BML lecture #3: variational Bayes

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

Rémi Bardenet

CNRS & CRIStAL, Univ. Lille, France





1 Introduction

2 Variational inference

1 Introduction

2 Variational inference



1 Introduction

2 Variational inference

When MCMC is intractable, variational inference comes to help

Turning integration into optimization over measures

Variational Bayesian inference (VB) consists in approximating

$$\int f(\theta)\pi(\theta)\mathrm{d}\theta \approx \int f(\theta)q(\theta)\mathrm{d}\theta$$

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

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

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,
$$L(q) = -J(q) \leq \log Z$$
. 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 $\mathcal Q$

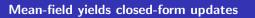
▶ The most common approach is the mean-field approximation

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

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.



1 Introduction

2 Variational inference

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

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

Back to LDA 2/2

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 $\mathcal{O}(NVK)$.

Reparametrization+SGD vs mean-field+closed updates

VB for deep networks

Automatic differentiation variational inference (Kucukelbir et al., 2017)

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.

References I

- [1] A. Kucukelbir et al. "Automatic differentiation variational inference". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 430–474.
- [2] K. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.