Advanced probabilistic methods

Lecture 9: Stochastic variational inference

Pekka Marttinen

Aalto University

March, 2019

Lecture 9 overview

- Recap of variational inference
- Black-box variational inference
- Stochastic variational inference (SVI)
- Lecture based on:
    - Ranganath, Gerrish, Blei (2014). Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 814-822.
- Also relevant:
    - Hoffman, Blei, Wang, Paisley (2013). Stochastic Variational Inference. *Journal of Machine Learning Research*, 14:1303-1347.

SVI, motivating example: topic models (1/2)

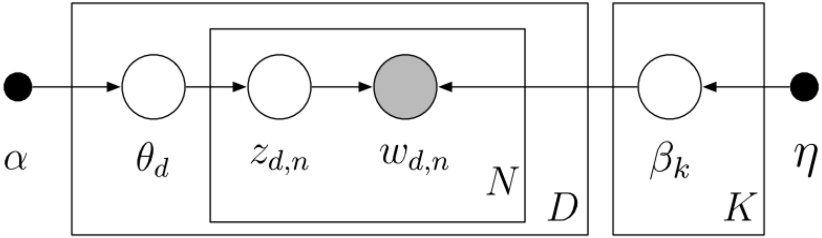
- Below are topics learned from 1.8M *New York Times* articles using **topic models** and **SVI** (from Hoffman et al., 2013)
- In topic models
  - Each document is a combination of topics.
  - Each topic defines a distribution of words.

music	book	art	game	show
band	life	museum	knicks	film
songs	novel	show	nets	television
rock	story	exhibition	points	movie
album	books	artist	team	series
jazz	man	artists	season	says
pop	stories	paintings	play	life
song	love	painting	games	man
singer	children	century	night	character
night	family	works	coach	know

theater	clinton	stock	restaurant	budget
play	bush	market	sauce	tax
production	campaign	percent	menu	governor
show	gore	fund	food	county
stage	political	investors	dishes	mayor
street	republican	funds	street	billion
broadway	dole	companies	dining	taxes
director	presidential	stocks	dinner	plan
musical	senator	investment	chicken	legislature
directed	house	trading	served	fiscal

SVI, motivating example: topic models (2/2)*

- d: documents, $d = 1, \ldots, D$.
- θ_d: proportions of topics in document d.
- $w_{d,n}$: n^{th} word in document d.
- $z_{d,n}$: assignment of word $w_{d,n}$ into a topic.
- k: topics, $k = 1, \ldots, K$.
- β_k: distribution of words in topic k.



Variational Bayes, recap (1/2)

- The simple model: observations $\mathbf{x} = (x_1, \ldots, x_N)$ i.i.d. from

$$p(x_n|\theta, \tau) = (1 - \tau)N(x_n|0, 1) + \tau N(x_n|\theta, 1)$$

- To approximate the posterior $p(\theta, \tau, \mathbf{z}|\mathbf{x})$, we assume

$$p(\theta, \tau, \mathbf{z}|\mathbf{x}) \approx q(\theta)q(\mathbf{z})q(\tau) \quad \text{(mean-field)}$$

and maximize the lower-bound (ELBO) $\mathcal{L}(q)$ in

$$\log p(\mathbf{x}) = \mathcal{L}(q) + KL(q||p),$$

by updating each factor $q(\theta), q(\mathbf{z}), q(\tau)$ in turn.

Variational Bayes, recap (2/2)

- To update each factor, we use the result ("important formula") from Lecture 6:

$$
\begin{aligned}
\log q^*(\mathbf{z}) &= E_{q(\tau)q(\theta)}\left[\log p(\theta, \tau, \mathbf{z}, \mathbf{x})\right] + \text{const.} \\
\log q^*(\theta) &= E_{q(\tau)q(\mathbf{z})}\left[\log p(\theta, \tau, \mathbf{z}, \mathbf{x})\right] + \text{const.} \\
\log q^*(\tau) &= E_{q(\mathbf{z})q(\theta)}\left[\log p(\theta, \tau, \mathbf{z}, \mathbf{x})\right] + \text{const.}
\end{aligned}
$$

- And exponentiate and normalize.

Variational Bayes, alternative view

- With conjugate priors, the distributions of the factors are known.
- For example, if $\tau \sim Beta(\alpha_0, \alpha_0)$, we know that

$$q^*(\tau) = Beta(\alpha_\tau, \beta_\tau).$$

- The 'important formula' tells what values of **variational parameters** α_τ, β_τ maximize the ELBO $\mathcal{L}(q)$, if other factors are kept fixed.
- For example

$$\alpha_\tau = N_2 + \alpha_0 \text{ and } \beta_\tau = N_1 + \alpha_0,$$

where

$$N_k = \sum_{n=1}^{N} r_{nk}.$$

Variational Bayes, alternative view

- In the simple model, the factors are thus (due to conjugacy)

$$\begin{aligned} q(z_n|r_n) &= Categorical(z_n|r_{n1}, r_{n2}) \quad n = 1, \dots, N \\ q(\tau|\alpha_\tau, \beta_\tau) &= Beta(\tau|\alpha_\tau, \beta_\tau) \\ q(\theta|\eta_1, \eta_2) &= N(\theta|\eta_1, \eta_2) \end{aligned}$$

  and mean-field VB corresponds to maximizing the ELBO $\mathcal{L}(q)$ w.r.t. the variational parameters of each factor in turn.

Black-box variational inference (1/2)

- Used in Edward
- Assuming a variational distribution $q(z|\lambda)$, where λ represents the variational parameters, the ELBO can be written as

$$\mathcal{L}(\lambda) = E_{q(z|\lambda)}[\log p(x,z) - \log q(z|\lambda)]$$

- **Goal**: to find variational parameters λ which maximize the ELBO.
- The maximization is done using **gradients**[1]

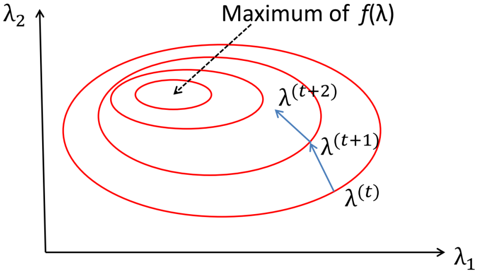
$$\nabla_\lambda \mathcal{L} = E_{q(z|\lambda)}\left[\nabla_\lambda \log q(z|\lambda)(\log p(x,z) - \log q(z|\lambda))\right]$$

[1] For the derivation of this formula, see the paper Ranganath et al. (2014).

Reminder: gradient ascent algorithm*

- Gradient ascent algorithm maximizes a given function f by taking steps of length ρ to the direction of the gradient ∇f.

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho \nabla_\lambda f(\lambda^{(t)}), \text{ where } \nabla_\lambda f = \left(\frac{\partial f}{\partial \lambda_1}, \ldots, \frac{\partial f}{\partial \lambda_D} \right)$$



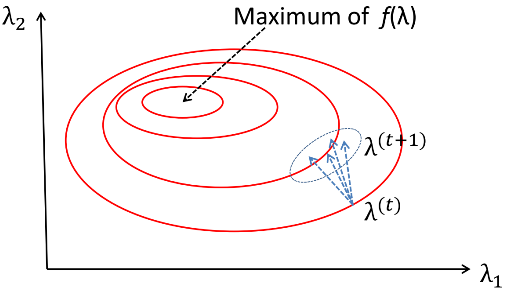
- $\lambda^{(t+1)} = \lambda^{(t)} - \rho \nabla_\lambda f(\lambda^{(t)})$ gives gradient descent.

Reminder: stochastic gradient ascent*

- Stochastic gradient ascent takes **random steps**, that are **on average to the correct direction**:

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho b_t(\lambda^{(t)}),$$

- $b_t(\lambda)$ is a random variable s.t. $E(b_t(\lambda)) = \nabla_\lambda f(\lambda)$.



cartoonstock.com

Reminder: SGA with a mini-batch*

- To find a maximum likelihood estimate $\widehat{\lambda}$, then

$$f(\lambda) = \frac{1}{N}\sum_{n=1}^{N} \log p(x_n|\lambda), \text{ and } \nabla_\lambda f(\lambda) = \frac{1}{N}\sum_{n=1}^{N} \nabla_\lambda \log p(x_n|\lambda)$$

  and we have to differentiate $\log p(x_n|\lambda)$ for all n.
- It is cheaper to sample a **minibatch** of S data points x_s and compute a noisy gradient

$$b(\lambda) = \frac{1}{S}\sum_s \nabla_\lambda \log p(x_s|\lambda),$$

  which points approximately to the direction of $\nabla_\lambda f(\lambda)$.

Black-box variational inference (2/2)

- In BBVI the gradients are approximated by sampling:

$$\nabla_\lambda \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z_s|\lambda)(\log p(x, z_s) - \log q(z_s|\lambda)), \quad z_s \sim q(z|\lambda)$$

- These **stochastic gradients** are used in SGA to maximize the ELBO.
- Need to be able to compute $\log p(x, z)$, no other model-specific derivations required!
- NB: BBVI involves also some tricks to reduce the variance of the stochastic gradients, details skipped.

Global and local parameters

- Many models can be represented using the generic model (left), where β are **global**, and z_n **local** hidden variables. (GMM, FA, topic models, mixture of FA models, ...)

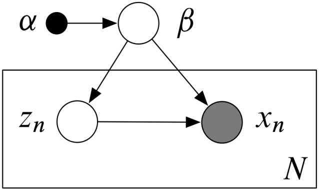
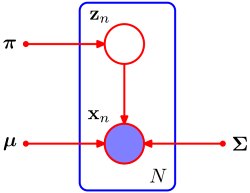
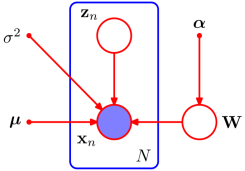


Figure: Hoffman et al. (2013), Fig. 2





Global and local parameters

- Mean-field approximation for $p(\beta, \mathbf{z}|\mathbf{x})$ is given by

$$q(z, \beta) = q(\beta|\lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{nj}|\phi_{nj}),$$

where λ is a **global variational parameter**, and ϕ_{nj} denote **local variational parameters**.

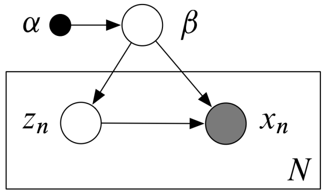


Figure: Hoffman et al. (2013), Fig. 2

Initialize global parameters λ
repeat
  for *each local* variational parameter ϕ_{nj} **do**
    Update ϕ_{nj} using the mean-field update[2]
  end for
  Update global variational parameters λ with mean-field update
until $\mathcal{L}(q)$ converges

- **Problem:** all local variational parameters are updated before the global parameters -> slow if N large.

[2] This is now for the regular VB, and not BBVI, i.e., we assume the closed-form VB updates.

Stochastic variational inference (SVI)

1: Initialize global parameters λ
2: Set step-size schedule ρ_t appropriately
3: **repeat**
4: Sample a data point x_i uniformly from the data set
5: Update the local parameter ϕ_i of the *sampled point only*
6: Form intermediate global parameters $\widehat{\lambda}$ as if x_i was observed N times
7: Update the global variational parameters using

$$\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \widehat{\lambda}$$

8: **until** ready

- Instead of a single data point, a mini-batch could also be used.

SVI in BBVI

- BBVI requires ability to evaluate the log joint: $\log p(\mathbf{x}, \mathbf{z}, \beta)$.
- Approximation using a mini-batch of size M:

$$\begin{aligned} \log p(\mathbf{x}, \mathbf{z}, \beta) &= \log p(\beta) + \sum_{n=1}^{N} [\log p(x_n | z_n, \beta) + \log p(z_n | \beta)] \\ &\approx \log p(\beta) + \frac{N}{M} \sum_{n=1}^{M} [\log p(x_m | z_m, \beta) + \log p(z_m | \beta)] \end{aligned}$$

- Each observation has a weight N/M
  - ensures the expectation of the second line equals the first line.

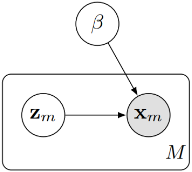
SVI for a mixture of Gaussians in Edward

```
beta = Normal(mu=tf.zeros([K, D]), sigma=tf.ones([K, D]))
z = Categorical(logits=tf.zeros([M, K]))
x = Normal(mu=tf.gather(beta, z), sigma=tf.ones([M, D]))

qbeta = Normal(mu=tf.Variable(tf.zeros([K, D])),
               sigma=tf.nn.softplus(tf.Variable(tf.zeros([K, D]))))
qz = Categorical(logits=tf.Variable(tf.zeros([M, D])))

inference = ed.VariationalInference({beta: qbeta, z: qz}, data={x: x_batch})
inference.initialize(scale={x: float(N)/M, z: float(N)/M})
```

M, not N (M<<N)

Use current mini-batch of size M as data

scale by N/M, to get the correct expectation



Tran et al. (2017) ICLR

Important points

- Black-box variational inference
  - Mean-field VB can be seen as an optimization problem: the variational parameters for each factor are updated in turn to maximize the variational lower bound $\mathcal{L}(q)$.
  - In BBVI the ELBO is maximized directly using stochastic gradient ascent.
  - Stochastic gradient of ELBO is approximated by sampling from the approximation q.
- Stochastic variational inference:
  - update only one (or a few) local variational parameters at each iteration, and update the global variational parameters on the basis of these few local factors.
  - Scales variational inference to massive data sets

Advertisement: summer internship

- One summer internship position still open in the Machine Learning for Health (Aalto-ML4H) research group.
- Conditions similar to the 'regular' summer internships at the CS department.
- Topic: developing and implementing Bayesian models for an application in genomics, together with top international collaborators.
- Selection criteria: stage of studies, study performance in machine learning courses, general skills (programming, math, ...), interest to work with a bioinformatics application.
- Topic is suitable for a Master's thesis, and can afterwards be continued towards a PhD.
- DL for applications Friday 29.3.
- Apply by sending a CV, transcript, and brief motivation letter by email to the lecturer.