Independent finite approximations for Bayesian nonparametric inference

Tin Nguyen

In collaboration with Jonathan Huggins, Lorenzo Masoero, Lester Mackey and Tamara Broderick

Motivating data analysis: topic modelling

"Barcelona has both a soccer and a basketball team."

"Bernie Sanders is an advocate for universal healthcare and taxing the rich. "

….

Documents (observed)

Soccer
Basketball
….

"Sports"

Healthcare
Tax
….

"Politics"

Opera
Stand-up
….

"Arts"

….

Topics (latent)

How to set the number of topics?

Efficient/simple algorithms to estimate the latent topics?

Nonparametric models are flexible, but computationally expensive

Outline

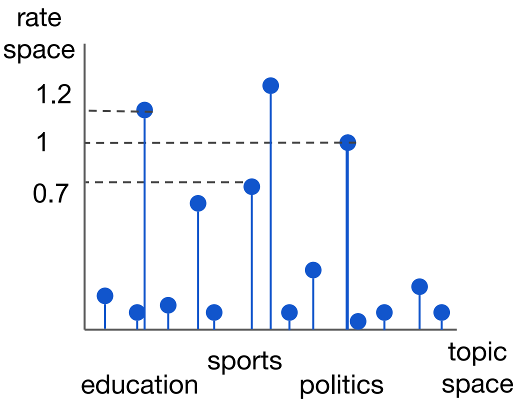
- BNP for flexible modeling
- Finite approximations allow us to use BNP in practice
- Which finite approximation to use? Independent (IFA) versus truncation (TFA)
    - How to construct general, arbitrarily accurate IFAs
    - IFAs are conceptually easier to use
    - Theoretical comparison of IFAs to TFAs
    - Empirical comparison of IFAs to TFAs

Why completely random measures (CRMs)?

- Number of communities is unknown and grows with number of observations
    - Topic modeling [Teh et al. 2006]: communities-topics, observations-documents
    - Dictionary learning [Zhou et al. 2009]: communities-low-level image features.
    - Interest groups [Palla et al. 2012], Speaker diarization [Fox et al. 2010] ...

- Postulate that the population has an infinite number of communities
    - Completely random measure = countably infinite collection of (rate, topic) tuples
    - Finitely many tuples appear in any finite data set

Illustration of (rate,topic).



Outline

- ~~BNP for flexible modeling~~

- Finite approximations allow us to use BNP in practice

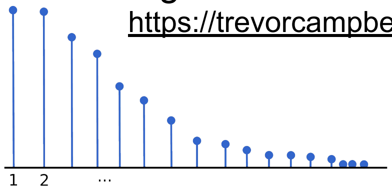
- Which finite approximation to use? Independent (IFA) versus truncation (TFA)
  - How to construct general, arbitrarily accurate IFAs
  - IFAs are conceptually easier to use
  - Theoretical comparison of IFAs to TFAs
  - Empirical comparison of IFAs to TFAs

Finite approximations for faster inference

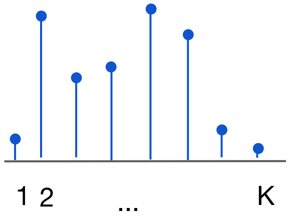
- Inference in infinite-dimensional models is hard/slow.
  - Can't update a countable infinity of parameters.
  - Collapsing [Griffiths et al. 2011] and slice sampling [Walker 2007] are slow.

- A practical alternative: finite-dimensional approximations.

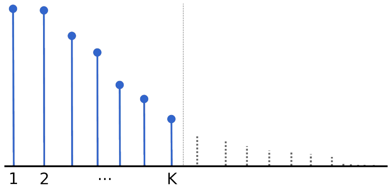
Figure taken from
https://trevorcampbell.me/



Target CRM



Independent finite approximation (IFA)



Truncated finite approximation (TFA)

Outline

- ~~BNP for flexible modeling~~

- ~~Finite approximations allow us to use BNP in practice~~

- Which finite approximation to use? Independent (IFA) versus truncation (TFA)
  - How to construct general, arbitrarily accurate IFAs
  - IFAs are conceptually easier to use
  - Theoretical comparison of IFAs to TFAs
  - Empirical comparison of IFAs to TFAs

General construction of IFA

$$\text{IFA}_K = \sum_{i=1}^{K} \xi_{K,i}\delta_{\psi_{K,i}} \text{ where } \xi_{K,i} \overset{i.i.d.}{\sim} g_K$$

- Past work: Showed converging (in distribution) approximations for special cases [Paisley et al. 2009, Acharya et al. 2015, Lee et al. 2016]

$$\text{For BP or GP } \nu(d\theta),\ \text{IFA}_K \overset{D}{\rightarrow} \text{CRM}(\nu) \text{ as K } \rightarrow \infty$$

- Our construction: Propose converging approximations for generic rate measures

$$\text{For general } \nu(d\theta),\ \text{IFA}_K \overset{D}{\rightarrow} \text{CRM}(\nu) \text{ as K } \rightarrow \infty$$

8

Outline

- ~~BNP for flexible modeling~~

- ~~Finite approximations allow us to use BNP in practice~~

- Which finite approximation to use? Independent (IFA) versus truncation (TFA)
    - ~~How to construct general, arbitrarily accurate IFAs~~
    - IFAs are conceptually easier to use
    - Theoretical comparison of IFAs to TFAs
    - Empirical comparison of IFAs to TFAs

IFA are conceptually easy to use

- For common CRMs, the atoms sizes of IFA are familiar exponential family distributions.

$$\text{For BP } \nu \text{ with d} = 0,\ \xi_{K,i} \overset{i.i.d.}{\sim} \text{Beta}$$

$$\text{If } x_{n,i} | \xi_{K,i} \overset{indep}{\sim} \text{Ber}(\xi_{K,i}) \text{ then } \xi_{K,i} | x_{1:n,i} \sim \text{Beta}$$

- TFA almost always have complicated dependencies in the prior that make incorporating observations difficult.

$$\text{For BP } \nu \text{ with d} = 0,\ \alpha = 1,\ \theta_i = \prod_{j=1}^{i} p_j \text{ where } p_i \overset{i.i.d.}{\sim} \text{Beta}$$

$$\text{If } x_{n,i} | \theta_i \overset{indep}{\sim} \text{Ber}(\theta_i) \text{ then } \theta_i | x_{1:n,i} \sim \text{Complicated}$$

Outline

- ~~BNP for flexible modeling~~

- ~~Finite approximations allow us to use BNP in practice~~

- Which finite approximation to use? Independent (IFA) versus truncation (TFA)
    - ~~How to construct general, arbitrarily accurate IFAs~~
    - ~~IFAs are conceptually easier to use~~
    - Theoretical comparison of IFAs to TFAs
    - Empirical comparison of IFAs to TFAs

Error bounds for finite approximations

- Past finite-sample theory: TFA requires a small number of atoms for good approximation [Campbell, Huggins et al. 2019].
    - But no finite-sample understanding of IFA quality.

- Our theoretical contributions: Finite-sample upper and lower bounds on IFA quality.
    - Worse performance of IFA in theory.

IFA have worse worst-case behavior

$$\Theta \sim \mathrm{CRM}(\nu), X_{1:N}|\Theta \overset{i.i.d.}{\sim} f(.|\Theta) \qquad \Theta_K \sim \mathrm{FA}_K, Y_{1:N}|\Theta_K \overset{i.i.d.}{\sim} f(.|\Theta_K)$$

$$\mathrm{Error}(\mathrm{FA}_K) = \frac{1}{2}\int_u |p_{X_{1:N}}(u) - p_{Y_{1:N}}(u)|du$$

Assumptions: CRM is exponential-like, with no power-law behavior (beta, gamma processes with discount = 0)

	$\mathrm{Error}(\mathrm{IFA}_K)$	$\mathrm{Error}(\mathrm{TFA}_K)$ [Campbell, Huggins et al. 2019]
Upper bound (typical f)	$(\ln^2 N)/K$	$N\eta^K$ for $\eta < 1$
Lower bound (bad f)	$1/K$	N/A

Outline

- ~~BNP for flexible modeling~~

- ~~Finite approximations allow us to use BNP in practice~~

- ~~Which finite approximation to use? Independent (IFA) versus truncation (TFA)~~
    - ~~How to construct general, arbitrarily accurate IFAs~~
    - ~~IFAs are conceptually easier to use~~
    - ~~Theoretical comparison of IFAs to TFAs~~
    - Empirical comparison of IFAs to TFAs

Performance of finite approximations

- Past empirics: TFA and IFA can have similar performance [Kurihara et al. 2007a, Doshi-Velez et al. 2009].

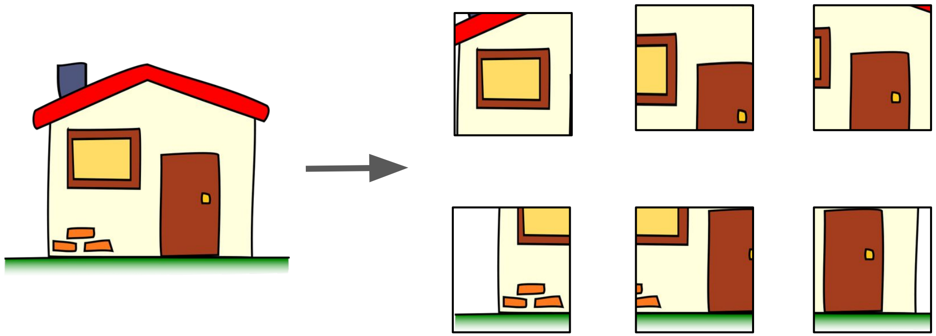
Our experimental contributions:

- Further confirmation of similar performance, for different models and larger problem sizes (N and K)
- The posterior modes of the approximations are similar to each other

Experimental details

Image denoising.

- Data: Patches from a noisy input image. Goal: Denoise input image.
- Metric: Peak signal-to-noise ratio.

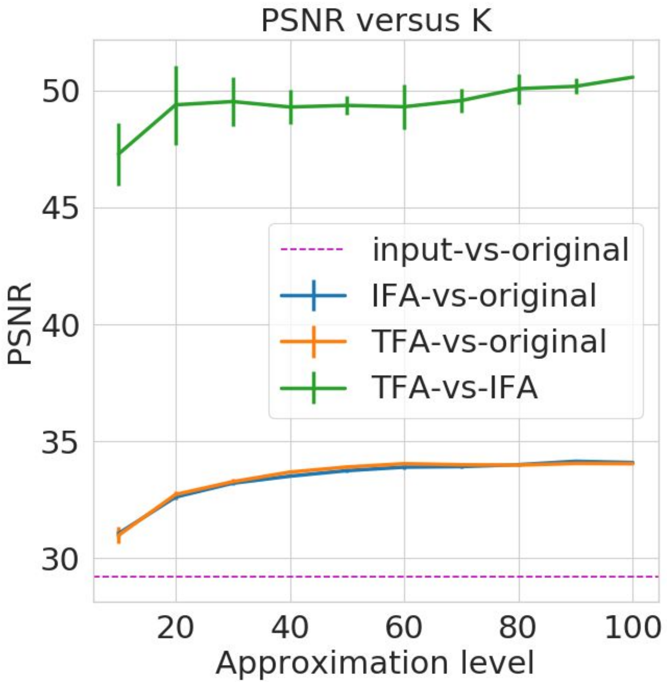


Topic modelling.

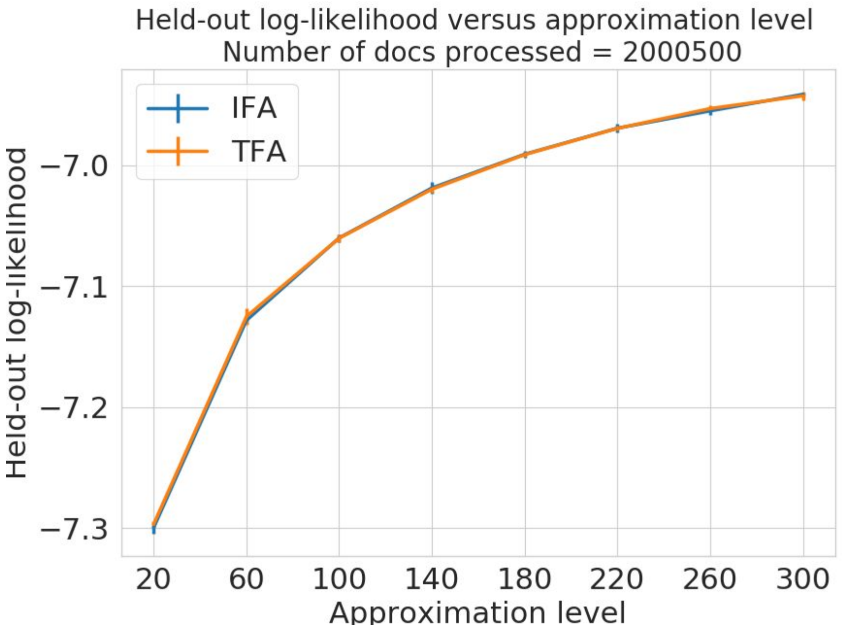
- Data: Wikipedia documents. Goal: Infer meaningful topics.
- Metric: predictive log-likelihood.

IFA and TFA have similar performance across K

Dictionary learning with beta-Bernoulli

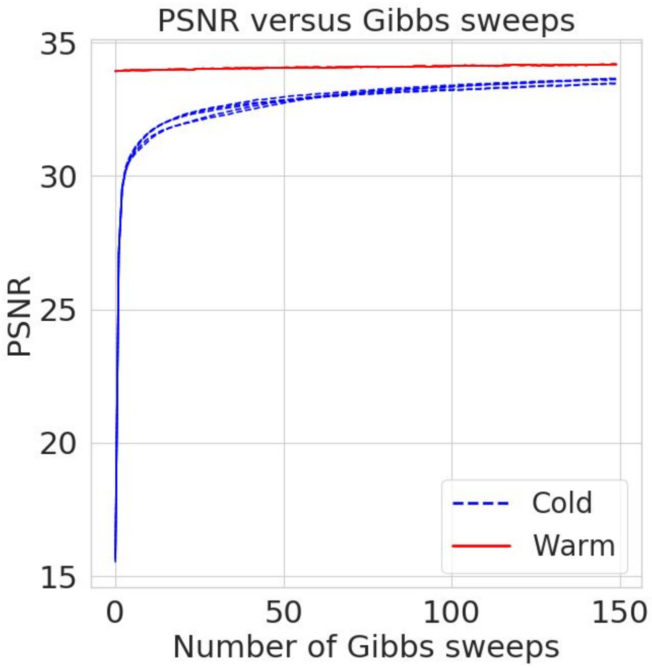


Topic modelling with modified HDP

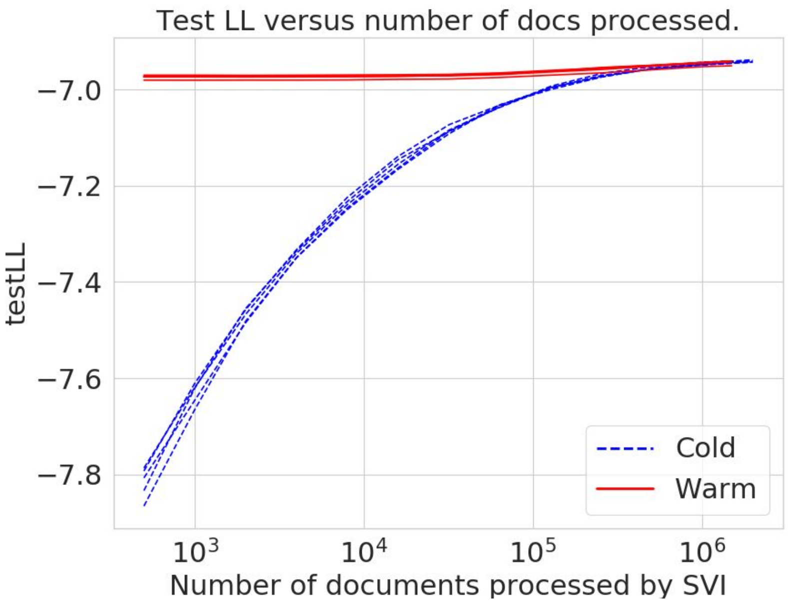


IFA and TFA have similar posterior modes

Dictionary learning with beta-Bernoulli



Topic modelling with modified HDP



Outline

- ~~BNP for flexible modeling~~

- ~~Finite approximations allow us to use BNP in practice~~

- ~~Which finite approximation to use? Independent (IFA) versus truncation (TFA)~~
  - ~~How to construct general, arbitrarily accurate IFAs~~
  - ~~IFAs are conceptually easier to use~~
  - ~~Theoretical comparison of IFAs to TFAs~~
  - ~~Empirical comparison of IFAs to TFAs~~

Conclusion

- Summary
  - Arbitrarily accurate IFAs exist for general CRMs and have simple form in many cases
  - TFAs are more component-efficient approximation than IFAs in the worst-case
  - Practically, IFAs and TFAs perform very similarly

- Links
  - arXiv: https://arxiv.org/abs/2009.10780
  - My contacts: https://www.mit.edu/~tdn/

References

Matthew D. Hoffman, David M. Blei, Chong Wang and John Paisley. Stochastic variational inference, JMLR 2013.

Jonathan Huggins, Lorenzo Masoero, Lester Mackey and Tamara Broderick. Generic finite approximations for practical Bayesian nonparametrics. BNP workshop at NeurIPS 2017.

Juho Lee, Lancelot F. James, and Seungjin Choi. Finite-dimensional BFRY priors and variational Bayesian inference for power law models. NeurIPS 2016.

Thomas L. Griffiths and Zoubin Ghahramani. The Indian buffet process: an introduction and review. JMLR 2011.

Trevor Campbell*, and Jonathan H. Huggins*, Jonathan P. How and Tamara Broderick. Truncated random measures. Bernoulli 2019. (* co-first authors)

Mingyuan Zhou, Haojun Chen, Lu Ren, Guillermo Sapiro,Lawrence Carin and John W. Paisley. Non-parametric Bayesian dictionary learning for sparse image representations. NeurIPS 2009.

References (continued)

Kenichi Kurihara, Max Welling and Yee Whye Teh. Collapsed variational Dirichlet process mixture models. IJCAI 2007.

Finale Doshi-Velez, Kurt Miller and Jurgen Van Gael and Yee Whye Teh. Variational inference for the Indian buffet process. AISTATS 2009.

Kenichi Kurihara, Max Welling and Nikos Vlassis. Accelerated variational Dirichlet process mixtures. NeurIPS 2007.

Stephen G. Walker. Sampling the Dirichlet mixture model with slices. Communications in Statistics—Simulation and Computation 2007.

A. Acharya, J. Ghosh, and M. Zhou. Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices. In AISTATS, 2015

John Paisley and L. Carin. Nonparametric factor analysis with beta process priors. ICML 2009.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. Journal of the American Statistical Association 2006.

References (continued)

H. Ishwaran and M. Zarepour. Exact and approximate sum representations for the Dirichlet process. Canadian Journal of Statistics, 2002.

C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical Dirichlet process. AISTATS 2011.

T. Broderick, M. I. Jordan, and J. Pitman. Beta Processes, Stick-Breaking and Power Laws. Bayesian Analysis 2012.

T. S. Ferguson. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1973.

E. B. Fox, E. Sudderth, M. I. Jordan, and A. S. Willsky. A Sticky HDP-HMM with Application to Speaker Diarization. The Annals of Applied Statistics. 2010

K. Palla, D. A. Knowles, and Z. Ghahramani. An Infinite Latent Attribute Model for Network Data. ICML. 2012

References (end)

N. L. Hjort. Nonparametric Bayes estimators based on beta processes in models for life history data. The Annals of Statistics 1990.

Y. W. Teh and D. Gorur. Indian buffet processes with power-law behavior. NeurIPS 2009.