BML lecture #3: variational Bayes

http://github.com/rbardenet/bml-course

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France



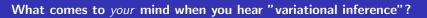


- 1 Introduction
- 2 Variational inference
- 3 Back to LDA
- 4 Generalizing VB

1 Introduction

2 Variational inference

- 3 Back to LDA
- 4 Generalizing VB



- 1 Introduction
- 2 Variational inference
- 3 Back to LDA
- 4 Generalizing VB

When MCMC is intractable, variational inference comes to help

Turning integration into optimization over measures

Variational Bayesian inference (VB) consists in approximating

$$\int f(\theta)\pi(\theta)\mathrm{d}\theta pprox \int f(\theta)q^{\star}(\theta)\mathrm{d}\theta$$

with $q^* \in \arg\min_{q \in \mathcal{Q}} \operatorname{distance}(\pi, q)$. Often we take

$$\mathsf{distance}(\pi,q) = \mathsf{KL}(q,\pi) := \int q(heta) \log rac{q(heta)}{\pi(heta)} \mathrm{d} heta.$$

for computational convenience.

But remember we can only evaluate $\pi_u = Z\pi...$

▶ Show that
$$J(q) := \int q(\theta) \log \frac{q(\theta)}{\pi_n(\theta)} d\theta = \mathsf{KL}(q,\pi) - \log Z$$
.

▶ In particular,
$$L(q) = -J(q) \le \log Z$$
. For

$$\pi_u(\theta) = p(\mathsf{data}|\theta)p(\theta),$$

L(q) is thus a lower bound for the evidence p(data) (ELBO).

Choosing the approximating family $\mathcal Q$

▶ The most common approach is the mean-field approximation

$$Q = \{\theta \mapsto \prod_{d=1}^{D} q_d(\theta_d)\}.$$

Include all variables over which you integrate, e.g.

$$q(\theta, z_{1:n}) = \prod_{d=1}^{D} q_d(\theta_d) \prod_{i=1}^{N} q_i(z_i).$$

- ▶ Try to keep some dependence if it is key in your application.
- If your original model has NEF conditionals, coordinate-wise maximization of $q \mapsto L(q)$ is easy.

Mean-field yields closed-form updates

1 Introduction

- 2 Variational inference
- 3 Back to LDA
- 4 Generalizing VB

Back to LDA 1/2

$$\begin{split} \log p(y, z, \pi, B) \\ &= \sum_{i=1}^{N} \left[\log p(\pi_{i} | \alpha) + \sum_{\ell=1}^{L_{i}} \left(\log p(z_{i\ell} | \pi_{i}) + \log p(y_{i\ell} | z_{i\ell}, B) \right) \right] + p(B | \gamma) \\ &\propto \sum_{i=1}^{N} \left[\sum_{k=1}^{K} \alpha_{k} \log \pi_{ik} + \sum_{\ell=1}^{L_{i}} \left(\sum_{k=1}^{K} \mathbf{1}_{z_{i\ell} = k} \log \pi_{ik} + \sum_{v=1}^{V} \sum_{k=1}^{K} \mathbf{1}_{y_{i\ell} = v} \mathbf{1}_{z_{i\ell} = k} \log b_{kv} \right) \right] \\ &+ \sum_{k=1}^{K} \sum_{v=1}^{V} \gamma_{k} \log b_{kv}. \end{split}$$

Lemma (exercise)

Let $\Psi(\cdot) := \Gamma'(\cdot)/\Gamma(\cdot)$ be the digamma function. Then

$$\mathbb{E}_{\mathsf{Dir}(\theta|\eta)}\log\theta_i = \Psi(\eta_i) - \Psi(\|\eta\|_1) =: \Psi_i(\eta).$$

VB for **LDA**: singling out π_i

$$\log p(y, z, \pi, B)$$

$$\propto \sum_{i=1}^{N} \left[\sum_{k=1}^{K} \alpha_k \log \pi_{ik} + \sum_{\ell=1}^{L_i} \left(\sum_{k=1}^{K} 1_{z_{i\ell}=k} \log \pi_{ik} + \sum_{v=1}^{V} \sum_{k=1}^{K} 1_{y_{i\ell}=v} 1_{z_{i\ell}=k} \log b_{kv} \right) \right] + \sum_{k=1}^{K} \sum_{v=1}^{V} \gamma_k \log b_{kv}.$$

VB for **LDA**: singling out $z_{i\ell}$

$$\log p(y, z, \pi, B)$$

$$\propto \sum_{i=1}^{N} \left[\sum_{k=1}^{K} \alpha_{k} \log \pi_{ik} + \sum_{\ell=1}^{L_{i}} \left(\sum_{k=1}^{K} 1_{z_{i\ell} = k} \log \pi_{ik} + \sum_{v=1}^{V} \sum_{k=1}^{K} 1_{y_{i\ell} = v} 1_{z_{i\ell} = k} \log b_{kv} \right) \right] + \sum_{k=1}^{K} \sum_{v=1}^{V} \gamma_{k} \log b_{kv}.$$

VB for **LDA**: singling out $B_{k:}$

$$\log p(y, z, \pi, B)$$

$$\propto \sum_{i=1}^{N} \left[\sum_{k=1}^{K} \alpha_k \log \pi_{ik} + \sum_{\ell=1}^{L_i} \left(\sum_{k=1}^{K} 1_{z_{i\ell} = k} \log \pi_{ik} + \sum_{\nu=1}^{V} \sum_{k=1}^{K} 1_{y_{i\ell} = \nu} 1_{z_{i\ell} = k} \log b_{k\nu} \right) \right] + \sum_{k=1}^{K} \sum_{\nu=1}^{V} \gamma_{\nu} \log b_{k\nu}.$$

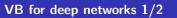
Using counts keeps space and time complexity low

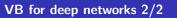
▶ Storing $\tilde{z}_{i\ell k}$ requires $\mathcal{O}(NK \sum_i L_i)$ space. In practice, one works with (sparse) count data

 n_{iv} = number of times word v appears in document i,

and variables c_{ivk} , thus reducing storage costs (and the dimension of the underlying integral!) to $\mathcal{O}(NVK)$.

Gradient-based VB





Lots of variants of VB exist (Murphy, 2012)

► For hidden variable models, EM is VB with

$$q(z,\theta) = \pi(z|\theta)\delta_{\tilde{\theta}}(\theta).$$

Variational EM is VB with

$$q(z,\theta) = q(z)\delta_{\tilde{\theta}}(\theta).$$

- ▶ VB for any PGMs with NEF arrows is variational message passing.
- Rather approximating

$$\pi(heta)pprox\prod_{f=1}^Fq_f(heta)$$

leads to expectation propagation.

► These days, ADVI with stochastic gradients is the default VI choice in probabilistic programming software like PyMC3, Stan, or PyRo.

1 Introduction

- 2 Variational inference
- 3 Back to LDA
- 4 Generalizing VB

"Optimization-centric" Bayesian inference1

Given a loss ℓ , a divergence D, and a set of distributions Q, consider

$$q^{\star} \in \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q} \sum_{i=1}^{N} \ell(\theta, x_{i}) + D(q(\theta) d\theta, p(\theta) d\theta).$$
 $(P(\ell, D, \mathcal{Q}))$

¹Knoblauch, Jewson, and Damoulas, 2022.

Conclusion on VB

- ► Computationally attractive alternative to MCMC.
- Lots of open questions on connecting VB and to the original SEU problem: is VB justified in itself or is it simply a computationally convenient backup option?
- Many partial answers, e.g. try to use VB in importance sampling, optimization-centric viewpoint.
- More meaningful alternatives to the KL: maximum mean discrepancy, etc.

References I

- J. Knoblauch, J. Jewson, and T. Damoulas. "An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference". In: *Journal of Machine Learning Research* 23.132 (2022), pp. 1–109.
- [2] K. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.