GENERALIZED TRANSFORMATION-BASED GRADIENT

Anonymous authors
Paper under double-blind review

ABSTRACT

The reparameterization trick has become one of the most useful tools in the field of variational inference. However, the reparameterization trick is based on the standardization transformation which restricts the scope of application of this method to distributions that have tractable inverse cumulative distribution functions or are expressible as deterministic transformations of such distributions. In this paper, we generalized the reparameterization trick by allowing a general transformation. Unlike other similar works, we develop the generalized transformation-based gradient model formally and rigorously. We discover that the proposed model is a special case of control variate indicating that the proposed model can combine the advantages of CV and generalized reparameterization. Based on the proposed gradient model, we propose a new polynomial-based gradient estimator which has better theoretical performance than the reparameterization trick under certain condition and can be applied to a larger class of variational distributions. In studies of synthetic and real data, we show that our proposed gradient estimator has a significantly lower gradient variance than other state-of-the-art methods thus enabling a faster inference procedure.

1 INTRODUCTION

Most machine learning objective function can be rewritten in the form of an expectation:

$$\mathcal{L} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}[f_{\boldsymbol{\theta}}(\boldsymbol{z})] \tag{1}$$

where $\boldsymbol{\theta}$ is a parameter vector. However, due to the intractability of the expectation, it's often impossible or too expensive to calculate the exact gradient w.r.t $\boldsymbol{\theta}$, therefore it's inevitable to estimate the gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}$ in practical applications. Stochastic optmization methods such as reparameterization trick and score function methods have been widely applied to address the stochastic gradient estimation problem. Many recent advances in large-scale machine learning tasks have been brought by these stochastic optimization tricks. Like in other stochastic optimzation related works, our paper mainly focus on variational inference tasks.

The primary goal of variational inference (VI) task is to approximate the posterior distribution in probabilistic models (Jordan et al., 1999; Wainwright & Jordan, 2008). To approximate the intractable posterior $p(\mathbf{z}|\mathbf{x})$ with the joint probability distribution $p(\mathbf{x}, \mathbf{z})$ over observed data $\mathbf{x}$ and latent random variables $\mathbf{z}$ given, VI introduces a parameteric family of distribution $q_{\boldsymbol{\theta}}(\mathbf{z})$ and find the best parameter $\boldsymbol{\theta}$ by optimizing the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(q(\mathbf{z}; \boldsymbol{\theta}) \| p(\mathbf{z}|\mathbf{x}))$. The performance of VI methods depends on the capacity of the parameteric family of distributions (often measured by Rademacher complexity) and the ability of the optimizer. In this paper, our method tries to introduce a better optimizer for a larger class of parameteric family of distributions.

The main idea of our work is to replace the parameter-independent transformation in reparameterization trick with generalized transformation and construct the generalized transformation-based (G-TRANS) gradient with the velocity field which is related to the characteristic curve of the sublinear partial differential equation associated with the generalized transformation. Our gradient model further generalizes the G-REP (Ruiz et al., 2016) and provides a more elegant and flexible way to construct gradient estimators. We mainly make the following contributions:

1. We develop a generalized transformation-based gradient model based on the velocity field related to the generalized transformation and explicitly propose the unbiasedness constraint

      on the G-TRANS gradient. The proposed gradient model provides a more poweful and flexible way to construct gradient estimators.

2. We show that our model is a generalization of the score function method and the reparameterization trick. Our gradient model can reduce to the reparameterization trick by enforcing a transport equation constraint on the velocity field. We also show our model's connection to control variate method.

3. We propose a polynomial-based gradient estimator that cannot be induced by any other existing generalized reparameterization gradient framework, and show its superiority over similar works on several experiments.

The rest of this paper is organized as follows. In Sec.2 we review the stochastic gradient variational inference (SGVI) and stochastic gradient estimators. In Sec.3 we propose the generalized transformation-based gradient. In Sec.4 we propose the polynomial-based G-TRANS gradient estimator. In Sec.5 we study the performance of our gradient estimator on synthetic and real data. In Sec.6 we review the related works. In Sec.7 we conclude this paper and discuss future work.

2 STOCHASTIC VARIATIONAL INFERENCE

To obtain the best variational parameter $\boldsymbol{\theta}$, rather than minimize the KL divergence $D_{\mathrm{KL}}(q(\mathbf{z};\boldsymbol{\theta})\|p(\mathbf{z}|\mathbf{x}))$, we usually choose to maximize the evidence lower bound (ELBO) (Jordan et al., 1999),

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\theta})}[f(\mathbf{z})] + \mathbb{H}[q(\mathbf{z};\boldsymbol{\theta})] \tag{2}$$

where $f(\mathbf{z}) = \log p(\mathbf{x},\mathbf{z})$ and $\mathbb{H}[q(\mathbf{z};\boldsymbol{\theta})] = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\theta})}[-\log q(\mathbf{z};\boldsymbol{\theta})]$. The entropy term $\mathbb{H}[q(\mathbf{z};\boldsymbol{\theta})]$ is often assumed to be available analytically and usually omitted in the procedure of stochastic optimization. This stochastic optimization problem is the basic setting for our method and experiments. Without extra description, we only consider the simplified version of the ELBO:

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\theta})}[f(\mathbf{z})] \tag{3}$$

Generally, this expectation is intractable to compute, let alone its gradient. Therefore, a common stochastic optimization method for VI task is to construct a Monte Carlo estimator for the exact gradient of the ELBO w.r.t $\boldsymbol{\theta}$. Among those gradient estimators, the score function method and the reparamterization trick are most popular and widely applied.

Score function method. The score function estimator, also called log-derivative trick or reinforce Glynn (1990); Williams (1992) is a general way to obtain unbiased stochastic gradients of the ELBO (Paisley et al., 2012; Ranganath et al., 2014; Mnih & Gregor, 2014). The simplest variant of the score function gradient estimator is defined as:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z})}[f(\mathbf{z})\nabla_{\boldsymbol{\theta}}\log q_{\boldsymbol{\theta}}(\mathbf{z})] \tag{4}$$

and then we can build the Monte Carlo estimator by drawing samples from the variational distribution $q_{\boldsymbol{\theta}}(\mathbf{z})$ independently.

Although the score function method is very general, the resulting gradient estimator suffers from high variance. Therefore, it's necessary to apply variance reduction (VR) methods such as Rao-Blackwellization (Casella & Robert, 1996) and control variates (Robert & Casella, 2013) in practice.

Reparameterization trick. In reparameterization trick, we assume that there is an invertible and continuously differentiable standardization function $\phi(\mathbf{z},\boldsymbol{\theta})$ that can transform the variational distribution $q(\mathbf{z};\boldsymbol{\theta})$ into a distribution $s(\boldsymbol{\rho})$ that don't depend on the variational parameter $\boldsymbol{\theta}$ as follows,

$$\phi(\mathbf{z},\boldsymbol{\theta}) = \boldsymbol{\rho} \sim s(\boldsymbol{\theta}), \qquad \mathbf{z} = \phi_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\rho})$$

Then the reparameterization trick can turn the computation of the gradient of the expectation into the expectation of the gradient:

$$\nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z})}[f(\mathbf{z})] = \mathbb{E}_{s(\boldsymbol{\rho})}[\nabla_{\boldsymbol{\theta}} f\left(\phi_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\rho})\right)]$$

Although this reparameterization can be done for many commonly used distributions, such as the Gaussian distribution, it's hard to find appropriate standardization functions for a number of standard distributions, such as Gamma, Beta or Dirichlet because the standardization functions will inevitably involve special functions. On the other hand, though the reparameterization trick is not as generally applicable as the score function method, it does result in a gradient estimator with lower variance.

3 THE GENERALIZED TRANSFORMATION-BASED GRADIENT MODEL

Define a random variable $\boldsymbol{\rho}$ by an invertible differentiable transformation $\boldsymbol{\rho} = \phi(\mathbf{z}, \boldsymbol{\theta})$, where ϕ is commonly called generalized standardization transformation (Ruiz et al., 2016) since it's dependent on the variational parameter $\boldsymbol{\theta}$.

Theorem 3.1. *Let θ be any component of the variational parameter $\boldsymbol{\theta}$, the probability density function of $\boldsymbol{\rho}$ be $w(\boldsymbol{\rho}, \boldsymbol{\theta})$ and $\mathcal{L} = \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\theta})}[f(\mathbf{z})]$. Then*

(1) $\frac{\partial w}{\partial \theta}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\phi(\mathbf{z}, \boldsymbol{\theta})} = \frac{1}{|\nabla_{\mathbf{z}}\phi(\mathbf{z}, \boldsymbol{\theta})|}\left(\frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + \nabla_{\mathbf{z}} \cdot (q(\mathbf{z}, \boldsymbol{\theta})\mathbf{v}^{\theta}(\mathbf{z}, \boldsymbol{\theta}))\right)$

(2) $\frac{\partial \mathcal{L}}{\partial \theta} = \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\theta})}\left[\frac{\partial \log q(\mathbf{z}, \boldsymbol{\theta})}{\partial \theta}f(\mathbf{z}) + \frac{1}{q(\mathbf{z}, \boldsymbol{\theta})}\nabla_{\mathbf{z}} \cdot (f(\mathbf{z})q(\mathbf{z}, \boldsymbol{\theta})\mathbf{v}^{\theta}(\mathbf{z}, \boldsymbol{\theta}))\right]$

where

$$\mathbf{v}^{\theta}(\mathbf{z}, \boldsymbol{\theta}) = \frac{d\mathbf{z}}{d\theta} = -(\nabla_{\mathbf{z}}\phi(\mathbf{z}, \boldsymbol{\theta}))\frac{\partial \phi(\mathbf{z}, \boldsymbol{\theta})}{\partial \theta} \tag{5}$$

$\mathbf{v}^{\theta}(\mathbf{z}, \boldsymbol{\theta})$ is often referred to as velocity field, and the Equ.5 above is called velocity field equation. The Monte Carlo estimator $\widehat{\frac{\partial \mathcal{L}}{\partial \theta}}$ for $\frac{\partial \mathcal{L}}{\partial \theta}$ is unbiased only when $\mathbf{v}^{\theta}(\mathbf{z}, \boldsymbol{\theta})$ satisfy the unbiasedness constraint:

$$\mathbb{E}_{q_{\boldsymbol{\theta}}}\left[\frac{1}{q_{\boldsymbol{\theta}}}\nabla_{\mathbf{z}} \cdot (f q_{\boldsymbol{\theta}}\mathbf{v}^{\theta})\right] = 0 \tag{6}$$

The proof details of the Theorem.3.1 are included in the Appendix.A.1.

We refer to the gradient $\frac{\partial \mathcal{L}}{\partial \theta}$ with $\mathbf{v}^{\theta}$ satisfying the unbiasedness constraint as generalized transformation-based (G-TRANS) gradient. We can construct the G-TRANS gradient estimator by choosing $\mathbf{v}^{\theta}$ of specific form. In the following, we demonstrate that the score function gradient and reparameterization gradient are special cases of our G-TRANS gradient model associating with special velocity fields.

Remark. *The score function method is a special case of the G-TRANS model when $f(\mathbf{z})q(\mathbf{z}, \boldsymbol{\theta})\mathbf{v}^{\theta} = const$.*

The standardization function ϕ doesn't depend on the parameter $\boldsymbol{\theta}$ when $\mathbf{v}^{\theta} = 0$ according to the velocity field equation (Equ.5). Conversely, for any ϕ that doesn't depend on $\boldsymbol{\theta}$, we have $\mathbf{v}^{\theta} = -(\nabla_{\mathbf{z}}\phi(\mathbf{z}, \boldsymbol{\theta}))\frac{\partial \phi(\mathbf{z}, \boldsymbol{\theta})}{\partial \theta} = 0$, thus the resulting gradient estimator has a same variance as the score function estimator.

Remark. *The reparameterization trick is a special case when $\frac{\partial w}{\partial \theta}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\phi(\mathbf{z}, \boldsymbol{\theta})} = 0$, that's to say*

$$\frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + \nabla_{\mathbf{z}} \cdot (q(\mathbf{z}, \boldsymbol{\theta})\mathbf{v}^{\theta}) = 0 \tag{7}$$

The detailed computation to obtain the transport equation (Equ.7) is included in the Appendix.A.1. The transport equation is firstly introduced by (Jankowiak & Obermeyer, 2018), however, their work derive this equation by an analog to the optimal transport theory. In 1-dimensional case, for any standardization distributions $w(\rho)$ that doesn't depend on the parameter $\boldsymbol{\theta}$, the variance of the resulting gradient estimator is some constant (for fixed $\boldsymbol{\theta}$) determined by the unique 1-dimensional solution of the transport equation.

Remark. Let $g(\mathbf{z}, \boldsymbol{\theta}) = -\nabla_{\mathbf{z}} \cdot (f(\mathbf{z})q(\mathbf{z}, \boldsymbol{\theta})\mathbf{v}^\theta(\mathbf{z}, \boldsymbol{\theta}))$, then $\frac{\partial \mathcal{L}}{\partial \theta} = \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\theta})}[\frac{\partial \log q(\mathbf{z}, \boldsymbol{\theta})}{\partial \theta} f(\mathbf{z}) - \frac{1}{q(\mathbf{z}, \boldsymbol{\theta})}g(\mathbf{z}, \boldsymbol{\theta})]$. $C(\mathbf{z}, \boldsymbol{\theta}) = \frac{1}{q(\mathbf{z}, \boldsymbol{\theta})}g(\mathbf{z}, \boldsymbol{\theta})$ can be viewed as a control variable with $\mu_C = \mathbb{E}[C(\mathbf{z}, \boldsymbol{\theta})] = 0$.

For the existence of the velocity field $\mathbf{v}^\theta$ and the generalized standardization transformation $\phi(\mathbf{z}, \boldsymbol{\theta})$, $g(\mathbf{z}, \boldsymbol{\theta})$ must satisfy some strong differential constraints (Evans, 2010). We can see that the G-TRANS model is a special case of the control variate method with a complex differential structure. This connection to CV means our gradient model can combine the advantages of CV and generalized reparameterization.

Theorem.3.1 transforms the generalized unbiased reparameterization procedure into finding the appropriate velocity field that satisfy the unbiasedness constraint. It's possible to apply variational optimization theory to find the velocity field with the least estimate variance, however, the solution to the Euler-Lagrange equation contains $f(\mathbf{z})$ in the integrand which makes it impractical to use in real-world model (See Appendix.A.2 for details).

By introducing the notion of velocity field, we provide a more elegant and flexible way to construct gradient estimator without the need to compute the Jacobian matrix for a specific transformation. In the next section, we introduce a polynomial-based G-TRANS gradient estimator that cannot be incorporated into any other existing generalized reparameterized gradient framework and is better than the reparameterization gradient estimator theoretically.

4 THE POLYNOMIAL-BASED G-TRANS GRADIENT ESTIMATOR

In this section, we always assume that the base distribution $q(\mathbf{z}, \boldsymbol{\theta})$ can be factorized as $\prod_{i=1}^{N} q_i(z_i, \boldsymbol{\theta_i})$ where N is the dimension of the random variable $\mathbf{z}$, $\boldsymbol{\theta_i}$ is a slice of $\boldsymbol{\theta}$ and $\boldsymbol{\theta_i}$ share no component with $\boldsymbol{\theta_j}$ if $i \neq j$. We consider an ad-hoc velocity field family:

$$\mathbf{v}_{ah}^\theta(\mathbf{z}, \boldsymbol{\theta}) = (-\frac{h(\mathbf{z}, \boldsymbol{\theta})}{f(\mathbf{z})}\frac{\delta_i(\theta)}{q(\mathbf{z}, \boldsymbol{\theta})}\int^{z_i}\frac{\partial q(\mathbf{z}', \boldsymbol{\theta})}{\partial \theta}dz_i') \tag{8}$$

with

$$\delta_i(\theta) = \begin{cases} 1 & \theta \text{ is a component of } \boldsymbol{\theta_i} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

We always assume $\mathbf{v}_{ah}^\theta$ to be continuous which guarantees the existence of the solution to the velocity field equation. We verify in the Appendix.A.3 that $\mathbf{v}_{ah}^\theta(\mathbf{z}, \boldsymbol{\theta})$ satisfy the unbiasedness constraint if $h(\mathbf{z}, \boldsymbol{\theta})$ is bounded.

It's easy to see that the gradient estimator that results from $\mathbf{v}_{ah}^\theta$ is more general than the score function method or reparameterization trick since they are two special cases when $h(\mathbf{z}, \boldsymbol{\theta}) = 0$ or $h(\mathbf{z}, \boldsymbol{\theta}) = f(\mathbf{z})$ respectively. In this paper, we mainly consider a more special family of the $\mathbf{v}_{ah}^\theta(\mathbf{z}, \boldsymbol{\theta})$:

$$\mathbf{v}_{poly}^\theta(\mathbf{z}, \boldsymbol{\theta}) = (-\frac{f(\mathbf{z}) - P_n(\mathbf{z}, \boldsymbol{\theta})}{f(\mathbf{z})}\frac{\delta_i(\theta)}{q(\mathbf{z}, \boldsymbol{\theta})}\int^{z_i}\frac{\partial q(\mathbf{z}', \boldsymbol{\theta})}{\partial \theta}dz_i') \tag{10}$$

where $P_k(\mathbf{z}, \boldsymbol{\theta}) = \sum_{i=0}^{k}\sum_{j_1+\cdots+j_N=i} C_i(\boldsymbol{\theta})z_1^{j_1}\cdots z_N^{j_N}$ is an polynomial of degree k. A dual form of $\mathbf{v}_{poly}^\theta(\mathbf{z}, \boldsymbol{\theta})$ is that $\mathbf{v}_{dp}^\theta(\mathbf{z}, \boldsymbol{\theta}) = (-\frac{P_n(\mathbf{z}, \boldsymbol{\theta})}{f(\mathbf{z})}\frac{\delta_i(\theta)}{q(\mathbf{z}, \boldsymbol{\theta})}\int^{z_i}\frac{\partial q(\mathbf{z}', \boldsymbol{\theta})}{\partial \theta}dz_i')$, but their properties are similar (we present some theoretical results of $\mathbf{v}_{dp}^\theta$ in the Appendix.A.4). Therefore we only consider $\mathbf{v}_{poly}^\theta(\mathbf{z}, \boldsymbol{\theta})$ here. We refer to $\mathbf{v}_{poly}^\theta$ as polynomial velocity field.

Proposition 4.1. *For distributions with analytical high order moments such as Gamma, Beta or Dirichlet distribution, the expectation $\mathbb{E}_{q(\mathbf{z}, \boldsymbol{\theta})}[P_k(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta}\log q]$ has an analytical expression.*

Proof. $\mathbb{E}_q[P_k(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta} \log q] = \frac{\partial}{\partial \theta}\mathbb{E}_q[P_k] - \mathbb{E}_q[\frac{\partial}{\partial \theta}P_k]$. Both $P_k(\mathbf{z}, \boldsymbol{\theta})$ and $\frac{\partial}{\partial \theta}P_k(\mathbf{z}, \boldsymbol{\theta})$ are polynomials of random variable $\mathbf{z}$. Therefore, for distribution with analytical high order moments, $\mathbb{E}_q[P_k]$ and $\mathbb{E}_q[\frac{\partial}{\partial \theta}P_k]$ both have analytical expressions, thus so does $\mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}[P_k(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta} \log q]$. $\square$

With Proposition.4.1, we can write the G-TRANS gradient for the polynomial velocity field as:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \theta} &= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}\left[\frac{\partial \log q(\mathbf{z}, \boldsymbol{\theta})}{\partial \theta}f(\mathbf{z}) + \frac{1}{q(\mathbf{z}, \boldsymbol{\theta})}\nabla_{\mathbf{z}} \cdot (f(\mathbf{z})q(\mathbf{z}, \boldsymbol{\theta})\mathbf{v}_{poly}^{\theta}(\mathbf{z}, \boldsymbol{\theta}))\right] \\
&= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}\left[P_k(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta} \log q\right] - \mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}\left[\frac{1}{q}\sum_i \frac{\partial(f - P_k)}{\partial z_i}\delta_i(\theta)\int^{z_i} \frac{\partial q(\mathbf{z}', \boldsymbol{\theta})}{\partial \theta}dz_i'\right]
\end{aligned}
\tag{11}
$$

Thus we can construct a G-TRANS gradient estimator based upon the polynomial velocity field with a sample $\tilde{\mathbf{z}}$ drawn from $q(\mathbf{z}, \boldsymbol{\theta})$:

$$
\widehat{\frac{\partial \mathcal{L}}{\partial \theta}} = \mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}[P_k(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta} \log q](\tilde{\mathbf{z}}) - \sum_i \frac{\partial(f - P_k)}{\partial z_i}\delta_i(\theta)\frac{\int^{\tilde{z}_i} \frac{\partial q(\mathbf{z}', \boldsymbol{\theta})}{\partial \theta}}{q(\tilde{\mathbf{z}}, \boldsymbol{\theta})}dz_i'
\tag{12}
$$

The polynomial-based G-TRANS gradient estimator has a form close to control variate, thus cannot be induced by any other existing generalized reparameterized gradient framework. In the following, we show that the polynomial-based G-TRANS gradient estimator performs better than the reparameterization gradient estimator under some condition.

Proposition 4.2. *If* $\text{Cov}(\frac{1}{q(\mathbf{z},\boldsymbol{\theta})}\Lambda(\mathbf{z}, \theta) \cdot \nabla_{\mathbf{z}}(P_k), \frac{1}{q(\mathbf{z},\boldsymbol{\theta})}\Lambda(\mathbf{z}, \theta) \cdot \nabla_{\mathbf{z}}(2f - P_k)) > 0$ *where* $\Lambda(\mathbf{z}, \theta) = (\delta_1(\theta)\int_{-\infty}^{z_1} \frac{\partial q(\mathbf{z}', \boldsymbol{\theta})}{\partial \theta}dz_1', ..., \delta_N(\theta)\int_{-\infty}^{z_N} \frac{\partial q(\mathbf{z}', \boldsymbol{\theta})}{\partial \theta}dz_N')$, *then the gradient estimator resulted from polynomial velocity field has a smaller variance than the reparameterization gradient estimator.*

Proof. Since $\mathbb{E}_q[P_k(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta} \log q]$ can be resolved analytically, we have

$$
\text{Var}(\frac{\partial \log q}{\partial \theta_i}f + \frac{1}{q}\nabla_{\mathbf{z}} \cdot (fq\mathbf{v}^{\theta_i})) = \text{Var}(-\frac{\partial f - P_k}{\partial z_i}\frac{\int^{z_i} q_i(z_i', \boldsymbol{\theta}_i)dz_i'}{q_i(z_i', \boldsymbol{\theta}_i)})
$$

then by reorganizing the expression $\text{Var}(-\frac{\partial f}{\partial z_i}\frac{\int^{z_i} q_i(z_i', \boldsymbol{\theta}_i)dz_i'}{q_i(z_i', \boldsymbol{\theta}_i)}) - \text{Var}(-\frac{\partial f - P_k}{\partial z_i}\frac{\int^{z_i} q_i(z_i', \boldsymbol{\theta}_i)dz_i'}{q_i(z_i', \boldsymbol{\theta}_i)})$, we can prove this proposition. $\square$

Remark. *As an example about how to choose a good polynomial, for* $P_1(\mathbf{z}, \boldsymbol{\theta}) = C_0(\boldsymbol{\theta}) + \sum_{i=1}^{N} C_i(\boldsymbol{\theta})z_i$, *by choosing* $C_i(\boldsymbol{\theta}) \in (0, \frac{2Cov(\frac{\int^{z_i} q_i(z_i', \boldsymbol{\theta}_i)dz_i'}{q_i(z_i, \boldsymbol{\theta}_i)}, \frac{\partial f}{\partial z_i}\frac{\int^{z_i} q_i(z_i', \boldsymbol{\theta}_i)dz_i'}{q_i(z_i, \boldsymbol{\theta}_i)})}{Cov(\frac{\int^{z_i} q_i(z_i', \boldsymbol{\theta}_i)dz_i'}{q_i(z_i, \boldsymbol{\theta}_i)}, \frac{\int^{z_i} q_i(z_i', \boldsymbol{\theta}_i)dz_i'}{q_i(z_i, \boldsymbol{\theta}_i)})})$, *we can obtain a polynomial-based G-TRANS gradient estimator that is better than the reparameterization gradient estimator according to the Proposition.4.2. And we can adjust the value of* $C_i(\boldsymbol{\theta})$ *to obtain better performance.*

According to the approximation theory, we can always find a polynomial $P_k(\mathbf{z}, \boldsymbol{\theta})$ that is close enough to $f(\mathbf{z})$, and in this case, we can dramatically reduce the variance of the resulting gradient estimator. For example, within the convergence radius, we can choose $P_k(\mathbf{z}, \boldsymbol{\theta})$ to be the k-th degree Taylor polynomial of $f(\mathbf{z})$ with the remainder $|f(\mathbf{z}) - P_k(\mathbf{z}, \boldsymbol{\theta})|$ being small. In the practical situation, however, it's often difficult to estimate the coefficients of the polynomial $P_k(\mathbf{z}, \boldsymbol{\theta})$. And when k is large, we need to estimate $\mathcal{O}(N^k)$ coefficients which is almost impossible in real-world applications. Therefore in the following experiments, we only consider $k < 2$.

5 EXPERIMENT

5.1 SYNTHETIC EXPERIMENT

In this section, we use a Dirichlet distribution to approximate the posterior distribution for a probilistic model which has a multinomial likelihood with a Dirichlet prior. We use Gamma distributions to simulate Dirichlet distributions. If $z_i \sim \text{Gamma}(\alpha_i, 1)$, then $\left(\frac{z_1}{\sum_{j=1}^{K} z_j}, \ldots, \frac{z_K}{\sum_{j=1}^{K} z_j}\right) \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_K)$. Then the problem we study here can be written as:

$$\mathbb{E}_{q(\mathbf{z}, \boldsymbol{\alpha})}[f(\mathbf{z})] = \int f(\frac{\mathbf{z}}{\sum_i z_i}) \prod_{k=1}^{K} \text{Gamma}(z_k; \alpha_k, 1) \mathrm{d}\mathbf{z} \tag{13}$$

with $f(\mathbf{z})$ being the multinomial log-likelihood.

We use shape parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ to parameterize the variational Dirichlet distribution. To construct polynomial-based G-TRANS gradient estimator for the factorized distribution $\prod_{k=1}^{K} \text{Gamma}(z_k; \alpha_k, 1)$, we need an accurate and fast way to approximate the derivative of the lower incomplete gamma function (part of the gamma CDF) w.r.t the shape parameter.

The lower incomplete gamma function $\gamma(\alpha, z)$ is a special function and does not admit analytical expression for derivative w.r.t. the shape parameter. However, for small α and z, we have

$$\frac{\partial \gamma(\alpha, z)}{\partial \alpha} = \sum_{k=0}^{\infty} \frac{(-1)^k z^{k+\alpha}}{k!} \left(\frac{\ln z}{\alpha + k} + \frac{(-1)}{(\alpha + k)^2}\right) \tag{14}$$

In practice, we take the first 200 terms from this power series. And the approximation error is smaller than 10^{-9} when $\alpha < 5$ and $z < 20$ with double precision floating point number. For large α, we use central finite difference to approximate the derivative. This approximation scheme for lower incomplete gamma function can also be used to construct polynomial-based G-TRANS gradient estimator for distributions that can be simulated by the Gamma distribution such as Beta distribution and Dirichlet distribution.

We follow the experiment setting in Naesseth et al. (2017). Fig.1 shows the resulting variance of the first component of the gradient based on samples simulated from a Dirichlet distribution with $K = 100$ components, and gradients are computed with $N = 100$ trials. We use $P_1(\mathbf{z}) = \mathbf{c} \cdot \mathbf{z}$ to construct the G-TRANS gradient estimator, and we assign **0.2,0** and $-\mathbf{0.1}$ to $\mathbf{c}$ successively as α_1 increases.

Results. From Fig.1 , we can see that the IRG (Figurnov et al., 2018) method and our G-TRANS gradient estimator has obviously lower gradient variance than the RSVI (even with the shape augmentation trick (Naesseth et al., 2017)) or G-REP (Ruiz et al., 2016) method. Further, our G-TRANS gradient estimator outperforms the IRG method when α_1 is large though there is no obvious difference between these two methods when α_1 is small.

5.2 REAL WORLD DATASET

In this section, we study the performance of our G-TRANS gradient estimator on the Sparse Gamma deep exponential family (DEF) model (Ranganath et al., 2015) with the Olivetti faces dataset that consists of 64×64 gray-scale images of human faces in 8 bits. We follow the Sparse Gamma DEF setting in Naesseth et al. (2017) where the DEF model is specified by:

$$
\begin{aligned}
z_{n,k}^{\ell} &\sim \text{Gamma}\left(\alpha_z, \frac{\alpha_z}{\sum_{k'} w_{k,k'}^{\ell} z_{n,k'}^{\ell+1}}\right) \\
x_{n,d} &\sim \text{Poisson}\left(\sum_k w_{k,d}^0 z_{n,k}^1\right)
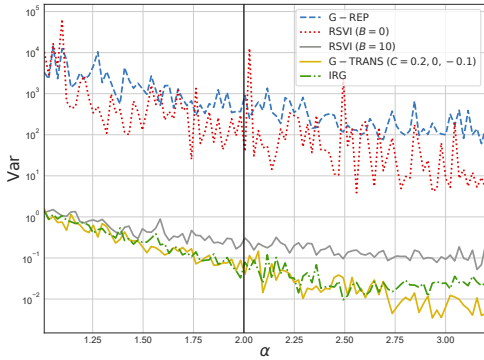\end{aligned}
\tag{15}
$$

Figure 1: Polynomial-bsaed G-TRANS gradient estimator (this paper) obtains lower variance compared to IRG, RSVI and G-REP. The estimated variance is for the first component of Dirichlet approximation to a multinomial likelihood with uniform Dirichlet prior (Naesseth et al., 2017). C is the polynomial coefficient, B denotes shape augmentation (Figurnov et al., 2018) and optimal concentration is $\alpha = 2$.
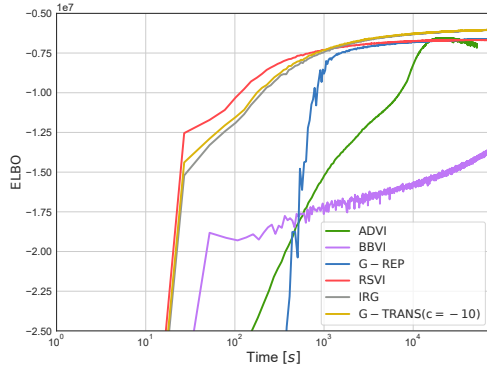
Figure 2: Polynomial-bsaed G-TRANS gradient estimator (this paper) achieves better accuracy than ADVI, BBVI (Ranganath et al., 2014), G-REP, RSVI. G-TRANS is faster than the IRG method with comparable accuracy. This experiment apply the sparse gamma DEF model to the Olivetti faces dataset. dataset

Here n is the number of observations, ℓ is the layer number, k denotes the k-th component in a specific layer and d is the dimension of the output layer (layer 0). $z_{n,k}^{\ell}$ is local random variable, $w_{k,k'}^{\ell}$ is global weight that connects different layers like deep neural networks, and $x_{n,d}$ denotes the set of observations.

We use the experiment setting in Naesseth et al. (2017). α_z is set to 0.1, all priors on the weights are set to Gamma$(0.1, 0.3)$, and the top-layer local variables priors are set to Gamma$(0.1, 0.1)$. The model consists of 3 layers, with 100, 40, and 15 components in each. All variational Gamma distributions are parameterized by the shape and mean. For non-negative variational parameters $\boldsymbol{\theta}$, the transfomration $\boldsymbol{\theta} = \log(1 + \exp(\boldsymbol{\vartheta}))$ is applied to avoid constrained optimization.

In this experiment, we use the step-size sequence ρ^n proposed by Kucukelbir et al. (2017):

$$\rho^n = \eta \cdot n^{-1/2+\delta} \cdot \left(1 + \sqrt{s^n}\right)^{-1}$$
$$s^n = t\left(\hat{g}^n\right)^2 + (1-t)s^{n-1} \tag{16}$$

$\delta = 10^{-16}, t = 0.1, \eta = 0.75$ is used in this experiment. The best result of RSVI is reproduced with $B = 4$ (Naesseth et al., 2017). We still use $P_1(\mathbf{z}) = \mathbf{c} \cdot \mathbf{z}$ to construct the G-TRANS gradient estimator and we use $\mathbf{c} = -\mathbf{10.0}$ for all time.

Results. From Fig.2, We can see that G-TRANS achieves significant improvements in the first 1000 runs and exceeds RSVI though with a slower initial improvement. G-TRANS achieves obviously better accuracy than ADVI, BBVI, G-REP and RSVI, and keeps improving the ELBO even after 75000 runs. G-TRANS is faster than the IRG in early training stage which means G-TRANS has a lower gradient variance. However, this speed advantage of G-TRANS gradually decreases as the step size goes down in the later training stage.

6 RELATED WORK

There are already some lines of research focusing on extending the reparameterization trick to a larger class of distributions. The G-REP (Ruiz et al., 2016) generalizes the reparameterization gradient by using a standardization transformation that allows the standardization distribution to depend

weakly on variational parameters. Our gradient model gives a more elegant expression of the generalized reparameterized gradient than that of G-REP which decomposes the gradient as $g_{\text{rep}} + g_{\text{cor}}$. Different from G-REP, our model hides the transformation behind the velocity field thus the expensive computation of the Jacobian matrix of the transformation is evaded. And it's more flexible to construct gradient estimator with the velocity field than the very detailed transformation. The RSVI (Naesseth et al., 2017) develops a similar generalized reparameterized gradient model with the tools from rejection sampling literatures. RSVI introduces a score function gradient term to compensate the gap that is caused by employing the proposal distribution of a rejection sampler as a surrogate distribution for reparameterization gradient, although the score function gradient term can often be ignored in practice to reduce the gradient variance at the cost of small bias. Unlike RSVI, our gradient estimator can be constructed with deterministic procedure which avoids the additional stochasticity introduced by the accept-reject steps thus lower gradient variance. The path-wise derivative (Jankowiak & Obermeyer, 2018) is closely related to our model. They obtain the transport equation by an analog to the displacement of particles, while we derive the transport euqation for reparameterization gradient by rigorous mathematical deduction.The path-wise gradient model can be seen as a special case of our G-TRANS gradient model. Their work only focus on standard reparameterization gradient while our model can admit generalized transformation-based gradient. The velocity field used in their work must conform to the transport equation while we only require the velocity field to satisfy the unbiasedness constraint. The implicit reparameterization gradient (IRG) (Figurnov et al., 2018) differentiates from the path-wise derivative only by adopting a different method for multivariate distributions.

There are also some other works trying to address the limitations of standard reparameterization. Graves (2016) applies implicit reparameterization for mixture distributions and Knowles (2015) uses approximations to the inverse CDF to derive gradient estimators. Both work involve expensive computation that cannot be extended to large-scale variational inference. Schulman et al. (2015) expressed the gradient in a similar way to G-REP and automatically estimate the gradient in the context of stochastic computation graphs, but their work is short of necessary details therefore cannot be applied to general variational inference task directly. ADVI (Kucukelbir et al., 2017) transforms the random variables such that their support are on the reals and then approximates transformed random variables with Gaussian variational posteriors. However, ADVI struggles to approximate probability densities with singularities as noted by Ruiz et al. (2016).

7 CONCLUSION

We proposed a generalized transformation-based (G-TRANS) gradient model which extends the reparameterization trick to a larger class of variational distributions. Our gradient model hides the details of transformation by introducing the velocity field and provides a flexible way to construct gradient estimators. Based on the proposed gradient model, we introduced a polynomial-based G-TRANS gradient estimator that cannot be induced by any other existing generalized reparameterization gradient framework. In practice, our gradient estimator provides a lower gradient variance than other state-of-the-art methods, leading to a fast converging process. For future work, We can consider how to construct G-TRANS gradient estimators for distributions that don't have analytical high-order moments. We can also utilize the results from the approximation theory to find certain kinds of high-order polynomial functions that can approximate the test function effectively with cheap computations for the coefficients. Constructing velocity fields with the optimal transport theory is also a promising direction.

REFERENCES

George Casella and Christian P Robert. Rao-blackwellisation of sampling schemes. *Biometrika*, 83 (1):81–94, 1996.

Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society, 2 edition, 2010. ISBN 0821849743.

Mikhail Figurnov, Shakir Mohamed, and Andriy Mnih. Implicit reparameterization gradients. In *NeurIPS*, pp. 439–450, 2018.

Peter W. Glynn. Likelihood ratio gradient estimation for stochastic systems. *Commun. ACM*, 33 (10):75–84, 1990.

Alex Graves. Stochastic backpropagation through mixture density distributions. *CoRR*, abs/1607.05690, 2016.

Martin Jankowiak and Fritz Obermeyer. Pathwise derivatives beyond the reparameterization trick. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2240–2249. PMLR, 2018.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

David A Knowles. Stochastic gradient variational bayes for gamma approximating distributions. *arXiv preprint arXiv:1509.01631*, 2015.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. *J. Mach. Learn. Res.*, 18:14:1–14:45, 2017.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 1791–1799. JMLR.org, 2014.

Christian A. Naesseth, Francisco J. R. Ruiz, Scott W. Linderman, and David M. Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pp. 489–498. PMLR, 2017.

John W. Paisley, David M. Blei, and Michael I. Jordan. Variational bayesian inference with stochastic search. In *ICML*. icml.cc / Omnipress, 2012.

Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black box variational inference. In *AISTATS*, volume 33 of *JMLR Workshop and Conference Proceedings*, pp. 814–822. JMLR.org, 2014.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David M. Blei. Deep exponential families. In *AISTATS*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

Francisco J. R. Ruiz, Michalis K. Titsias, and David M. Blei. The generalized reparameterization gradient. In *NIPS*, pp. 460–468, 2016.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *NIPS*, pp. 3528–3536, 2015.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

A APPENDIX

A.1 PROOF OF THEOREM.3.1

We assume that transformed random variable $\boldsymbol{\rho} = \boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})$ is of the same dimension as $\mathbf{z}$. And we assume that there exists $\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})$ that satisfy the constraint $\mathbf{z} = \boldsymbol{\psi}(\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}), \boldsymbol{\theta})$.

Firstly, by the change-of-variable technique, we have

$$w(\boldsymbol{\rho}, \boldsymbol{\theta}) = q(\mathbf{z}, \boldsymbol{\theta})|\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))| \tag{17}$$

Take derivative w.r.t θ (any component of $\boldsymbol{\theta}$) at both sizes, we have

$$\frac{\partial}{\partial \theta}w(\boldsymbol{\rho}, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta}|\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))| * q(\mathbf{z}, \boldsymbol{\theta}) + \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) * |\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))|$$
$$+(\nabla_{\mathbf{z}}q(\mathbf{z}, \boldsymbol{\theta}) \cdot \frac{\partial \mathbf{z}}{\partial \theta}) * |\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))| \tag{18}$$

With the rule of determinant derivation, we have

$$\frac{\partial}{\partial \theta}|\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))| = |\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))| * \mathrm{tr}((\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))^{-1}\frac{\partial}{\partial \theta}\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})) \tag{19}$$

Substitute the Equ.19 into Equ.18, we have

$\frac{\partial}{\partial \theta}w(\boldsymbol{\rho}, \boldsymbol{\theta}) = |\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))|(\mathrm{tr}((\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))^{-1}\frac{\partial}{\partial \theta}\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})) * q(\mathbf{z}, \boldsymbol{\theta})$
$+ \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + (\nabla_{\mathbf{z}}q(\mathbf{z}, \boldsymbol{\theta}) \cdot \frac{\partial \mathbf{z}}{\partial \theta}))$
$= |\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))|(\mathrm{tr}((\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))^{-1}\frac{\partial}{\partial \theta}\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})) * q(\mathbf{z}, \boldsymbol{\theta})$
$+ \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + (\nabla_{\mathbf{z}}q(\mathbf{z}, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))$
$= |\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))|(\mathrm{tr}(\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta}\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})) * q(\mathbf{z}, \boldsymbol{\theta})$
$+ \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + (\nabla_{\mathbf{z}}q(\mathbf{z}, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))$

Since we have $\nabla_{\mathbf{z}}\boldsymbol{\psi}(\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}), \boldsymbol{\theta}) = \nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})}\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})$, then

$\frac{\partial}{\partial \theta}w(\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}), \boldsymbol{\theta}) = \frac{1}{|\det(\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}))|}(\mathrm{tr}(\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta}\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})}) * q(\mathbf{z}, \boldsymbol{\theta})$
$+ \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + (\nabla_{\mathbf{z}}q(\mathbf{z}, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})})$
$= \frac{1}{|\det(\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}))|}(\mathrm{tr}(\frac{\partial}{\partial \theta}\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})}\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})) * q(\mathbf{z}, \boldsymbol{\theta})$
$+ \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + (\nabla_{\mathbf{z}}q(\mathbf{z}, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})})$
$= \frac{1}{|\det(\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}))|}(\mathrm{tr}(\frac{\partial}{\partial \theta}(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})}\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}))) * q(\mathbf{z}, \boldsymbol{\theta})$
$+ \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + (\nabla_{\mathbf{z}}q(\mathbf{z}, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})})$
$= \frac{1}{|\det(\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}))|}(\mathrm{tr}(\frac{\partial}{\partial \theta}\nabla_{\mathbf{z}}\boldsymbol{\psi}(\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}), \boldsymbol{\theta})) * q(\mathbf{z}, \boldsymbol{\theta})$
$+ \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + (\nabla_{\mathbf{z}}q(\mathbf{z}, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})})$
$= \frac{1}{|\det(\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}))|}(\mathrm{tr}(\nabla_{\mathbf{z}}\frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})}) * q(\mathbf{z}, \boldsymbol{\theta})$
$+ \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + (\nabla_{\mathbf{z}}q(\mathbf{z}, \boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})})$
$= \frac{1}{|\det(\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}))|}(\mathrm{tr}(\nabla_{\mathbf{z}}(q(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})})) + \frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}))$
$= \frac{1}{|\det(\nabla_{\mathbf{z}}\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta}))|}(\frac{\partial}{\partial \theta}q(\mathbf{z}, \boldsymbol{\theta}) + \nabla_{\mathbf{z}} \cdot (q(\mathbf{z}, \boldsymbol{\theta})\frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})}))$

Then let $\mathbf{v}^{\theta}(\mathbf{z}, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta}\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z}, \boldsymbol{\theta})}$, we obtain the first conclusion of the Theorem.3.1.

As for the second part, we have

$$\mathbb{E}_{w(\boldsymbol{\rho}, \boldsymbol{\theta})}[f(\boldsymbol{\psi}(\boldsymbol{\rho}, \boldsymbol{\theta}))] = \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\theta})}[f(\mathbf{z})]$$

Then

$$\frac{\partial}{\partial\theta}\mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}[f(\mathbf{z})] = \frac{\partial}{\partial\theta}\mathbb{E}_{w(\boldsymbol{\rho},\boldsymbol{\theta})}[f(\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta}))]$$

$$= \frac{\partial}{\partial\theta}\mathbb{E}_{w(\boldsymbol{\rho},\boldsymbol{\theta})}[\nabla_{\mathbf{z}}f(\mathbf{z})|_{\mathbf{z}=\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta})}\frac{\partial}{\partial\theta}\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta}) + f(\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta}))\frac{\partial}{\partial\theta}\log w(\boldsymbol{\rho},\boldsymbol{\theta})] \tag{20}$$

Since $\frac{\partial}{\partial\theta}\log w(\boldsymbol{\rho},\boldsymbol{\theta}) = \frac{\frac{\partial}{\partial\theta}w(\boldsymbol{\rho},\boldsymbol{\theta})}{w(\boldsymbol{\rho},\boldsymbol{\theta})} = \frac{\frac{\partial}{\partial\theta}w(\boldsymbol{\rho},\boldsymbol{\theta})}{q(\mathbf{z},\boldsymbol{\theta})|\det(\nabla_{\boldsymbol{\rho}}\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta}))|}$

Then we have

$\frac{\partial}{\partial\theta}\mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}[f(\mathbf{z})] = \mathbb{E}_{w(\boldsymbol{\rho},\boldsymbol{\theta})}[\nabla_{\mathbf{z}}f(\mathbf{z})|_{\mathbf{z}=\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta})}\frac{\partial}{\partial\theta}\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta}) + f(\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta}))\frac{\partial}{\partial\theta}\log w(\boldsymbol{\rho},\boldsymbol{\theta})]$

$= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}[\nabla_{\mathbf{z}}f(\mathbf{z})\frac{\partial}{\partial\theta}\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z},\boldsymbol{\theta})} + f(\mathbf{z})\frac{\partial}{\partial\theta}\log w(\boldsymbol{\rho},\boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z},\boldsymbol{\theta})}]$

$= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}[\nabla_{\mathbf{z}}f(\mathbf{z})\frac{\partial}{\partial\theta}\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z},\boldsymbol{\theta})} + f(\mathbf{z})\frac{1}{q(\mathbf{z},\boldsymbol{\theta})}\nabla_{\mathbf{z}}\cdot(q(\mathbf{z},\boldsymbol{\theta})\frac{\partial}{\partial\theta}\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z},\boldsymbol{\theta})}) + f(\mathbf{z})\frac{\frac{\partial q(\mathbf{z},\boldsymbol{\theta})}{\partial\theta}}{q(\mathbf{z},\boldsymbol{\theta})}]$

$= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}[f(\mathbf{z})\frac{\frac{\partial q(\mathbf{z},\boldsymbol{\theta})}{\partial\theta}}{q(\mathbf{z},\boldsymbol{\theta})} + \frac{1}{q(\mathbf{z},\boldsymbol{\theta})}\nabla_{\mathbf{z}}\cdot(f(\mathbf{z})q(\mathbf{z},\boldsymbol{\theta})\frac{\partial}{\partial\theta}\boldsymbol{\psi}(\boldsymbol{\rho},\boldsymbol{\theta})|_{\boldsymbol{\rho}=\boldsymbol{\phi}(\mathbf{z},\boldsymbol{\theta})})]$

$= \mathbb{E}_{q(\mathbf{z},\boldsymbol{\theta})}[f(\mathbf{z})\frac{\frac{\partial q(\mathbf{z},\boldsymbol{\theta})}{\partial\theta}}{q(\mathbf{z},\boldsymbol{\theta})} + \frac{1}{q(\mathbf{z},\boldsymbol{\theta})}\nabla_{\mathbf{z}}\cdot(f(\mathbf{z})q(\mathbf{z},\boldsymbol{\theta})\mathbf{v}^{\theta})]$

Thus we obtain the second part of the Theorem.3.1. Proof ends.

As a by-product, if we make $\frac{\partial}{\partial\theta}w(\boldsymbol{\phi}(\mathbf{z},\boldsymbol{\theta}),\boldsymbol{\theta}) = 0$, we can obtain the transport equation for the reparameterization trick:

$$\frac{\partial}{\partial\theta}q(\mathbf{z},\boldsymbol{\theta}) + \nabla_{\mathbf{z}}\cdot(q(\mathbf{z},\boldsymbol{\theta})\frac{\partial}{\partial\theta}\mathbf{v}^{\theta}) = 0 \tag{21}$$

And $\frac{\partial}{\partial\theta}w(\boldsymbol{\phi}(\mathbf{z},\boldsymbol{\theta}),\boldsymbol{\theta}) = 0$ also means that the standardization distribution is independent with $\boldsymbol{\theta}$ which is the core of the reparameterization trick.

A.2   VARIANCE VARIATIONAL OPTIMIZATION

For the simplicity of the proof, we only consider the 1-dimensional here.   And denote $\frac{\partial\log q(z,\theta)}{\partial\theta}f(z) + \frac{1}{q(z,\theta)}\frac{\mathrm{d}}{\mathrm{d}z}(f(z)q(z,\theta)v^{\theta}(z,\theta))$ as $r(z,\theta)$.

The variance of the G-TRANS gradient with N independent samples is defined as:

$\mathrm{Var}(\widehat{\frac{\partial\mathcal{L}}{\partial\theta}}) = \frac{1}{N}\mathrm{Var}(r(z,\theta)) = \frac{1}{N}(\mathbb{E}_{q(z,\theta)}[(r(z,\theta))^2] - (\mathbb{E}_{q(z,\theta)}[r(z,\theta)])^2)$

where with the unbiased constraint, we have $\mathbb{E}_{q(z,\theta)}[r(z,\theta)] = \mathbb{E}_{q(z,\theta)}[f(z)\frac{\frac{\partial q(z,\theta)}{\partial\theta}}{q(z,\theta)}] = \text{const}$, so we need to consider the term $\mathbb{E}_{q(z,\theta)}[(r^{\theta}(z,\theta))^2]$ only.

$$\mathbb{E}_{q(z,\theta)}[(r^{\theta}(z,\theta))^2] = \int_{-\infty}^{+\infty}(f\frac{\frac{\partial q}{\partial\theta}}{q} + \frac{1}{q}\frac{\mathrm{d}}{\mathrm{d}z}(fqv^{\theta}))^2qdz \tag{22}$$

According to the Euler–Lagrange equation, we have

$$\frac{\mathrm{d}}{\mathrm{d}v^{\theta}}[(f\frac{\frac{\partial q}{\partial\theta}}{q} + \frac{1}{q}\frac{\mathrm{d}}{\mathrm{d}z}(fqv^{\theta}))^2q] = \frac{\mathrm{d}}{\mathrm{d}z}\frac{\mathrm{d}}{\mathrm{d}v'^{\theta}}[(f\frac{\frac{\partial q}{\partial\theta}}{q} + \frac{1}{q}\frac{\mathrm{d}}{\mathrm{d}z}(fqv^{\theta}))^2q] \tag{23}$$

Simplify it, we have

$$(f\frac{\frac{\partial q}{\partial\theta}}{q} + \frac{1}{q}\frac{\mathrm{d}}{\mathrm{d}z}(fqv^{\theta}))\frac{\mathrm{d}}{\mathrm{d}z}(fq) = \frac{\mathrm{d}}{\mathrm{d}z}(fq(f\frac{\frac{\partial q}{\partial\theta}}{q} + \frac{1}{q}\frac{\mathrm{d}}{\mathrm{d}z}(fqv^{\theta}))) \tag{24}$$

Then we have

$$fq\frac{\mathrm{d}}{\mathrm{dz}}((f\frac{\frac{\partial q}{\partial\theta}}{q} + \frac{1}{q}\frac{\mathrm{d}}{\mathrm{dz}}(fqv^\theta))) = 0 \tag{25}$$

which means

$$(f\frac{\frac{\partial q}{\partial\theta}}{q} + \frac{1}{q}\frac{\mathrm{d}}{\mathrm{dz}}(fqv^\theta)) = C(\theta) \tag{26}$$

Thus we have

$$v^\theta = \frac{1}{fq}\int^z qC(\theta) - f\frac{\partial q}{\partial\theta}dz \tag{27}$$

which is usually intractable in real world practice.

A.3 UNBIASEDNESS CONSTRAINT VERIFICATION

Here we verify that $\mathbf{v}_{ah}^\theta(\mathbf{z},\boldsymbol{\theta})$ satisfy the unbiasedness constraint if $h(\mathbf{z},\boldsymbol{\theta})$ is bounded. Let θ_i be any component of $\boldsymbol{\theta_i}$, then

$$\begin{aligned}
\int\frac{1}{q}\nabla_{\mathbf{z}}\cdot(qf\mathbf{v}_{ah}^{\theta_i})qd\mathbf{z} &= \int\frac{1}{q}\frac{\partial}{\partial z_i}(-h(\mathbf{z},\boldsymbol{\theta})\int^{z_i}\frac{\partial q(\mathbf{z}',\boldsymbol{\theta})}{\partial\theta_i}dz_i')qd\mathbf{z} \\
&= \int\frac{\partial}{\partial z_i}(-h(\mathbf{z},\boldsymbol{\theta})\int^{z_i}\frac{\partial q_i(z_i',\boldsymbol{\theta}_i)}{\partial\theta_i}dz_i')dz_i\prod_{j\neq i}q_j(z_j,\boldsymbol{\theta}_j)dz_j \\
&= \int\prod_{j\neq i}q_j(z_j,\boldsymbol{\theta}_j)dz_j\int\frac{\partial}{\partial z_i}(-h(\mathbf{z},\boldsymbol{\theta})\int^{z_i}\frac{\partial q_i(z_i',\boldsymbol{\theta}_i)}{\partial\theta_i}dz_i')dz_i
\end{aligned} \tag{28}$$

If $h(\mathbf{z},\boldsymbol{\theta})$ is bounded, we have

$$\int\frac{\partial}{\partial z_i}(-h(\mathbf{z},\boldsymbol{\theta})\int^{z_i}\frac{\partial q_i(z_i',\boldsymbol{\theta}_i)}{\partial\theta_i}dz_i')dz_i = -(h(\mathbf{z},\boldsymbol{\theta})\int^{z_i}\frac{\partial q_i(z_i',\boldsymbol{\theta}_i)}{\partial\theta_i}dz_i')\Big|_{-\infty}^{+\infty} = 0 \tag{29}$$

Therefore, $\mathbb{E}_{q_\theta}[\frac{1}{q_\theta}\nabla_{\mathbf{z}}\cdot(fq_\theta\mathbf{v}_{ah}^{\theta_i})] = \int\frac{1}{q}\nabla_{\mathbf{z}}\cdot(qf\mathbf{v}_{ah}^{\theta_i})qd\mathbf{z} = 0$.

A.4 DUAL POLYNOMIAL VELOCITY FIELD

If we take the dual polynomial velocity field $\mathbf{v}_{dp}^\theta$ in the G-TRANS framework, we can reach a dual result to the Proposition.4.2:

Proposition A.1. *If* $\mathrm{Cov}(P_k\frac{\partial\log q(\mathbf{z},\boldsymbol{\theta})}{\partial\theta}, (2f - P_k)\frac{\partial\log q(\mathbf{z},\boldsymbol{\theta})}{\partial\theta}) > 0$, *then the gradient estimator resulted from dual polynomial velocity field has a smaller gradient variance than the score function gradient estimator.*

The proof is similar to that of Proposition.4.2.