Online but Accurate Inference for Latent Variable Models with Local Gibbs Sampling

Christophe Dupuy
INRIA - Technicolor
christophe.dupuy@inria.fr

Francis Bach
INRIA - ENS
francis.bach@inria.fr

Abstract

We study parameter inference in large-scale latent variable models. We first propose a unified treatment of online inference for latent variable models from a non-canonical exponential family, and draw explicit links between several previously proposed frequentist or Bayesian methods. We then propose a novel inference method for the frequentist estimation of parameters, that adapts MCMC methods to online inference of latent variable models with the proper use of local Gibbs sampling. Then, for latent Dirichlet allocation,we provide an extensive set of experiments and comparisons with existing work, where our new approach outperforms all previously proposed methods. This work is currently under review for JMLR [1] (submitted on July, 27 2016).

1 Introduction

Probabilistic graphical models provide general modelling tools for complex data, where it is natural to include assumptions on the data generating process by adding latent variables in the model. Such latent variable models are adapted to a wide variety of unsupervised learning tasks [2, 3]. In this paper, we focus on parameter inference in such latent variable models where the main operation needed for the standard expectation-maximization (EM) algorithm is intractable, namely dealing with conditional distributions over latent variables given the observed variables; latent Dirichlet allocation (LDA) [4] is our motivating example, but many hierarchical models exhibit this behavior, e.g., ICA with heavy-tailed priors. For such models, there exist two main classes of methods to deal efficiently with intractable exact inference in large-scale situations: sampling methods or variational methods.

Sampling methods can handle arbitrary distributions and lead to simple inference algorithms while converging to exact inference. However it may be slow to converge and non scalable to big datasets in practice. In particular, although efficient implementations have been developed, for example for LDA [5, 6], MCMC methods may not deal efficiently yet with continuous streams of data for our general class of models.

On the other hand, *variational inference* builds an approximate model for the posterior distribution over latent variables—called variational—and infer parameters of the true model through this approximation. The fitting of this variational distribution is formulated as an optimization problem where efficient (deterministic) iterative techniques such as gradient or coordinate ascent methods apply. This approach leads to scalable inference schemes [7], but due to approximations, there always remains a gap between the variational posterior and the true posterior distribution, inherent to algorithm design, and that will not vanish when the number of samples and the number of iterations increase.

Beyond the choice of approximate inference techniques for latent variables, parameter inference may be treated either from the *frequentist* point of view, e.g., using maximum likelihood inference, or a *Bayesian* point of view, where the posterior distribution of the parameter given the observed data is approximated. With massive numbers of observations, this posterior distribution is typically

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

peaked around the maximum likelihood estimate, and the two inference frameworks should not differ much [8].

In this paper, we focus on methods that make a single pass over the data to estimate parameters. We make the following contributions: (1) We review and compare existing methods for online inference for latent variable models from a non-canonical exponential family and draw explicit links between several previously proposed frequentist or Bayesian methods; (2) We propose a novel inference method for the frequentist estimation of parameters, that adapts MCMC methods to online inference of latent variable models with the proper use of "local" Gibbs sampling; (3) We provide an extensive set of experiments for LDA, where our new approach outperforms all previously proposed methods.

2 Online EM

We consider an *exponential family* model on random variables (X, h) with parameter $\eta \in \mathcal{E} \subseteq \mathbb{R}^d$ and with density [9]:

$$p(X, h|\eta) = a(X, h) \exp\big[\langle \phi(\eta), S(X, h)\rangle - \psi(\eta)\big]. \tag{1}$$

We assume that h is hidden and X is observed. The vector $\phi(\eta) \in \mathbb{R}^d$ represents the natural parameter, $S(X, h) \in \mathbb{R}^d$ is the vector of sufficient statistics, $\psi(\eta)$ is the log-normalizer, and $a(X, h)$ is the underlying base measure. We consider a *non-canonical* family as in many models (such as LDA), the natural parameter $\phi(\eta)$ does not coincide with the model parameter η, that is, $\phi(\eta) \not\equiv \eta$; we however assume that ϕ is injective.

We consider N *i.i.d.* observations $(X_i)_{i=1,...,N}$ from a distribution $t(X)$, which may be of the form $P(X|\eta^*) = \int_h p(X, h|\eta^*)\mathrm{d}h$ for our model above and a certain $\eta^* \in \mathcal{E}$ (well-specified model) or not (misspecified model). Our goal is to obtain a predictive density $r(X)$ built from the data and using the model defined in (1), with the maximal expected log-likelihood $\mathbb{E}_{t(X)} \log r(X)$.

Maximum likelihood estimation. In the frequentist perpective, the predictive distribution $r(X)$ is of the form $p(X|\hat{\eta})$, for a well-defined estimator $\hat{\eta} \in \mathcal{E}$. The most common method is the EM algorithm [10], which aims at maximizing the likelihood of the observed data, that is,

$$\max_{\eta \in \mathcal{E}} \ \textstyle\sum_{i=1}^N \log p(X_i|\eta). \tag{2}$$

More precisely, the EM algorithm is an iterative process to find the maximum likelihood (ML) estimate given observations $(X_i)_{i=1,...,N}$ associated to hidden variables $(h_i)_{i=1,...,N}$. It may be seen as the iterative construction of lower bounds of the log-likelihood function [11]. In the exponential family setting (1), we have, by Jensen's inequality, given the model defined by $\eta' \in \mathcal{E}$ from the previous iteration, and for any parameter $\eta \in \mathcal{E}$:

$$\log p(X_i|\eta) \geq \langle \phi(\eta), \mathbb{E}_{p(h_i|X_i,\eta')}\left[S(X_i, h_i)\right]\rangle - \psi(\eta) - C_i(\eta'),$$

for a certain constant $C_i(\eta')$, with equality if $\eta' = \eta$. Thus, EM-type algorithms build locally tight lower bounds of the log-likelihood in (2), which are equal to $\langle \phi(\eta), \sum_{i=1}^N s_i\rangle - N\psi(\eta) + \text{cst}$, for appropriate values of $s_i \in \mathbb{R}^d$ obtained by computing conditional expectations with the distribution of h_i given X_i for the current model defined by η' (E-step), i.e., $s_i = \mathbb{E}_{p(h_i|X_i,\eta')}\left[S(X_i, h_i)\right]$. Then this function of η is maximized to obtain the next iterate (M-step). In standard EM applications, these two steps are assumed tractable. In Section 3, we will only assume that the M-step is tractable while the E-step is intractable.

Standard EM will consider $s_i = \mathbb{E}_{p(h_i|X_i,\eta')}\left[S(X_i, h)\right]$ for the previous value of the parameter η for all i, and hence, at every iteration, all observations X_i, $i = 1, \ldots, N$ are considered for latent variable inference, leading to a slow "batch" algorithm for large N.

Stochastic approximation. Given our frequentist objective $\mathbb{E}_{t(X)} \log p(X|\eta)$ to maximize defined as an expectation, we may consider two forms of stochastic approximation [12], where observations X_i sampled from $t(X)$ are processed only once. The first one is stochastic gradient ascent, of the form $\eta_i = \eta_i + \gamma_i \frac{\partial \log p(X_i|\eta)}{\partial \eta}$, or appropriately renormalized version thereof, i.e., $\eta_i = \eta_i + \gamma_i H^{-1} \frac{\partial \log p(X_i|\eta)}{\partial \eta}$, with several possibilities for the $d \times d$ matrix H, such as the negative

Hessian of the partial or the full log-likelihood, or the negative covariance matrix of gradients, which can be seen as versions of natural gradient—see [13, 14, 15]. This either leads to slow convergence (without H) or expensive iterations (with H), with the added difficulty of choosing a proper scale and decay for the step-size γ_i.

A key insight from [14, 15] is to use a different formulation of stochastic approximation, *not explicitly based on stochastic gradient ascent*. Indeed, they consider the stationary equation $\mathbb{E}_{t(X)}\big[\frac{\partial \log p(X|\eta)}{\partial \eta}\big] = 0$ and expand it using the exponential family model (1), which leads to [see 1, 15, for details]:

$$\mathbb{E}_{t(X)}\left[\mathbb{E}_{p(h|X,\eta)}\left[S(X,h)\right]\right] = \mathbb{E}_{p(h,X|\eta)}\left[S(X,h)\right].$$

This stationary equation states that at optimality the sufficient statitics have the same expectation for the full model $p(h, X|\eta)$ and the joint "model/data" distribution $t(X)p(h|X, \eta)$.

Another important insight of [14, 15] is to consider the change of variable on sufficient statistics $s(\eta) = \mathbb{E}_{p(h,X|\eta)}\left[S(X,h)\right]$, which is equivalent to $\eta = \eta^*(s) \in \arg\max\, \langle \phi(\eta), s\rangle - \psi(\eta)$, (which is the usual M-step update). See [15] for detailed assumptions allowing this inversion. We may then rewrite the equation above as

$$\mathbb{E}_{t(X)}\big(\mathbb{E}_{p(h|X,\eta^*(s))}\left[S(X,h)\right]\big) = s.$$

This is a non-linear equation in $s \in \mathbb{R}^d$, with an expectation with respect to $t(X)$ which is only accessed through i.i.d. samples X_i, and thus a good candidate for the Robbins-Monro algorithm to solve stationary equations (and not to minimize functions) [12], which takes the simple form:

$$s_i = s_{i-1} - \gamma_i\big(s_{i-1} - \mathbb{E}_{p(h_i|X_i,\eta^*(s_{i-1}))}\left[S(X_i,h_i)\right]\big),$$

with a step-size γ_i. It may be rewritten as

$$\left\{\begin{array}{rcl} s_i &=& (1-\gamma_i)s_{i-1} + \gamma_i \mathbb{E}_{p(h_i|X_i,\eta_{i-1})}[S(X_i,h_i)] \\ \eta_i &=& \eta^*(s_i), \end{array}\right. \tag{3}$$

which has a particularly simple interpretation: instead of computing the expectation for all observations as in full EM, this stochastic version keeps tracks of old sufficient statistics through the variable s_{i-1} which is updated towards the current value $\mathbb{E}_{p(h_i|X_i,\eta_{i-1})}[S(X_i,h_i)]$. The parameter η is then updated to the value $\eta^*(s_i)$. [15] shows that this update is asymptotically equivalent to the natural gradient update with three main improvements: (a) no matrix inversion is needed, (b) the algorithm may be accelerated through Polyak-Ruppert averaging [16], i.e., using the average $\bar{\eta}_N$ of all η_i instead of the last iterate η_N, and (c) the step-size is particularly simple to set, as we are taking *convex* combinations of sufficient statistics, and hence only the decay rate of γ_i has to be chosen, i.e., of the form $\gamma_i = i^{-\kappa}$, for $\kappa \in (0, 1]$, without any multiplicative constant.

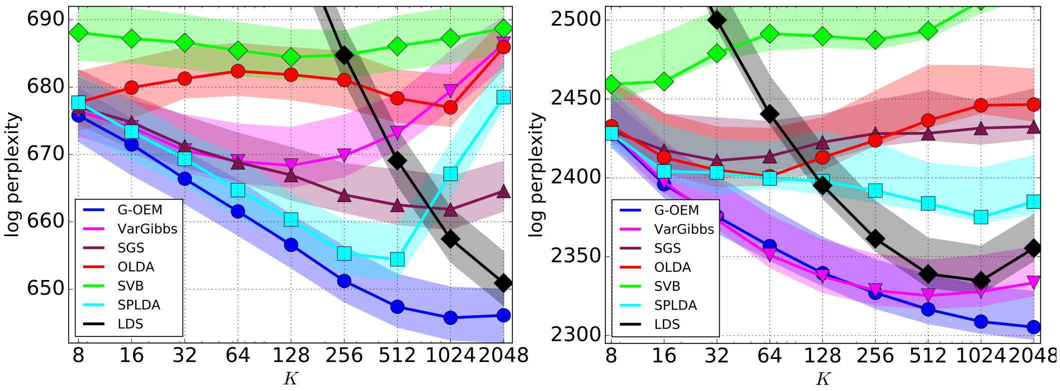
For the stepsize $\gamma_i = 1/i$, online EM (3) corresponds exactly to the incremental EM presented above [17].

3 Online EM with intractable models

The online EM updates in (3) lead to a scalable algorithm for optimization when the local E-step is tractable. However, in many latent variable models—e.g., LDA, hierarchical Dirichlet processes [19], or ICA [20]—it is intractable to compute the conditional expectation $\mathbb{E}_{p(h|X,\eta)}[S(X,h)]$.

Following [21], we propose to leverage the scalability of online EM updates (3) and locally approximate the conditional distribution $p(h|X, \eta)$ in the case this distribution is intractable to compute.

Sampling methods: **`G-OEM`.** MCMC methods to approximate the conditional distribution of latent variables with online EM have been considered by [21], who apply locally the Metropolis-Hasting (M-H) algorithm [22, 23], and show results on simple synthetic datasets. While Gibbs sampling is widely used for many models such as LDA due to its simplicity and lack of external parameters, M-H requires a proper proposal distribution with frequent acceptance and fast mixing, which may be hard to find in high dimensions. We provide a different simpler local scheme based on Gibbs sampling (thus adapted to a wide variety of models), and propose a thorough favorable comparison on synthetic and real datasets with existing methods.



(a) IMDB

(b) New York Times

Figure 1: Perplexity on 11 different test sets as a function of K, the number of topics inferred.

The Gibbs sampler is used to estimate posterior distributions by alternatively sampling parts of the variables given the other ones [see 24, for details], and is standard and easy to use in many common latent variable models. In the following, the online EM method with Gibbs estimation of the conditional distribution $p(h|X, \eta)$ is denoted `G-OEM`.

As mentioned above, the online EM updates correspond to a stochastic approximation algorithm and thus are robust to random noise in the local E-step. As a result, our sampling method is particularly adapted as it is a random estimate of the E-step—see a theoretical analysis by [21], and thus we only need to compute a few Gibbs samples for the estimation of $p(h|X_i, \eta_{i-1})$. A key contribution of our paper is to reuse sampling techniques that have proved competitive in the batch set-up and to compare them to existing variational approaches.

“Boosted” inference. As the variational and MCMC estimations of $p(h|X_i, \eta_{i-1})$ are done with iterative methods, we can boost the inference of the proposed algorithm by applying the update in the parameter η in (3) after each iteration of the estimation of $p(h|X_i, \eta_{i-1})$. In the context of LDA, this was proposed by [25] for incremental EM and we extend it to all versions of online EM. With this boost, we expect that the global parameters η converge faster, as they are updated more often. We empirically show that this boost only leads to significantly better results for variational method [see 1, for details].

4 Experiments for LDA model

We derive our algorithm for the particular model of LDA [4] [see 1, for complete derivation]. We evaluate our method by computing the likelihood on held-out documents, that is $p(X|\eta)$ for any test document X. For LDA, the likelihood is intractable to compute. We approximate $p(X|\eta)$ with the “left-to-right” evaluation algorithm [26] applied to each test document. In the following, we present results in terms of log-perplexity, defined as the opposite of the log-likelihood $-\log p(X|\eta)$. The lower the log-perplexity, the better the corresponding model. In our experiments, we compute the average test log-perplexity over test documents.

We compare seven different methods: `G-OEM` (our main algorithm): Gibbs online EM with step-size $\gamma_i = 1/\sqrt{i}$; `OLDA`: online LDA [27]; `VarGibbs`: Sparse stochastic inference for LDA [28]; `SPLDA`: single pass LDA [25]; `SGS`: streaming Gibbs sampling [29]; `LDS`: Stochastic gradient Riemannian Langevin dynamics sampler [30].

Results for IMDB (600,000 reviews from [31]) and New York Times (300,000 documents from UCI dataset [32]) are presented in Figure 1 [see 1, for further experiments].

All methods for the same problem are similar (in fact a few characters away from each other); ours is based on a proper stochastic approximation maximum likelihood framework, is empirically the most robust as it performs better for all experiments.

References

[1] C. Dupuy and F. Bach. Online but accurate inference for latent variable models with local gibbs sampling. *arXiv preprint arXiv:1603.02644*, 2016.

[2] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[3] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[5] H. Zhao, B. Jiang, and J. Canny. Same but different: Fast and high-quality Gibbs parameter estimation. *arXiv preprint arXiv:1409.5402*, 2014.

[6] F. Yan, N. Xu, and Y. Qi. Parallel inference for latent Dirichlet allocation on graphics processing units. In *Adv. NIPS*, 2009.

[7] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *JMLR*, 14(1):1303–1347, 2013.

[8] A. W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[9] E. L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 39(1):1–38, 1977.

[11] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[12] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Springer-Verlag, second edition, 2003.

[13] M. D. Titterington. Recursive parameter estimation using incomplete data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):257–267, 1984.

[14] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

[15] O. Cappé and E. Moulines. Online EM algorithm for latent data models. *Journal of the Royal Statistical Society*, 71(3):593–613, 2009.

[16] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[17] R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[18] J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *arXiv preprint arXiv:1402.4419*, 2014.

[19] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

[20] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[21] D. Rohde and O. Cappé. Online maximum-likelihood estimation for latent factor models. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE*. IEEE, 2011.

[22] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[23] K. W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[24] G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[25] I. Sato, K. Kurihara, and H. Nakagawa. Deterministic single-pass algorithm for LDA. *Adv. NIPS*, 2010.

[26] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. *Proc. ICML*, 2009.

[27] M. D. Hoffman, D. M. Blei, and F. R. Bach. Online learning for latent Dirichlet allocation. *Adv. NIPS*, 2010.

[28] D. Mimno, M. Hoffman, and D. Blei. Sparse stochastic inference for latent Dirichlet allocation. *Proc. ICML*, 2012.

[29] Y. Gao, J. Chen, and J. Zhu. Streaming gibbs sampling for LDA model. *arXiv preprint arXiv:1601.01142*, 2016.

[30] S. Patterson and Y. W. Teh. Stochastic gradient Riemannian langevin dynamics on the probability simplex. *Adv. NIPS*, 2013.

[31] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS). In *Proc. ACM SIGKDD*, 2014.

[32] M. Lichman. UCI machine learning repository, 2013.