Bayesian multimodeling: variational inference

MIPT

2021

Variational calculus

Variational calculus problem is to find maxima and minima of functionals: mappings from a set of functions to the real numbers.

Example

Find a PDF p that gives maximum of entropy $H = -\int_w \log p(w)p(w)dw$.

- p — function to find
- H — functional

If a function is set from a predefined set of functions, we can consider the variational calculus problem as an approximation problem.

Model selection: coherent Bayesian inference

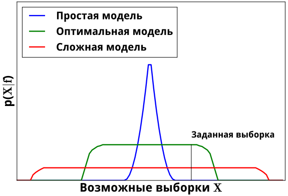
First level: find optimal parameters:

$$\mathbf{w} = \arg\max \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathfrak{D}|\mathbf{h})},$$

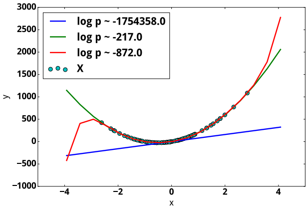
Second level: find optimal model:

Evidence:

$$p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$



Model selection scheme



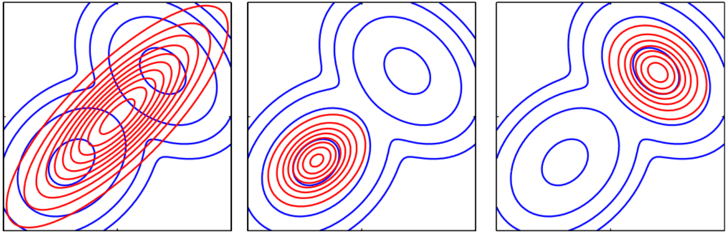
Polynomial regression example

Evidence lower bound, ELBO

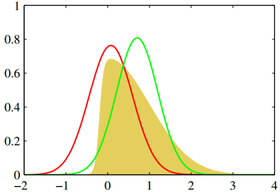
Evidence lower bound is a method of approximation of intractable distribution $p(\mathrm{w}|\mathfrak{D}, \mathrm{h})$ with a distribution $q(\mathrm{w}) \in \mathfrak{Q}$.

Evidence lower bound estimation often reduces to optimization problem

$$\log p(\mathfrak{D}|\mathrm{h}) \geq \mathrm{KL}(q(\mathrm{w})||p(\mathrm{w}|\mathfrak{D})) = -\int_{\mathrm{w}} q(\mathrm{w}) \log \frac{p(\mathrm{w}|\mathfrak{D})}{q(\mathrm{w})} d\mathrm{w} = \mathsf{E}_{\mathrm{w}} \log p(\mathfrak{D}|\mathrm{w}) - \mathrm{KL}(q(\mathrm{w})||p(\mathrm{w}|\mathrm{h}))$$



Variational inference vs. expectation propogation (Bishop)



Laplace Approximation vs Variational inference

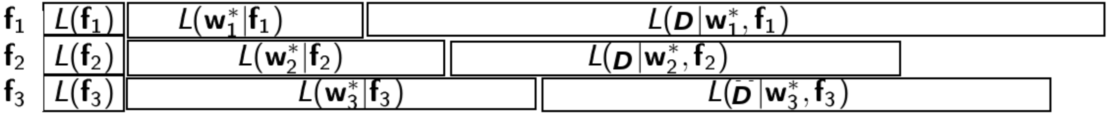
Minimum description length principle

$$\mathrm{MDL}(\mathrm{f}, \mathfrak{D}) = L(\mathrm{f}) + L(\mathfrak{D}|\mathrm{f}),$$

where f is a model, $\mathfrak{D}$ is a dataset, L is a description length in bits.

$$\mathrm{MDL}(\mathrm{f}, \mathfrak{D}) \sim L(\mathrm{f}) + L(\mathrm{w}^*|\mathrm{f}) + L(\mathfrak{D}|\mathrm{w}^*, \mathrm{f}),$$

w^* — optimal parameters.



ELBO estimation

ELBO maximization

$$\int_{\mathrm{w}} q(\mathrm{w})\log \frac{p(\mathrm{y},\mathrm{w}|\mathrm{X},\mathrm{h})}{q(\mathrm{w})} d\mathrm{w}$$

is equivalent to KL-divergence minimization between $q(\mathrm{w}) \in \mathfrak{Q}$ and posteriod distribution $p(\mathrm{w}|\mathrm{y},\mathrm{X},\mathrm{h})$:

$$\hat{q} = \arg\max_{q\in\mathfrak{Q}} \int_{\mathrm{w}} q(\mathrm{w})\log \frac{p(\mathrm{y},\mathrm{w}|\mathrm{X},\mathrm{h})}{q(\mathrm{w})} d\mathrm{w} \Leftrightarrow$$

$$\hat{q} = \arg\min_{q\in\mathfrak{Q}} \mathrm{D}_{\mathrm{KL}}\big(q(\mathrm{w})||p(\mathrm{w}|\mathrm{y},\mathrm{X},\mathrm{h})\big),$$

$$\mathrm{D}_{\mathrm{KL}}\big(q(\mathrm{w})||p(\mathrm{w}|\mathrm{y},\mathrm{X},\mathrm{h})\big) = \int_{\mathrm{w}} q(\mathrm{w}) \log\left(\frac{q(\mathrm{w})}{p(\mathrm{w}|\mathrm{y},\mathrm{X},\mathrm{h})}\right) d\mathrm{w}.$$

ELBO and sample size

Statement

Let $m \gg 0$, $\lambda > 0$, $\frac{m}{\lambda} \in \mathbb{N}$, $\frac{m}{\lambda} \gg 0$. Then optimization

$$\mathsf{E}_q \log p(\mathsf{y}|\mathsf{X}, \mathsf{w}) - \lambda \mathsf{D}_{\mathsf{KL}}\big(q(\mathsf{w})||p(\mathsf{w}|\mathsf{y}, \mathsf{X}, \mathsf{h})\big)$$

is equivalent to optimization of ELBO for a random subsample $\hat{\mathsf{y}}, \hat{\mathsf{X}}$ with size $\frac{m}{\lambda}$.

See also, [β-VAE, Fixing Broken ELBO].

ELBO usage

ELBO: when to use?

- Evidence estimation;
- Latent distribution estimation (topic modeling, dimmension reduction).

Why ELBO?

- reduces the problem of ELBO estimation to optimization;
- scales easily (compare with Laplace approximati;
- easy to use in comparison to MC-based methods.

ELBO can give a very biased evidence estimation.

ELBO: normal distribution

Let $q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathrm{A}_q)$.
Then ELBO equals to:

$$\int_{\mathrm{w}} q(\mathrm{w}) \log p(\mathrm{Y}|\mathrm{X}, \mathrm{w}, \mathrm{h}) d\mathrm{w} - D_{\mathrm{KL}}\big(q(\mathrm{w}) || p(\mathrm{w}|\mathrm{h})\big) \simeq$$

$$\sum_{i=1}^{m} \log p(\mathrm{y}_i|\mathrm{x}_i, \hat{\mathrm{w}}) - D_{\mathrm{KL}}\big(q(\mathrm{w}) || p(\mathrm{w}|\mathrm{h})\big) \to \max_{\mathrm{A}_q, \boldsymbol{\mu}_q}, \quad \hat{\mathrm{w}} \sim q.$$

If prior $p(\mathrm{w}|\mathrm{h})$ is normal:

$$p(\mathrm{w}|\mathrm{h}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathrm{A}),$$

KL-divergence $D_{\mathrm{KL}}\big(q(\mathrm{w}) || p(\mathrm{w}|\mathrm{h})$ is computed analytically:

$$D_{\mathrm{KL}}\big(q(\mathrm{w}) || p(\mathrm{w}|\mathrm{h})\big) = \frac{1}{2}\big(\mathrm{tr}(\mathrm{A}^{-1}\mathrm{A}_q) + (\boldsymbol{\mu} - \boldsymbol{\mu}_q)^{\mathsf{T}}\mathrm{A}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_q) - n + \ln |\mathrm{A}| - \ln |\mathrm{A}_q|\big).$$

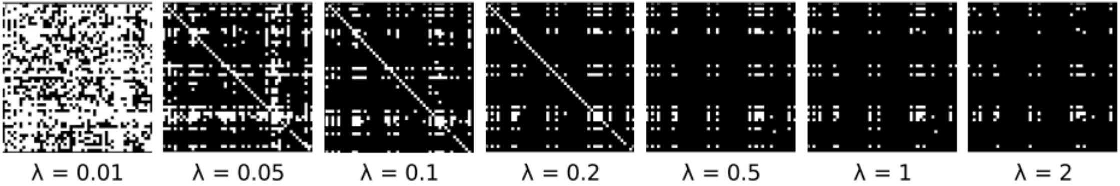
Graves, 2011

Prior: $p(\mathrm{w}|\sigma) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma \mathrm{I})$.
Variational distribution: $q(\mathrm{w}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \sigma_q \mathrm{I})$.
Greedy hyperparameter optimization:

$$\boldsymbol{\mu} = \hat{\mathsf{E}}\mathrm{w}, \quad \sigma = \hat{\mathsf{D}}\mathrm{w}.$$

Parameter pruning w_i using relative PDF:

$$\lambda = \frac{q(0)}{q(\boldsymbol{\mu}_{i,q})} = \exp(-\frac{\mu_i^2}{2\sigma_i^2}).$$



ELBO: normal distribution

“Common” loss function:

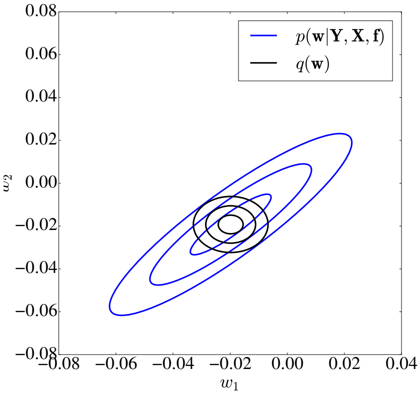
$$L = \sum_{x,y \in \mathfrak{D}} -\log p(y|x, w) + \lambda ||w||_2^2.$$

Variational inference with $(p(w|h) \sim \mathcal{N}(0, 1))$:

$$L = \sum_{x,y} \log\, p(y|x, \hat{w}) +$$

$$+ \frac{1}{2}\left(\mathrm{tr}(A_q) + \boldsymbol{\mu}_q^T A^{-1} \boldsymbol{\mu}_q - \ln |A_q|\right).$$

Poor approximation example q



Local reparametrization

How to calculate $\mathrm{E}_q \log p(\mathrm{y}|\mathrm{X}, \mathrm{w})$?

- Graves, 2011: 1 sample per iteration. Use the following properties:

$$w \sim \mathcal{N}(\mu, \sigma^2) \to w \sim \varepsilon\sigma + \mu, \quad \varepsilon \sim \mathcal{N}(0, 1).$$

  - Poor expectation approximation
- Naive solution: sample 1 iteration per element in batch
  - BackProp will be very slow

Local reparametrization, Kingma et al., 2015

Let $y = \text{ReLU}(XW)$ and parameter matrix W be distributed normally: $w_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2_{i,j})$.
Then XW is a Gaussian matrix:

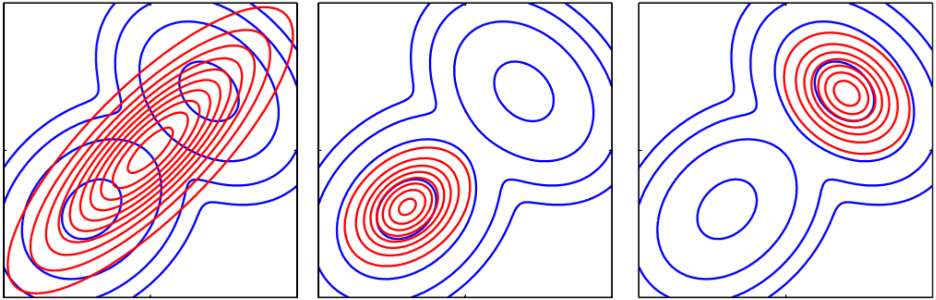
$$G = XW, \quad G_{i,j} \sim \mathcal{N}(\sum_k x_{i,k}\mu_{k,j}, \sum_k x^2_{i,k}\sigma^2_{k,j}).$$

Instead of sampling parameters, sample elements from G (units after activation).

Example

Batch size = 64, matrix W dim is 64×64.

- Graves: one sample, $64 \times 64 = 4096$ elemets. Poor approximation.
- Naive solution: sample parameters 64 times, $64 \times 64 \times 64 = 262144$ elements. Better approximation (in theory).
- Local reparametrization: sample G, $64 \times 64 = 4096$ elements. Better approximation.



Expectation propagation

Minka, 2001: represent prior and approximation distribution via multiplication of factors: $p(\mathrm{w}|\mathfrak{D}) = \prod_i f_i$, $q(\mathrm{w}) = \prod_i \tilde{f}_i$.
Main idea — minimize $KL(p(\mathrm{w}|\mathfrak{D})|q(\mathrm{w}))$.

- Select factor $\tilde{f}_i$ to approximate, «removing» it from consideration, changing into real factor value:
$$q^i \propto f_i \prod_{j \neq i} \tilde{f}_j$$
- Set moments of q^i equal to monents to distribution to approximate (correct, if q is from exponential distribution)
- Repeat until convergence

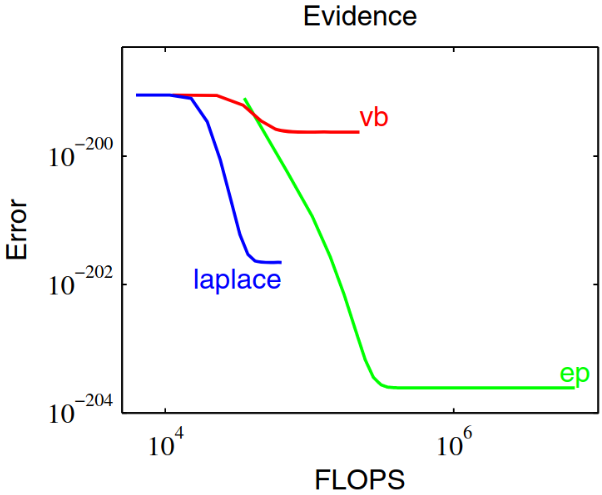
Expectation propagation: pros and cons

Cons:

- Assumption about posteriod distribution (rather slight)
- Original version works only for q from exponential distribution
- No convergencce guarantee

Pros:

- Minimizes KL, not it's lower bound



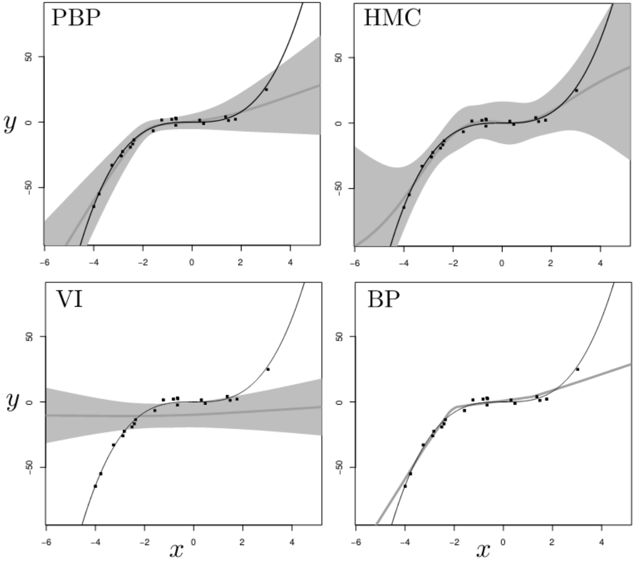
Plot for the 2-component Gaussian mixture.

Probabilistic backpropagation

Combination of Expectiation propagation and backpropagation.
Backward pass:
Update parameters using Bayes rule:

$$p(w_i|\mathfrak{D}) = Z^{-1} p(\mathfrak{D}|w_i, \mathrm{w}^i) p(w) \to p(w_i|\mathfrak{D}) = Z^{-1} p(\mathfrak{D}|w_i, \mathrm{w}^i) \mathcal{N}(w|\mu, \sigma^2).$$

Problem is to calculate Z.
Forward pass:
Compute Z approximately with ab assumption $\mathrm{f}(\mathrm{x}, \mathrm{w}) \sim \mathcal{N}(\mathrm{m}, \mathrm{v})$.
For m, v with ReLU activation there exists an iterative algorithm.


PBP
HMC
VI
BP
y
y
x
x
50
0
-50
-6
-4
-2
0
2
4

Reference

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- Бахтеев О. Ю., Стрижов В. В. Выбор моделей глубокого обучения субоптимальной сложности //Автоматика и телемеханика. – 2018. – №. 8. – С. 129-147.
- Graves A. Practical variational inference for neural networks //Advances in neural information processing systems. – 2011. – Т. 24.
- Louizos C., Ullrich K., Welling M. Bayesian compression for deep learning //arXiv preprint arXiv:1705.08665. – 2017.
- Kingma D. P., Salimans T., Welling M. Variational dropout and the local reparameterization trick //Advances in neural information processing systems. – 2015. – Т. 28. – С. 2575-2583.
- Higgins I. et al. beta-vae: Learning basic visual concepts with a constrained variational framework. – 2016.
- Alemi A. et al. Fixing a broken ELBO //International Conference on Machine Learning. – PMLR, 2018. – С. 159-168.
- Minka T. P. Expectation propagation for approximate Bayesian inference //arXiv preprint arXiv:1301.2294. – 2013.
- Hernández-Lobato J. M., Adams R. Probabilistic backpropagation for scalable learning of bayesian neural networks //International conference on machine learning. – PMLR, 2015. – С. 1861-1869.