Straight-Through Estimator as Projected Wasserstein Gradient Flows

Pengyu Cheng[1], Chang Liu[2], Chunyuan Li[3], Dinghan Shen[1], Ricardo Henao[1] and Lawrence Carin[1]

[1]Duke University, [2]Tsinghua University, [3]Microsoft Research

Abstract

To back-propagate gradients through discrete random variables, the Straight-Through (ST) estimator is widely used, but it lacks theoretical justification. In this paper, we interpret ST as the simulation of projected Wasserstein gradient flow (pWGF). Further, a new pWGF estimator variant is proposed, which exhibits superior performance on distributions with infinite support, *e.g.*, Poisson distributions. Empirically, we show that ST and our proposed estimator, while applied to different types of discrete structures (including both Bernoulli and Poisson latent variables), exhibit comparable or even better performances relative to other state-of-the-art methods.

Problem Description

We aim to minimize the expected cost

$$L(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{\theta}}(\boldsymbol{z})}[f(\boldsymbol{z})], \tag{1}$$

where $\boldsymbol{z}$ is a d-dimensional discrete random vector, $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ is a discrete distribution with parameter $\boldsymbol{\theta}$, and $f(\boldsymbol{z})$ is a cost function.

Straight-Through Estimator

In Bernoulli cases, the distribution parameter $\boldsymbol{\theta}$ is the Bernoulli parameter $\boldsymbol{p} = (p^1, p^2, \ldots, p^d)$. Aim to calculate $\nabla_{\boldsymbol{p}}\mathbb{E}_{\boldsymbol{z}\sim\text{Bern}(\boldsymbol{p})}[f(\boldsymbol{z})]$.

In i-th dimension, $z^i \sim \text{Bern}(p^i)$ can be interpreted as $z^i = h(p^i, \epsilon^i) = \mathbf{1}_{p^i>\epsilon^i}$, where $\epsilon^i \sim \text{U}(0,1)$, and $h(p^i, \epsilon^i)$ is a hard threshold function. With the reparameterization trick:

$$\nabla_{\boldsymbol{p}}\mathbb{E}_{\boldsymbol{z}\sim\text{Bern}(\boldsymbol{p})}[f(\boldsymbol{z})] = \nabla_{\boldsymbol{p}}\mathbb{E}_{\boldsymbol{\epsilon}\sim\text{U}(0,1)}[f(h(\boldsymbol{p},\boldsymbol{\epsilon}))] = \mathbb{E}_{\boldsymbol{\epsilon}}[\nabla_{\boldsymbol{p}} f(h(\boldsymbol{p},\boldsymbol{\epsilon}))]$$

When applying chain rule:

$$\frac{\partial f(h(p^i,\epsilon^i))}{\partial p^i} = \frac{\partial f(h(p^i,\epsilon^i))}{\partial h(p^i,\epsilon^i)}\frac{\partial h(p^i,\epsilon^i)}{\partial p^i} = \frac{\partial f}{\partial z^i}\frac{\partial h(p^i,\epsilon^i)}{\partial p^i} \overset{ST}{\approx} \frac{\partial f}{\partial z^i},$$

ST directly sets $\frac{\partial h}{\partial p^i} = 1$, which lacks mathematical justification.

Proposed pWGF Framework

- Denote $\mathcal{M}$ as the d-dimensional discrete distribution family parameterized by $\boldsymbol{\theta}$.

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{\theta}}}[f(\boldsymbol{z})] = \min_{\mu\in\mathcal{M}} \mathbb{E}_{\boldsymbol{z}\sim\mu}[f(\boldsymbol{z})] =: \min_{\mu\in\mathcal{M}} F[\mu].$$

- If the gradient of F on $\mathcal{M}$ as $\nabla_{\mathcal{M}}F$ available, gradient descent can be apply. However, discrete condition on $\mathcal{M}$ makes too many constraints on $\nabla_{\mathcal{M}}F$. if we relax the discrete constraint and perform updates in an appropriate larger space $\tilde{\mathcal{M}}$, the calculation of the gradient $\nabla_{\tilde{\mathcal{M}}}F$ can be much easier.

- We propose a 3-step updating scheme: In k-th iteration,
  1. A: Draw samples $\{\boldsymbol{z}_n\}$ from current distribution μ_k;
  2. B: Update $\{\boldsymbol{z}_n\}$ to $\{\tilde{\boldsymbol{z}}_n\} \sim \tilde{\mu}_k$ via Wasserstein gradient flow in 2-Wasserstein space $\tilde{\mathcal{M}}$;
  3. C: Project $\tilde{\mu}_k$ back to μ_{k+1} by minimizing Wasserstein distance $W(\mu, \tilde{\mu}_k)$.

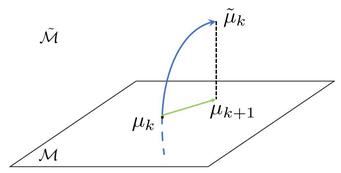

Figure: Updating scheme

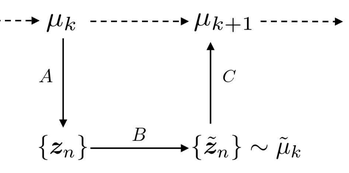

Figure: Algorithm outline

Mathematical Justification to ST

- Straight-Through (ST) estimator is a special case of our projected Wasserstein Gradient Flow (pWGF) updating scheme.
- When projecting $\tilde{\mu}_k$ back to μ_{k+1}, $\mu_{k+1} = \arg\min_{\mu\in\mathcal{M}} W(\mu, \tilde{\mu}_k)$, ST approximates 2-Wasserstein distance via its lower bound, the absolute value of expectation $|\mathbb{E}_{z\sim\mu}[z] - \mathbb{E}_{\tilde{z}\sim\tilde{\mu}_k}[\tilde{z}]|$.
- For one-dimensional Bernoulli distribution, $\mu \in \mathcal{M}$ can be parameterized by p, $\mu = \text{Bern}(p)$. Then we can calculate $\frac{\partial}{\partial p}W(\mu, \tilde{\mu}_k)$ as the direction to minimize $W(\mu, \tilde{\mu}_k)$:

$$\frac{\partial}{\partial p}W(\mu,\tilde{\mu}_k)^2 \approx \frac{\partial}{\partial p}|\mathbb{E}_{z\sim\mu}[z] - \mathbb{E}_{\tilde{z}\sim\tilde{\mu}_k}[\tilde{z}]|^2 = \frac{\partial}{\partial p}(p - \frac{1}{N}\sum_{n=1}^{N}\tilde{z}_n)^2$$
$$= 2(p - \frac{1}{N}\sum_{n=1}^{N}\tilde{z}_n) \approx \frac{2}{N}\sum_{n=1}^{N}(z_n - \tilde{z}_n) = \frac{2\varepsilon}{N}\sum_{n=1}^{N}\nabla f(z_n),$$

which is a multi-sample version ST estimator.

Proposed Estimator

A more principled way to approximate the Wasserstein distance is to use Maximum Mean Discrepancy (MMD): $\Delta^2(\mu,\nu) = \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2\sim\mu}[K(\boldsymbol{x}_1,\boldsymbol{x}_2)] + \mathbb{E}_{\boldsymbol{y}_1,\boldsymbol{y}_2\sim\nu}[K(\boldsymbol{y}_1,\boldsymbol{y}_2)] - 2\mathbb{E}_{\boldsymbol{x}\sim\mu,\boldsymbol{y}\sim\nu}[K(\boldsymbol{x},\boldsymbol{y})]$, where $K(\cdot,\cdot)$ is a selected kernel. In practice, instead of minimizing $W(\mu,\tilde{\mu}_k)$, we can minimize the empirical expectation $\Delta^2(\mu,\tilde{\mu}) \approx \mathbb{E}_{z_1,z_2\sim\mu}[K(z_1,z_2)] + \frac{1}{N^2}\sum_{n,n'=1}^{N}K(\tilde{z}_n,\tilde{z}_{n'}) - 2\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{z\sim\mu}K(z,\tilde{z}_n)$.

Experiment Results

pWGF shows improvement on parameter inference task of Poisson distributions.

- Real data $\{z_n\} \sim p(z) = \text{Poisson}(\lambda_0 = 5)$.
- Fake data $\{z'_n\} \sim q_\lambda(z) = \text{Poisson}(\lambda)$
- Discriminator $D_\omega(z)$ gives probability that z comes from real data.
- $\max_\lambda \min_\omega\{\mathbb{E}_{z\sim p}[\log D_\omega(z)] + \mathbb{E}_{z'\sim q_\lambda}[\log(1 - D_\omega(z'))]\}$

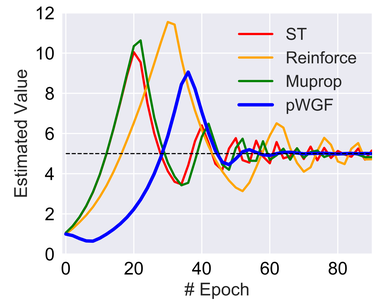

Figure: Learning Curves of Poisson parameter

	pWGF	ST	Muprop	Reinforce
Mean	5.0076	5.1049	5.0196	4.9452
Std	0.013	0.161	0.159	0.173

Table: Mean and Standard Derivation of Inference

Conclusion

We explain the origin of the widespread adoption of ST estimator, and represent a helpful step towards exploring alternative gradient estimators for discrete variables.