Generalized Variational Inference via Optimal Transport

Jinjin Chi[1,2], **Zhichao Zhang**[1,2], **Zhiyao Yang**[1,2], **Jihong Ouyang**[1,2], **Hongbin Pei**[3*]

[1] College of Computer Science and Technology, Jilin University, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[3] MOE KLINNS Lab, School of Cyber Science and Engineering, Xi'an Jiaotong University, China
chijinjin616@gmail.com, zhangzc0929@163.com, yangzy9529@gmail.com, ouyj@jlu.edu.cn, peihongbin@xjtu.edu.cn

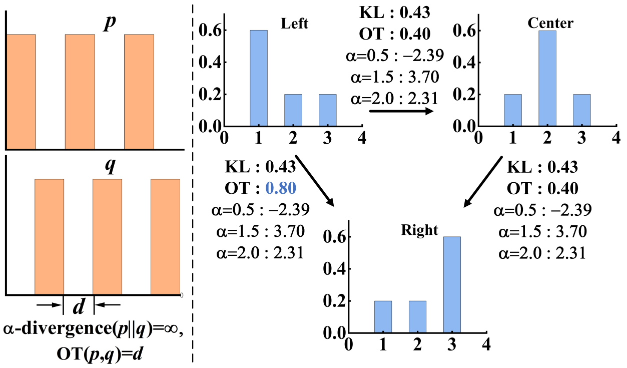
Abstract

Variational Inference (VI) has gained popularity as a flexible approximate inference scheme for computing posterior distributions in Bayesian models. Original VI methods use Kullback-Leibler (KL) divergence to construct variational objectives. However, KL divergence has zero-forcing behavior and is completely agnostic to the metric of the underlying data distribution, resulting in bad approximations. To alleviate this issue, we propose a new variational objective by using Optimal Transport (OT) distance, which is a metric-aware divergence, to measure the difference between approximate posteriors and priors. The superior performance of OT distance enables us to learn more accurate approximations. We further enhance the objective by gradually including the OT term using a hyperparameter λ for over-parameterized models. We develop a Variational inference method with OT (VOT) which presents a gradient-based black-box framework for solving Bayesian models, even when the density function of approximate distribution is not available. We provide the consistency analysis of approximate posteriors and demonstrate the practical effectiveness on Bayesian neural networks and variational autoencoders.

Introduction

Variational **I**nference (VI) (Jordan et al. 1999; Blei, Kucukelbir, and McAuliffe 2017; Nazaret and Blei 2022) is a powerful tool in modern probabilistic machine learning for approximating intractable posterior distributions. The idea behind VI is to posit a family of distributions over the latent variables and then find the closest member as the approximation of the true posterior by minimizing a divergence objective function. VI has many elegant and favorable properties such as the fact that it tends to be fast and easy to scale to large data (Blei, Kucukelbir, and McAuliffe 2017). Therefore, VI is widely used to approximate posterior distributions for Bayesian deep learning (Shi, Titsias, and Mnih 2020; Rudner et al. 2021; Pei et al. 2022, 2020), deep generative models (Okada and Taniguchi 2019), among many others.

A dominating factor for successful VI relies on the choice of a proper divergence metric (Wang, Liu, and Liu 2018). A different divergence results in the approximation having different properties. The most commonly used divergence is the Kullback-Leibler (KL) divergence in standard VI (Ranganath, Gerrish, and Blei 2014; Blei, Kucukelbir, and McAuliffe 2017), which measures the dissimilarity between the approximate distribution q and the true posterior p, i.e., $\mathrm{KL}(q \parallel p)$. *However, KL divergence has some limitations*. For example, it is not a proper metric, due to its non-symmetry and violation of the triangle inequality. Notably, it exhibits zero-forcing behavior, that is, $p(x) = 0$ must imply $q(x) = 0$, leading to a severe underestimation of the posterior variance and an inability to capture the multimodality within the posterior distribution (Blei, Kucukelbir, and McAuliffe 2017). This becomes particularly problematic if $p(x) = 0$ and $q(x) > 0$, since $\mathrm{KL}(q \parallel p)$ is infinite.



(a) (b)

Figure 1: (a) An example of distribution q not exactly match distribution p. (b) Three ordinal discrete distributions: Left: (0.6,0.2,0.2). Center: (0.2,0.6,0.2). Right: (0.2,0.2,0.6). "→" denotes a transfer with distances computed by the standard KL divergence, OT distance and α-divergences ($\alpha = 0.5,\ 1.5,\ 2.0$).

To overcome the zero-forcing behavior, some efforts turn to more general families of divergences such as the α-divergence due to its zero-avoiding behavior when $\alpha > 0$ (Hernandez-Lobato et al. 2016; Dieng et al. 2017; Li and Turner 2016). By choosing different α, one can get some divergences as special cases, including reverse KL divergence,

[*] Corresponding author

χ^2-divergence, and others. However, α-divergences suffer from the issue of being infinite when the approximation q does not exactly match the true posterior p as depicted in Figure 1(a). This makes it problematic to use α-divergences, despite the zero-avoiding behavior they promise. Unfortunately, many practical problems of interest involve scenarios where q is very different from p, see Example 3.1 in (Wang, Liu, and Liu 2018). Furthermore, α-divergences, including the standard KL divergence, do not accurately measure the difference between two distributions in terms of the actual distance between the data samples. Figure 1(b) shows an example with three simple distributions: Left: (0.6,0.2,0.2), Center: (0.2,0.6,0.2) and Right: (0.2,0.2,0.6). The distances between Left to Center and Center to Right should be the same, and the distance between Left to Right should be about twice as much. But with α-divergences, including the standard KL divergence, all three distances are the same. Therefore, it is desirable to find a powerful tool to measure the dissimilarity between distributions in VI.

Optimal **T**ransport (OT) distance (Villani 2008) has recently displayed promising performance in the comparison of probability distributions (Genevay et al. 2016; Seguy et al. 2018; Chi et al. 2023), such as in Wasserstein GAN (WGAN) (Arjovsky, Chintala, and Bottou 2017). Compared to existing divergences, OT distance has several advantageous properties (Peyré, Cuturi et al. 2019). It serves as a valid metric that is symmetric and satisfies triangular inequality. Notably, it exhibits the unique ability to capture the geometric structure of distributions, making it applicable even to distributions with non-overlapping supports (as shown in Figure 1(a)). Furthermore, it provides an actual distance measurement, as demonstrated in Figure 1(b). These favorable properties motivate us to employ OT distance in VI to achieve more accurate approximations.

Our contributions. In this work, we try to extend the generalized VI framework to OT distances. To achieve this, we design a novel variational objective by using OT distances to measure the distance between approximate posterior and priors. The superior performance of OT distances enables us to learn more accurate approximations. For over-parameterized models, we further enhance the variational objective by using a single hyperparameter λ to balance the contributions of model fitting and OT terms. We develop a **V**ariational inference method with **O**ptimal **T**ransport (**VOT**), which provides a stable gradient-based black-box algorithm for solving Bayesian inference problems, even when the density function of approximate distribution is not available. On the theoretical side, we prove the consistency of the approximate posterior. Finally, we demonstrate both qualitatively and quantitatively that our method can achieve state-of-the-art inference performance compared to various baselines on Bayesian neural network and variational autoencoders.

Related Work

Some recent works try to extend the standard VI framework to other statistical divergences to mitigate limitations of standard KL divergence. Many of these divergences are special cases of f-divergence (Csiszár, Shields et al. 2004):

$$D_f(p||q) \triangleq \int f\left(\frac{p(x)}{q(x)}\right) q(x)dx,$$

where $f(\cdot)$ is a convex function. The most commonly used class of f-divergence is α-divergence, i.e., $f(t) = t^\alpha/\alpha(\alpha-1)$ for $\alpha \in \mathbb{R}\backslash\{0,1\}$, due to its zero-avoiding behavior. By choosing different α, one can get some well-known divergences as special cases, including the standard KL divergence ($\alpha \to 0$), the reverse KL divergence $\mathrm{KL}(p \parallel q)$ ($\alpha \to 1$) and the χ^2-divergence ($\alpha = 2$). The reverse KL divergence is used in expectation propagation (Minka 2001a; Li, Hernández-Lobato, and Turner 2015; Minka 2001b), and χ^2-divergence is mainly studied in importance sampling (Dieng et al. 2017; Kuleshov and Ermon 2017). Additionally, the general α-divergence is discussed in VI (Li and Turner 2016). When $\alpha > 0$, the approximation q tends to cover more modes of p. In general, larger α enforces stronger zero-avoiding behavior. However, it is important to note that using large α values may lead to high or infinite variance (Wang and van Hoof 2020). This is due to the involvement of the α-th power of the ratio $p(x)/q(x)$, which is likely to have a fat-tailed distribution. In fact, when q is very different from p, the α-divergence becomes infinite, rendering it ineffective as a divergence metric. To address this issue, f-divergence as a more inclusive statistical divergence is employed for VI. To achieve tail-adaptiveness, (Wang, Liu, and Liu 2018) proposes an adaptive f-divergence variational inference where different f functions are adjusted based on the tail distribution of the density ratio $p(x)/q(x)$. However, obtaining such adaptive function f that satisfies the requirement of convexity is a challenging task. Besides, (Wan, Li, and Hovakimyan 2020) proposes the f-VI framework which generalizes variational inference to all f-divergences unifying a number of existing VI methods, including KL-VI, χ-VI, α-VI.

Recently, OT distances have shown several advantages over the traditional KL divergence in various applications. Inspired by this, there is a growing interest in exploring the use of OT distances in VI. To the best of our knowledge, rarely few works focus on this challenging topic. For example, (Ambrogioni et al. 2018) proposes a pseudo-OT distance which includes f-divergence as a special case, and uses it in a special variational inference, i.e., *joint-contrastive* VI, where the variational distribution $q(x,z)$ is defined as a joint form about latent variables z and observed variables x. This method constructs the objective by using the pseudo OT distance between two joint distributions and solves it through Sinkhorn algorithm. However, the Sinkhorn algorithm becomes unstable when the regularization parameter approaches 0 (Cuturi 2013). In contrast, our method focuses on the generalized VI and uses a general OT distance to measure the distance between variational distribution and priors.

Background

In this section, we briefly review the essential background of variational inference and optimal transport.

Variational Inference

In a Bayesian model, we consider a set of n i.i.d samples $X = \{x_i\}_{i=1}^n$ observed from a conditional distribution $p(X|z)$ parameterized with a latent variable z which is drawn from a prior $p_0(z)$. The goal of VI is to obtain the posterior distribution $p(z|X)$ given observed data,

$$p(z|X) = \frac{p(X|z)p_0(z)}{p(X)} \tag{1}$$

In general, this posterior is intractable since the evidence $p(X) = \int p(z, X)dz$ is difficult to compute. Therefore, VI introduces a family of tractable distributions $\mathscr{Q}$ and finds the member $q_\theta(z) \in \mathscr{Q}$ that is closest to the true posterior $p(z|X)$, where θ is the variational parameter. The variational distribution $q_\theta(z)$ is constructed by solving an optimization problem that minimizes two terms that do not interact (Knoblauch, Jewson, and Damoulas 2022):

1) *The loss* $\sum_{i=1}^{n} - \log p(x_i|z)$ *to be expected under* $q_\theta(z)$.

2) *The deviation of the posterior from the prior* $p_0(z)$ *as measured by a divergence metric* D.

Thus, a versatile and modular representation of the generalized VI objective is:

$$\min_{q_\theta \in \mathscr{Q}} \left\{ \mathbb{E}_{q_\theta(z)}[\sum_{i=1}^{n} - \log p(x_i|z)] + D(q_\theta, p_0) \right\} \tag{2}$$

The first term acts as a model fitting term and the second one is a regularizer penalizing the solution where $q_\theta(z)$ is far away from $p_0(z)$. The divergence D is also known as the uncertainty quantifier. When $D =$ KL divergence, Eq.2 is the objective of standard VI.

Optimal Transport

Let $\mathcal{Z}$ be an arbitrary space equipped with a ground cost $c(z_1, z_2) : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, for two probability distributions $q(z_1)$ and $p(z_2)$, the OT distance seeks an optimal joint distribution of $q(z_1)$ and $p(z_2)$ that minimizes the total transport cost (Villani 2008). It is formulated as

$$D_{ot}(q, p) \triangleq \min\nolimits_{\pi \in \Pi(q,p)} \ \mathbb{E}_{\pi(z_1,z_2)}[c(z_1, z_2)],$$

where $\Pi(q, p)$ is the set of all joint distributions with prescribed marginals $q(z_1)$ and $p(z_2)$. OT distances capture the geometry of distributions via $c(z_1, z_2)$, which is the cost to move a unit of mass from z_1 to z_2.

The OT problem admits an equivalent dual form (Villani 2008),

$$D_{ot}(q, p) = \max_{\psi \oplus \phi \leq c} \ \mathbb{E}_{q(z_1)}[\psi(z_1)] + \mathbb{E}_{p(z_2)}[\phi(z_2)] \tag{3}$$

where ψ and ϕ are dual variables, which can be represented by vectors or neural networks. The constraint $\psi \oplus \phi \leq c$ means $\psi(z_1) + \phi(z_2) \leq c(z_1, z_2)$ for all (z_1, z_2).

Our Method

The goal of this paper is to introduce the use of OT distances in generalized VI to obtain more accurate and robust approximations. We begin by presenting a new generalized variational objective and then develop a black-box method VOT based on automatic differentiation through stochastic optimization. Our method does not require tractable density functions of variational distributions, thereby enabling the use of neural networks as variational distributions, since we only have access to independent samples.

Variational Objective

Inspired by OT distance, denoted by D_{ot}, we use it in the generalized variational objective Eq.2 to measure the distance between the prior $p_0(z)$ and the variational distribution $q_\theta(z)$. As a result, we obtain the following objective [1]:

$$\mathcal{L} \triangleq \min_{q_\theta \in \mathscr{Q}} \left\{ \underbrace{\mathbb{E}_{q_\theta(z_1)}[\sum\nolimits_{i=1}^{n} - \log p(x_i|z_1)]}_{\text{model fitting}} + \lambda \underbrace{D_{ot}(q_\theta, p_0)}_{\text{regularizer}} \right\} \tag{4}$$

where λ is a regularization parameter to effectively balance the contributions of the model fitting and regularizer terms. When $\lambda = 1$, it corresponds to the original VI framework. However, in practice, we have observed that using other values of λ may provide better performance, especially for over-parameterized models. This is because, without λ, the regularizer term dominates the overall objective in these models. For instance, a mean-field posterior approximation turns the regularizer term into a sum of as many regularizer terms as the number of model parameters, which can dominate the overall objective when the number of model parameters is large. Consequently, the optimization tends to prioritize keeping the approximation close to the prior, neglecting the important model fitting term. To address this issue, we propose enhancing the original VI framework by gradually including the regularizer term using a hyperparameter $\lambda = 0.1(1 + \exp(-a(iter - b)))^{-1}$, where a and b are constants, and $iter$ denotes the $iter$th iteration.

Having specified priors, even when the constructions of p_0 and q_θ do not satisfy that q_θ is dominated by p_0, the OT distance still satisfies $D_{ot}(q_\theta, p_0) \geq 0$ and $D_{ot}(q_\theta, p_0) = 0 \Rightarrow q_\theta = p_0$ while other divergences are infinite in this situation. Therefore, the proposed objective can provide a desirable approximation. Additionally, the approximate posterior computed by our objective, i.e., $q_\theta^* = \arg \mathcal{L}$ is **consistent**, meaning that it concentrates around the true value. We discuss the theoretical guarantee for consistency in the following theorem (proof in the Appendix).

Theorem 1 (Consistency). *Let* $p_t(x)$ *be the true probability distribution of random variable* x *and suppose that observations* $x_{1:n}$ *are i.i.d samples from* x. *Assume the cost* c *in OT is a lower semi-continuous function on* $\mathcal{Z} \times \mathcal{Z}$. *If the model fitting term is correctly specified and the prior* p_0 *is not infinitely bad for the population of* x, *i.e.,*

[1] To avoid ambiguity, we denote $z_1 \sim q_\theta(z_1), z_2 \sim p_0(z_2)$.

$\mathbb{E}_{p_0}[\mathbb{E}_{p_t}[\sum_{i=1}^{n} - \log p(x_i|z_1)]] < \infty$, *then the approximate posterior* q_θ^* *computed by Eq.4 is consistent and concentrates at the true parameters.*

Optimization Framework

We solve the proposed objective Eq.4 by using an alternating optimization algorithm. For each iteration, we first fix the variational parameter θ to explore the way to compute the OT distance between the variational distribution q_θ and the prior p_0, i.e., $D_{ot}(q_\theta, p_0)$, which is essential to our solution. Then, we learn the variational parameter θ based on updated OT distance through stochastic optimization.

Fixing θ, computing $D_{ot}(q_\theta, p_0)$. When fixing the variational parameter θ, we proceed to compute OT distance $D_{ot}(q_\theta, p_0)$ relying on the dual formulation, as given in Eq.3. However, directly optimizing the dual formulation has cubic time complexity. To alleviate this issue, we relax the dual problem and get a fast computation by adding an entropy regularization with parameter ε (Cuturi 2013),

$$F_\varepsilon(\psi(z_1), \phi(z_2)) = \varepsilon \exp(\frac{1}{\varepsilon}(\psi(z_1) + \phi(z_2) - c(z_1, z_2)))$$

Thus, the regularized dual version of $D_{ot}(q_\theta, p_0)$ is

$$D_{ot}^{\varepsilon}(q_\theta, p_0) \triangleq \max_{\psi,\phi} \mathbb{E}_{q_\theta(z_1)p_0(z_2)}[\psi(z_1) + \phi(z_2) - F(\psi(z_1), \phi(z_2)] \quad (5)$$

There are two dual variables ψ, ϕ that need to be optimized. The optimal ψ^*, ϕ^* satisfy $\phi^* = (\psi^*)^c$, where $(\psi^*)^c$ is the c-transform of ψ^* (Villani 2008). For ψ, we define its entropically c-transform,

$$\psi_\varepsilon^c(z_2) \triangleq -\varepsilon \log\left(\int_{\mathcal{Z}} \exp(\frac{\psi(z_1) - c(z_1, z_2)}{\varepsilon}) q_\theta(z_1) dz_1\right)$$

We then rewrite $D_{ot}^{\varepsilon}(q_\theta, p_0)$ as a *semi-dual* formulation

$$D_{ot}^{\varepsilon}(q_\theta, p_0) = \max_{\psi} \mathbb{E}_{q_\theta(z_1)}[\psi(z_1)] + \mathbb{E}_{p_0(z_2)}[\psi_\varepsilon^c(z_2)] - \varepsilon \quad (6)$$

A key advantage of the semi-dual formulation is that one of the dual variables is eliminated and is computed in close form. When using empirical samples, Eq.6 is a finite dimensional concave maximization problem.

Our aim is now to solve semi-dual OT problem, i.e., Eq.6 by using averaged stochastic gradient methods based on empirical samples. Given an empirical distribution described by $q_\theta^S(z_1) = \sum_{s=1}^{S} q'_s \delta_{z_1^s}$, the variable ψ is a S-dimensional vector $(\psi_s)_{s=1,...S}$. The gradient of $D_{ot}^{\varepsilon}(q_\theta, p_0)$ with respect to ψ reads $\nabla_\psi D_{ot}^{\varepsilon}(q_\theta, p_0) = q' - \mathbb{E}_{p_0(z_2)}[\pi(z_2)]$, where $\pi(z_2)_i = \exp(\frac{\psi_i - c(z_1^i, z_2)}{\varepsilon})\left(\sum_{s=1}^{S} \exp(\frac{\psi_s - c(z_1^s, z_2)}{\varepsilon})\right)^{-1}$. Then we can directly form a noisy gradient using the Monte Carlo samples drawn from $p_0(z_2)$. We summarize the procedure of computing dual variable ψ in *Algorithm 1*.

Algorithm 1: Computation of OT distance

```
Input: $q_\theta(z_1)$, $p_0(z_2)$, $c(z_1, z_2)$, $\varepsilon$ (the regularization parameter), $L$ (the number of iterations), $S$ (batch size)
$\hat{\psi} \leftarrow \mathbf{0}_J, \psi \leftarrow \hat{\psi}$
for $l = 1, 2, ..., L$ do
    sample a batch $(z_2^1, ..., z_2^S)$ from $p_0(z_2)$
    $\hat{\psi} \leftarrow \hat{\psi} + \frac{c}{\sqrt{l}} \nabla_\psi D_{ot}^{\varepsilon}(q_\theta, p_0)$
    $\psi \leftarrow \frac{1}{l}\hat{\psi} + \frac{l-1}{l}\psi$
end for
Output: $\psi$
```

Fixing ψ, computing θ. Assuming that the variable ψ in OT distance is fixed, we now derive a black-box Monte Carlo estimate of the gradient of objective Eq.4 with respect to parameters θ that can be used together with stochastic optimization methods. For simplicity, here we use the regularized dual formulation of OT distance, i.e., Eq.5 which can be recovered from the semi-dual problem as $\phi = \psi_\varepsilon^c$. Thus, the gradient of $\mathcal{L}$ with respect to θ can be represented as an expectation form by using the score function method:

$$\begin{aligned}\nabla_\theta \mathcal{L} = \nabla_\theta \Bigg[& \mathbb{E}_{q_\theta(z_1)}[\sum_{i=1}^{n} - \log p(x_i|z_1)] + \\ & \lambda \mathbb{E}_{q_\theta(z_1)p_0(z_2)}[\psi(z_1) + \phi(z_2) - F(\psi(z_1), \phi(z_2))]\Bigg] \\ = \mathbb{E}_{q_\theta(z_1)} \Bigg[& \nabla_\theta \log q_\theta(z_1) \Bigg(\sum_{i=1}^{n} - \log p(x_i|z_1) + \\ & \lambda \mathbb{E}_{p_0(z_2)}[\psi(z_1) - F(\psi(z_1), \phi(z_2))] \Bigg) \Bigg], \quad (7)\end{aligned}$$

where $\nabla_\theta \log q_\theta(z_1)$ is called the score function (Cox and Hinkley 1979), and $\nabla_\theta[q_\theta(z_1)] = \nabla_\theta[\log q_\theta(z_1)]q_\theta(z_1)$.

However, since the resulting estimator Eq.7 often suffers from high variance resulting in worse performance, the score function gradient is usually employed along with variance reduction methods such as reparameterization trick.

Variance reduction by reparameterization. An alternative to the score function gradient is the reparameterization gradient, which works well in reducing the sampling variance (Kingma and Welling 2013). The reparameterization trick assumes the existence of a noise variable $\epsilon \sim \hat{q}(\epsilon)$ and a mapping function $g_\theta(\cdot)$ such that $z_1 = g_\theta(\epsilon)$. Instead of sampling $\{z_1^s\}_{s=1}^S$ from $q_\theta(z_1)$, the reparameterization estimators rely on the samples $\{\epsilon^s\}_{s=1}^S$ drawn from $\hat{q}(\epsilon)$. One prevalent example is the Gaussian reparameterization: $z_1 \sim q_\theta(z_1) = \mathcal{N}(\mu, \Sigma)$ can be reparameterized with a standard Gaussian variable $\epsilon \sim \mathcal{N}(0, \mathrm{I})$ and a mapping function $z_1 = g_\theta(\epsilon) = \mu + \Sigma^{1/2}\epsilon$.

Following this reparameterization, the gradient with re-

Algorithm 2: Optimization of VOT

```
Input: $X = \{x_i\}_{i=1}^n$ (the real data), $S$ (batch size), $\rho_t$ (learning rate)
Setting: (1) Choose the cost function $c(\cdot)$; (2) set the parameter $\lambda$; (3) construct a distribution $\hat{q}(\epsilon)$ for reparameterization; (4) and initialize $\theta$ randomly
repeat
    // fix $\theta$, and optimize over OT distances.
    compute the dual variable $\psi$ using Algorithm 1.
    // fix $\psi$, and optimize over $\theta$.
    for $s = 1$ to $S$
        draw $\epsilon^{(s)} \sim \hat{q}(\epsilon)$
        draw $z_2^{(s)} \sim p_0(z_2)$
    end for
    form the noisy gradient using Eq.9
    update $\theta$ using the Adam method
until Convergence
Output: The variational distribution $q_\theta$
```

spect to θ, i.e., Eq.7 can be re-written as:

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{\hat{q}(\epsilon)}\Bigg[\nabla_\theta\Bigg(\sum_{i=1}^{n} - \log p(x_i|g_\theta(\epsilon)) + \lambda \mathbb{E}_{p_0(z_2)}[\psi(g_\theta(\epsilon)) \; - F(\psi(g_\theta(\epsilon)), \phi(z_2))]\Bigg)\Bigg] \quad (8)$$

The unbiased Monte Carlo estimator for Eq.8 is:

$$\nabla_\theta \mathcal{L} \approx \frac{1}{S}\sum_{s=1}^{S} \nabla_\theta \Bigg(\sum_{i=1}^{n} - \log p(x_i|g_\theta(\epsilon^s)) + \lambda\psi(g_\theta(\epsilon^s)) - F(\psi(g_\theta(\epsilon^s)), \phi(z_2^s))\Bigg)$$
$$\epsilon^s \sim \hat{q}(\epsilon), z_2^s \sim p_0(z_2) \quad (9)$$

In practice, those gradients in Eq.9 can be calculated via automatic differentiation tools (Team 2015).

Then, at each iteration t, the parameter of interest θ can be updated as follows:

$$\theta_t \leftarrow \theta_{t-1} - \rho_t \nabla_\theta \mathcal{L}, \quad (10)$$

where ρ_t is the learning rate.

Full algorithm. In summary, we outline the full optimization process of VOT in *Algorithm 2*.

Convergence analysis. In the optimization process, we use the entropy-regularized OT in variational objective. Though it is the approximation solution of the primal OT problem, it can converge exponentially fast to a solution of the non-regularized OT problem when $\varepsilon \to 0$ and $S \to \infty$ (Cominetti and San Martín 1994). Besides, the optimization over variational parameter θ is not convex on the space of probability distributions. However, the optimization process over θ strictly follows the spirit of stochastic optimization, where the expectation of the noisy gradient Eq.9 is equivalent to the true gradient Eq.8. Therefore, it guarantees to converge to a local optimum if the learning rate ρ_t satisfies the Robbins-Monro condition:

$$\sum_{t=1}^{\infty} \rho_t = \infty, \quad \sum_{t=1}^{\infty} \rho_t^2 < \infty.$$

Dataset	**Instance**	**Attributes**
Concrete Compressive Strength (concrete)	1030	8
Real Estate Valuation (estate)	414	6
Yacht Hydrodynamics (yacht)	308	6
QSAR Fish Toxicity (fish)	908	6
QSAR Aquatic Toxicity (aquatic)	546	8
Airfoil Self-Noise (airfoil)	1503	5
Combined Cycle Power Plant (ccpp)	9568	4

Table 1: Characterisics of the UCI data sets

Experiments

In this section, we present qualitative and quantitative results that showcase the effectiveness of our method VOT on both synthetic and real data.

Experimental Setup

Though the ground cost c in OT distances is data-dependent, there are few studies to guide how to choose it for different data (Peyré, Cuturi et al. 2019). In practice, the Euclidean distance is commonly used. Therefore, we adopt the Euclidean distance as the ground cost in VOT, i.e., $c(z_1, z_2) = \|z_1 - z_2\|_2^2$, which is sufficient to examine whether VOT can accurately approximate posterior distributions.

We compare VOT against the following well-known variational inference methods with different α-divergences ($\alpha \in \{0.5, \to 1 \text{ (KL divergence)}, \ 1.5, \ 2 \ (\chi^2 - \text{divergence})\}$), which are implemented using publicly available code [2]. In all methods, Adam optimizer is employed to adjust the learning rate with parameters β_1=0.9, β_2=0.999 and α=0.001 (Kingma and Ba 2015). The sample number S is set to 128 and the training epoch is set to 500. The entropy regularization parameter ε is set to 0.1. The constants a and b in λ are set to $2 * 10^{-3}$ and $2 * 10^4$, respectively. More details can be found in Appendix.

Synthetic Example

We first verify the theoretical result that VOT mitigates the issue of underestimating posterior variance through a synthetic example. We consider a simple Bayesian linear regression model with two highly correlated predictors: $\sigma^2 \sim \mathcal{IG}(a_0, b_0)$, $\omega|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2 V_0)$, $y_i|\omega, \sigma^2 \sim \mathcal{N}(X_i\omega, \sigma^2)$. $\mathcal{IG}(\cdot,\cdot)$ denotes an inverse gamma distribution, and $\mathcal{N}(\cdot,\cdot)$ denotes a normal distribution. Additional details are available in Appendix. Our goal is to estimate posterior distribution of $\omega = (\omega_1, \omega_2)$. The approximation results of ω_1 under different divergences are shown in Figure 2. We can see that our method VOT with $\lambda = 1$ achieves marginal variances that more closely correspond to the exact posterior than other methods. This result demonstrates: 1. Keeping

[2] https://github.com/YingzhenLi/VRbound

Dataset	**α=0.5**	**$\alpha \to$ 1 (KL divergence)**	**α=1.5**	**α=2 (χ^2-divergence)**	**VOT (Ours)**
concrete	5.3740±0.23	5.8052±0.14	5.5913±0.23	5.5670±0.08	**4.7659±0.05**
estate	6.0342±0.35	5.8615±0.10	5.8390±0.07	6.1237±0.16	**5.8276±0.13**
yacht	1.0951±0.32	0.7842±0.04	0.8847±0.04	0.8406±0.04	**0.6212±0.06**
fish	1.004±0.03	1.3285±0.07	1.3620±0.01	1.3627±0.01	**0.9668±0.01**
aquatic	0.9887±0.04	1.1779±0.03	1.0872±0.05	1.1189±0.04	**0.8706±0.01**
airfoil	2.0978±0.13	1.8322±0.05	1.9504±0.07	2.0588±0.07	**1.7033±0.04**
ccpp	4.2586±0.08	4.2071±0.02	4.2479±0.03	4.2372±0.03	**4.0579±0.02**

Table 2: Average Test RMSE. Results in the format ”mean ± std”. Lower is better.

Dataset	**α=0.5**	**$\alpha \to$ 1 (KL divergence)**	**α=1.5**	**α=2 (χ^2-divergence)**	**VOT (Ours)**
concrete	**3.1112±0.03**	3.1705±0.02	3.1497±0.03	3.1476±0.01	3.4900±0.02
estate	3.2910±0.03	3.3281±0.01	3.3149±0.01	3.3360±0.01	**3.1848±0.03**
yacht	1.8259±0.05	1.7677±0.01	2.0108±0.01	2.0267±0.01	**1.3956±0.02**
fish	**1.4240**±0.04	1.7178±0.04	1.7517±0.01	1.7508±0.01	1.6817±0.02
aquatic	1.4248±0.05	1.6138±0.02	1.5141±0.04	1.5344±0.03	**1.4219±0.02**
airfoil	2.1680±0.03	2.0635±0.02	2.1429±0.03	2.1842±0.03	**1.9541±0.03**
ccpp	2.8540±0.02	2.8579±0.01	2.8679±0.01	2.8653±0.01	**2.8257±0.01**

Table 3: Average test NLL. Results in the format ”mean ± std”. Lower is better.

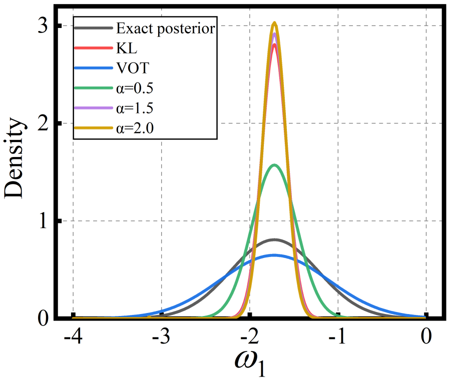


Figure 2: Marginal posterior for ω_1 of a Bayesian linear model under different divergences. In VOT, we set $\lambda = 1$. Prior specification: $\sigma^2 \sim \mathcal{IG}(2,5)$, $\omega_1 \sim \mathcal{N}(0, 0.16^2)$, $\omega_2 \sim \mathcal{N}(0, 0.16^2)$.

priors p_0 and variational family $\mathcal{Q}$ unchanged, the choice of divergence D significantly impacts the performance. 2. OT distances exhibit a desirable behavior in VI.

Bayesian Neural Network Regression

We now evaluate our method on a Bayesian neural network, where a single-layer neural network with 100 hidden units (ReLU) for all data sets is used. We choose a default prior, i.e., a fully factorized Gaussian prior $p_0(z) = \mathcal{N}(z; 0, \mathrm{I})$ for the network weights, which is often an inappropriate prior in practice (Fortuin et al. 2022; Wenzel et al. 2020). We assume the variational distribution to be a fully factorized Gaussian distribution i.e., $q_\theta(z) = \mathcal{N}(z; \mu_\theta, diag(\sigma_\theta^2))$, where variational parameters μ_θ and σ_θ are to be optimized. The linear regression task is performed on seven widely-used benchmark data sets from the UCI dataset repository [3]. The statistics of the data sets are shown in Table 1. Each data set is randomly split into 90% for training and 10% for testing. We compare our method with four well-known variational inference methods with different α-divergences, i.e., $\alpha \in \{0.5, \to 1 \text{ (KL divergence)}, 1.5, 2 \text{ } (\chi^2 - \text{divergence})\}$. All methods are evaluated on the test sets using the average **R**oot **M**ean **S**quare **E**rror (**RMSE**) and the average **N**egative **L**og **L**ikelihood (**NLL**).

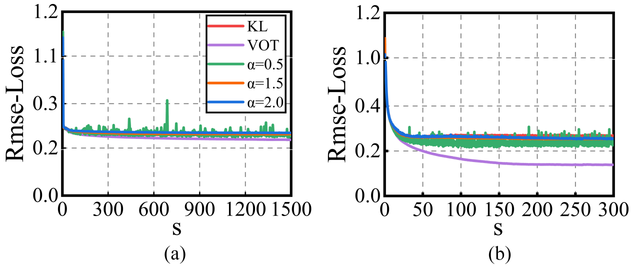


Figure 3: Convergence curves of variational inference methods on two data sets: (a) ccpp and (b) concrete.

The results are shown in Tables 2 and 3. We see that both test RMSE and test NLL values of VOT are significantly lower than those computed by other methods in most cases. These results imply that our method VOT can compute more accurate approximations to the true posterior, since it takes advantage of OT distances to produce desirable uncertainty quantification under misspecified prior distribution.

Convergence. We present the convergence of VOT on training sets using RMSE evaluation method in Figure 3. Despite the non-convex nature of the optimization over θ and the presence of regularization in the OT distance, VOT can

[3] http://archive.ics.uci.edu/ml/datasets.html

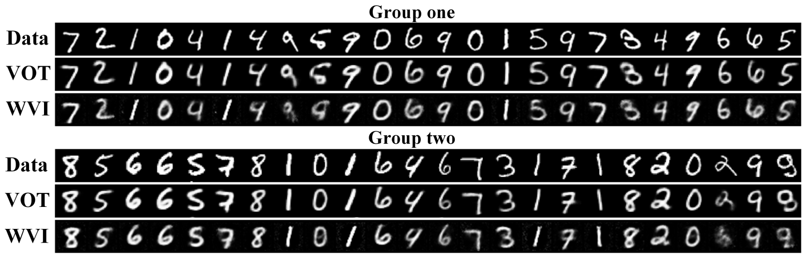


Figure 4: MNIST dataset reconstruction. The first line is the real number in MNIST dataset. The second line is generated by our method VOT, and the third line is generated by WVI.

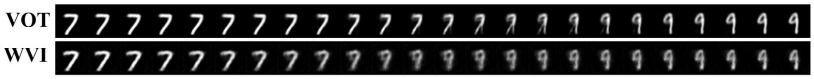


Figure 5: MNIST interpolations. The first line shows the interpolation behavior for VOT. The second line shows the interpolation behavior for WVI.

Dataset	VOT		VOT($\lambda = 1$)	
	RMSE	NLL	RMSE	NLL
concrete	**4.9419**	**3.6076**	5.1378	3.6230
estate	**8.5780**	**3.6791**	10.6277	3.9733
aquatic	**1.2768**	**1.6782**	1.3901	1.7101

Table 4: Effectiveness of λ on over-parameterized models.

empirically converge to a local optimum. Due to the complex computation of OT distances, VOT converges relatively slowly in some cases, but this is acceptable.

Effectiveness of λ. To evaluate the effectiveness of the hyperparameter λ on over-parameterized models, we set the network to have two hidden layers and 128 features with ReLU activations. The linear regression task is performed on three data sets. We give the result of the ablative version of VOT, i.e., $\lambda = 1$ in Table 4. We see that VOT with $\lambda = 1$ performs worse than $\lambda \neq 1$. This directly indicates the positive impact of λ in VI.

Variational Autoencoder

We present the qualitative performance of our method for image reconstruction and generation on MNIST dataset[4], a collection of handwritten digits from zero to nine. The sizes of training and testing are 60000 and 10000, respectively. Similarly to (Ambrogioni et al. 2018), we use a three-layered fully connected network which has 100-300-500-1568 units with ReLu nonlinearities in the hidden layers for generative model and a three-layered ReLu network with 748-500-300-100 units for a variational model. We compare our method with the pseudo OT-based method, referred to as *WVI*(Ambrogioni et al. 2018), which is implemented upon the publicly available code [5].

Some reconstructed and generated images from VOT and WVI are presented in Figure 4. More results can be found in Appendix. We clearly see that the visual quality of these samples from VOT is identical to those from real data in most cases. Additionally, Figure 5 offers an illustration of transforming a digit 7 into a digit 9, highlighting the interpolation behavior of VOT on the MNIST dataset. We find that the images generated by VOT exhibit remarkable clarity and completeness, implying that VOT is capable of producing realistic and smooth interpolations. These results demonstrate that using OT distance to measure the difference between the prior and variational distribution in VOT can yield improved behavior of the variational objective, thereby resulting in accurate approximations for inference.

Conclusion

In this paper, we propose a novel method VOT for variational inference, which employs OT distance to measure the distance between the prior and the variational distribution. We further enhance the objective by gradually including the OT term using a hyperparameter λ for over-parameterized models. The proposed variational objective can be iteratively optimized by a gradient-based black-box algorithm with the reparameterization trick. We demonstrate the effectiveness of our proposed method on Bayesian neural network for linear regression and variational autoencoder for image reconstruction and interpolation.

Acknowledgments

We would like to acknowledge support for this project from the National Natural Science Foundation of China (NSFC) (No. 62006094, No. 62202369, No.62276113).

References

Ambrogioni, L.; Güçlü, U.; Güçlütürk, Y.; Hinne, M.; van Gerven, M. A. J.; and Maris, E. 2018. Wasserstein Variational Inference. In *Neural Information Processing Systems*, 2478–2487.

[4] http://yann.lecun.com/exdb/mnist/

[5] https://github.com/zqkhan/wvi_pytorch

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 214–223.

Blei, D. M.; Kucukelbir, A.; and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518): 859–877.

Chi, J.; Yang, Z.; Li, X.; Ouyang, J.; and Guan, R. 2023. Variational Wasserstein Barycenters with C-cyclical Monotonicity Regularization. In *AAAI Conference on Artificial Intelligence*.

Cominetti, R.; and San Martín, J. 1994. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1): 169–187.

Cox, D. R.; and Hinkley, D. V. 1979. *Theoretical statistics*. CRC Press.

Csiszár, I.; Shields, P.; et al. 2004. Information Theory and Statistics: A Tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4): 417–528.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Neural Information Processing Systems*, 2292–2300.

Dieng, A. B.; Tran, D.; Ranganath, R.; Paisley, J. W.; and Blei, D. M. 2017. Variational Inference via \chi Upper Bound Minimization. In *Neural Information Processing Systems*, 2732–2741.

Fortuin, V.; Garriga-Alonso, A.; Ober, S. W.; Wenzel, F.; Ratsch, G.; Turner, R. E.; van der Wilk, M.; and Aitchison, L. 2022. Bayesian Neural Network Priors Revisited. In *International Conference on Learning Representations*.

Genevay, A.; Cuturi, M.; Peyré, G.; and Bach, F. 2016. Stochastic optimization for large-scale optimal transport. In *Neural Information Processing Systems*, 3440–3448.

Hernandez-Lobato, J.; Li, Y.; Rowland, M.; Bui, T.; Hernández-Lobato, D.; and Turner, R. 2016. Black-box alpha divergence minimization. In *International Conference on Machine Learning*, 1511–1520.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine Learning*, 37(2): 183–233.

Kingma, D. P.; and Ba, J. 2015. Adam: a Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.

Knoblauch, J.; Jewson, J.; and Damoulas, T. 2022. An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference. *Journal of Machine Learning Research*, 23(132): 1–109.

Kuleshov, V.; and Ermon, S. 2017. Neural variational inference and learning in undirected graphical models. In *Neural Information Processing Systems*, 6737–6746.

Li, Y.; Hernández-Lobato, J. M.; and Turner, R. E. 2015. Stochastic Expectation Propagation. In *Neural Information Processing Systems*, 2323–2331.

Li, Y.; and Turner, R. E. 2016. Rényi divergence variational inference. In *Neural Information Processing Systems*, 1081–1089.

Minka, T. P. 2001a. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 362–369.

Minka, T. P. 2001b. *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, Massachusetts Institute of Technology.

Nazaret, A.; and Blei, D. 2022. Variational inference for infinitely deep neural networks. In *International Conference on Machine Learning*, 16447–16461.

Okada, M.; and Taniguchi, T. 2019. Variational Inference MPC for Bayesian Model-based Reinforcement Learning. In *Conference on Robot Learning*, volume 100, 258–272. PMLR.

Pei, H.; Wei, B.; Chang, K. C.-C.; Lei, Y.; and Yang, B. 2020. Geom-GCN: Geometric Graph Convolutional Networks. In *International Conference on Learning Representations*.

Pei, H.; Yang, B.; Liu, J.; and Chang, K. C.-C. 2022. Active Surveillance via Group Sparse Bayesian Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3): 1133–1148.

Peyré, G.; Cuturi, M.; et al. 2019. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6): 355–607.

Ranganath, R.; Gerrish, S.; and Blei, D. M. 2014. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, 814–822.

Rudner, T. G. J.; Key, O.; Gal, Y.; and Rainforth, T. 2021. On Signal-to-Noise Ratio Issues in Variational Inference for Deep Gaussian Processes. In *International Conference on Machine Learning*, volume 139, 9148–9156. PMLR.

Seguy, V.; Damodaran, B. B.; Flamary, R.; Courty, N.; Rolet, A.; and Blondel, M. 2018. Large-scale optimal transport and mapping estimation. In *International Conference on Learning Representations*.

Shi, J.; Titsias, M. K.; and Mnih, A. 2020. Sparse Orthogonal Variational Inference for Gaussian Processes. In *International Conference on Artificial Intelligence and Statistics*, volume 108, 1932–1942. PMLR.

Team, S. D. 2015. Stan: A c++ library for probability and sampling, version 2.8.0. https://mc-stan.org/. Accessed: 2023-05-10.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer.

Wan, N.; Li, D.; and Hovakimyan, N. 2020. f-Divergence Variational Inference. In *Neural Information Processing Systems*.

Wang, D.; Liu, H.; and Liu, Q. 2018. Variational Inference with Tail-adaptive f-Divergence. In *Neural Information Processing Systems*, 5742–5752.

Wang, Q.; and van Hoof, H. 2020. Doubly Stochastic Variational Inference for Neural Processes with Hierarchical Latent Variables. In *International Conference on Machine Learning*, volume 119, 10018–10028. PMLR.

Wenzel, F.; Roth, K.; Veeling, B.; Swiatkowski, J.; Tran, L.; Mandt, S.; Snoek, J.; Salimans, T.; Jenatton, R.; and Nowozin, S. 2020. How Good is the Bayes Posterior in Deep Neural Networks Really? In *International Conference on Machine Learning*, 10248–10259.