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

Christophe Dupuy, Francis Bach

1 – Motivations

- More & more observed text data
- Topic models are suited for this type of data
- Existing inference methods are either:
  - Global and accurate (Gibbs sampling)
  - Online and approximate (variational inference)

Question: Can we go **online** and **accurate**?

2 – Contributions

- Draw explicit links between existing frequentist and Bayesian approaches,
- Adapt MCMC methods for online inference of latent variable models,
- Compare our approach on LDA to existing methods.
- Extensive set of experiments on synthetic and real datasets , where our new approach outperforms all existing methods.

3 – Latent variable models and online inference

Observations: $\mathbf{X} = (X_i)_{i=1,\dots,N}$; Hidden variables: $\mathbf{h} = (h_i)_{i=1,\dots,N}$

Parameters: $\eta \in \mathcal{E}$

Non-canonical exponential family models:

$$\boxed{p(X, h|\eta) = \exp\left[\langle \phi(\eta), S(X,h)\rangle - \psi(\eta)\right]}$$

Bayesian approach: $\eta \sim p(\eta) \quad \rightarrow \quad$ Approximation of $p(\eta|(X_i)_{i=1,\dots,N})$

Frequentist approach: $\eta \in \mathcal{E} \quad \rightarrow \quad$ ML: $\max_{\eta \in \mathcal{E}} \sum_{i=1}^{N} \log p(X_i|\eta)$

EM (Dempster et al., 1977):

$$\log p(\mathbf{X}|\eta) \geq \int_{\mathbf{h}} p(\mathbf{h}|\mathbf{X}, \eta_t) \log \frac{p(\mathbf{X},\mathbf{h}|\eta)}{p(\mathbf{h}|\mathbf{X},\eta_t)} dh = \mathbb{E}_{\mathbf{h}|\mathbf{X},\eta_t}[\log p(\mathbf{X},\mathbf{h}|\eta)] - C(\mathbf{X},\eta_t)$$

- E-step: $Q(\eta|\eta_t) = \mathbb{E}_{\mathbf{h}|\mathbf{X},\eta_t}[\log p(\mathbf{X},\mathbf{h}|\eta)] = \left\langle \phi(\eta), \sum_{i=1}^{N} \mathbb{E}_{h_i|X_i,\eta_t}[S(X_i,h_i)] \right\rangle - N\psi(\eta)$
- M-step: $\eta_{t+1} = \arg\max_{\eta} Q(\eta|\eta_t)$

Online EM (Cappé and Moulines, 2008):

- Assumption: $\eta^*(s) = \arg\max_{\eta \in \mathcal{E}} \left[\langle \phi(\eta), s\rangle - \psi(\eta)\right]$ exists.
- Robbins-Monro (1951) on:

$$\mathbb{E}_{t(X)}\left[\frac{\partial \log p(X|\eta)}{\partial \eta}\right] = 0 \Leftrightarrow \mathbb{E}_{t(X)}\left[\mathbb{E}_{h|X,\eta^*(s)}[S(X,h)]\right] = s$$

- Online EM algorithm:

$$\boxed{\begin{cases} s_{t+1} &= (1-\rho_t)s_t + \rho_t \mathbb{E}_{h_{i_{t+1}}|X_{i_{t+1}},\eta_t}[S(X_{i_{t+1}}, h_{i_{t+1}})] \\ \eta_{t+1} &= \eta^*(s_{t+1}) \end{cases}}$$

What to do when $\mathbb{E}_{h_{i_{t+1}}|X_{i_{t+1}},\eta_t}[S(X_{i_{t+1}}, h_{i_{t+1}})]$ is intractable?

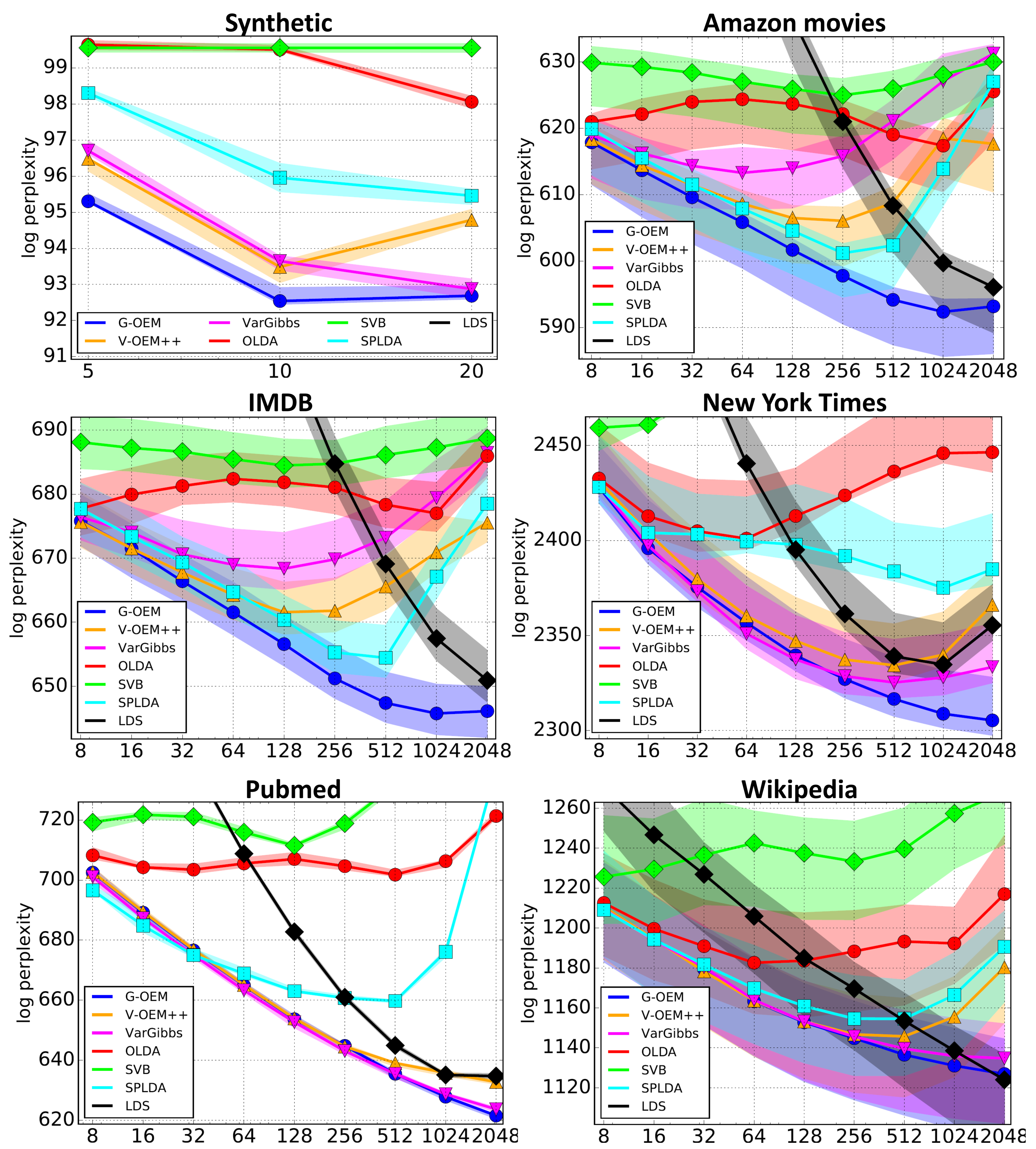
1. Variational EM: Approximate $p(h|X,\eta)$ with variational $q(h,\eta)$
   a) Solve: $\hat{q} = \arg\min_q \mathrm{KL}\left[q(h,\eta)||p(h,\eta|X)\right]$
   b) Approximation: $\mathbb{E}_{p(h|X,\eta)}[S(X,h)] \approx \mathbb{E}_{\hat{q}(h|\eta)}[S(X,h)]$
2. Gibbs Online EM:
   a) Draw P samples $h|X,\eta$ with Gibbs sampling
   b) Approximation: $\hat{p}(h|X,\eta)$ with the P samples

3 – Evaluation for LDA (Blei et al., 2003)

- Perplexity: $LP(X) = -\log p(X|\eta)$
  - Intractable to compute for LDA
  - Estimation with left-to-right algorithm (Wallach et al., 2009).
  - Evaluation measure on test set: $\boxed{TP = \frac{1}{N_t}\sum_{i=1}^{N_t} LP(X_i)}$
- Existing Methods:
  - Stochastic variational inference (Hoffman et al., 2013):
    - **OLDA** (Hoffman et al., 2010); **SVB** (Broderick et al., 2013);
  - Variational online EM:
    - **SPLDA** (Sato et al., 2010); **V-OEM++** (our method);
  - Gibbs online EM: **G-OEM** (our method);
  - Hybrid / other:
    - **VarGibbs** (Mimno et al., 2012); **LDS** (Patterson and Teh, 2013).
- Datasets:

Dataset	#documents	N_X	#words
Synthetic	1,000,000	60	1,000
Wikipedia	1,010,000	162.3	7702
IMDB	614,589	82.2	10,000
Amazon movies	338,565	75.4	10,000
New York Times	299,877	287.4	44,228
Pubmed	2,100,000	82.0	113,568

4 – Results



G-OEM

#	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
1	violence	comedy	romantic	bad	brilliant	didn't	narrative	town
2	violent	funny	comedy	worst	perfect	i've	ultimately	local
3	disturbing	laugh	love	waste	beautiful	wasn't	cinematic	mountain
4	brutal	joke	romance	boring	masterpiece	isn't	approach	village
5	murder	hilarious	funny	awful	amazing	we're	protagonist	location
6	graphic	comic	charming	poor	superb	i'll	seemingly	road
7	killer	comedic	chemistry	dialogue	stunning	couldn't	nature	journey
8	torture	steve	sweet	worse	wonderful	wouldn't	tone	group
9	victim	amusing	enjoy	dull	absolutely	pretty	filmmaker	travel
10	rape	fun	heart	fail	best	bad	whose	country

OLDA

#	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
1	horror	hilarious	love	bad	great	wrong	visual	young
2	night	romance	enjoy	action	best	didn't	focus	family
3	dead	clever	pretty	interesting	star	i've	reality	child
4	twist	smart	old	original	long	wait	difficult	father
5	scary	intriguing	funny	far	john	catch	filmmaker	son
6	effective	comedic	comedy	special	excellent	exactly	image	age
7	mysterious	funniest	fun	fight	classic	wasn't	narrative	whose
8	bloody	progress	hard	hero	beautiful	huge	intelligent	tale
9	ghost	sport	perfect	entire	drama	i'll	accept	discover
10	haunt	site	laugh	half	wonderful	choice	impression	easily

5 – Conclusion & Future Work

- How to deal with intractable latent variable models?
  - Approximation by sampling or variational
  - Apply online EM
- Results on LDA:
  - Gap between frequentist and Bayesian methods
  - Accurate estimation (Gibbs) > variational approximation

- Explore other models (e.g., HDP)
- Explore distributed settings (Yan et al., 2009)
- Explore larger (constant?) step-sizes (Bach et Moulines, 2013)