Incremental Variational Inference

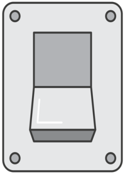
Applied to Latent Dirichlet Allocation

Cédric Archambeau
cedrica@amazon.com

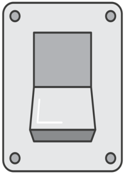
Joint work with Beyza Ermiş (Bogazici University).

NIPS workshop on Advances in Approximate Bayesian Inference
Montréal, December 2015.

Democratising (probabilistic) machine learning

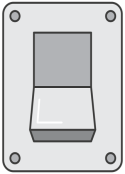


Democratising (probabilistic) machine learning



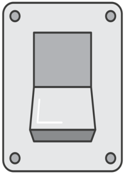
- Abstract away algorithms (scalable & practical)

Democratising (probabilistic) machine learning



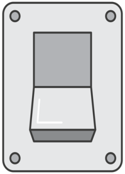
- Abstract away algorithms (scalable & practical)
  - Probabilistic programming language
  - Bayesian optimisation

Democratising (probabilistic) machine learning



- Abstract away algorithms (scalable & practical)
  - Probabilistic programming language
  - Bayesian optimisation
- Abstract away feature engineering

Democratising (probabilistic) machine learning



- Abstract away algorithms (scalable & practical)
  - Probabilistic programming language
  - Bayesian optimisation
- Abstract away feature engineering

- Abstract away memory constraints
- Abstract away network constraints
- Abstract away computing infrastructure

Variational inference (Beal, 2003)

Variational inference (Beal, 2003)

- The goal is to maximise the evidence: $p(\mathbf{X})$.

Variational inference (Beal, 2003)

- The goal is to maximise the evidence: $p(\mathbf{X})$.
- Using *Jensen's inequality*, we get for any distribution $q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\theta})$:

$$
\begin{aligned}
\ln p(\mathbf{X}) &= \ln \iint p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})\, d\mathbf{Z}\, d\boldsymbol{\theta} \\
&\geqslant \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\theta})}\, d\mathbf{Z}\, d\boldsymbol{\theta} \\
&= \ln p(\mathbf{X}) - \mathrm{KL}[q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\theta}) \| p(\mathbf{Z}, \boldsymbol{\theta} | \mathbf{X})] \triangleq -\mathcal{F}(\mathbf{w}).
\end{aligned}
$$

Variational inference (Beal, 2003)

- The goal is to maximise the evidence: $p(\mathbf{X})$.
- Using *Jensen's inequality*, we get for any distribution $q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\theta})$:

$$\begin{aligned}
\ln p(\mathbf{X}) &= \ln \iint p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})\, d\mathbf{Z}\, d\boldsymbol{\theta} \\
&\geqslant \iint q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})}{q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\theta})}\, d\mathbf{Z}\, d\boldsymbol{\theta} \\
&= \ln p(\mathbf{X}) - \mathrm{KL}[q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\theta}) \| p(\mathbf{Z}, \boldsymbol{\theta} | \mathbf{X})] \triangleq -\mathcal{F}(\mathbf{w}).
\end{aligned}$$

- A tractable solution is found by assuming $q_{\mathbf{w}}$ factorises given the data:

$$q_{\mathbf{w}}(\mathbf{Z}, \boldsymbol{\theta}) = \prod_n q(\mathbf{z}_n; \mathbf{w}_n) \times \prod_m q(\boldsymbol{\theta}_m; \mathbf{w}_m).$$

Mean field variational inference (MVI)

$$\mathbf{w}_n \leftarrow \arg\max_{\mathbf{w}_n} \quad \langle \ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})\rangle - \mathrm{KL}\left[q(\mathbf{z}_n; \mathbf{w}_n) \| p(\mathbf{z}_n)\right],$$

$$\mathbf{w}_m \leftarrow \arg\max_{\mathbf{w}_m} \quad \sum_n \langle \ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})\rangle - \mathrm{KL}\left[q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right].$$

Mean field variational inference (MVI)

$$\mathbf{w}_n \leftarrow \arg\max_{\mathbf{w}_n} \quad \langle \ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})\rangle - \mathrm{KL}\left[q(\mathbf{z}_n; \mathbf{w}_n) \| p(\mathbf{z}_n)\right],$$

$$\mathbf{w}_m \leftarrow \arg\max_{\mathbf{w}_m} \quad \sum_n \langle \ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta})\rangle - \mathrm{KL}\left[q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right].$$

- Monotonic increase of the bound; converges to local maximum.
- Priors are conjugate to the likelihood; updates are similar to Gibbs.
- Batch method; not suitable for large data sets.
- Block-coordinate ascent.

Stochastic variational inference (SVI) (Hoffman, et al., NIPS 2010)

Let $\ell_n(\mathbf{w}) = \langle \ln p(\mathbf{x}_n | \mathbf{z}_n, \boldsymbol{\theta}) \rangle$:

$$\mathbf{w}_m \leftarrow \mathbf{w}_m + \rho_t \ \arg\max_{\mathbf{w}_m} \ N\ell_n(\mathbf{w}) - \mathrm{KL}\left[q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m) \right],$$

where $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$.

Stochastic variational inference (SVI) (Hoffman, et al., NIPS 2010)

Let $\ell_n(\mathbf{w}) = \langle \ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta}) \rangle$:

$$\mathbf{w}_m \leftarrow \mathbf{w}_m + \rho_t \ \arg\max_{\mathbf{w}_m} \ N\ell_n(\mathbf{w}) - \mathrm{KL}\left[q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right],$$

where $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$.

- Noisy, but unbiased estimates of the gradients wrt $\mathbf{w}_m$.
- Monotonic increase of bound is lost – no sanity check
- Small memory footprint; sequential method.
- Requires adjusting the learning rate.
- Natural gradients wrt $q_{\mathbf{w}_m}$

Incremental variational inference (IVI)

Let $\ell_N(\mathbf{w}) = \sum_n \langle \ln p(\mathbf{x}_n | \mathbf{z}_n, \boldsymbol{\theta}) \rangle$ and $\mathbf{s}(\mathbf{X}, \mathbf{Z}) = \sum_n \mathbf{s}_n(\mathbf{x}_n, \mathbf{z}_n)$ be the vector of sufficient statistics:

$$\mathbf{w}_m \leftarrow \arg\max_{\mathbf{w}_m} \; \ell_N(\mathbf{s}, \mathbf{w}) - \ell_n(\mathbf{s}_n, \mathbf{w}) + \ell_n(\mathbf{s}_n^*, \mathbf{w}) - \mathrm{KL}\left[q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right].$$

where $\mathbf{s}_n^*(\mathbf{x}_n, \mathbf{z}_n)$ is the new vector of sufficient statistics.

Incremental variational inference (IVI)

Let $\ell_N(\mathbf{w}) = \sum_n \langle \ln p(\mathbf{x}_n | \mathbf{z}_n, \boldsymbol{\theta}) \rangle$ and $\mathbf{s}(\mathbf{X}, \mathbf{Z}) = \sum_n \mathbf{s}_n(\mathbf{x}_n, \mathbf{z}_n)$ be the vector of sufficient statistics:

$$\mathbf{w}_m \leftarrow \arg\max_{\mathbf{w}_m} \; \ell_N(\mathbf{s}, \mathbf{w}) - \ell_n(\mathbf{s}_n, \mathbf{w}) + \ell_n(\mathbf{s}_n^*, \mathbf{w}) - \mathrm{KL}\left[q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right].$$

where $\mathbf{s}_n^*(\mathbf{x}_n, \mathbf{z}_n)$ is the new vector of sufficient statistics.

- Monotonic increase of bound is recovered!
- Need for storing the sufficient statistics.
- Sequential, but maintains a batch estimate of $\mathbf{s}(\mathbf{X}, \mathbf{Z})$.
- No parameters to tune.

Incremental variational inference (IVI)

Let $\ell_N(\mathbf{w}) = \sum_n \langle \ln p(\mathbf{x}_n|\mathbf{z}_n, \boldsymbol{\theta}) \rangle$ and $\mathbf{s}(\mathbf{X}, \mathbf{Z}) = \sum_n \mathbf{s}_n(\mathbf{x}_n, \mathbf{z}_n)$ be the vector of sufficient statistics:

$$\mathbf{w}_m \leftarrow \arg\max_{\mathbf{w}_m} \ \ell_N(\mathbf{s}, \mathbf{w}) - \ell_n(\mathbf{s}_n, \mathbf{w}) + \ell_n(\mathbf{s}_n^*, \mathbf{w}) - \mathrm{KL}\left[q(\boldsymbol{\theta}_m; \mathbf{w}_m) \| p(\boldsymbol{\theta}_m)\right].$$

where $\mathbf{s}_n^*(\mathbf{x}_n, \mathbf{z}_n)$ is the new vector of sufficient statistics.

- Monotonic increase of bound is recovered!
- Need for storing the sufficient statistics.
- Sequential, but maintains a batch estimate of $\mathbf{s}(\mathbf{X}, \mathbf{Z})$.
- No parameters to tune.
- Can be interpretted as stochastic average gradient descent (SAG).

Relation to incremental EM

(Neal & Hinton, 1993)

$$\begin{aligned}\ln p(\mathbf{X}) &= \ln \iint p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})\, d\mathbf{Z}\, d\boldsymbol{\theta} \\ &\geqslant \iint q(\mathbf{Z})q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})}{q(\mathbf{Z})q(\boldsymbol{\theta})}\, d\mathbf{Z}\, d\boldsymbol{\theta} \\ &= \ln p(\mathbf{X}) - \mathrm{KL}[q(\mathbf{Z})q(\boldsymbol{\theta}) \| p(\mathbf{Z}, \boldsymbol{\theta} | \mathbf{X})] \triangleq -\mathcal{F}(\mathbf{w}).\end{aligned}$$

- MVI updates can be re-written as follows:

$$\begin{aligned} q(\mathbf{z}_n; \mathbf{w}_n) &\propto \exp\left(\langle \ln p(\mathbf{s}_n | \boldsymbol{\theta}) \rangle\right), \\ q(\boldsymbol{\theta}_m; \mathbf{w}_m) &\propto \exp\left(\langle \ln p(\mathbf{s} | \boldsymbol{\theta}) \rangle_{\neg \boldsymbol{\theta}_m}\right) p(\boldsymbol{\theta}_m). \end{aligned}$$

Relation to incremental EM (Neal & Hinton, 1993)

$$\begin{aligned}\ln p(\mathbf{X}) &= \ln \iint p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})\; d\mathbf{Z}\; d\boldsymbol{\theta} \\ &\geqslant \iint q(\mathbf{Z})q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\theta})}{q(\mathbf{Z})q(\boldsymbol{\theta})}\; d\mathbf{Z}\; d\boldsymbol{\theta} \\ &= \ln p(\mathbf{X}) - \mathrm{KL}[q(\mathbf{Z})q(\boldsymbol{\theta}) \| p(\mathbf{Z}, \boldsymbol{\theta} | \mathbf{X})] \triangleq -\mathcal{F}(\mathbf{w}).\end{aligned}$$

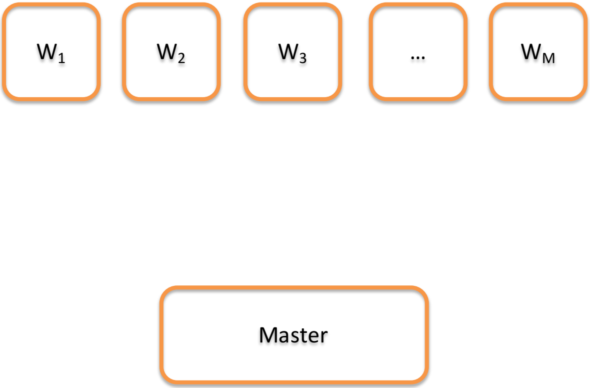
- MVI updates can be re-written as follows:

$$\begin{aligned}q(\mathbf{z}_n; \mathbf{w}_n) &\propto \exp\left(\langle \ln p(\mathbf{s}_n | \boldsymbol{\theta}) \rangle\right), \\ q(\boldsymbol{\theta}_m; \mathbf{w}_m) &\propto \exp\left(\langle \ln p(\mathbf{s} | \boldsymbol{\theta}) \rangle_{\neg \boldsymbol{\theta}_m}\right) p(\boldsymbol{\theta}_m).\end{aligned}$$

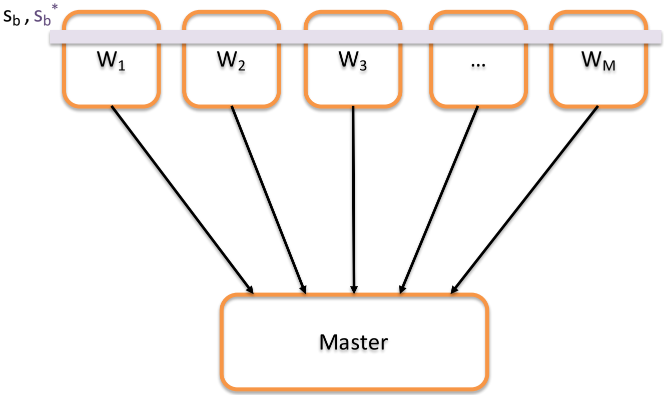
- IVI updates can be re-written as follows:

$$\begin{aligned}q(\mathbf{z}_n; \mathbf{w}_n) &\propto \exp\left(\langle \ln p(\mathbf{s}_n^* | \boldsymbol{\theta}) \rangle\right), \\ q(\boldsymbol{\theta}_m; \mathbf{w}_m) &\propto \exp\left(\langle \ln p(\mathbf{s} - \mathbf{s}_n + \mathbf{s}_n^*, \boldsymbol{\theta}) \rangle_{\neg \boldsymbol{\theta}_m}\right).\end{aligned}$$

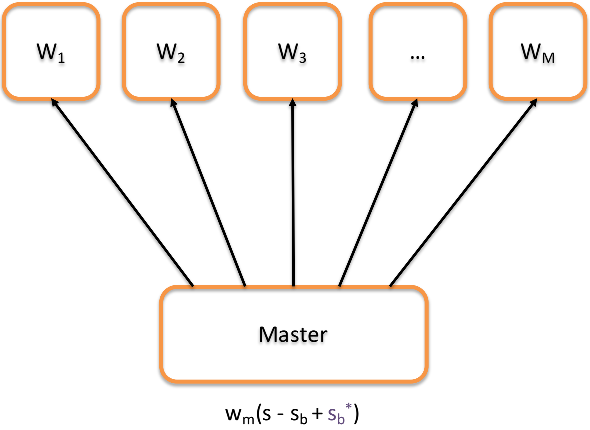
Distributed version



Distributed version



Distributed version

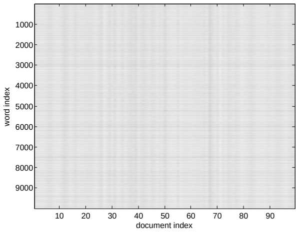


Latent Dirichlet allocation (LDA) (Blei, et al., JMLR 2003)

Simple generative model for text, based on a bag-of-words representation:

Latent Dirichlet allocation (LDA) (Blei, et al., JMLR 2003)

Simple generative model for text, based on a bag-of-words representation:

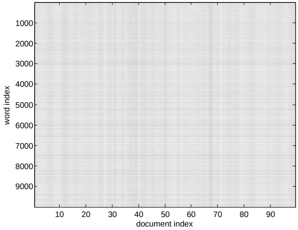


Observations are word counts per document. LDA assumes an admixture model:

$$\mathbf{X} \in \mathbb{N}^{V \times D}.$$

Latent Dirichlet allocation (LDA) (Blei, et al., JMLR 2003)

Simple generative model for text, based on a bag-of-words representation:



Observations are word counts per document. LDA assumes an admixture model:

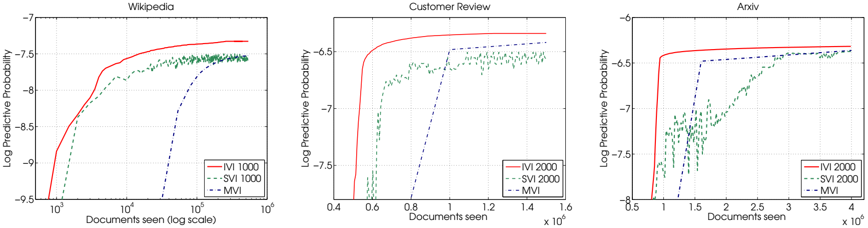
$$\mathbf{X} \in \mathbb{N}^{V \times D}.$$

LDA infers a low-rank approximation of the matrix of counts:

$$\mathrm{E}(\mathbf{X}) \approx \mathbf{\Phi}\mathbf{\Theta}^\top, \qquad \mathbf{x}_d \sim \mathrm{Multinomial}(\mathbf{\Phi}\boldsymbol{\theta}_d, N_d)$$

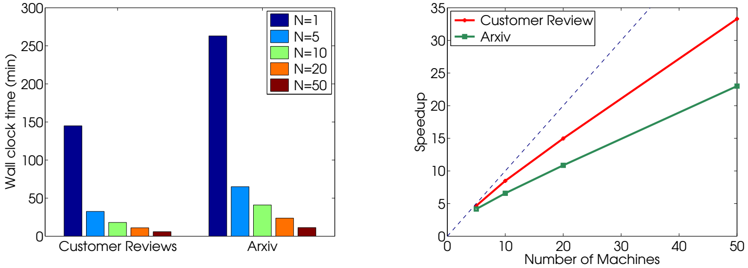
where $\mathbf{\Phi} \in \mathbb{R}_+^{V \times K}$, $\mathbf{\Theta} \in \mathbb{R}_+^{D \times K}$ and K is small.

Log-predictive probability for LDA as a function of the number of processed documents



IVI converges faster and to a higher value on all considered datasets. (K=100, $\alpha_0 = 0.5$ and $\beta_0 = 0.05$)

Wall-clock time comparisons and speed-up at MVI performance



Left: Wall-clock time (in minutes) comparisons for D-IVI for different number of machines on Arxiv and Customer Review. *Right:* Speed-up results of D-IVI for varying number of machines with respect to single machine.

Effect of the number of topics

Table 3: Log-prediction-probability (LPP) and runtime (in terms of minutes per iteration) of the IVI for different number of topics and number of processors (mini-batch size = 2000).

Datasets	Customer Review						Arxiv					
Number of Topics		Number of Machines						Number of Machines				
		1	5	10	20	50		1	5	10	20	50
25	LPP	-6.46	-6.46	-6.46	-6.46	-6.46	LPP	-6.57	-6.57	-6.57	-6.57	-6.57
	Time	138	31.6	16.7	10.8	5.3	Time	224	61	37	21.6	10.1
50	LPP	-6.33	-6.33	-6.33	-6.33	-6.33	LPP	-6.42	-6.42	-6.42	-6.42	-6.42
	Time	145	32.5	18	11	5.9	Time	263	65	41	23.7	11.3
100	LPP	-6.29	-6.29	-6.29	-6.29	-6.29	LPP	-6.33	-6.33	-6.33	-6.33	-6.33
	Time	148	33.2	18.6	11.5	6.1	Time	268	68	43	24.5	11.7
200	LPP	-6.49	-6.49	-6.49	-6.49	-6.49	LPP	-6.46	-6.46	-6.46	-6.46	-6.46
	Time	159	35.4	19.5	11.9	6.3	Time	297	73.7	46.2	26.8	12.8
1000	LPP	-6.84	-6.84	-6.84	-6.84	-6.84	LPP	-6.97	-6.97	-6.97	-6.97	-6.97
	Time	167	37.3	21.2	12.4	6.7	Time	306	78	49	28.2	13.4

Conclusion

- Distributed inference framework
- Monotonic increase of the bound
- Free of learning parameters
- Memory requirements scale linearly with the number of mini-batches
- Applicable to other data models

http://arxiv.org/abs/1507.05016