Zijing Ou web: https://j-zin.github.io email: ouzj@mail2.sysu.edu.cn

# Research Proposal

of Zijing Ou (CS DPhil applicant for Fall, 2022)

Most machine learning (ML) methods depend on gradient descent optimization, which raises a concern of the differentiability in discrete cases. On one hand, discreteness appears more naturally from the pointview of human mind, which is unavoidable in use and helps improve interpretability; on the other hand, discrete values requires fewer computational resources and memory footprints, making ML algorithms more scalable. In this circumstance, approximating gradient for discrete variables with low variance and unbiased estimators appeals to researchers. At present, research in discrete gradient estimation centers around two aspects: developing methods for estimating gradient and utilizing approximated gradient for various ML tasks.

## 1 Gradient Estimation through Discreteness

In statistic inference, we generally focus on maximizing an expected reward of certain quantities $f(z)$ of interest over the distribution $p_\theta(z)$ with parameters $\theta$, denoted by $\mathbb{E}_p[f(z)]$. To optimize parameters, we are acquired to compute $\nabla_\theta \mathbb{E}_p[f(z)]$. if $z \sim p_\theta$ is *continuous* and can be generated via a probability transformation $z = \mathcal{T}_\theta(\epsilon), \epsilon \sim q(\epsilon)$, then the gradient can be calculated as

$$\nabla_\theta \mathbb{E}_{p_\theta(z)}[f(z)] = \mathbb{E}_{q(\epsilon)}\left[\nabla_\theta f(\mathcal{T}_\theta(\epsilon))\right]. \tag{1}$$

This trick is known as reparameterization [1, 2, 3], which have been generalized to a variety of variational distributions [4]. However, these methods cannot be applied to *discrete* random variables since they require a differentiable density distribution with respect to the variable. To recover efficient back-propagation, recently many researchers focus on how to estimate gradient through discrete variables.

REINFORCE algorithm [5] is an practicable strategy with the log-derivative trick, *i.e.*, $\nabla_\theta \mathbb{E}_p[f(z)] = \mathbb{E}_p[f(z)\nabla_\theta \log p_\theta(z)]$, while it suffers from high variance which limits its practical use. One simple trick to reduce the variance is subtracting a baseline from the objectives, which does not alter the expectation [6, 7, 8, 9]. To leverage the information of gradients through objectives, reparameterization tricks [10, 11] for the discrete random variables are proposed for the sake of lower variance, though at the cost of introducing biases. Grathwohl et al. [8] and Tucker et al. [9] proposed to eliminate biases and reduce variance by exploiting conditional Gumbel relaxation as baseline for the REINFORCE estimator. Recently, a more general estimator GO gradient [12] makes reparameterization applicable for both continuous and discrete cases. Beyond that, antithetic coupling [13] is another method to reduce variance in Monte Carlo sampling. It was first exploited in ARM estimator [14] for binary variables, and subsequently expanded to categorical cases in ARSM [15]. Although ARM reduces variance via antithetic sampling, it also increase variance due to the reparameterization. Dong et al. [16] addresses this issue by marginalizing out the continuous augmentation. Moreover, the Rao-Blackwellization [17] seems to be very popular in recent years. Kool et al. [18] derived a novel estimator based on sampling without replacement [19], which is equivalent to Rao-Blackwellizing and thus can reduce variance. Paulus et al. [20] also proposed a excellent estimator that equips Gumbel-softmax with the Rao-Blackwellization augmentation. Though successful so far, these methods are generally difficult to implement, which undermines the popularity of discrete variables in ML fields.

## 2 Applications of Gradient Estimation

As the above mentioned, the discrete gradient estimate has remained an open problem, which means there still lies space for me to delve deeper. Besides, I am keen on this line of research because it has many applications in ML such as energy based models (EBM), generative models, discrete representation learning, etc.

**Energy-based Models.** The EBMs, *i.e.* $p_\theta(x) = \frac{1}{Z(\theta)} exp(-E(x;\theta))$, is a distribution family with unknown normalizing constant. By noting that $\log p_\theta(x) = -E(x;\theta)$, one can train EBMs by estimating the gradient of log probability with respect to data. Specifically, we can minimize the Stein discrepancy [21]

$$D(q,p) := \sup_{f\in\mathcal{F}} \mathbb{E}_q[(\nabla_x \log p_\theta(x) - \nabla_x \log q(x))f(x)], \tag{2}$$

where $q(x)$ is the empirical distribution of data. It turns out that (2) has close form by restricting $\mathcal{F}$ to be the unit ball of reproducing kernel Hilbert space [22] and the well-known Fisher divergence [23] is a special case of (2) by carefully restricting the space $\mathcal{F}$. Such gradient estimation methods, as well as its nonparametric version [24], have been applied in various applications such as goodness-of-fit [22], posterior inference [25] etc. However, these methods apply exclusively to distributions with smooth density functions. Only very few literature focuses on discrete scenarios [26, 27], leaving a large research gap to explore.

**Generative Modelling.** The estimated gradient can be used for generating samples, which is also known as gradient-based Markov chain Monte Carlo. For example, we can iteratively generate samples via

$$x_{t+1} = x_t + \epsilon\nabla_x \log p(x_t) + \sqrt{2\epsilon}z_t, \quad t = 1, 2, ... \tag{3}$$

where $\epsilon$ is the step size and $z_t \sim \mathcal{N}(0,1)$. Known as Langiven dynamics [28], this method has been shown to generate samples competitive with GAN [29]. However, due to the undifferentiability property, it has not been widely used to generate discrete data like graphical, textual and sequential data.

**Discrete Representation Learning.** Discrete representation learning is attractive in recent years, thanks to its fewer computational resources and memory footprint. Beyond generative models [1], mutual information theorem [30] recently inspired a line of work to learn informative and robust representation. Moreover, Stratos and Wiseman [31] attempted to learn discrete representation $b$ of the given data $x$ by

$$\max_\psi \min_\theta \mathbb{E}_{p_\theta(b,x)}[\log p_\theta(b|x) - \log q_\phi(b)], \tag{4}$$

where $p_\theta(b|x)$ is the encoder with parameter $\theta$, $q_\phi(z)$ is a variational distribution, and $\min_\phi \mathbb{E}_{p_\theta}[\log \frac{p_\theta(b|x)}{q_\phi(z)}] = \mathbb{I}(b;x)$ holds for fix $\theta$ with $\mathbb{I}(\cdot;\cdot)$ denoting mutual information. This work bridges the gap between information theory and discrete representation, while the adversarial training process is unstatable.

# 3 Future Research Plans

In my previous research, we adopted latent variable models to learn the structured representation for document hashing [32], then extended it to image hashing via contrastive information bottleneck [33]. Recently, my research interests lie in energy based models and mainly focus on EBM inference through discreteness. Thanks to my supervisor, who assigned me the first research topic on discrete gradient estimation of generative hashing [34], I became fascinated with approximate inference in my undergraduate years. In this sequel, I aim at building a more advanced algorithm for Bayesian inference with discrete variables, then apply them for various applications as the aforementioned. I especially dream of becoming a qualified Bayesian who can describe the world from the viewpoints of probability, uncertainty, and rationality. To more specific, my Ph.D. research attempts to solve the following sub-objectives:

- To derive a more scalable and efficient gradient estimation for discrete probability models.
- To adopt this technique in representation learning and discrete data (*e.g.,* text, graph) generation.
- To further explore better approximate inference and sampling methods for Bayesian computation.

I strongly believe that the above research directions can advance the machine learning research and applications, and will have impacts in both academic and industry.

## References


[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[2] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[3] Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *International conference on machine learning*, pages 1971–1979. PMLR, 2014.

[4] Francisco R Ruiz, Titsias RC AUEB, David Blei, et al. The generalized reparameterization gradient. *Advances in neural information processing systems*, 29:460–468, 2016.

[5] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[6] Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, pages 1791–1799. PMLR, 2014.

[7] Shixiang Gu, Sergey Levine, Ilya Sutskever, and Andriy Mnih. Muprop: Unbiased backpropagation for stochastic neural networks. *arXiv preprint arXiv:1511.05176*, 2015.

[8] Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *arXiv preprint arXiv:1711.00123*, 2017.

[9] George Tucker, Andriy Mnih, Chris J Maddison, Dieterich Lawson, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. *arXiv preprint arXiv:1703.07370*, 2017.

[10] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[11] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[12] Yulai Cong, Miaoyun Zhao, Ke Bai, and Lawrence Carin. Go gradient for expectation-based objectives. *arXiv preprint arXiv:1901.06020*, 2019.

[13] Art B Owen. Monte carlo theory, methods and examples. 2013.

[14] Mingzhang Yin and Mingyuan Zhou. Arm: Augment-reinforce-merge gradient for stochastic binary networks. *arXiv preprint arXiv:1807.11143*, 2018.

[15] Mingzhang Yin, Yuguang Yue, and Mingyuan Zhou. Arsm: Augment-reinforce-swap-merge estimator for gradient backpropagation through categorical variables. In *International Conference on Machine Learning*, pages 7095–7104. PMLR, 2019.

[16] Zhe Dong, Andriy Mnih, and George Tucker. Disarm: An antithetic gradient estimator for binary latent variables. *arXiv preprint arXiv:2006.10680*, 2020.

[17] C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in statistics*, pages 235–247. Springer, 1992.

[18] Wouter Kool, Herke van Hoof, and Max Welling. Estimating gradients for discrete random variables by sampling without replacement. *arXiv preprint arXiv:2002.06043*, 2020.

[19] Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International Conference on Machine Learning*, pages 3499–3508. PMLR, 2019.

[20] Max B Paulus, Chris J Maddison, and Andreas Krause. Rao-blackwellizing the straight-through gumbel-softmax gradient estimator. *arXiv preprint arXiv:2010.04838*, 2020.

[21] Jackson Gorham. *Measuring sample quality with Stein's method*. Stanford University, 2017.

[22] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In

*International conference on machine learning*, pages 276–284. PMLR, 2016.

[23] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[24] Yingzhen Li and Richard E Turner. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

[25] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

[26] Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR, 2018.

[27] Jun Han, Fan Ding, Xianglong Liu, Lorenzo Torresani, Jian Peng, and Qiang Liu. Stein variational inference for discrete distributions. In *International Conference on Artificial Intelligence andowen2013monte Statistics*, pages 4563–4572. PMLR, 2020.

[28] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[29] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019.

[30] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[31] Karl Stratos and Sam Wiseman. Learning discrete structured representations by adversarially maximizing mutual information. In *International Conference on Machine Learning*, pages 9144–9154. PMLR, 2020.

[32] Zijing Ou, Qinliang Su, Jianxing Yu, Bang Liu, Jingwen Wang, Ruihui Zhao, Changyou Chen, and Yefeng Zheng. Integrating semantics and neighborhood information with graph-driven generative models for document retrieval. *arXiv preprint arXiv:2105.13066*, 2021.

[33] Zexuan Qiu, Qinliang Su, Zijing Ou, Jianxing Yu, and Changyou Chen. Unsupervised hashing with contrastive information bottleneck. *arXiv preprint arXiv:2105.06138*, 2021.

[34] Dinghan Shen, Qinliang Su, Paidamoyo Chapfuwa, Wenlin Wang, Guoyin Wang, Lawrence Carin, and Ricardo Henao. Nash: Toward end-to-end neural architecture for generative semantic hashing. *arXiv preprint arXiv:1805.05361*, 2018.