Variational Inference - 2021년 P-sat 논문스터디 4회차

김원구, 오정민, 유경민

2021년 2월 16일

'Variational Inference: A Review for Statisticians' 논문을 바탕으로 Varitaional Inference에 대한 전반적인 내용을 소개하는 자료입니다.

1. Introduction

1.1 등장배경

- 베이즈 통계의 주된 목적중 하나는 사후분포를 구하는 것이다. 베이지안 추론과 관련하여 사후분포를 근사하는 대표적인 방법에는 MCMC(Markov chain Monte Carlo) sampling 방법이 있다. MCMC방법에 대해 다루고 싶지만 내용이 상당하기 때문에 따라 설명하지는 않겠다(양해부탁)!
- 해당 논문에 따르면 Variational Inferenc는 MCMC의 어떤 대안 느낌으로 베이지안 모델들에 있어 posterior density들을 근사하는데 사용된다. MCMC와 비교하여 변분추론은 large data에 대해서 더 빠르고 쉽게 적용할수 있는 경향이 있다고 한다. 결국 복잡한 모델을 다루거나 large data를 사용하는데 있어 MCMC가 가지는 단점들을 보완하는 접근법으로 등장한 개념이라고 생각해주면 될 것 같다.

1.2 Variational Inference

- 변분 추론이란 결국 간단하게 정리하자면 사후확률(posterior)분포 $p(z|x)$를 다루기 쉬운 확률분포 $q(z)$로 근사(approximation)하는 것을 말한다. 여기서 근사를 하는 이유는 사후확률 분포를 계산하는게 불가능에 가까울 정도로 어려운 경우가 많기 때문이다. 따라서 q(z)에 대해 더 제한된 종류의 분포들만 고려할 것이며, 이들 중 쿨백 라이블러 발산이 최소가 되는 분포를 찾을 것이다. 분포의 종류를 충분히 제한함으로써 다루는 것이 가능한 분포들만을 남기는 동시에, 분포의 종류가 충분히 크고 유연해서 실제 사후분포에 대해 충분히 좋은 근삿값을 제공할 수 있도록 하는 것이다.
- 변분 추론의 메인 아이디어는 optimization이다. 변분적 방법론 자체에는 본질적으로 근사하는 성질이 없다고 하는데, 최적화가 시행되는 함수들의 범위를 제한하는 방식으로 이루어져 자연스럽게 근사 해를 찾는 방향으로 나아가게 된다고 한다. 위에서 언급한 것처럼 이 과정에서 KL-divergence를 사용하게 되는데, 식은 다음과 같다.

$$q^*(\mathsf{z}) = \underset{q(\mathsf{z})\in\mathcal{Q}}{\operatorname{argmin}}\, KL(q(\mathsf{z}) \,||\, p(\mathsf{z}|\mathsf{x}))$$

우리는 KL divergence값을 최소로 만들어주는 $q^(\mathsf{z})$를 찾는 것이 목적이다. 여기서 $\mathcal{Q}$는 위해서 말한 제한된 종류의 분포들이다. (EM알고리즘 때도 언급을 하였지만 KL-divergence는 두 분포가 얼마나 닮았는지를 비교하는 척도로 사용된다.)

2. Variational Inference

2.1 Evidence Low BOund (ELBO)

확률 분포 $p(x)$가 잠재 변수 z로 대부분 설명될 수 있다고 하면, 그 때의 z에 대한 사후분포 $p(z|x)$를 알아야 한다. 그런데 Posterior는 대부분 계산이 closed form이 아니거나 시간복잡도가 지수 시간($O(2^n)$ 등)이기에 구하는 것이 불가능하다. 따라서 Posterior $p(z|x)$를 다루기 쉬운 분포 $q(z)$로 바꿔 근사하는 것이 변분추론이다.

Notation $q(z)$: Variational density. $q(z) \in Q$.
$p(x)$: Evidence
$p(z)$: Prior
$p(x|z)$: Likelihood $p(z|x)$: Posterior

$p(x)$와 $q(x)$에 대한 KL Divergence

$$\begin{aligned} KL\left(q\left(z\right)||p\left(z|x\right)\right) &= \int q\left(z\right) \log \frac{q\left(z\right)}{p\left(z|x\right)} dz \\ &= \int q\left(z\right) \log \frac{q\left(z\right) p\left(x\right)}{p\left(x|z\right) p\left(z\right)} dz \\ &= \int q\left(z\right) \log \frac{q\left(z\right)}{p\left(z\right)} dz + \int q\left(z\right) \log p\left(x\right) dz - \int q\left(z\right) \log p(x|z) dz \\ &= KL\left(q\left(z\right)||p\left(z\right)\right) + \log p\left(x\right) - E\left[\log p(x|z)\right] \end{aligned}$$

Posterior p(z|x)와 다루기 쉬운 variational density q(z) 사이의 KLD는 위 마지막 세 개 항으로 분해할 수 있다. 그런데 KLD 값은 0과 1사이, 즉 항상 0보다 크기 때문에 아래처럼 표현할 수 있다.

$$0 \leq KL(q(z)||p(z)) + logp(x) - E[logp(x|z)]$$

$$logp(x) \geq E[logp(x|z)] - KL(q(z)||p(z))$$

위 식에서 right-hand side를 ELBO라고 하며, Evidence의 하한이 된다. 이를 다시 정리하면,

$$KL(q(z)||p(z|x)) = KL(q(z)||p(z)) + logp(x) - E[logp(x|z)]$$

$$logp(x) = ELBO + KL(q(z)||p(z|x))$$

위 식을 보면, Evidence는 ELBO와 KLD의 합임을 알 수 있다. 이 식의 의미를 살펴보면, KLD를 최소화하는 것은 곧 ELBO를 최대화하는 것과 의미가 같다.
다만 ELBO와 KLD 모두 함수 $q(\cdot)$에 dependent하기에, 단순히 한 쪽을 최소화하는 $q(\cdot)$이 반드시 Evidence의 값을 최소화한다고 말하기는 어렵다.

2.2 Mean field variational family

VI에서는 $q(z)$가 mean-field variational family임을 전제한다. 즉, $z_j, j = 1, ..., m$인 잠재 변수 z_j가 independent할 때, 전체 $q(z)$ 또한 각 부분집합의 $q(z_i)$로 factorization 된다고 가정한다.
따라서, $q(z)$는 다음과 같이 분해될 수 있다.

$$q(\mathbf{z}) = \prod_j q_j(z_j)$$

2.3 Coordinate Ascent mean-field Variational Inference (CAVI)

위 처럼 Variational 분포 $q(z)$를 각각의 곱으로 factorization하면, 각 factor에 대해 Coordinate Ascent Optimization을 적용해 VI를 수행할 수 있게 된다.
CAVI는 한 쪽을 고정한 채, $q(z)$의 각 factor를 반복적으로 최적화하는데, 결과는 ELBO의 Local Optimum으로 이끈다.

CAVI의 핵심 아이디어는 $q(\cdot)$ 함수가 분해될 수 있다는 사실을 이용하는 것이다. 그리고 우리는 x와 j번째 잠재 변수 z_{-j}가 주어졌을 때의 z_j의 조건부 확률을 알아야 한다. 이는 아래와 같이 표현할 수 있다.

$$logp(z_j|\mathbf{z}_{-j}, \mathbf{x})$$

그런데 앞서 했던 mean-field varational family 가정에 따라 모든 잠재 변수는 독립이다. 따라서 위 식은 다시 아래처럼 바뀔 수 있다.

$$logp(z_j, \mathbf{z}_{-j}, \mathbf{x})$$

이를 바탕으로 ELBO 식을 q_j의 관점에서 풀어쓴다면, 아래와 같다. (l은 j번째 잠재 변수가 아닌 나머지 variational factors: $q_l(z_l)$를 의미)

$$\begin{aligned} ELBO = & \quad E[logp(\mathbf{x}|z)] - KL(q(z)||p(z|x)) \\ = & \int_z q(z)logp(\mathbf{x}, z) - q(z)logq(z)dz \quad \because assumption \\ = & \quad E[logp(\mathbf{x}, z)] - E[logq(z)] \\ = & \quad E_q[logp(\mathbf{x}, z_j, \mathbf{z}_{-j})] - E_{q_l}[logq_l(\mathbf{z}_l)] \end{aligned}$$

Using iterated expectation that $E[X] = E[E[X|Y]]$,

$$E_j[E_{-j}[logp(\mathbf{x}, z_j, \mathbf{z}_{-j})|z_j]] - E_{q_j}[logq_j] + Const$$

첫 항의 안쪽 부분은 기댓값의 정의에 따라 다음과 같이 식을 전개할 수 있다.

$$\begin{aligned} E_{-j}[logp(\mathbf{x}, z_j, \mathbf{z}_{-j})|z_j] = & \int_{-j} logp(\mathbf{x}, z_j, \mathbf{z}_{-j})q(\mathbf{z}_{-j}|z_j)dq_{-j} \\ = & \int_{-j} logp(\mathbf{x}, z_j, \mathbf{z}_{-j})q(\mathbf{z}_{-j})dq_{-j} \\ = & \quad E_{-j}[logp(\mathbf{x}, z_j, z_{-j})] \end{aligned}$$

최종적으로 q_j에 대한 ELBO는 아래와 같다.

$$ELBO = E_j[E_{-j}[logp(\mathbf{x}, z_j, \mathbf{z}_{-j})]] - E_j[logq_j] + Const$$

우변 첫 번재 항을 최대로 하는 것이 q_j에 대한 ELBO를 최대화하는 길이다. 따라서, q_j에 대한 Optimal Solution은 아래와 같이 표현할 수 있다.

$$q_j{}^* z_j \propto exp(E_{-j}[logp(\mathbf{x}, z_j, \mathbf{z}_{-j})])$$

KLD의 형태를 고려할 때,

$$KL(Q(x)||P(x)) = E_{X\sim P}[-log\frac{Q(x)}{P(x)}]$$

q_j에 대한 ELBO 식은 $q_j{}^* z_j$와 $q_j(z_j)$ 사이의 Negative KLD 값을 의미한다. 따라서 j번째 잠재변수의 분포를 j번째 잠재변수의 최적화된 잠재 분포와 유사하게 만드는 것이 q_j의 ELBO를 최대화하는 것이다. 그러므로, 이러한 과정을 모든 j에 대해 ELBO가 수렴할 때까지 반복한다면 우리가 원하는 $q(\cdot)$ 를 얻을 수 있다.

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 $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] + \mathbb{E}[\log q(\mathbf{z})]$
end
return $q(\mathbf{z})$

CAVI의 일반적인 절차는 아래와 같다.
1) Variational 분포 q를 초기화
2) 각 q_j를 최적화 (Biship(2006)에서는 각 잠재 변수의 Gradient를 잡아 최적화, 본 논문에서는 iterated expectation 사용)
3) ELBO를 계산
4) ELBO가 수렴할 때 까지 위 과정을 반복

CAVI는 고전적이고 좋은 VI 방법론이지만, Non-Convex optim. prob.에서 Global Optimum에 도달할 것이라고 보장하지 못한다. 즉, 충분히 KLD 값을 최소화하지 못할 수도 있다.
또한 MCMC 같은 Posterior Estimation 보다는 (최적화 방법이기에) 속도가 상대적으로 빠르지만 한 쪽을 고정하고 다른 쪽을 교대로 계산하기 때문에 속도가 아주 빠르지는 않다.

3. Gaussian Mixture Models 적용

위에서 변분추론에 대한 전반적인 과정들을 설명해 주었는데, 이번에는 예시를 통해 그 과정에 대해서 보는 시간을 가져보도록하자. Gaussian Mixture Models을 가지고 진행을 할것이다.

우리가 예시로 사용할 분포는 다음과 같다.

$$\begin{aligned}
&\mu_k \sim \mathcal{N}(0, \sigma^2) \quad k = 1, ..., K \\
&c_i \sim categorical(1/K, ..., 1/k) \quad i = 1, ..., n \\
&x_i | c_i, \mu \sim \mathcal{N}(c_i^T \mu, 1) \quad i = 1, ..., n
\end{aligned}$$

여기서 우리의 잠재변수들은 $\{\mathbf{c}, \mu\}$가 된다.
잠재변수와 관측변수의 joint density는 다음과 같다.

$$p(\mu, \mathbf{c}, \mathbf{x}) = p(\mu) \prod_{i=1}^{n} p(c_i) p(x_i | c_i, \mu)$$

위에서 언급했던 것처럼, 우리가 근사에 사용하는 $q(z)$의 형태는 다음과 같다.

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

$q(\mu_k; m_k, s_k^2)$는 kth mixture component로 Gaussian 분포를 따르며 m_k,s_k^2가 variational parameter로 존재하고, $q(c_i; \varphi_i)$는 ith 관측치의 mixture assignment로 categorical 분포를 따르고 φ_i가 variational parameter로 존재한다.

- 우리가 이번예시에서 구하고자하는 ELBO는 다음과 같다.

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

이제 CAVI 알고리즘을 통해 각각의 variational parameter를 업데이트 하는 과정을 거치게 될 것이다.

3.1 Variational update for Mixture assignment

먼저 cluster assignment인 c_i를 업데이트 할것이다.

위에서 언급했던 $q^*(z_j)$에 대한 식을 적용하면 다음과 같이 표현을 할 수 있다.

$$q^*(c_i; \varphi_i) \propto exp\{log\, p(c_i) + E[log\, p(x_i|c_i; \mu); \mathbf{m}, \mathbf{s^2}]\}$$

여기서 첫 번째 term은 prior of c_i가 되는데, 모든 c_i에 대해서 확률이 같기 때문에 첫 번째 term은 다음과 같이 표현이 가능하다.

$$log\, p(c_i) = log\frac{1}{K} = log1 - logK = -logK$$

두 번째 term은 c_i번째 Gaussian density의 로그의 Expectation이다. 여기서 c_i는 indicator vector기 때문에 우리는 다음과 같이 표현이 가능하다.

$$p(x_i|c_i, \mu) = \prod_{k=1}^{K} p(x_i|\mu_k)^{c_{ik}} \quad c_{ik} = \{0, 1\}$$

그래서 이제 두 번째 term을 전개해보면 다음과 같이 나타낼수 있다.

$$
\begin{aligned}
&E[log\, p(x_i|c_i; \mu); \mathbf{m}, \mathbf{s^2}] \\
&= \sum_k c_{ik}\, E[log\, p(x_i|\mu_k); m_k, s_k^2] \\
&= \sum_k c_{ik}\, E[-(x_i - \mu_k)^2/2; m_k, s_k^2] + const. \\
&= \sum_k c_{ik} (E[\mu_k; m_k, s_k^2] x_i - E[\mu_k^2; m_k, s_k^2]/2) + const.
\end{aligned}
$$

결국 이 계산 식은 $E[\mu_k]$와 $E[\mu_k^2]$를 필요로 한다. 그래서 결국 정리를 해보면 다음과 같다.

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

3.2 Variational update for Mixture-Component mean

다음으로는 Mixture-Component mean인 μ_k에 대해서 업데이트를 할 것이다.
마찬가지로 위에서 언급했던 $q^*(z_j)$에 대한 식을 적용하면 다음과 같이 표현을 할 수 있다.

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

해당 식을 전개해보면 다음과 같이 나타낼수 있다.

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

위와 같은 과정에서 $q(\mu_k)$를 업데이트 하는데 필요한 m_k와 s_k^2를 구할 수 있는데, 다음과 같이 표현된다.

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

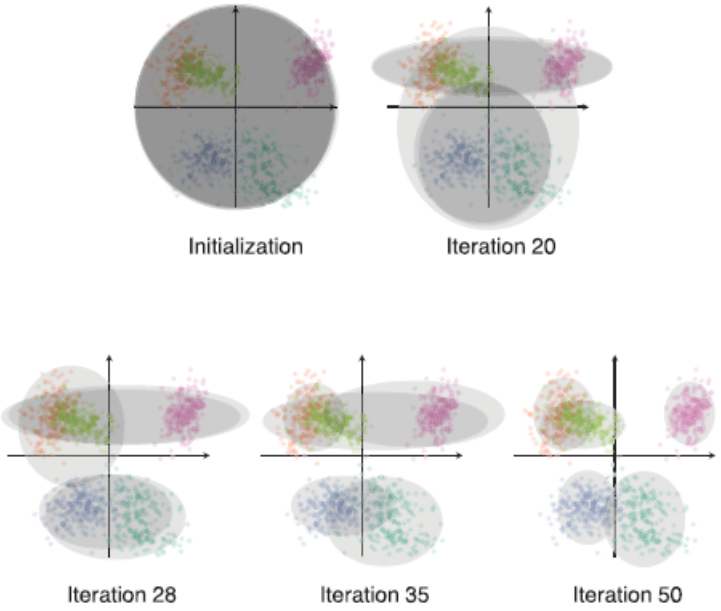
M-step

위 과정을 정리해보면 다음과 같다.

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 $q(z_i; \varphi_i)$ (K-categorical)
Initialize: Variational parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s}^2 = s^2_{1:K}$, 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**
    Set $m_k \longleftarrow \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}$
    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 $q(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$

Gaussian 모델에 대한 CAVI 알고리즘 진행 과정을 그림으로 표현하면 다음과 같다.



4. Variational Infernce with Exponential Family

4.1 General Case

각각의 완전 조건부 분포가 지수족에 해당하는 $p(\mathbf{z}, \mathbf{x})$가 있다고 가정하자.

$$p(z_j|\mathbf{z}_{-j}, \mathbf{x}) = h(z_j)\,\exp\{\eta_j(\mathbf{z}_{-j}, \mathbf{x})^T z_j - a(\eta_j(\mathbf{z}_{-j}, \mathbf{x}))\}$$

지수족의 가정은 앞서 다루었던 coordinate update를 간단하게 만든다.

$$\begin{aligned} q(z_j) &\propto \exp\{\mathbb{E}[\log\{p(z_j|\mathbf{z}_{-j}, \mathbf{x})]\} \\ &= \exp\{\log\{h(z_j) + \mathbb{E}[\eta_j(\mathbf{z}_{-j}, \mathbf{x})]^T z_j - \mathbb{E}[a(\eta_j(\mathbf{z}_{-j}, \mathbf{x}))]\} \\ &\propto h(z_j)\,\exp\{\mathbb{E}[\eta_j(\mathbf{z}_{-j}, \mathbf{x})]^T z_j\} \end{aligned}$$

이러한 update는 최적의 variational factor에 대한 parametric form을 보여준다. 여기서 j번째 variational factor에 대한 variational parameter를 ν_j로 나타내면, $\nu_j = \mathbb{E}[\eta_j(\mathbf{z}_{-j}, \mathbf{x})]$로 정리할 수 있다. 이러한 표현 방식은 CAVI 알고리즘을 효율적으로 진행할 수 있도록 한다.

4.2 Conditionally conjugate model

사후 분포 $p(\theta|\mathbf{x})$와 사전 분포 $p(\theta)$의 분포가 동일한 distributional family에 속하는 분포를 켤레 분포(conjugate distribution)이라고 한다. 지수족의 특별한 케이스 중 하나는 local, global variable이 있는 conditionally conjugate model의 경우이다.

β를 global latent variable, z를 각각의 i번째 component와 관련있는 local latent variable이라하자. Gaussian mixture의 예에서 mixture component는 global variable, 각 데이터에 대한 cluster assignment와 관련된 부분을 local variable이라 할 수 있다.

$q(\beta|\lambda)$는 β에 대한 variational approximation이고, λ는 global variational parameter이다.$q(z_i|\varphi_i)$는 각각의 local variable z_i에 대한 variational approximation이고, φ_i는 local variational parameter이다. local, global variable이 있는 경우 CAVI 알고리즘은 각각의 local variational parameter에 대해 업데이트하고 global variational paramter에 대해 업데이트 하는 것을 반복한다. local variable update는 $\varphi_i = \mathbb{E}_\lambda[\eta(\beta, x_i)]$로 이루어지는데, 이는 general case에서 언급한 variational paramter의 형태임을 알 수 있다. 이를 이용하여 global variable의 update는 $\lambda = [\alpha_1 + \sum_{i=1}^{n} \mathbb{E}_{\varphi_i}[t(z_i, x_i)]], \alpha_2 + n]$로 이루어진다.

4.3 Stochastic Variational Inference

CAVI 알고리즘은 매 iteration마다 모든 데이터셋에대해 coordinate ascent를 반복하는데, 이는 매우 계산량이 많다. 따라서 coordinate ascent를 반복하는 대신 ELBO의 gradient를 이용하여 최적화를 진행하는 방법이 제안되었다. 여기서는 확률 파라미터들의 기하학적 구조를 반영하는 natural gradient를 적용하여 ELBO의 gradient를 계산한다. 이를 통해 natural gradient $g(\lambda)$가 다음과 같이 구해진다.

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

이것은 coordinate update $\mathbb{E}_{\varphi_i}[\hat{\alpha}]$와 variational parameter λ의 차이를 의미한다. 이 natural gradient를 이용하여 매 iteration마다 global parameter를 다음과 같이 update한다. ϵ_t는 step size를 의미한다.

$$\lambda_t = \lambda_{t-1} + \epsilon_t \, g(\lambda_{t-1})$$

$g(\lambda_{t-1}) = \mathbb{E}_{\varphi_i}[\hat{\alpha}] - \lambda_{t-1}$을 두번째 term에 대입하여 정리하면 다음과 같다.

$$\begin{aligned} & \lambda_t \\ &= \lambda_{t-1} + \epsilon_t \left(\mathbb{E}_{\varphi_i}[\hat{\alpha}] - \lambda_{t-1}\right) \\ &= (1 - \epsilon_t)\lambda_{t-1} + \mathbb{E}_{\varphi_i}[\hat{\alpha}] \end{aligned}$$

즉, 우리는 매 iteration마다 coordinate update를 계산하고, 현재의 추정치를 coordinate update와 현재의 variational parameter의 가중평균을 내는 방식으로 variational paramter를 업데이트 하게 된다. 이는 계산이 쉽지만, 결국 모든 데이터에대하여 합해줘야 하기에 기존의 CAVI 알고리즘과 동일한 연산량이 요구된다. 이 문제는 데이터를 확률적으로 샘플링하여($t \sim Unif(1, .., n)$) 최적화하는 방식으로 해결된다. SVI의 진행과정을 다음과 같이 정리할 수 있다.

1. 전체 데이터셋에서 subsampling, $t \sim Unif(1, .., n)$
2. 현재의 global parameter를 이용하여 최적의 local parameter 계산
3. coordinate update $\mathbb{E}_{\varphi_i}[\hat{\alpha}]$계산, 이를 이용하여 global parameter update $(1 - \epsilon_t)\lambda_{t-1} + \mathbb{E}_{\varphi_i}[\hat{\alpha}]$

5. Conclusion

5.1 Theoretical guarantees

- mean-field variational posterior parameter는 빈도주의자들의 관점에서도 consistent 했고, asymptotic normality를 보였다.
- stochastic block model의 희박하고 제한된 variants를 추정하는데 다른 ml 접근에 비해 계산적, 이론적 이점들이 있었다.

5.2 Open Problems

- 변분추론의 목표는 사후확률분포 $p(z|x)$를 다루기 쉬운 $q(z)$로 근사하는것이다. 이를 위해 우리는 $p(z|x)$와$q(z)$간의 KL divergence를 최적화하는 것에 집중하였지만, α-divergence같은 새로운 지표로 최적화하려는 연구가 진행중이며, ELBO보다 더 tight한 하한이 발전되어있기도 하다.
- mean-field family는 강한 독립의 가정으로 최적화를 확장시키지만, local optima에 더욱 취약하게 만들고 사후 분포를 저평가하는 원인이 되는 한계가 존재하여 이에 대한 다양한 연구들도 존재한다.
- MCMC와 변분추론의 접점에 대해 다루어지지 않았지만, MCMC에 변분추론의 구조를 포함하고자 하는 시도들이 존재한다. MCMC와 변분추론을 결합하여 사용하는 방법은 큰 영향력이 있을 것으로 생각된다.

- 변분추론의 통계적 특성들에 대한 연구가 MCMC와 비교하여 잘 이루어지지 않았다.