Bayesian Methods in Machine Learning, Seminar: 4,5

Evgenii Egorov, Skoltech, 2020

Approximation Inference for Non-Conjugate Models

Wide range of models of the interest have the posterior p(w) in the following form →

$$p(w) = \frac{1}{Z}\mathcal{N}(w|\mu, \Sigma)\prod_{n=1}^{N}\phi_n(w),$$

$$Z = \int \mathcal{N}(w|\mu, \Sigma)\prod_{n=1}^{N}\phi_n(w)\; dw.$$

Examples of such models:

- Sparse Linear Models:
  Gaussian Likelihood,
  Non-Conjugate Prior for Sparsity
- Logistic Regression (GLM):
  Gaussian Prior,
  Non-Conjugate Likelihood
- … , including generative models (ICA)

How could we approximate posterior?

Approximation Inference for Non-Conjugate Models

Posterior p(w) approximation:

$$p(w) = \frac{1}{Z}\mathcal{N}(w|\mu, \Sigma)\prod_{n=1}^{N} \phi_n(w),$$

$$Z = \int \mathcal{N}(w|\mu, \Sigma) \prod_{n=1}^{N} \phi_n(w) \, dw.$$

1. Laplace approximation
   (see the Lecture material)
2. Gaussian Kullback-Leibler Approximation
   [Edward Challis et. al]
3. Boosting Variational Inference
   [Fangjian Guo et. al]
4. MaxEnt Variational Inference
   [Evgenii Egorov et. al]

- All this methods uses as base family for approximation Gaussian distribution
- (3, 4) approaches allows to obtain approximation as mixture
- We will see a lot of similarities with Laplace Approximation
- Important conceptual distinguish from Laplace:
  - Laplace approximation (1) based on the concentration of the posterior (Ref. for connection with Bernstein-von Mises theorem)
  - Other approaches (2, 3, 4) based on optimization of the variational bound on KL divergence

Gaussian Kullback-Leibler Approximation

We consider an approximation in the family of Gaussian distribution:

$$q(w) = \mathcal{N}(w|\mu, S)$$

In what sense we would like to optimize? Let's use KL divergence:

$$KL[q(w)||p(w)] = \int q(w) \log \frac{q(w)}{p(w)} \, dw,$$

$$KL \geq 0, \;\; \forall q(w),$$

$$KL[q(w)||p(w)] = 0 \text{ iff } p = q.$$

Problem: Using non-negative property of the KL, get lower bound to the evidence of model, Z.

Gaussian Kullback-Leibler Approximation

$$\mathcal{B}_{KL}(\mathbf{m},\mathbf{S}) := \underbrace{-\langle \log q(\mathbf{w})\rangle_{q(\mathbf{w})}}_{\text{entropy}} + \underbrace{\langle \log \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})\rangle_{q(\mathbf{w})}}_{\text{Gaussian potential}} + \underbrace{\sum_{n=1}^{N} \langle \log \phi_n(\mathbf{w})\rangle_{q(\mathbf{w})}}_{\text{site potentials}}, \qquad (4)$$

1. Entropy is closed form and concave
2. Gaussian potential leads to the closed form quadratic form, which is concave
3. The all difficulty lies at the last terms
   a. We need to approximate it
   b. So we need to make some assumption →

$$q(w) = \mathcal{N}(w|\mu, S)$$

$$\langle \phi_n(w)\rangle_{q(w)} = \ldots\ ?$$

$$\text{Assumption: } \phi_n(w) = \phi_n(w^T h_n).$$

Problem:

- **Find the distribution of the w^T h_n (i.e. expectation and variance)**
- **Rewrite expectation of the last term as under one dimensional standard normal distribution**

Gaussian Kullback-Leibler Approximation

$$\mathcal{B}_{KL}(\mathbf{m},\mathbf{S}) = \underbrace{\frac{1}{2}\log\det(2\pi e\mathbf{S})}_{\text{entropy}} + \underbrace{\sum_{n=1}^{N}\langle\log\phi_n(m_n+zs_n)\rangle_{\mathcal{N}(z|0,1)}}_{\text{site projection potentials}}$$
$$\underbrace{-\frac{1}{2}\left[\log\det(2\pi\boldsymbol{\Sigma}) + (\mathbf{m}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{m}-\boldsymbol{\mu}) + \operatorname{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\right]}_{\text{Gaussian potential}}. \quad (7)$$

$$m_n := \mathbf{m}^{\mathsf{T}}\mathbf{h}_n \text{ and } s_n^2 := \mathbf{h}_n^{\mathsf{T}}\mathbf{S}\mathbf{h}_n$$

- This objective contains only 1-dim integrals
    - They could be efficiently estimated:
        - by quadratures or we could take stochastic gradient for optimization
- We should perform optimization for covariance over the Cholesky factor, S = CC^T, so objective:

$$\mathcal{B}_{KL}(\mathbf{m},\mathbf{C}) \stackrel{c.}{=} \sum_{d=1}^{D}\log C_{dd} - \frac{1}{2}\mathbf{m}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m} + \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m} - \frac{1}{2}\operatorname{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{C}\mathbf{C}^{\mathsf{T}}\right) + \langle\log\phi(\mathbf{w}^{\mathsf{T}}\mathbf{h})\rangle. \quad (8)$$

Gaussian Kullback-Leibler Approximation

$$\mathcal{B}_{KL}(\mathbf{m}, \mathbf{C}) \stackrel{\text{c.}}{=} \sum_{d=1}^{D} \log C_{dd} - \frac{1}{2}\mathbf{m}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m} + \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{m} - \frac{1}{2}\operatorname{trace}\left(\boldsymbol{\Sigma}^{-1}\mathbf{C}\mathbf{C}^{\mathsf{T}}\right) + \left\langle \log\phi(\mathbf{w}^{\mathsf{T}}\mathbf{h})\right\rangle . \qquad (8)$$

- The all terms besides the last are clearly concave
- For arbitrary potential phi it is not the case
- So, let's assume that phi is concave

Problem:

- **Prove that last term is concave, given that function ϕ is a concave function**
- **Hint: use definition of the concavity by inequality**

Boosting Variational Inference

1. Laplace or G-KL is good
2. But the approximation power is limited
3. Could we improve it with mixture of approximation of this kind?
4. Yep.

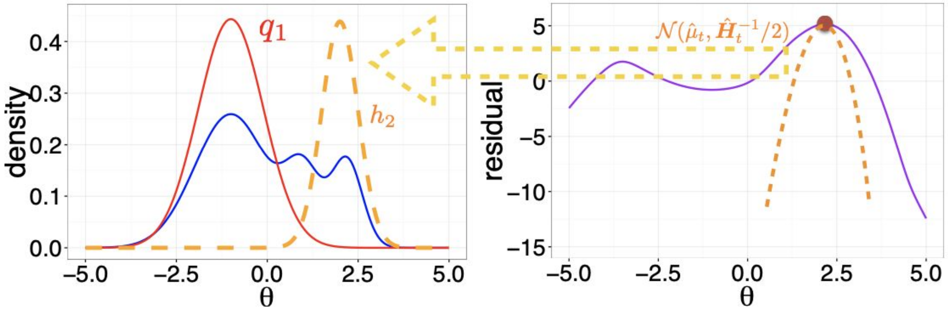

FIGURE 1. Algorithm 2 identifies new component h_2 by finding a (local) peak of the log residual and its corresponding Hessian.

Boosting Variational Inference

Again, we would like to minimize the KL divergence.

$$KL[q(w)\|p(w)] = \int q(w)\log\frac{q(w)}{p(w)}\,dw,$$
$$KL \geq 0,\ \forall q(w),$$
$$KL[q(w)\|p(w)] = 0 \text{ iff } p = q.$$

In each step, we consider to learn new component to the current approximation:

$$q_t = (1-\alpha_t)q_{t-1} + \alpha_t h_t.$$

1. We need to optimized over new component and its weight
    a. It is hard, so let's break the problem on 2 steps: find component, than given it, find the weight
2. We could assume that weight is small, so we could make linearization

$$\begin{aligned}\mathcal{D}((1-\epsilon)q + \epsilon h) &= \mathcal{D}(q + \epsilon\ (h-q)) \\ &= \mathcal{D}(q) + \epsilon\ \langle h - q, g\rangle + o(\epsilon^2),\end{aligned}$$

Problem:

- **Find approximation of the KL divergence, keeping only first order terms of alpha**
- **Hint: just use first term for taylor expansion of the logarithm**

Boosting Variational Inference

Approximation of the KL divergence:

Functional gradient

$$\tilde{\mathcal{D}}_{\mathrm{KL}}(q_t) = \tilde{\mathcal{D}}_{\mathrm{KL}}(q_{t-1}) + \boxed{\alpha_t \langle h_t, \log(q_{t-1}/f)\rangle} - \alpha_t \langle q_{t-1}, \log(q_{t-1}/f)\rangle + o(\alpha_t^2). \quad (14)$$

Make new component the same as negative gradient:

$$h_t = \arg\max_{h_t} \langle h_t, -\log \frac{q_{t-1}}{f} \rangle = \arg\min_{h_t} \langle h_t, \log \frac{q_{t-1}}{f} \rangle$$

Boosting Variational Inference

Approximation of the KL divergence:

$$\tilde{\mathcal{D}}_{\mathrm{KL}}(q_t) = \tilde{\mathcal{D}}_{\mathrm{KL}}(q_{t-1}) + \underbrace{\boxed{\alpha_t \langle h_t, \log(q_{t-1}/f) \rangle}}_{\text{Functional gradient}} - \alpha_t \langle q_{t-1}, \log(q_{t-1}/f) \rangle + o(\alpha_t^2). \quad (14)$$

Make new component the same as negative gradient:

Ill posed problem with degenerate solution

$$h_t = \arg\max_{h_t} \langle h_t, -\log \frac{q_{t-1}}{f} \rangle = \boxed{\arg\min_{h_t} \langle h_t, \log \frac{q_{t-1}}{f} \rangle}$$

Boosting Variational Inference

Approximation of the KL divergence:

Functional gradient

$$\tilde{D}_{\mathrm{KL}}(q_t) = \tilde{D}_{\mathrm{KL}}(q_{t-1}) + \boxed{\alpha_t \langle h_t, \log(q_{t-1}/f) \rangle} - \alpha_t \langle q_{t-1}, \log(q_{t-1}/f) \rangle + o(\alpha_t^2). \quad (14)$$

Make new component the same as negative gradient:

Ok problem, but complex → more approximation

$$h_t = \arg\max_{h_t} \langle h_t, -\log \frac{q_{t-1}}{f} \rangle = \boxed{\arg\min_{h_t} \langle h_t, \log \frac{q_{t-1}}{f} \rangle \quad + \frac{\lambda}{2} \log \|h\|_2^2}$$

Boosting Variational Inference

Approximation of the KL divergence:

Functional gradient

$$\tilde{\mathcal{D}}_{\mathrm{KL}}(q_t) = \tilde{\mathcal{D}}_{\mathrm{KL}}(q_{t-1}) + \boxed{\alpha_t \langle h_t, \log(q_{t-1}/f)\rangle} - \alpha_t \langle q_{t-1}, \log(q_{t-1}/f)\rangle + o(\alpha_t^2). \quad (14)$$

Make new component the same as negative gradient:

Ok problem, but complex → more approximation

$$h_t = \arg\max_{h_t} \langle h_t, -\log \frac{q_{t-1}}{f} \rangle = \boxed{\arg\min_{h_t} \langle h_t, \log \frac{q_{t-1}}{f} \rangle \quad + \frac{\lambda}{2} \log \|h\|_2^2}$$

Ok problem, but complex → more approximation → (local) Laplace style:

$$\log(f(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta})) \approx -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\eta})^T \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\eta}) + \text{const.}$$

$$h_\phi(\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\theta})$$

Boosting Variational Inference

1. Given that component Gaussian, find expression for regularization term: $\log \|h\|_2^2 = \dots\ ?$ $h_\phi(\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\theta})$

2. Find the solution of the problem, using quadratic approximation:

$$h_t = \arg\max_{h_t} \langle h_t, -\log \frac{q_{t-1}}{f} \rangle = \boxed{\arg\min_{h_t} \langle h_t, \log \frac{q_{t-1}}{f} \rangle + \frac{\lambda}{2} \log \|h\|_2^2}$$

$$\log(f(\boldsymbol{\theta})/q_{t-1}(\boldsymbol{\theta})) \approx -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\eta})^T \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\eta}) + \text{const.}$$

$$h_\phi(\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\theta})$$