Stochastic Variational Inference

Amir Mohammad Esmaieeli Sikaroudi, amesmaieeli@arizona.edu

University of Arizona, CSC696

September 26, 2022

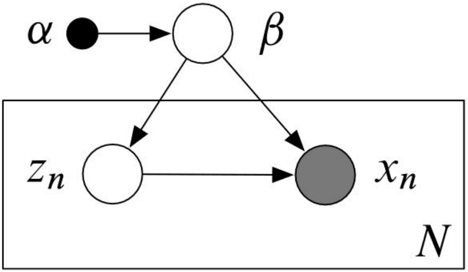
Outline

- Variational Inference
- Mean-Field Variational Inference
- Stochastic Variational Inference
- Topic model with latent Dirichlet allocation (LDA)
- Topic model with hierarchical Dirichlet process (HDP)
- Results

Variational Inference

- Local hidden variables (z_n)
- Global hidden variables (β)
- Observations x_n

The distinction between local and global hidden variables is determined by the conditional dependencies. The n-th observation x_n and the n-th local variable z_n are conditionally independent, given global variables β, of all other observations and local hidden variables.



Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Variational Inference

The joint distribution is factorized below

$$p(x, z, \beta|\alpha) = p(\beta|\alpha) \prod_{n=1}^{N} p(x_n, z_n|\beta)$$

The goal is to approximate the posterior distribution of the hidden variables given observations

$$p(\beta, z|x)$$

Each observation is also independent from other observations

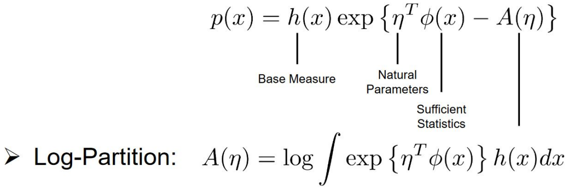
$$p(x_n, z_n|x_{-n}, z_{-n}, \beta, \alpha) = p(x_n, z_n|\beta, \alpha)$$

Variational Inference

Exponential family distribution of hidden variables

$$p(\beta|x, z, \alpha) = h(\beta)exp\{\eta_g(x, z, \alpha)^T t(\beta) - a_g(\eta_g(x, z, \alpha))\}$$
$$p(z_{nj}|x_n, z_{n,-j}, \beta) = h(z_{nj})exp\{\eta_\ell(x_n, z_{n,-j}, \beta)^T t(z_{nj}) - a_\ell(\eta_\ell(x_n, z_{n,-j}, \beta))\}$$

Slide from lecture



Prior is also from exponential family

$$p(\beta) = h(\beta)exp\{\alpha^T t(\beta) - a_g(\alpha)\}$$

Mean-Field Variational Inference

- Approximate posterior which hidden variables are independent
- Minimizing Kullback-Leibler (KL) divergence. Why KL divergence?
  - $\mathcal{N}(0, 10000)$ and $\mathcal{N}(10, 10000)$ have 10 difference in parameter but almost 0 difference in probability
  - $\mathcal{N}(0, 0.001)$ and $\mathcal{N}(0.1, 0.001)$ have 0.1 difference in parameter but almost 0 overlap
- Maximizing Evidence Lower Bound (ELBO) (derived by introducing distribution q and Jensen's inequality)

$$KL(q(z,\beta)||p(z,\beta|x)) = \mathbb{E}_q[log\, q(z,\beta)] - \mathbb{E}_q[log\, p(x,z,\beta)] + log\, p(x)$$
$$= -\mathcal{L}(q) + const$$

$\mathcal{L}(q)$ is ELBO
Assuming the natural parameters for global and local variables are λ and ϕ_{nj}. By independence assumption of Mean-Field Variational Inference we have

$$q(z,\beta) = q(\beta|\lambda)\prod_{n=1}^{N}\prod_{j=1}^{J} q(z_{nj}|\phi_{nj})$$
$$q(\beta|\lambda) = h(\beta)exp\{\lambda^T t(\beta) - a_g(\lambda)\}$$
$$q(z_{nj}|\phi_{nj}) = h(z_{nj})exp\{\phi_{nj}^T t(z_{nj}) - a_\ell(\phi_{nj})\}$$

Mean-Field Variational Inference

Rewriting ELBO by λ and applying abbreviations of $q(z_{nj})$ instead of $q(z_nj|\phi_{nj})$ and $q(\beta)$ instead of $q(\beta|\lambda)$

$$\mathcal{L}(\lambda) = \mathbb{E}_q[log\, p(\beta|x,z)] - \mathbb{E}_q[log\, q(\beta)] + const$$

Based on exponential family properties, the expectation of the sufficient statistics is the gradient of log normalizer

$$\mathbb{E}_q[t(\beta)] = \nabla_\lambda a_g(\lambda)$$

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\eta_g(x,z,\alpha)]^T \nabla_\lambda a_g(\lambda) - \lambda^T \nabla_\lambda a_g(\lambda) + a_g(\lambda) + const$$

Mean-Field Variational Inference

Now we get derivative of ELBO by λ

$$\nabla_\lambda \mathcal{L} = \nabla_\lambda{}^2 a_g(\lambda)(\mathbb{E}_q[\eta_g(x, z, \alpha)] - \lambda)$$

After setting the derivative to zero we have

$$\lambda = \mathbb{E}_q[\eta_g(x, z, \alpha)] \text{ (M step in EM)}$$

Same derivative is taken for ϕ_{nj} (skipping the construction of $\mathcal{L}(\phi_{nj})$)

$$\nabla_{\phi_{nj}} \mathcal{L} = \nabla^2_{\phi_{nj}} a_\ell(\phi_{nj})(\mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)] - \phi_{nj})$$

Setting the derivative to zero we have

$$\phi_{nj} = \mathbb{E}_q[\eta_\ell(x_n, z_{n,-j}, \beta)] \text{ (E step in EM)}$$

Mean-Field Variational Inference

To update the natural parameters, coordinate ascent is used to optimize ELBO

```
Initialize λ^(0) randomly.
repeat
    for each local variational parameter φ_nj do
        Update φ_nj, φ_nj^(t) = E_{q^(t-1)}[η_{ℓ,j}(x_n, z_{n,-j}, β)].
    end for
    Update the global variational parameters, λ^(t) = E_{q^(t)}[η_g(z_{1:N}, x_{1:N})]
until the ELBO converges
```

Initialize $\lambda^{(0)}$ randomly.

repeat

  for each local variational parameter ϕ_{nj} **do**

    Update ϕ_{nj}, $\phi_{nj}^{(t)} = \mathbb{E}_{q^{(t-1)}}[\eta_{\ell,j}(x_n, z_{n,-j}, \beta)]$.

  end for

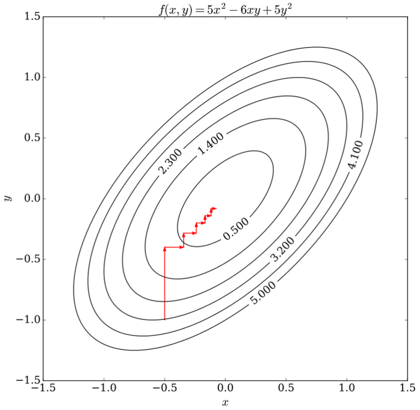
  Update the global variational parameters, $\lambda^{(t)} = \mathbb{E}_{q^{(t)}}[\eta_g(z_{1:N}, x_{1:N})]$

until the ELBO converges

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Mean-Field Variational Inference

To update the natural parameters, coordinate ascent is used to optimize ELBO



Source: Wikipedia, Coordinate descent, 2022, `https://en.wikipedia.org/wiki/Coordinate_descent`

Stochastic Variational Inference

Instead of evaluating the local variables (E step) for all dataset, one data is sampled uniformly from the dataset and the representative local variable is updated and for the global variable it's like the data point is repeated N times. The general idea is to apply Robbins-Monro algorithm on the M step (global variable λ) which the E step (local variables π_{nj}) is noisy.

$$\lambda^{(t)} = \lambda^{t-1} + \rho_t b_t(\lambda^{(t-1)})$$

p_t is the step size and b_t is an independent draw from noisy gradient.
Based on Robbins-Monro algorithm, the step size must satisfy two conditions to guarantee convergence

- $\sum_{t=0}^{\infty} \rho_t = \infty$
- $\sum_{t=0}^{\infty} \rho_t^2 < \infty$
- The updating steps (ρ) follows the following equation with forgetting rate κ and delay factor of τ which down-weights early iterations

$$\rho_t = (t + \tau)^{-\kappa}$$

Stochastic Variational Inference

The algorithm

1: Initialize $\lambda^{(0)}$ randomly.
2: Set the step-size schedule ρ_t appropriately.
3: **repeat**
4: Sample a data point x_i uniformly from the data set.
5: Compute its local variational parameter,

$$\phi = \mathbb{E}_{\lambda^{(t-1)}}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

6: Compute intermediate global parameters as though x_i is replicated N times,

$$\hat{\lambda} = \mathbb{E}_{\phi}[\eta_g(x_i^{(N)}, z_i^{(N)})].$$

7: Update the current estimate of the global variational parameters,

$$\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \rho_t\hat{\lambda}.$$

8: **until** forever

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Stochastic Variational Inference

Extensions

Using a batch instead of one data point will perform better (also shown in results)

$$\lambda^{(t)} = (1 - \rho_t)\lambda^{t-1} + \frac{\rho_t}{S} \sum_s \hat{\lambda}_s$$

Estimation of hyperparameters can be done in ELBO update step simultaneously with λ (global variables)

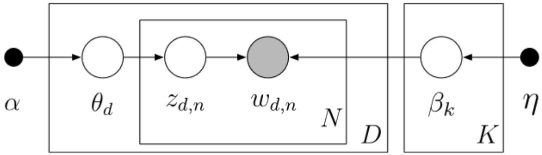
$$\alpha^{(t)} = \alpha^{(t-1)} + \rho_t \nabla_\alpha \mathcal{L}_t(\lambda^{(t-1)}, \phi, \alpha^{(t-1)})$$

Topic model with latent Dirichlet allocation (LDA)

Notations

- n-th word and d-th documents is w_{dn}
- there are V vocabulary terms
- β_k is a distribution over the vocabulary. β_{kw} is the w-th entry in k-th topic.
- there are K topics
- θ_d is a distribution over topics in a K-1 simplex
- each word is assumed to be drawn from a single topic with assignment variable z_{dn}

Topic model with latent Dirichlet allocation (LDA)



Var	Type	Conditional	Param	Relevant Expectations
z_{dn}	Multinomial	$\log\theta_{dk} + \log\beta_{k,w_{dn}}$	ϕ_{dn}	$\mathbb{E}[Z_{dn}^k] = \phi_{dn}^k$
θ_d	Dirichlet	$\alpha + \sum_{n=1}^{N} z_{dn}$	γ_d	$\mathbb{E}[\log\theta_{dk}] = \Psi(\gamma_{dk}) - \sum_{j=1}^{K}\Psi(\gamma_{dj})$
β_k	Dirichlet	$\eta + \sum_{d=1}^{D}\sum_{n=1}^{N} z_{dn}^k w_{dn}$	λ_k	$\mathbb{E}[\log\beta_{kv}] = \Psi(\lambda_{kv}) - \sum_{y=1}^{V}\Psi(\lambda_{ky})$

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Topic model with latent Dirichlet allocation (LDA)

Generative process of LDA

- Draw topics $\beta_k \sim Dirichlet(\eta ... \eta)$ for $k \in 1 ... K$
- For each document $d \in 1 ... D$
  - Draw topic proportions $\theta \sim Dirichlet(\alpha, ..., \alpha)$
  - For each word $w \in 1 ... N$
    - Draw topic assignment $z_{dn} \sim Multinomial(\theta_d)$
    - Draw word $w_{dn} \sim Multinomial(\beta_{z_{dn}})$

Topic model with latent Dirichlet allocation (LDA)

Initialize $\lambda^{(0)}$ randomly.
Set the step-size schedule ρ_t appropriately.
repeat
  Sample a document w_d uniformly from the data set.
  Initialize $\gamma_{dk} = 1$, for $k \in \{1, \ldots, K\}$.
  repeat
    For $n \in \{1, \ldots, N\}$ set

$$\phi_{dn}^{k} \propto \exp\left\{\mathbb{E}[\log \theta_{dk}] + \mathbb{E}[\log \beta_{k,w_{dn}}]\right\}, k \in \{1, \ldots, K\}.$$

    Set $\gamma_d = \alpha + \sum_n \phi_{dn}$.
  until local parameters ϕ_{dn} and γ_d converge.
  For $k \in \{1, \ldots, K\}$ set intermediate topics

$$\hat{\lambda}_k = \eta + D \sum_{n=1}^{N} \phi_{dn}^{k} w_{dn}.$$

  Set $\lambda^{(t)} = (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$.
until forever

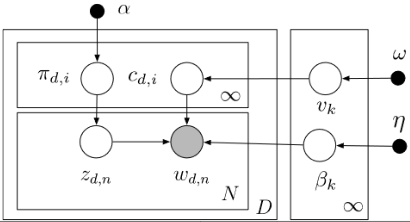
Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Topic model with hierarchical Dirichlet process (HDP)

A limitation of LDA is that the number of topics needs to be known in advance. HDP creates new topics (potentially infinite number of topics) by combination of predefined topics.
Dirichlet Process is a Bayesian nonparametric model used to create topics.
Dirichlet process is a distribution of distributions. In HDP the distribution of topics (θ_d) is a point in an infinite simplex
Generative process of HDP

- Draw an infinite number of topics, $\beta_k \sim Dirichlet(\eta) for k \in 1, 2, 3...$
- Draw corpus breaking proportions, $v_k \sim Beta(1, \omega) for k \in 1, 2, 3, ...$
- For each document d:
  - Draw document-level topic indices, $c_{di} \sim Multinomial(\sigma(v)) for i \in 1, 2, 3, ...$
  - Draw document breaking proportions, $\pi_{di} \sim Beta(1, \alpha) for i \in 1, 2, 3, ...$
  - For each word n:
    - Draw topic assignment $z_{dn} \sim Multinomial(\sigma(\pi_d))$
    - Draw word $w_n \sim Multinomial(\beta_{c_{d,z_{dn}}})$

Topic model with hierarchical Dirichlet process (HDP)



Var	Type	Conditional	Param	Relevant expectation
z_{dn}	Multinomial	$\log\sigma_i(\pi_d)+\sum_{k=1}^{\infty} c_{di}^k \log\beta_{k,w_{dn}}$	ϕ_{dn}	$\mathbb{E}[Z_{dn}^i]=\phi_{dn}^i$
π_{di}	Beta	$(1+\sum_{n=1}^N z_{dn}^i,\ \alpha+\sum_{n=1}^N\sum_{j=i+1}^{\infty} z_{dn}^j)$	$(\gamma_{di}^{(1)},\gamma_{di}^{(2)})$	(Expectations are similar to those for v_k.)
c_{di}	Multinomial	$\log\sigma_k(V)+\sum_{n=1}^N z_{dn}^i \log\beta_{k,w_{dn}}$	ζ_{di}	$\mathbb{E}[c_{di}^k]=\zeta_{di}^k$
v_k	Beta	$(1+\sum_d\sum_i c_{di}^k,\ \omega+\sum_d\sum_i\sum_{\ell=k+1}^{\infty} c_{di}^\ell)$	$(a_k^{(1)},a_k^{(2)})$	$\mathbb{E}[\log V_k]=\Psi(a_k)-\Psi(a_k+b_k)$ $\mathbb{E}[\log(1-V_k)]=\Psi(b_k)-\Psi(a_k+b_k)$ $\mathbb{E}[\log\sigma_k(V)]=\mathbb{E}[\log V_k]+\sum_{\ell=1}^{k-1}\mathbb{E}[\log(1-V_\ell)]$
β_k	Dirichlet	$\eta+\sum_{d=1}^D\sum_{i=1}^{\infty} c_{di}^k\sum_{n=1}^N z_{dn}^i w_{dn}$	λ_k	$\mathbb{E}[\log\beta_{kv}]=\Psi(\lambda_{kv})-\Psi(\sum_{v'}\lambda_{kv'})$

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Topic model with hierarchical Dirichlet process (HDP)

For latent variables z_{dn}^i and c_{dn}^i the conditional distributions are
$p(z_{dn}^i = 1|\pi_d, \beta_{1:K}, w_{dn}, c_d) \propto exp\{log\,\sigma_i(\pi_d) + \sum_{k=1}^{\infty} c_{di}^k log\,\beta_{k,w_{dn}}\}$

$p(c_{di}^k = 1|v, \beta_{1:K}, w_d, z_d) \propto exp\{log\,\sigma_k(v) + \sum_{n=1}^{N} z_{dn}^i log\,\beta_{k,w_{dn}}\}$

β variable follows the Dirichlet Process conditional distribution
$p(\beta_k|z, c, w) = Dirichlet(\eta + \sum_{d=1}^{D} \sum_{i=1}^{\infty} c_{di}^k \sum_{n=1}^{N} z_{dn}^i w_{dn})$

Variables v_k and π_{di} follow Beta distribution
$p(v_k|c) = Beta(1 + \sum_{d=1}^{D} \sum_{i=1}^{\infty} c_{di}^k, \omega + \sum_{d=1}^{D} \sum_{i=1}^{\infty} \sum_{j>k} c_{di}^j)$

$p(\pi_d i|z_d) = Beta(1 + \sum_{n=1}^{N} z_{dn}^i, \alpha + \sum_{n=1}^{N} \sum_{j>i} z_{dn}^i)$

Based on Mean-Field Variational model, distribution q will be
$q(\beta, v, z, \pi) =$
$(\prod_{k=1}^{K} q(\beta_k|\lambda_k) q(v_k|a_k))(\prod_{d=1}^{D} \prod_{i=1}^{T} q(c_{di}|\zeta_{di}) q(\pi_{di}|\gamma_{di}) \prod_{n=1}^{N} q(z_{dn}|\phi_{dn}))$

Topic model with hierarchical Dirichlet process (HDP)

Initialize $\lambda^{(0)}$ randomly. Set $a^{(0)} = 1$ and $b^{(0)} = \omega$.

Set the step-size schedule ρ_t appropriately.

repeat

  Sample a document w_d uniformly from the data set.

  For $i \in \{1, \ldots, T\}$ initialize

$$\zeta_{di}^k \propto \exp\{\textstyle\sum_{n=1}^N \mathbb{E}[\log \beta_{k,w_{dn}}]\}, k \in \{1, \ldots, K\}.$$

  For $n \in \{1, \ldots, N\}$ initialize

$$\phi_{dn}^i \propto \exp\left\{\textstyle\sum_{k=1}^K \zeta_{di}^k \mathbb{E}[\log \beta_{k,w_{dn}}]\right\}, i \in \{1, \ldots, T\}.$$

  repeat

    For $i \in \{1, \ldots, T\}$ set

$$\begin{aligned}
\gamma_{di}^{(1)} &= 1 + \textstyle\sum_{n=1}^N \phi_{dn}^i, \\
\gamma_{di}^{(2)} &= \alpha + \textstyle\sum_{n=1}^N \sum_{j=i+1}^T \phi_{dn}^j \\
\zeta_{di}^k &\propto \exp\left\{\mathbb{E}[\log \sigma_k(V)] + \textstyle\sum_{n=1}^N \phi_{dn}^i \mathbb{E}[\log \beta_{k,w_{dn}}]\right\}, k \in \{1, \ldots, K\}.
\end{aligned}$$

    For $n \in \{1, \ldots, N\}$ set

$$\phi_{dn}^i \propto \exp\left\{\mathbb{E}[\log \sigma_i(\pi_d)] + \textstyle\sum_{k=1}^K \zeta_{di}^k \mathbb{E}[\log \beta_{k,w_{dn}}]\right\}, i \in \{1, \ldots, T\}.$$

  until local parameters converge.

  For $k \in \{1, \ldots, K\}$ set intermediate topics

$$\begin{aligned}
\hat{\lambda}_{kv} &= \eta + D \textstyle\sum_{i=1}^T \zeta_{di}^k \sum_{n=1}^N \phi_{dn}^i w_{dn}, \\
\hat{a}_k &= 1 + D \textstyle\sum_{i=1}^T \zeta_{di}^k, \\
\hat{b}_k &= \omega + D \textstyle\sum_{i=1}^T \sum_{\ell=k+1}^K \zeta_{di}^\ell.
\end{aligned}$$

  Set

$$\begin{aligned}
\lambda^{(t)} &= (1 - \rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}, \\
a^{(t)} &= (1 - \rho_t) a^{(t-1)} + \rho_t \hat{a}, \\
b^{(t)} &= (1 - \rho_t) b^{(t-1)} + \rho_t \hat{b}.
\end{aligned}$$

until forever

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Results

The proposed stochastic variational inference is tested on "Nature journal" (350K docs, 58M words), "New York Times" (1.8M docs, 461M words), and "Wikipedia" (3.8M docs, 482M words) datasets. 10,000 documents are used for testing the model.

The measure of performance is Predictive Probability with $\mathcal{D}$ as training data. The test data is separated into hold-out w_{ho} and observed data (w_{obs}). Finally, the better models should output higher probability for w_{obs} from the predictive distribution obtained below:

$$\begin{aligned} p(w_{\text{new}} \mid \mathcal{D}, w_{\text{obs}}) &= \int\int \left(\sum_{k=1}^{K} \theta_k \beta_{k,w_{\text{new}}}\right) p(\theta \mid w_{\text{obs}}, \beta) p(\beta \mid \mathcal{D}) d\theta d\beta \\ &\approx \int\int \left(\sum_{k=1}^{K} \theta_k \beta_{k w_{\text{new}}}\right) q(\theta) q(\beta) d\theta d\beta \\ &= \sum_{k=1}^{K} \mathbb{E}_q[\theta_k] \mathbb{E}_q[\beta_{k,w_{\text{new}}}], \end{aligned}$$

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Results

It is tested how long it takes for the Topic Models to run and what Log Predictive Probability is achieved considering the batch size and the forgetting rate

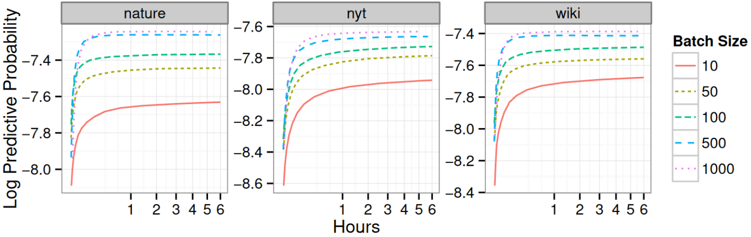


Figure 15: 100-topic LDA inference: Holding the learning rate κ fixed at 0.9, we varied the batch size. Bigger batch sizes are preferred.

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Results

It can be seen that larger batch size performs better

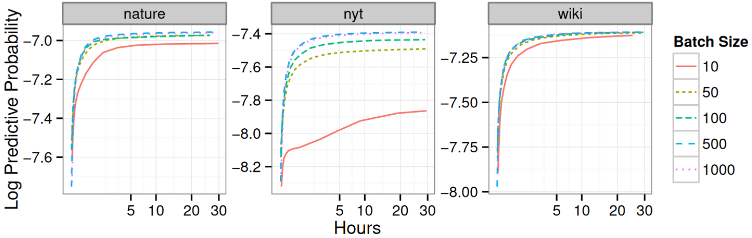


Figure 13: HDP inference: Holding the forgetting rate κ fixed at 0.9, we varied the batch size. Batch sizes may be set too small (e.g., ten documents) but the difference in performance is small once set high enough.

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Results

Higher forgetting rate performs better for LDA

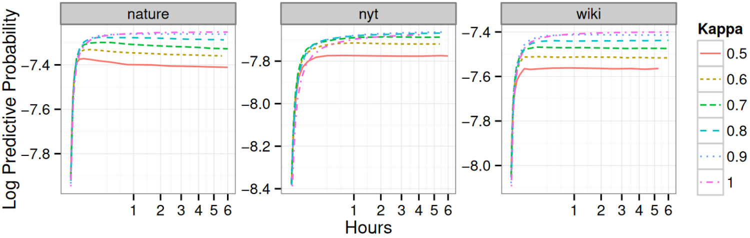


Figure 14: 100-topic LDA inference: Holding the batch size fixed at 500, we varied the forgetting rate κ. Slower forgetting rates are preferred.

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Results

Higher forgetting rate performs better for HDP

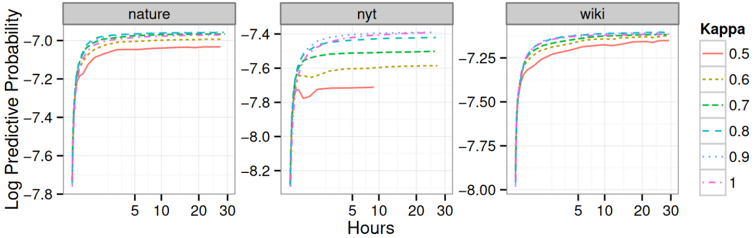


Figure 12: HDP inference: Holding the batch size fixed at 500, we varied the forgetting rate κ. Slower forgetting rates are preferred.

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Results

HDP outperforms LDA. HDP avoids overfitting (not shown explicitly).

	Nature	*New York Times*	*Wikipedia*
LDA 25	-7.24	-7.73	-7.44
LDA 50	-7.23	-7.68	-7.43
LDA 100	-7.26	-7.66	-7.41
LDA 200	-7.50	-7.78	-7.64
LDA 300	-7.86	-7.98	-7.74
HDP	**-6.97**	**-7.38**	**-7.07**

Figure 11: Stochastic inference lets us compare performance on several large data sets. We fixed the forgetting rate $\kappa = 0.9$ and the batch size to 500 documents. We find that LDA is sensitive to the number of topics; the HDP gives consistently better predictive performance. Traditional variational inference (on subsets of each corpus) did not perform as well as stochastic inference.

Source: Stochastic Variational Inference, Journal of Machine Learning Research (2013)

Thank you!