Reliable and scalable variational inference for the hierarchical Dirichlet process

Michael C. Hughes, Dae Il Kim, & Erik B. Sudderth
Dept. of Computer Science, Brown University

HDP topic model
- HDP prior: data-driven learning of number of topics $K$
- Our direct assignment representation better than alternatives

New variational inference
Goal: Find approximate factorized posterior
$q(\phi)q(\beta)q(\pi)q(z) \approx p(\phi, \beta, \pi, z|x)$

Algorithm template
Initialize global factors $q(\phi) q(\beta)$
Loop until converged:
1) Local step
2) Summary step
3) Global step
b) Try merge proposals
c) Try delete proposals

Scalable variational inference
Memoized algorithm
- As scalable as stochastic, without pesky learning rate.
- Requires tracking statistics for each batch & topic.

Reliable inference
- Algorithm should recover similar compact set of topics, regardless of initialization.
- Algorithm should avoid local optima & remove useless junk topics.

Model selection
- Chosen form of $q(\beta)$ is important.
- MAP Point Estimate: $q(\beta) = \delta_{\beta}^{	ext{true}}$
  - Zang et al. '07
- Full distribution: $q(\beta) = \text{StickBreaking}(\rho, \omega)$
  - Integrate away all parameters that grow with $K$.
- 2) Full distribution: $q(\beta) = \text{StickBreaking}(\rho, \omega)$
  - Integrate away all parameters that grow with $K$.
- 3) Global step gives new batch value
- Must also track summaries $S_{k1}, T_{k1}$

Updateing tracked statistics
Incremental update to whole-dataset value:

Stochastic algorithm
- Natural gradient descent for global step update.
- Less effective for merges/deletes. Can’t easily check whole-dataset objective.

Experiments
- Memoized alg. with merges/deletes rapidly finds small set of high-quality topics.
- Other algorithms get stuck quickly or improve very slowly.

Toy bars
- 1000 docs, 900 vocab types
- num pass thru data
- memo + delete & merge initial $K=100$, final $K=10$
- Gibbs
- memoized
- num pass thru data
- initial $K=100$, final $K=10$

Wiki articles
- 7961 docs
- 6131 vocab types
- num pass thru data
- num pass thru data
- NY Times articles
- 1.8 million docs
- 8000 vocab
- num pass thru data
- num pass thru data

Image patches
- 3 million $64 \times 64$ patches from 400 images
- Patch samples from trained model

Model comparison:
- image-specific frequencies (HDP admixture)
- universal frequencies (DP mixture)

Surrogate objective
New function lower bounds intractable ideal objective.
Penalizes junk topics, key to merge/delete moves.

Nested truncation
Only assign tokens to first $K$ topics of infinite set.

Easy to contract truncation level.

Track probability of all inactive topics ($k > K$).

Need not be represented during inference.

Topics > $K$ are conditionally independent of data.

HDP exact
HDP point est

c) Try
Initialize global factors
Algorithm template
Goal: Find approximate factorized posterior
$q(\phi)q(\beta)q(\pi)q(z) \approx p(\phi, \beta, \pi, z|x)$

Algorithm template
Initialize global factors $q(\phi) q(\beta)$
Loop until converged:
1) Local step
2) Summary step
3) Global step
b) Try merge proposals
c) Try delete proposals

Scalable variational inference
Memoized algorithm
- As scalable as stochastic, without pesky learning rate.
- Requires tracking statistics for each batch & topic.

Reliable inference
- Algorithm should recover similar compact set of topics, regardless of initialization.
- Algorithm should avoid local optima & remove useless junk topics.

Model selection
- Chosen form of $q(\beta)$ is important.
- MAP Point Estimate: $q(\beta) = \delta_{\beta}^{	ext{true}}$
  - Zang et al. '07
- Full distribution: $q(\beta) = \text{StickBreaking}(\rho, \omega)$
  - Integrate away all parameters that grow with $K$.

Updateing tracked statistics
Incremental update to whole-dataset value:

Stochastic algorithm
- Natural gradient descent for global step update.
- Less effective for merges/deletes. Can’t easily check whole-dataset objective.

Experiments
- Memoized alg. with merges/deletes rapidly finds small set of high-quality topics.
- Other algorithms get stuck quickly or improve very slowly.

Toy bars
- 1000 docs, 900 vocab types
- num pass thru data
- memo + delete & merge initial $K=100$, final $K=10$
- Gibbs
- memoized
- num pass thru data
- initial $K=100$, final $K=10$

Wiki articles
- 7961 docs
- 6131 vocab types
- num pass thru data
- num pass thru data
- NY Times articles
- 1.8 million docs
- 8000 vocab
- num pass thru data
- num pass thru data

Image patches
- 3 million $64 \times 64$ patches from 400 images
- Patch samples from trained model

Model comparison:
- image-specific frequencies (HDP admixture)
- universal frequencies (DP mixture)

Surrogate objective
New function lower bounds intractable ideal objective.
Penalizes junk topics, key to merge/delete moves.

Nested truncation
Only assign tokens to first $K$ topics of infinite set.

Easy to contract truncation level.

Track probability of all inactive topics ($k > K$).

Need not be represented during inference.

Topics > $K$ are conditionally independent of data.

HDP exact
HDP point est

c) Try
Initialize global factors
Algorithm template
Goal: Find approximate factorized posterior
$q(\phi)q(\beta)q(\pi)q(z) \approx p(\phi, \beta, \pi, z|x)$

Algorithm template
Initialize global factors $q(\phi) q(\beta)$
Loop until converged:
1) Local step
2) Summary step
3) Global step
b) Try merge proposals
c) Try delete proposals

Scalable variational inference
Memoized algorithm
- As scalable as stochastic, without pesky learning rate.
- Requires tracking statistics for each batch & topic.

Reliable inference
- Algorithm should recover similar compact set of topics, regardless of initialization.
- Algorithm should avoid local optima & remove useless junk topics.

Model selection
- Chosen form of $q(\beta)$ is important.
- MAP Point Estimate: $q(\beta) = \delta_{\beta}^{	ext{true}}$
  - Zang et al. '07
- Full distribution: $q(\beta) = \text{StickBreaking}(\rho, \omega)$
  - Integrate away all parameters that grow with $K$.

Updateing tracked statistics
Incremental update to whole-dataset value:

Stochastic algorithm
- Natural gradient descent for global step update.
- Less effective for merges/deletes. Can’t easily check whole-dataset objective.

Experiments
- Memoized alg. with merges/deletes rapidly finds small set of high-quality topics.
- Other algorithms get stuck quickly or improve very slowly.

Toy bars
- 1000 docs, 900 vocab types
- num pass thru data
- memo + delete & merge initial $K=100$, final $K=10$
- Gibbs
- memoized
- num pass thru data
- initial $K=100$, final $K=10$

Wiki articles
- 7961 docs
- 6131 vocab types
- num pass thru data
- num pass thru data
- NY Times articles
- 1.8 million docs
- 8000 vocab
- num pass thru data
- num pass thru data

Image patches
- 3 million $64 \times 64$ patches from 400 images
- Patch samples from trained model

Model comparison:
- image-specific frequencies (HDP admixture)
- universal frequencies (DP mixture)

Surrogate objective
New function lower bounds intractable ideal objective.
Penalizes junk topics, key to merge/delete moves.

Nested truncation
Only assign tokens to first $K$ topics of infinite set.

Easy to contract truncation level.

Track probability of all inactive topics ($k > K$).

Need not be represented during inference.

Topics > $K$ are conditionally independent of data.

HDP exact
HDP point est

c) Try
Initialize global factors
Algorithm template
Goal: Find approximate factorized posterior
$q(\phi)q(\beta)q(\pi)q(z) \approx p(\phi, \beta, \pi, z|x)$

Algorithm template
Initialize global factors $q(\phi) q(\beta)$
Loop until converged:
1) Local step
2) Summary step
3) Global step
b) Try merge proposals
c) Try delete proposals

Scalable variational inference
Memoized algorithm
- As scalable as stochastic, without pesky learning rate.
- Requires tracking statistics for each batch & topic.

Reliable inference
- Algorithm should recover similar compact set of topics, regardless of initialization.
- Algorithm should avoid local optima & remove useless junk topics.

Model selection
- Chosen form of $q(\beta)$ is important.
- MAP Point Estimate: $q(\beta) = \delta_{\beta}^{	ext{true}}$
  - Zang et al. '07
- Full distribution: $q(\beta) = \text{StickBreaking}(\rho, \omega)$
  - Integrate away all parameters that grow with $K$.

Updateing tracked statistics
Incremental update to whole-dataset value:

Stochastic algorithm
- Natural gradient descent for global step update.
- Less effective for merges/deletes. Can’t easily check whole-dataset objective.