Inference for Indian Buffet Processes

October 20, 2018

Contents

1 Taxonomy of Inference for IBP

This is in rough temporal order of papers published about these methods. By (perhaps?) coincidence, this also corresponds to an order of decrease in theoretical quality of posterior inference and an increase in the scalability or speed.

1.1 Gibbs sampling

The main reference is Ghahramani and Griffiths, (2006), however Doshi et al., (2009a) gives a clear explanation of both the collapsed and uncollapsed Gibbs samplers and provides derivations useful for implementation.

There are two major variants of Gibbs: **collapsed** and **uncollapsed**. Collapsed sampling marginalizes out model or latent variables that you are not interested in obtaining samples of or a distribution for. For example, in the linear Gaussian IBP model, you may want Z but not A. Uncollapsed sampling explicitly samples each model variable and latent variable. There are many choices of "partial collapsed" samplers between fully collapsed and fully uncollapsed. Each have their own trade-offs, depending on the size of the dataset and other considerations.

There are a number of tweaks and improvements worth noting:

1. MH split/merge proposals (Meeds et al., 2007) (Rachit, by this do you mean the Reverse Jump MCMC?)
2. slice sampling (Teh, Görür, and Ghahramani, 2007)
3. particle filtering (Wood and Griffiths, 2007)
4. accelerated collapsed Gibbs sampling (Doshi-Velez and Ghahramani, 2009)

1.2 Coordinate Ascent Variational Inference (CAVI)

By CAVI we mean mean-field variatonal inference (Beal et al., 2003; Wainwright, Jordan, et al., 2008). In this method, we factor the variational posterior into one component for each variable, write out the ELBO in terms of these factors, take the derivative with respect to the parameters to optimize over, set it equal to 0 and solve. This yields coordinate-wise updates for the parameters of the variational distributions. These variational distributions are necessarily finite dimensional, which is referred to as the "truncation" of infinite models.

The major reference here is Doshi et al., (2009b), which implements CAVI using a truncated posterior approximation to both (1) finite approximation to the IBP and (2) the infinite/untruncated IBP, for linear Gaussian models.

1.3 Stochastic Variatonal Inference (SVI)

Let D be the data and θ be some model parameters. Say we are interested in optimizing a function $f(D, \theta)$ with respect to θ. Stochastic optimization replaces this objective with a noisy function $\hat{f}$ such that $E[\hat{f}] = f$. The randomness of $\hat{f}$ is usually that of drawing random subsamples (batches) x from the data and evaluating the function f on the batch rather than the full dataset. See Robbins & Munro.

Stochastic Variational Inference (Hoffman et al., (2013)) takes the CAVI objective and applies stochastic gradient descent. The CAVI obbjective is re-written in terms of data batches. This allows one to learn and optimize the global parameters (e.g. topics in a Topic Model) after just a few data points, improving inference for the rest of the data set in that iteration instead of using stale parameters.

The main reference here (one of many) is Shah, Knowles, and Ghahramani, (2015), which does a (remarkably thorough) investigation into the different kinds of SVI algorithms that can be applied (especially looking into different kind of approximations, structured or not).

1.4 Autograd on exact ELBO VI (ADVI-exact)

Consider again the ELBO without using the stochastic objective involving data subsamples. In the traditional CAVI approach, one must derive all of the variational parameter updates by writing out the ELBO, taking derivatives, setting equal to 0, and solving. Another approach is to take advantage of autograd. Note that this still requires you to be able to write out the objective in closed form (CAVI also required this). This also requires that your autograd system implement derivatives for any function used in the ELBO. This does save you from derivative derivations!

ADVI (citation goes here) does use autograd, but they consider the specific case of using autograd in a setting that may be called "doubly stochastic optimization of the ELBO":

- ADVI uses SGD, so it's stochastic with respect to the data subsamples as described above for SVI.

- ADVI assumes that the objective cannot be written in closed form and they optimize a stochastic approximation as follows. Usually, it's the expectation of the likelihood term with respect to the variational distributions over model parameters that cannot be written in closed form (corresponding to a conjugate model). So they sample particular instantiations of the model parameters and evaluate the likelihood given those particular parameters and average across samples. They use the reparameterization gradient to optimize through the sampling step. This sampling/reparameterization gradient step is what adds the second level of stochasticity.

We consider the use of autograd to be relevant and preferable in the non-stochastic case, the SVI case, and the doubly stochastic ADVI case.

From Doshi et al., (2009a) we have an exact form of the ELBO for the finite approximation to the IBP and an *almost closed form* of the ELBO for the infinite IBP ($\mathbb{E}[\log \text{stick}]$ term needs approximation, but once it is approximated, we can take it as part of the closed-form objective). We can do gradient ascent on ELBO in either case directly via autograd with minimal intervention to approximate the log stick term. Moreover, if we can approximate the log-stick term in a differentiable way, then we can use gradient ascent without any intervention to optimize slightly-inexact ELBO.

Note that this is different from the usual method employed in VAEs, where the ELBO is sampled - here the ELBO is exact.

1.5 Autograd on sampled ELBO (ADVI)

AS mentioned, this approach uses autograd on a stochastic approximation to the ELBO that samples model parameters from the variational distribution instead of integrating the likelihood with respect to the variational distributions over model parameters. One still has the additional choice here to be stochastic over data as well. Approximating the likelihood integral allows you to work with non-conjugate models.

1.6 Variational Autoencoders (VAE)

In this method, the main difference is the addition of *amortization*, i.e. a map from $x \to \lambda(x)$ for the variational distribution $q(z, \pi; \lambda)$. Note that this limits the variational posterior, so in theory this method should be less effective than regular CAVI as a mean-field method.

The two main references for this are Singh, Ling, and Doshi-Velez, (2017) and Chatzis, (2018) (note: the former is an earlier version of this work, and the latter has changed many times since its original publication in 2014). However, it recently came to our attention that Fan and Heller, (2015) also attempts to apply the same method[1].

1.7 Semi-amortized Variational Autoencoders (SA-VAE)

This is a reference to Kim et al., (2018), which essentially blurs the line between SVI and VAEs by "semi-amortizing"[2]. Note that Cremer, Li, and Duvenaud, (2018), Krishnan, Liang, and Hoffman, (2018), and Hjelm et al., (2016) are also recent works that consider the amortization gap - Kim et al., (2018) is perhaps the most successful but maybe the most difficult to implement.

A much newer reference still under review proposes something much simpler: at the beginning of training, just update the inference network much more than the generative model. The

[1] TODO: we need to get a comparison of the quality of each of these methods

[2] Since the line is, I guess, amortization

decision of when to stop “aggressive“ updates is based on the mutual information between $\boldsymbol{z}$ and $\boldsymbol{x}$ in $q(\boldsymbol{z}|\boldsymbol{x})$.

2 What does inference mean?

We have to make a clear distinction between the two types of inference[3] that we care about: (1) inference on **local** variables (i.e. z_{nk}) and (2) inference on **global** variables (i.e. A, π).

3 Structured vs. Mean Field

Several papers tackle which should be used. Shah, Knowles, and Ghahramani, (2015) found that using a variational approximation that maintains local-local dependencies was more important than the difference between mean-field and structured posteriors. See also Maaløe et al., (2016) for a ’VAE’-like construction of the same idea.

4 Research Questions

Feel free to add more questions here as you come up with them. Let’s try our best to answer each question with a set of experiments, each in a separate Python script.

1. `data_efficiency`: Which inference algorithm is the most data efficient? Again, theoretically speaking the farther down we go, the less ’flexible’ the approximate posterior, in some ways (though there are significant exceptions, like in the ’SVI’ algorithms in Shah, Knowles, and Ghahramani, (2015) the variational approximations are *more rich* than in the mean-field CAVI of Doshi et al., (2009b), and SA-VAE is certainly more rich than the VAE), but there are other tradeoffs that might effect data efficiency - for example, amortizing the inference might restrict the *flexibility* but we might learn faster because information is shared between data points. However for the purpose of this question we only consider the *data efficiency*, not the speed.

2. `advi_exact_cavi`: Is it reasonable to assume that ADVI-exact is *just as good* as CAVI? Technically if the loss is convex they should be, however we know from Doshi et al., (2009a)[4] that the loss *is not convex*, so we don’t know which is a better algorithm. In fact this may vary a lot based on hyperparameters (e.g. learning rate schedule, initialization[5]) *or* the choice of optimization algorithm. Luckily in the case of CAVI there aren’t many choices, besides update schedules, so it shouldn’t be too hard to find the ”best” version.

3. `speed`: Do we care about the speed of each algorithm? Finale seemed to care about it, but hopefully modern computers (and PyTorch) are fast enough that we don’t really care about this anymore for reasonably sized datasets.

4. `conjugacy`: One of the selling points of the IBP-VAE is that we could build *non-conjugate* generative models, i.e. models where the prior and the posterior don’t match. However, we found that using non-conjugate generative models led to strange ”shelf-like” behavior in the posterior over z. The methods above that allow for non-conjugate generative models

[3] We take it as given (since this is a nonparametric Bayesian method) that what we care about is the (approximate) posterior, whether it’s a parametrized distribution or samples from that distribution.

[4] See Section 3

[5] At least in the case of learning rate schedules we can use the same initialization! So we can check the trajectories and compare them over time. This will make for a neat plot.

are: ADVI (sometimes ADVI-exact), VAE, SA-VAE. There are two subquestions I'm interested in:

(a) Are the methods that allow for non-conjugate inference *much worse* than the ones that don't?[6]

(b) If they are, what's going on with these models? Do we just not have enough data to get started fitting them? One possibility (which I consider likely) is that *gradient descent* wasn't the problem, but *amortization* was.

References

[1] Matthew James Beal et al. *Variational algorithms for approximate Bayesian inference.* university of London London, 2003.

[2] Sotirios P Chatzis. "Indian Buffet Process Deep Generative Models for Semi-Supervised Classification". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 2456–2460.

[3] Chris Cremer, Xuechen Li, and David Duvenaud. "Inference suboptimality in variational autoencoders". In: *arXiv preprint arXiv:1801.03558* (2018).

[4] Finale Doshi et al. *Variational inference for the Indian buffet process.* Tech. rep. CBL-2009-001. 2009. URL: http://mlg.eng.cam.ac.uk/pub/pdf/DosMilVanTeh09b.pdf.

[5] Finale Doshi et al. "Variational inference for the Indian buffet process". In: *Artificial Intelligence and Statistics.* 2009, pp. 137–144.

[6] Finale Doshi-Velez and Zoubin Ghahramani. "Accelerated sampling for the Indian buffet process". In: *Proceedings of the 26th annual international conference on machine learning.* ACM. 2009, pp. 273–280.

[7] Kai Fan and Katherine Heller. "Scalable Non-linear Beta Process Factor Analysis". In: (2015).

[8] Zoubin Ghahramani and Thomas L Griffiths. "Infinite latent feature models and the Indian buffet process". In: *Advances in neural information processing systems.* 2006, pp. 475–482.

[9] Devon Hjelm et al. "Iterative refinement of the approximate posterior for directed belief networks". In: *Advances in Neural Information Processing Systems.* 2016, pp. 4691–4699.

[10] Matthew D Hoffman et al. "Stochastic variational inference". In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 1303–1347.

[11] Yoon Kim et al. "Semi-Amortized Variational Autoencoders". In: *arXiv preprint arXiv:1802.02550* (2018).

[12] Rahul Krishnan, Dawen Liang, and Matthew Hoffman. "On the challenges of learning with inference networks on sparse, high-dimensional data". In: *International Conference on Artificial Intelligence and Statistics.* 2018, pp. 143–151.

[13] Lars Maaløe et al. "Auxiliary deep generative models". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*. JMLR. org. 2016, pp. 1445–1454.

[14] Edward Meeds et al. "Modeling dyadic data with binary latent factors". In: *Advances in neural information processing systems.* 2007, pp. 977–984.

[6] In which case, it doesn't matter if we can 'fit' non-conjugate models, if we're not fitting them at all, right?

[15] Amar Shah, David Knowles, and Zoubin Ghahramani. "An empirical study of stochastic variational inference algorithms for the beta Bernoulli process". In: *International Conference on Machine Learning*. 2015, pp. 1594–1603.

[16] Rachit Singh, Jeffrey Ling, and Finale Doshi-Velez. "Structured Variational Autoencoders for the Beta-Bernoulli Process". In: (2017).

[17] Yee Whye Teh, Dilan Görür, and Zoubin Ghahramani. "Stick-breaking construction for the Indian buffet process". In: *Artificial Intelligence and Statistics*. 2007, pp. 556–563.

[18] Martin J Wainwright, Michael I Jordan, et al. "Graphical models, exponential families, and variational inference". In: *Foundations and Trends® in Machine Learning* 1.1–2 (2008), pp. 1–305.

[19] Frank Wood and Thomas L Griffiths. "Particle filtering for nonparametric Bayesian matrix factorization". In: *Advances in neural information processing systems*. 2007, pp. 1513–1520.