(\mathbf{z}_{-j}, \mathbf{x}\right)\right]$

5-2. Conditional conjugacy and Bayesian models

Special case of exponential family : "conditional conjugate models" with local & global variables

Conditionally conjugate models

- β : global latent variables
- $\mathbf{z}$: local latent variables
- joint pdf : $p(\beta, \mathbf{z}, \mathbf{x})=p(\beta) \prod_{i=1}^{n} p\left(z_{i}, x_{i} \mid \beta\right)$

Assumption 1) joint density of each $\left(x_{i}, z_{i}\right)$ pair, conditional on β = exponential family

$p\left(z_{i}, x_{i} \mid \beta\right)=h\left(z_{i}, x_{i}\right) \exp \left\{\beta^{\top} t\left(z_{i}, x_{i}\right)-a(\beta)\right\}$.........................(a)

Assumption 2) prior (on global variables) to be conjugate prior

$p(\beta)=h(\beta)\exp\left\{\alpha^{\top}[\beta,-a(\beta)]-a(\alpha)\right\}$ .... (b)

- natural (hyper) parameter $\alpha=\left[\alpha_1, \alpha_2\right]^{\top}$

(a) and (b) : conjugate $\rightarrow$ $\hat{\alpha}=\left[\alpha_1+\sum_{i=1}^n t\left(z_i, x_i\right), \alpha_2+n\right]^{\top}$

Complete conditional of local variable z_i :

- (given β and x_i) z_i is conditionally independent!

  $\rightarrow \quad p\left(z_i \mid x_i, \beta, \mathbf{z}_{-i}, \mathbf{x}_{-i}\right)=p\left(z_i \mid x_i, \beta\right)$
- assumption) exponential family

  $\rightarrow \quad p\left(z_i \mid x_i, \beta\right)=h\left(z_i\right) \exp \left\{\eta(\beta, x_i)^{\top} z_i-a\left(\eta\left(\beta, x_i\right)\right)\right\}$

VI in conditionally conjugate models

describe CAVI for general class of models

notation

- λ : **global** variational parameter

  $q(\beta \mid \lambda)$: variational posterior approximation on β
- ϕ : **local** variational parameter

  $q(z_i \mid \phi)$: variational posterior on each local variable z_i

Local variational update : $\varphi_i=\mathbb{E}_\lambda\left[\eta\left(\beta, x_i\right)\right]$ .... by

- (1) $v_j=\mathbb{E}\left[\eta_j\left(\mathbf{z}_{-j}, \mathbf{x}\right)\right]$
- (2) $p\left(z_i \mid x_i, \beta, \mathbf{z}_{-i}, \mathbf{x}_{-i}\right)=p\left(z_i \mid x_i, \beta\right)$

Global variational update : $\lambda=\left[\alpha_1+\sum_{i=1}^n \mathbb{E}_{\varphi_i}\left[t\left(z_i, x_i\right)\right], \alpha_2+n\right]^{\top}$

- expectation of $\hat{\alpha}=\left[\alpha_1+\sum_{i=1}^n t\left(z_i, x_i\right), \alpha_2+n\right]^{\top}$

CAVI optimizes the ELBO by iterating "local updates" and "global updates"

To assess convergence, compute ELBO at each iteration!

ELBO :

- $\operatorname{ELBO}(q)=\mathbb{E}[\log p(\mathbf{z}, \mathbf{x})]-\mathbb{E}[\log q(\mathbf{z})]$

  $p(\beta, \mathbf{z}, \mathbf{x})=p(\beta) \prod_{i=1}^n p\left(z_i, x_i \mid \beta\right)$
- Therefore, $\operatorname{ELBO}=\left(\alpha_1+\sum_{i=1}^n \mathbb{E}_{\varphi_i}\left[t\left(z_i, x_i\right)\right]\right)^{\top} \mathbb{E}_\lambda[\beta]-\left(\alpha_2+n\right) \mathbb{E}_\lambda[a(\beta)]-\mathbb{E}[\log q(\beta, \mathbf{z})]$
  - $\mathbb{E}[\log q(\beta, \mathbf{z})]=\lambda^{\top} \mathbb{E}_\lambda[t(\beta)]-a(\lambda)+\sum_{i=1}^n \varphi_i^{\top} \mathbb{E}_{\varphi_i}\left[z_i\right]-a\left(\varphi_i\right)$

5-3. SVI (Stochastic Variational Inference)

Modern problems : require analyizing massive data

but, most do not easily scale (ex. CAVI)

CAVI, not scalable!

- requires iteration through entire data at each iteration
- alternative : **gradient-based optimization**

  (= Key to SVI)

SVI focuses on optimizing the global variational params λ of conditionally conjugate model

- step 1) subsample data
- step 2) use current global param to compute **optimal local params** for the subsampled data
- step 3) adjust the current **global params**

Natural gradient of ELBO

SVI focuses on optimizing the global variational params λ

Euclidan gradient of ELBO :

- $\nabla_\lambda \text{ELBO} = a''(\lambda)\left(\mathbb{E}_{\varphi}[\hat{\alpha}] - \lambda\right)$ (Hoffman et al. 2013)

Natural gradient : (premultiply by inverse Fisher info)

- $g(\lambda) = \mathbb{E}_{\varphi}[\hat{\alpha}] - \lambda$

Update param, using **natural gradient** in gradient-based optimization method

- at each iteration, update $\lambda_t = \lambda_{t-1} + \epsilon_t g\left(\lambda_{t-1}\right)$

  (= $\lambda_t = \left(1 - \epsilon_t\right)\lambda_{t-1} + \epsilon_t \mathbb{E}_{\varphi}[\hat{\alpha}]$)

Stochastic Optimization of the ELBO

goal : construct a cheaply computed, noisy, unbiiased natural gradient

- $g(\lambda) = \mathbb{E}_{\varphi}[\hat{\alpha}] - \lambda$
- $\hat{\alpha} = \left[\alpha_1 + \sum_{i=1}^n t\left(z_i, x_i\right), \alpha_2 + n\right]^\top$

$\rightarrow g(\lambda) = \alpha + \left[\sum_{i=1}^n \mathbb{E}_{\varphi_i^*}\left[t\left(z_i, x_i\right)\right], n\right]^\top - \lambda$

(φ_i^* : optimized local variational param)

Construct noisy natural gradient, by **sampling an index from the data** & rescaling

- $t \sim \text{Unif}(1, \ldots, n)$

$\hat{g}(\lambda) = \alpha + n\left[\mathbb{E}_{\varphi_t^*}\left[t\left(z_t, x_t\right)\right], 1\right]^\top - \lambda$

- $\hat{g}(\lambda)$ is unbiased ($\mathbb{E}_t[\hat{g}(\lambda)] = g(\lambda)$) & ceheap to compute

  (not only sample 1 data, but can also use mini-batch)

Step-size sequence

- should follow conditions of Robbins and Monro,

  $\sum_t \epsilon_t = \infty; \sum_t \epsilon_t^2 < \infty$

Algorithm 3: SVI for conditionally conjugate models

Input: Model $p(\mathbf{x}, \mathbf{z})$, data $\mathbf{x}$, and step size sequence ϵ_t

Output: Global variational densities $q_\lambda(\beta)$

Initialize: Variational parameters λ_0

while *TRUE* **do**

  Choose a data point uniformly at random, $t \sim \text{Unif}(1, \ldots, n)$

  Optimize its local variational parameters $\varphi_t^* = \mathbb{E}_\lambda[\eta(\beta, x_t)]$

  Compute the coordinate update as though x_t was repeated n times,

$$\hat{\lambda} = \alpha + n\mathbb{E}_{\varphi_t^*}[f(z_t, x_t)]$$

  Update the global variational parameter, $\lambda_t = (1 - \epsilon_t)\lambda_t + \epsilon_t \hat{\lambda}_t$

end

return λ

6. Discussion

Summary

- MFVI
- CAVI
- Bayesian Mixture of Gaussians
- special case of exponential families & conditional conjugacy
- SVI