Training Discrete Deep Generative Models via Gapped Straight-Through Estimator

Ting-Han Fan[*1], Ta-Chung Chi[*2], Alexander I. Rudnicky[2], Peter J. Ramadge[1]

[1] Princeton University, [2] Carnegie Mellon University
[*] Equal Contribution

ICML | 2022

Training Discrete Deep Generative Models

- Neural network model for a discrete r.v. $\boldsymbol{D}$:

$$\mathbb{P}(\boldsymbol{D} = e_i) = [p_\theta]_i, \qquad p_\theta = \mathrm{Softmax}_1(\mathrm{logit}_\theta).$$

  - θ: trainable parameters.
  - e_i: one-hot vector with 1 at the ith entry.

- Objective:

$$\min_\theta \mathbb{E}_{\boldsymbol{D}\sim p_\theta}[g(\boldsymbol{D})], \qquad g: \text{loss function}.$$

  - If $\boldsymbol{D}$ admits a reparameterization model $D(\theta, \xi)$, update θ by:

$$\nabla_\theta g(D(\theta, \xi)), \qquad \xi: \text{random source}.$$

- Appear in many scenarios:
  - VAE, GAN, Natural Language Processing, Reinforcement Learning

How to reparameterize discrete random variables?

The Family of Gumbel-Softmax Estimators

- Gumbel-Softmax and its Straight-Through Variant[1]:

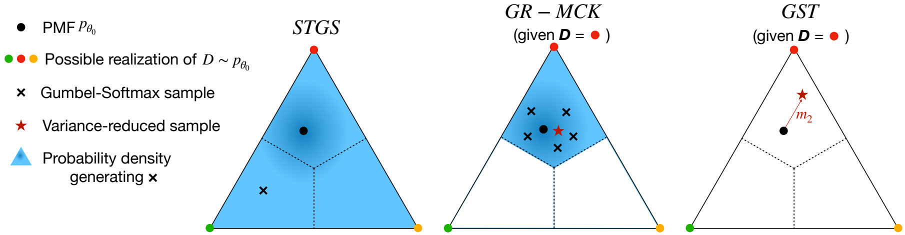
$$D_{\mathsf{GS}}(\theta, \boldsymbol{\xi}) = \mathrm{Softmax}_\tau(\mathrm{logit}_\theta + \boldsymbol{G})$$
$$D_{\mathsf{STGS}}(\theta, \boldsymbol{\xi}) = D(\theta_0, \boldsymbol{\xi}) - D_{\mathsf{GS}}(\theta_0, \boldsymbol{\xi}) + D_{\mathsf{GS}}(\theta, \boldsymbol{\xi})$$

  - $\boldsymbol{G}$: Gumbel(0,1) random vector
  - $\theta_0 = \mathrm{stop\_grad}(\theta)$ is the NN parameter during forward pass.
  - $D(\theta_0, \boldsymbol{\xi}) = \mathrm{sample\_onehot\_from}(p_{\theta_0})$; $p_{\theta_0} = \mathrm{Softmax}_1(\mathrm{logit}_{\theta_0})$
- GR-MCK: Variance reduction of STGS by conditioning and averaging.[2]
- GST: We improve upon the STGS paradigm and propose a method to reduce variance without resampling.

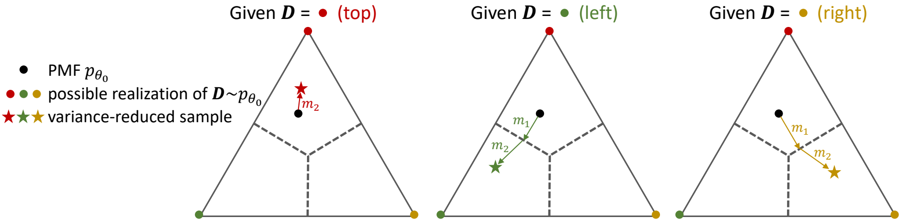
[1] Jang, Gu, and Poole, "Categorical Reparameterization with Gumbel-Softmax". 2017

[2] Paulus, Maddison, and Krause, "Rao-Blackwellizing the Straight-Through Gumbel-Softmax Gradient Estimator". 2021

Illustration of Estimators



Inner Workings of Gapped Straight-Through



- m_1 ensures the sampled category of D is the largest.
- m_2 further ensures the margin of difference.
- Made possible by the three observations and proofs detailed in the paper.

Gapped Straight-Through: Algorithm

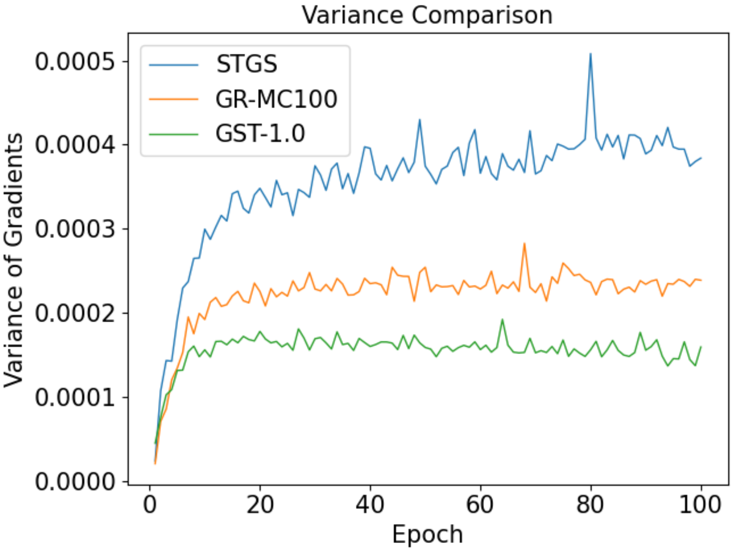
1. Sample a $\boldsymbol{D} = D(\theta_0, \xi) \sim p_{\theta_0}$.
2. Construct perturbation functions

$$m_1(\theta_0, \boldsymbol{D}) = \left(\max_{1 \le j \le N} [\text{logit}_{\theta_0}]_j - \langle \text{logit}_{\theta_0}, \boldsymbol{D} \rangle \right) \cdot \boldsymbol{D}$$

$$m_2(\theta_0, \boldsymbol{D}, g) = \left(\text{logit}_{\theta_0} + g - \max_{1 \le j \le N} [\text{logit}_{\theta_0}]_j \right)_+ \cdot (1 - \boldsymbol{D})$$

   - $g \ge 0$: the gap parameter. Can be set as $g \approx 1$.
3. $h(\theta, \boldsymbol{D}) = \text{Softmax}_\tau(\text{logit}_\theta + m_1 - m_2)$.
4. If hard sample, return $\boldsymbol{D}$ - stop_gradient$(h(\theta, \boldsymbol{D})) + h(\theta, \boldsymbol{D})$.
5. If soft sample, return $h(\theta, \boldsymbol{D})$.

Empirical Verification of Variance on MNIST-VAE



Experiments on MNIST-VAE and ListOps

Temp.	Estimator	Neg. ELBO	Std.
1.0	STGS	122.96	3.08
	GR-MC100	120.65	2.95
	GST-1.0	113.63	1.48
	GST-1.2	**112.58**	**1.11**
0.5	STGS	118.96	2.51
	GR-MC100	117.88	3.01
	GST-1.0	108.43	1.08
	GST-1.2	**107.33**	**0.69**

Temp.	Estimator	Acc.	Std.
1.0	STGS	0.659	0.006
	GR-MC100	0.651	0.009
	GST-1.0	**0.662**	**0.005**
	GST-1.2	0.660	0.011
0.1	STGS	0.645	0.014
	GR-MC100	0.637	0.049
	GST-1.0	**0.664**	**0.012**
	GST-1.2	0.660	0.018

Table: **MNIST-VAE (left, 10 seeds) and ListOps[3] (right, 5 seeds).**

[3] Nangia and Bowman, "Listops: A diagnostic dataset for latent tree learning". 2018

Conclusion

- We propose GST, a low variance gradient estimator for discrete random variables in a neural network.
- Experiment results demonstrate reduced gradient variance and improved task performance.
- Code released at: `https://github.com/chijames/GST`