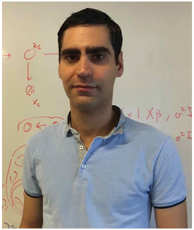
Augment and Reduce:

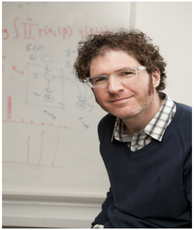
Stochastic Inference for Large Categorical Distributions

Adji Bousso Dieng

Columbia University
IN THE CITY OF NEW YORK

Collaborators

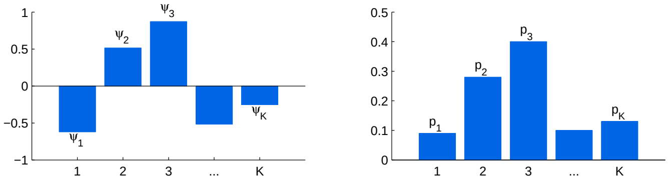




- Francisco J. R. Ruiz
- Michalis Titsias
- David M. Blei

Augment and Reduce: Stochastic Inference for Large Categorical Distributions
F. J. R. Ruiz, M. Titsias, A. B. Dieng, and D. M. Blei
Under review at ICML, 2018.

Categorical Distributions: Applications



Categorical distributions are ubiquitous in Statistics and Machine Learning

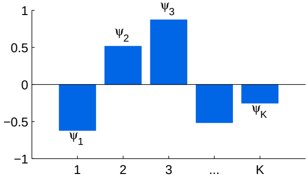
→ discrete choice models

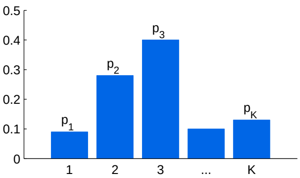
→ language models

→ recommendation systems

→ reinforcement learning

Categorical Distributions: Example Parameterization





→ One widely applied parameterization of a categorical is the softmax,

$$p(y = k \mid \psi) = \mathrm{softmax}(\psi)|_k = \frac{e^{\psi_k}}{\sum_{k'} e^{\psi_{k'}}}$$

→ Transforms reals into probabilities

→ Can be costly because of normalization ... $\mathcal{O}(K)$

→ A computational burden when learning with categorical distributions

A Closer Look at Softmax

- → Draw random standard Gumbel errors i.i.d.,

$$\varepsilon_k \sim \mathrm{Gumbel}(\varepsilon \mid 0, 1)$$

- → Define a *utility* for each outcome k,

$$\psi_k + \varepsilon_k$$

- → Choose the outcome with the largest utility,

$$y = \arg\max_k (\psi_k + \varepsilon_k)$$

- → Integrate out the error terms (ε_k's) to find the marginal $p(y \mid \psi)$

Softmax is the marginal!!

The Augmented Model

- → The augmented model is

$$p(y = k, \varepsilon \mid \psi) = \phi(\varepsilon) \prod_{k' \neq k} \Phi(\varepsilon + \psi_k - \psi_{k'})$$

- → Nice property: The log-joint has a summation over the categories,

$$\log p(y = k, \varepsilon \mid \psi) = \log \phi(\varepsilon) + \sum_{k' \neq k} \log \Phi(\varepsilon + \psi_k - \psi_{k'})$$

- → This enables fast unbiased estimates,
  - Sample a subset of outcomes $\mathcal{S} \subseteq \{1, \ldots, K\} \setminus \{k\}$
  - Compute an estimate of the log-joint

$$\log \phi(\varepsilon) + \frac{K-1}{|\mathcal{S}|} \sum_{k' \in \mathcal{S}} \log \Phi(\varepsilon + \psi_k - \psi_{k'})$$

- → This has $\mathcal{O}(|\mathcal{S}|)$ complexity

The Inference Algorithm: Variational EM

- We are not interested in the log-joint, but in the log-marginal

- Variational inference relates both quantities,

$$\log p(y \mid \psi) \geq \mathbb{E}_{q(\varepsilon)} \left[\log p(y, \varepsilon \mid \psi) - \log q(\varepsilon) \right]$$

- Maximize the bound using *variational EM*
  - E step: Optimize w.r.t. the distribution $q(\varepsilon)$
  - M step: Take a gradient step w.r.t. ψ

- The complexity is controlled by the user (via $|\mathcal{S}|$)

Things are Prettier with Softmax

$\rightarrow$ We can compute the optimal $q(\varepsilon)$ distribution,

$$q^\star(\varepsilon) = \mathrm{Gumbel}(\log \eta^\star, 1), \quad \eta^\star = 1 + \sum_{k' \neq k} e^{\psi_{k'} - \psi_k}$$

$\rightarrow$ This is $\mathcal{O}(K)$. Instead, set

$$q(\varepsilon) = \mathrm{Gumbel}(\log \eta, 1)$$

$\rightarrow$ Estimate the optimal natural parameter in $\mathcal{O}(|\mathcal{S}|)$,

$$\widetilde{\eta} = 1 + \frac{K-1}{|S|} \sum_{k' \in S} e^{\psi_{k'} - \psi_k}$$

(to update η, take a step in the direction of the natural gradient)

Scale All Categorical Distributions!

- → Choose other distributions for ε to get other models,
  - Gaussian for multinomial probit
  - Logistic for multinomial logistic
- → Form Monte Carlo gradient estimators using reparameterization
- → Useful for both E and M steps

Empirical Evidence

→ Baselines:
  - Exact Softmax for MNIST and Bibtex
  - OVE – Also a lower bound but only applicable to softmax

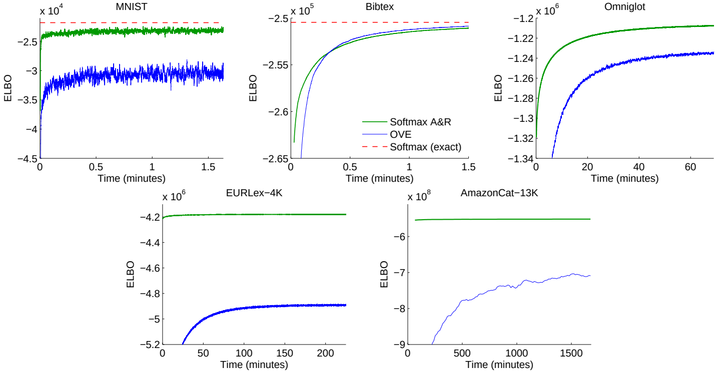
→ Time complexity (top) and Predictive performance (bottom)

dataset	OVE (Titsias, 2016)	A&R [this paper] softmax	A&R [this paper] multi. probit	A&R [this paper] multi. logistic
MNIST	0.336 s	0.337 s	0.431 s	0.511 s
Bibtex	0.181 s	0.188 s	0.244 s	0.246 s
Omniglot	4.47 s	4.65 s	5.63 s	5.57 s
EURLex-4K	5.54 s	5.65 s	6.46 s	6.23 s
AmazonCat-13K	2.80 h	2.80 h	2.82 h	2.91 h

	softmax model						multi. probit		multi. logistic	
	exact		OVE (Titsias, 2016)		A&R [this paper]		A&R [this paper]		A&R [this paper]	
dataset	log lik	acc	log lik	acc	log lik	acc	log lik	acc	log lik	acc
MNIST	−0.261	0.927	−0.276	0.919	**−0.271**	**0.924**	−0.302	0.918	−0.287	0.917
Bibtex	−3.188	0.361	−3.300	0.352	**−3.036**	**0.361**	−4.184	0.346	−3.151	0.353
Omniglot	–	–	−5.667	0.179	**−5.171**	**0.201**	−7.350	0.178	−5.395	0.184
EURLex-4K	–	–	**−4.241**	**0.247**	−4.593	0.207	−4.193	0.263	−4.299	0.226
AmazonCat-13K	–	–	−3.880	0.388	**−3.795**	**0.420**	−3.593	0.411	−4.081	0.350

Empirical Evidence

→ Quality of the bound



Take Home: The A&R Recipe

- → Choose a distribution for ε
- → Augment your model with ε to get an *augmented model*—

$$\mathcal{L} = \log p(y = k, \varepsilon \mid \psi) = \log \phi(\varepsilon) + \sum_{k' \neq k} \log \Phi(\varepsilon + \psi_k - \psi_{k'})$$

- → Reduce cost to $\mathcal{O}(|\mathcal{S}|)$ with an estimate of the log-joint,

$$\mathcal{L} \approx \tilde{\mathcal{L}} = \log \phi(\varepsilon) + \frac{K-1}{|\mathcal{S}|} \sum_{k' \in \mathcal{S}} \log \Phi(\varepsilon + \psi_k - \psi_{k'})$$

- → Use stochastic variational EM with the bound

$$\log p(y \mid \psi) \geq \mathbb{E}_{q(\varepsilon)} \left[\mathcal{L} - \log q(\varepsilon) \right]$$

A&R is a principled method that scales up training for models involving large categorical distributions using latent variable augmentation and stochastic variational inference.