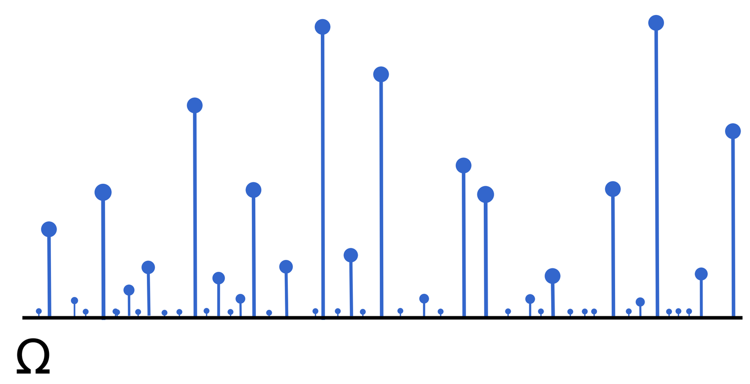
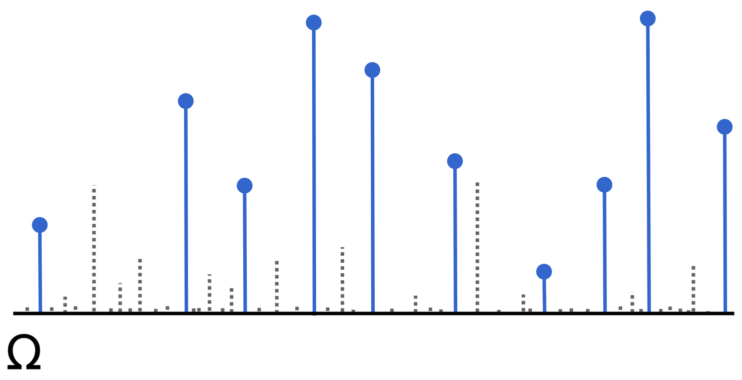
Generic finite approximations for practical Bayesian nonparametrics

Jonathan Huggins[1], Lorenzo Masoero[1], Lester Mackey[2], Tamara Broderick[1] • [1]CSAIL, MIT [2]Microsoft Research New England

Why Bayesian nonparametrics?

Bayesian nonparametrics (BNP) provides flexible models which **adapt their complexity** as the amount of data grows by allowing (in principle) infinitely many latent parameters.





Crucial to develop **generic**, **fast** and **practical** inferential schemes via **finite approximations** which retain the desirable properties of these complex models.

For example, using finite approximations has been considered in applications such as speaker diarization [**?**] and factor analysis for population genetics [**?**]

Our contribution: a general recipe to construct approximations for a large class of nonparametric priors, amenable to easy implementation in probabilistic programming languages

Inference and approximations in BNP models

Generative model

$$\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$$
$$Z_n \mid \Theta \overset{\text{i.i.d.}}{\sim} \mathrm{LP}(h, \Theta)$$
$$Y_n \mid Z_n \overset{\text{indep}}{\sim} f(\cdot \mid Z_n)$$

Example application: topic modeling

→ Generate *a priori* countably many topics
→ For each doc, sample topic counts
→ For each doc, sample words given topics

Question: *how can we perform inference in an infinite model?*

Integrate out infinite parameter

- ▷ Yields a combinatorial process
- ▷ Requires developing **ad hoc algorithms**
- ▷ Samplers often mix slowly

Use finite approximation

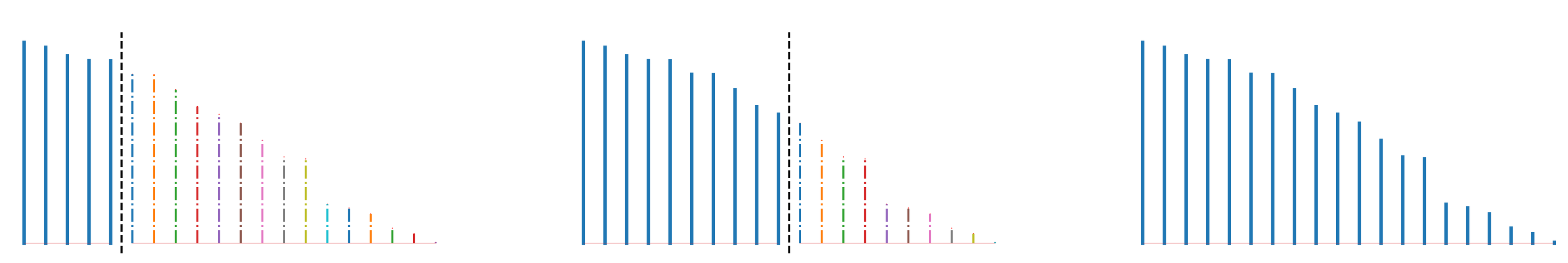
- ▷ Finite amount of data only needs finite model capacity
- ▷ Amenable to **generic** and **parallelizable** inference
- ▷ Can lead to faster mixing

Finite approximations of BNP priors

Truncated finite approximations (TFAs) and **non-nested finite approximations** (NNFAs) are different approaches to construct finite dimensional approximations to infinite dimensional priors.

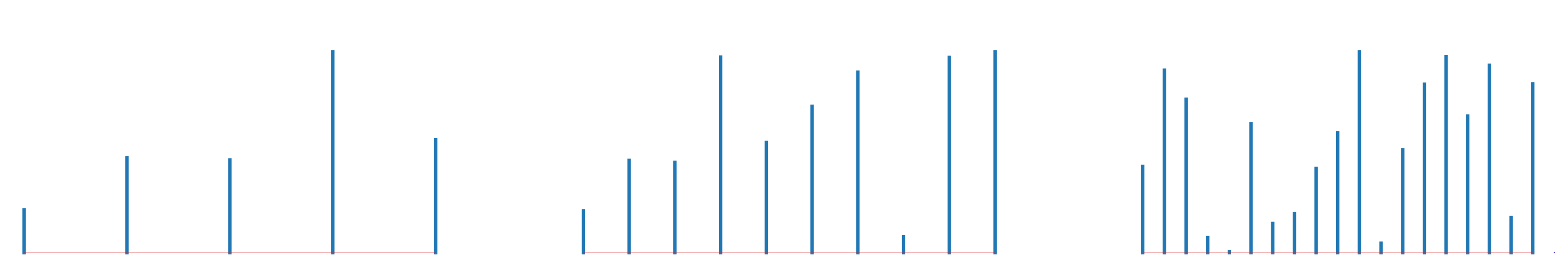
Truncated Finite Approximations

$$\Theta^{\text{TCRM}}_{K_1} = \sum_{k=1}^{K_1} \theta_k \delta_{\psi_k} \qquad \Theta^{\text{TCRM}}_{K_2} = \sum_{k=1}^{K_2} \theta_k \delta_{\psi_k} \qquad \Rightarrow \Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$$



Non-nested Finite Approximations

$$\Theta^{\text{NNFA}}_{K_1} = \sum_{k=1}^{K_1} \theta_{K_1,k} \delta_{\psi_k} \qquad \Theta^{\text{NNFA}}_{K_2} = \sum_{k=1}^{K_2} \theta_{K_2,k} \delta_{\psi_k} \qquad \Rightarrow \Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k}$$



$$\theta_{K_1,k} \overset{i.i.d.}{\sim} \nu_{K_1} \qquad \theta_{K_2,k} \overset{i.i.d.}{\sim} \nu_{K_2} \qquad \{\theta_k\}_{k=1}^{\infty} \sim \mathrm{PP}(\nu)$$

TFAs

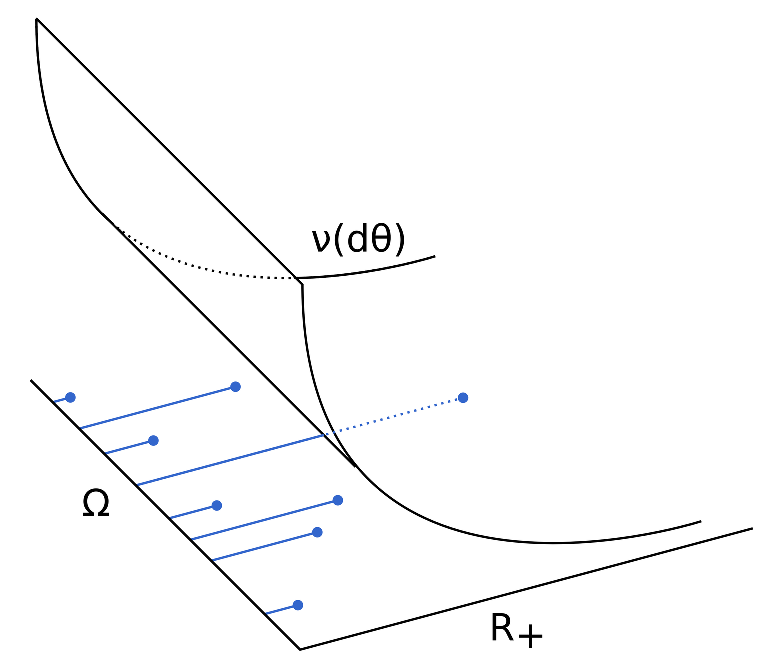
- − **Complex dependences** between atoms $(\theta_k)_{k=1}^{\infty}$ make inference challenging
- \+ Approximation level K doesn't need to be chosen ahead of time

NNFAs

- \+ **i.i.d. atoms** make inference potentially easier and parallelizable
- − A completely new approximation must be constructed if K changes

References

Background: Completely Random Measures



Most BNP priors are constructed using Poisson processes to obtain **completely random measures**: these are random measures which couple (random) **rates** with (random) **traits**

- ▷ Draw **rates** $(\theta_k)_{k=1}^{\infty}$ from a Poisson process Π on $[0, +\infty)$
- ▷ Draw **traits** from a base measure H on a trait space Ω, $(\psi_k)_{k=1}^{\infty} \overset{iid}{\sim} H$
- ▷ Form the measure

$$\Theta = \sum_{k=1}^{\infty} \theta_k \delta_{\psi_k} \sim \mathrm{CRM}(H, \nu)$$

Constructing NNFAs

- ▷ Methods for constructing **truncated finite approximations** are well understood (see **?**)
- ▷ **Goal**: provide a recipe to construct **non-nested finite approximations** of the form

$$\Theta_K = \sum_{k=1}^{K} \theta_{K,k} \delta_{\psi_{K,k}} \qquad \theta_{K,k} \overset{\text{indep}}{\sim} \nu_K \qquad \psi_{K,k} \overset{\text{i.i.d.}}{\sim} H,$$

- ▷ **Intuition:** Choose distribution ν_K such that $K\nu_K(\theta) \approx \nu(\theta)$ and $K\nu_K \to \nu$.

Main Theorem: Let

$$\nu(\mathrm{d}\theta) = \gamma \theta^{-1} h(\theta; \eta) Z(1, \eta)^{-1} \mathrm{d}\theta$$
$$\nu_K(\mathrm{d}\theta) := \theta^{-1+cK^{-1}} h(\theta; \eta) Z_K^{-1} \mathrm{d}\theta \approx K^{-1} \theta^{cK^{-1}} \nu(\mathrm{d}\theta),$$

where $Z(\xi, \eta) := \int \theta^{\xi-1} g(\theta)^{\xi} h(\theta; \eta) \mathrm{d}\theta$ and $c := \gamma \frac{h(0;\eta)}{Z(1,\eta)}$.
Under mild assumptions on h,

$$\Theta_K \overset{\mathcal{D}}{\Longrightarrow} \mathrm{CRM}(H, \nu).$$

Corollary: If ν is in an (improper) exponential family, then ν_K is in the same EF. For example:

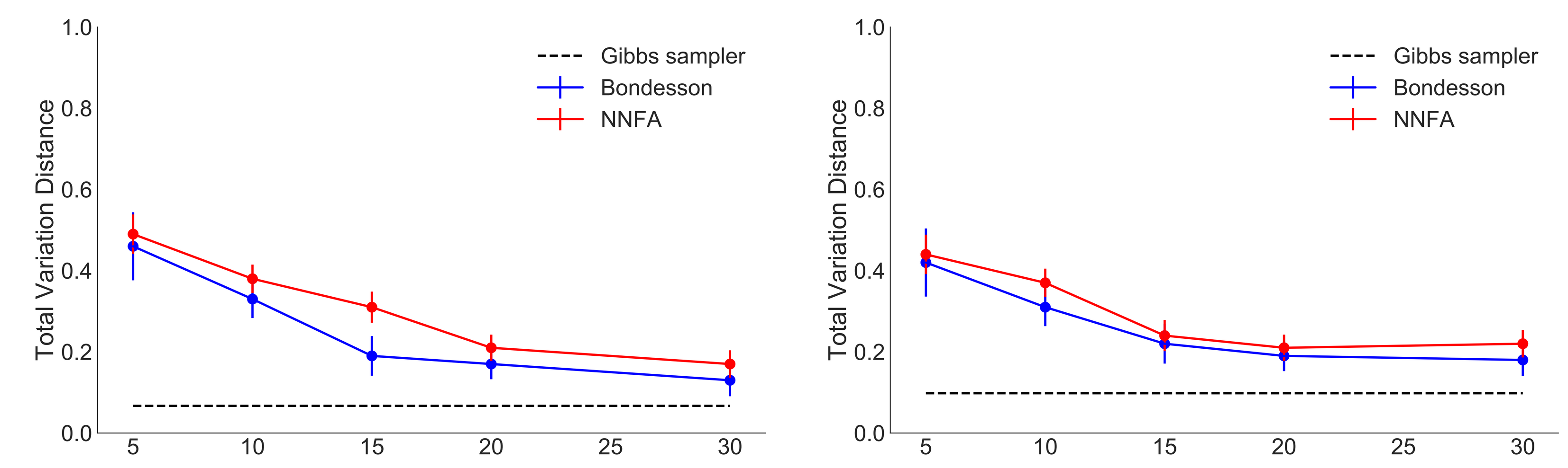
- ▷ If $\Theta \sim \Gamma\mathrm{P}(\gamma, \lambda)$, then $\nu_k = \mathrm{Gam}(\gamma\lambda/K, \lambda)$
- ▷ If $\Theta \sim \mathrm{BP}(\gamma, \alpha)$, then $\nu_k = \mathrm{Beta}(\gamma\alpha/K, \alpha)$

Experiments: Dirichlet process mixture models

To test the performance of TFAs and NNFAs we consider a Dirichlet process mixture model:

$$\Xi = \sum_{k \geq 1} \xi_k \delta_{\psi_k} \sim \mathrm{DP}(\alpha, H), \quad X_n \mid \Xi \sim \sum_{k \geq 1} \xi_k \mathcal{N}(\psi_k, I_d), \quad H = \mathcal{N}(0, \sigma_0^2 I_d)$$

- ▷ We build our model using the **gamma process**:
  - ▷ if $\Theta \sim \Gamma\mathrm{P}(\gamma, \lambda)$, then $\Xi := \Theta / \Theta(\Psi) \sim \mathrm{DP}(\alpha, H)$
- ▷ We compare a TFA approximation (the Bondesson representation) and the NNFA approximation derived from the main theorem
- ▷ We implement both approximations using Stan [**?**]
- ▷ Use Gibbs sampler as gold-standard reference



Example of metric for posterior quality evaluation:

- ▷ Let $p_{\mathcal{D}}(k) = \mathbb{P}[\#\{z_1, \dots, z_M\} = k \mid \mathcal{D}]$ be the probability that the data is in k clusters
- ▷ We calculate the total variation distance between this distribution under true and approximate model:

$$d_{TV}(p_{\mathcal{D},true}, p_{\mathcal{D},approx}) = \frac{1}{2} \sum_{k=1}^{\infty} |p_{\mathcal{D},true}(k) - p_{\mathcal{D},approx}(k)|$$

Ongoing work and future directions:

- ▷ Investigate more complex modeling scenarios and cases where parallelism is of interest.
- ▷ Obtain error bounds for NNFAs (like those available for TFAs).

ONR grant N00014-17-1-2072, ONR MURI grant N00014-11-1-0688, and a Google Faculty Research Award

E-Mail: {jhuggins@, lom@, tbroderick@csail.} mit.edu, lmackey@microsoft.com