A New Distribution on the Simplex with Auto-Encoding Applications

Andrew Stirn, Tony Jebara, David A Knowles
{andrew.stirn,jebara,daknowles}@cs.columbia.edu

Columbia University

New York Genome Center

Contributions

We develop a surrogate distribution for the Dirichlet that offers explicit, tractable reparameterization, the ability to capture sparsity, and has barycentric symmetry (exchangeability) properties equivalent to the Dirichlet.

Stick Breaking Process

Algorithm 1 Ordered Stick-Breaking

Require: $K \geq 2$
Require: base dist. $p_i(v; a_i, b_i)\ \forall\, i \in [K]$
Require: ordering (permutation) o
  Sample: $v_{o_1} \sim p_{o_1}(v; a_{o_1}, b_{o_1})$
  Assign: $x_{o_1} \leftarrow v_{o_1}, i \leftarrow 2$
  while $i < K$ **do**
    Sample: $v_{o_i} \sim p_{o_i}(v; a_{o_i}, b_{o_i})$
    Assign: $x_{o_i} \leftarrow v_{o_i}\left(1 - \sum_{j=1}^{i-1} x_{o_j}\right)$
    Assign: $i \leftarrow i + 1$
  end while
  Assign: $x_{o_K} \leftarrow 1 - \sum_{j=1}^{K-1} x_{o_j}$
  return x

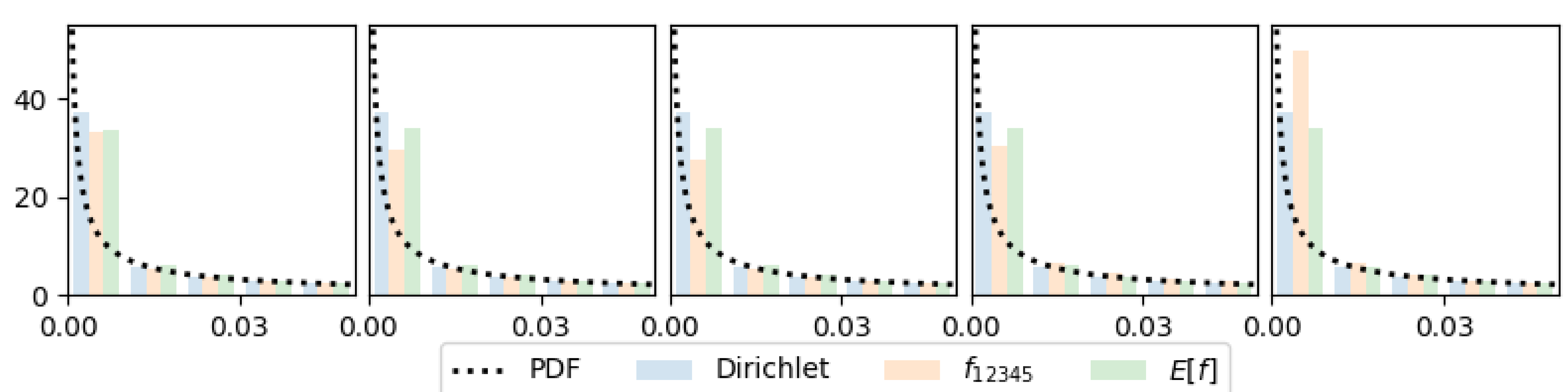
A New Dirichlet Surrogate

Let $f_o(x_{o_1:o_K}; \alpha_{o_1:o_K})$ be the joint density of K random variables returned from algorithm 1 with $p_i(v; a_i, b_i) \equiv \text{Kumaraswamy}(x; \alpha_i, \sum_{j=i+1}^{K} \alpha_j)$ and an ordering o, then our proposed distribution for the $(K-1)$-simplex is MV-Kumaraswamy$(x; \alpha) =$

$$\mathop{\mathbb{E}}_{o \sim \text{Uniform}(O)} [f_o(x_{o_1:o_K}; \alpha_{o_1:o_K})]$$

Corollary 1 *Let $S \subseteq \{1, \ldots, K\}$ be the set of indices i where for $i \neq j$ we have $\alpha_i = \alpha_j$. Define $A = \{1, \ldots, K\} \setminus S$. Then, $\mathbb{E}_{o \sim \text{Uniform}(O)}[f_o(x_{o_1:o_K}; \alpha_{o_1:o_K})]$ is symmetric across barycentric axes $x_a\ \forall\, a \in A$ (i.e. it is exchangeable).*

Sampling Bias with Fixed-Order Kumarswamy Stick Breaks



Sampling bias for a 5-dimensional Dirichlet approximation with $\alpha = \frac{1}{5}(1, 1, 1, 1, 1)$. We maintain histograms for three methods: Dirichlet, fixed-order Kumaraswamy stick-breaks, random-order Kumaraswamy stick-breaks. Note the bias on the last dimension (last subplot) when using a fixed order. Randomizing order eliminates this bias.

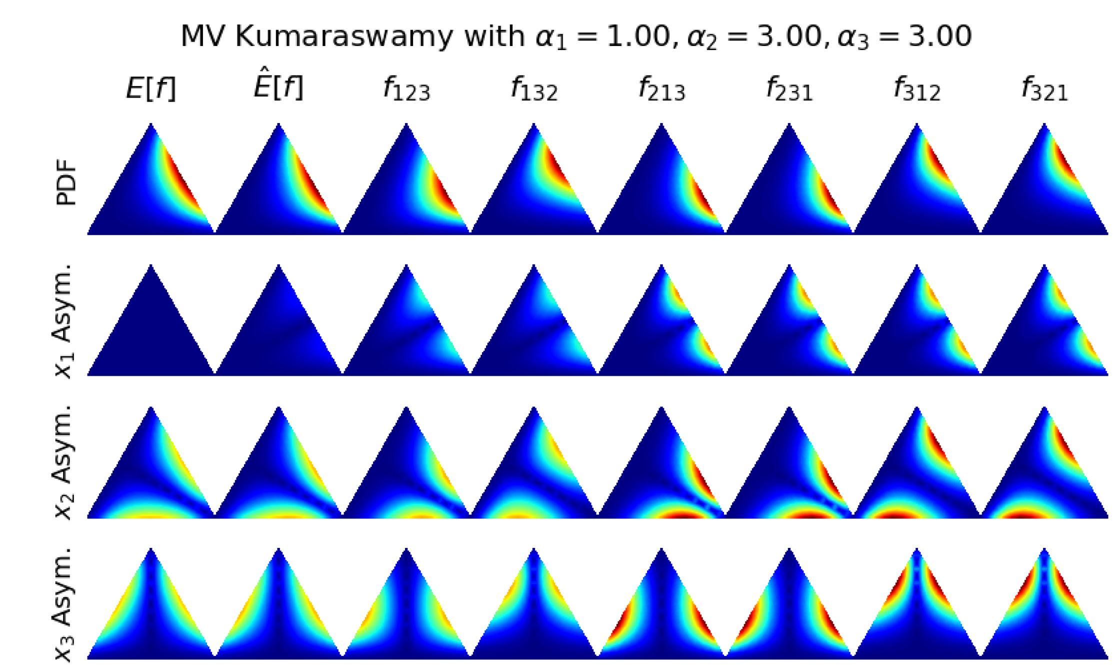
Semi-supervised Variational Auto-encoding Tasks

We specify the a generative process with partially observed labels y. We fit this model with a VAE. Each method varies in its treatment of the variational posterior $q(\pi; \alpha_\phi(x))$

$$\pi_i \stackrel{iid}{\sim} \text{Dirichlet}(\pi; \alpha), \qquad z_i \stackrel{iid}{\sim} \mathcal{N}(z; 0, I),$$
$$y_i | \pi_i \sim \text{Discrete}(y; \pi_i), \qquad x_i | y_i, z_i \sim p(x; f_\theta(y_i, z_i)),$$

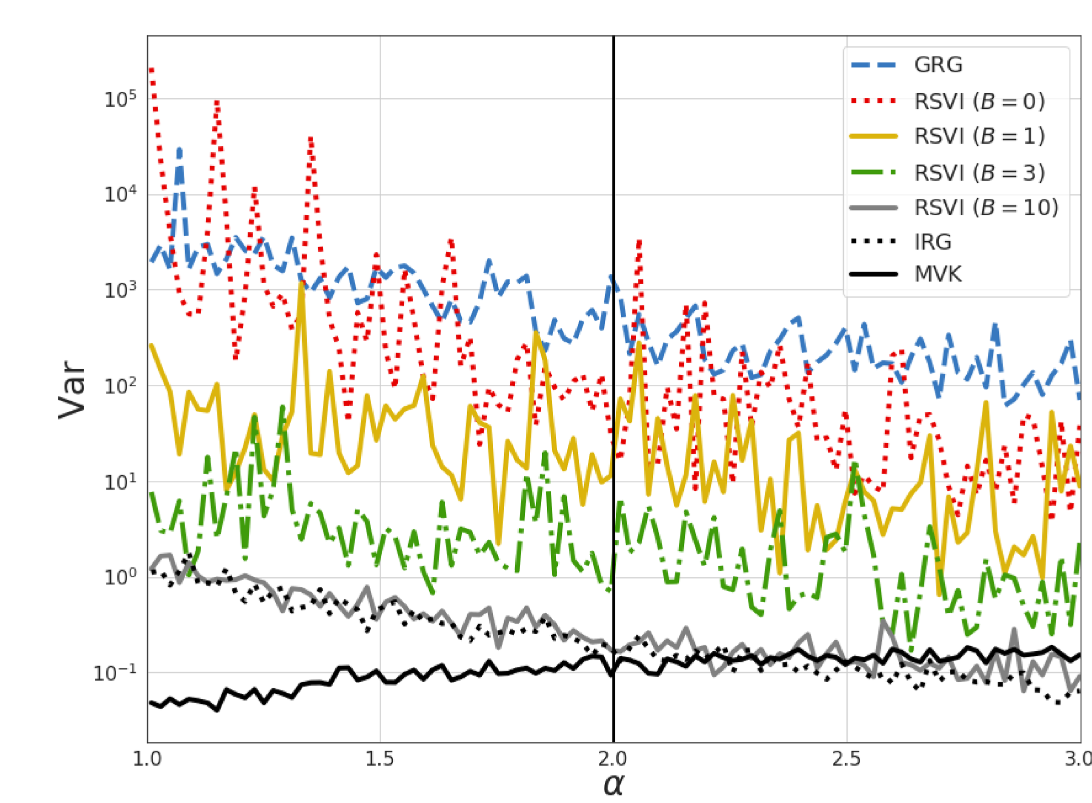
Experiment	Method	Error	p-value	Log Likelihood	p-value
MNIST	MV-Kum.	0.099 ± 0.011	$--$	-6.4 ± 6.3	$--$
10 trials	IRG[1]	0.097 ± 0.008	0.72	-7.8 ± 7.1	0.64
600 labels	Kumar-SB[2]	0.248 ± 0.009	1.05×10^{-17}	-6.5 ± 6.3	0.95
$\dim(z) = 0$	Softmax	0.093 ± 0.009	0.24	-6.5 ± 6.2	0.95
MNIST	MV-Kum.	0.043 ± 0.005	$--$	45.06 ± 0.92	$--$
10 trials	IRG[1]	0.044 ± 0.006	0.89	45.69 ± 0.38	0.06
600 labels	M2 (ours)	0.098 ± 0.014	5.37×10^{-10}	Not collected	$--$
$\dim(z) = 2$	Kumar-SB[2]	0.138 ± 0.015	1.65×10^{-13}	44.33 ± 1.65	0.24
	Softmax	0.042 ± 0.003	0.40	45.14 ± 0.73	0.82
MNIST	MV-Kum.	0.018 ± 0.004	$--$	116.58 ± 0.68	$--$
10 trials	IRG[1]	0.018 ± 0.004	0.98	116.57 ± 0.43	0.97
600 labels	M2 (ours)	0.020 ± 0.003	0.32	Not collected	$--$
$\dim(z) = 50$	Kumar-SB[2]	0.071 ± 0.008	2.58×10^{-13}	116.22 ± 0.33	0.15
	Softmax	0.018 ± 0.003	0.87	116.24 ± 0.45	0.21
	M2†[3]	0.049 ± 0.001	$--$	Not reported	$--$
	M1 + M2†[3]	0.026 ± 0.005	$--$	Not reported	$--$
SVHN	MV-Kum.	0.288 ± 0.025	$--$	669.69 ± 0.37	$--$
4 trials	IRG[1]	0.291 ± 0.017	0.85	668.93 ± 0.53	0.06
1000 labels	M2 (ours)	0.396 ± 0.010	1.86×10^{-04}	Not collected	$--$
$\dim(z) = 50$	Kumar-SB[2]	0.707 ± 0.012	8.10×10^{-08}	669.03 ± 0.43	0.06
	Softmax	0.332 ± 0.009	0.02	669.55 ± 0.11	0.49
	M1 + M2†[3]	0.360 ± 0.001	$--$	Not reported	$--$

Dirichlet Approximation



2-simplex with Kumaraswamy sticks

Gradient Variance



Variance of the ELBO's gradient's first dimension for Categorical data with 100 dimensions and a Dirichlet prior. Others fit a Dirichlet. We fit a MV-Kumaraswamy using $K = 100$ samples (linear complexity) from Uniform(O) to Monte-Carlo approximate the full expectation.

References & Code

Paper and references available at:
`arxiv.org/abs/1905.12052`

Source code available at:
`github.com/astirn/MV-Kumaraswamy`

References

[1] M. Figurnov, S. Mohamed, and A. Mnih, “Implicit reparameterization gradients,” in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 441–452. [Online]. Available: http://papers.nips.cc/paper/7326-implicit-reparameterization-gradients.pdf

[2] E. Nalisnick and P. Smyth, “Stick-breaking variational autoencoders,” *International Conference on Learning Representations (ICLR)*, Apr 2017. [Online]. Available: http://par.nsf.gov/biblio/10039928

[3] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, “Semi-supervised learning with deep generative models,” in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 3581–3589. [Online]. Available: http://papers.nips.cc/paper/5352-semi-supervised-learning-with-deep-generative-models.pdf