GAD-PVI : A General Accelerated Dynamic-Weight Particle-Based Variational Inference Framework

Fangyikang Wang[1], **Huminhao Zhu**[1], **Chao Zhang**[*1,2], **Hanbin Zhao**[1,2], **Hui Qian**[1,3]
[1]College of Computer Science and Technology, Zhejiang University
[2]Advanced Technology Institute, Zhejiang University
[3]State Key Lab of CAD&CG, Zhejiang University
{wangfangyikang, zhuhuminhao, zczju, zhaohanbin, qianhui}@zju.edu.cn

Abstract

Particle-based Variational Inference (ParVI) methods approximate the target distribution by iteratively evolving finite weighted particle systems. Recent advances of ParVI methods reveal the benefits of accelerated position update strategies and dynamic weight adjustment approaches. In this paper, we propose the first ParVI framework that possesses both accelerated position update and dynamical weight adjustment simultaneously, named the General Accelerated Dynamic-Weight Particle-based Variational Inference (GAD-PVI) framework. Generally, GAD-PVI simulates the semi-Hamiltonian gradient flow on a novel Information-Fisher-Rao space, which yields an additional decrease on the local functional dissipation. GAD-PVI is compatible with different dissimilarity functionals and associated smoothing approaches under three information metrics. Experiments on both synthetic and real-world data demonstrate the faster convergence and reduced approximation error of GAD-PVI methods over the state-of-the-art.

Introduction

Particle-based Variational Inference (ParVI) methods have gained significant attention in the Bayesian inference literature owing to their effectiveness in providing approximations of the target posterior distribution π (Liu and Wang 2016; Zhu, Liu, and Zhu 2020; Dong et al. 2020; Shen, Heinonen, and Kaski 2021; Dong et al. 2021; Zhang et al. 2022; Li et al. 2023). The essence of ParVI lies in deterministically evolving a system of finite weighted particles by simulating the probability space gradient flow of certain dissimilarity functional $\mathcal{F}(\mu) := \mathcal{D}(\mu|\pi)$ vanishing at $\mu = \pi$ (Liu et al. 2019). Since the seminal work Stein Variational Gradient Descent (SVGD) (Liu and Wang 2016), classical ParVI focuses on simulating the *first-order* gradient flow in *Wasserstein* space. By using different dissimilarity and associated smoothing approaches, various effective ParVI methods have been proposed, including the BLOB method(Chen et al. 2018a), the GFSD method(Liu et al. 2019), and the KSDD method(Korba et al. 2021).

To improve the efficiency of ParVIs, recent works explore different aspects of the underlying geometry structures in the probability space and design two types of refined particle systems with either *accelerated position update* or *dynamic weight adjustment*.

- *Accelerated position update.* By considering the second-order Riemannian information of the Wasserstein probability space, different accelerated position update strategies have been proposed(Liu et al. 2019; Taghvaei and Mehta 2019): Liu et al. (2019) follows the accelerated gradient descend methods in the Wasserstein probability space (Liu et al. 2017; Zhang and Sra 2018) and derives the WNES and WAG methods, which update the particles' positions with an extra momentum; the ACCEL method (Taghvaei and Mehta 2019) directly discretizes the Hamiltonian gradient flow in the Wasserstein space and update the position with the damped velocity field, which effectively decrease the Hamiltonian potential of the particle system. Later, Wang and Li (2022) consider the Hamiltonian gradient flow for general information probability space (Lafferty 1988), and derive novel accelerated position update strategies according to the Kalman-Wasserstein/Stein Hamiltonian flow. They theoretically show that the Hamiltonian flow usually has a faster convergence to the equilibrium compared with the original first-order counterpart under mild condition. Numerous experimental studies demonstrate that these accelerated position update strategies usually drift the particle system to the target distribution more efficiently (Liu et al. 2019; Wang and Li 2022).
- *Dynamic weight adjustment.* Delving into the orthogonality structure of the Wasserstein-Fisher-Rao(WFR) space, Zhang et al. (2022) developed the first dynamic-weight ParVI (DPVI) methods. Specifically, they derive effective dynamical weight adjustment approaches by mimicing the reaction variational step in a JKO splitting scheme of first-order WFR gradient flow (Gallouët and Monsaingeon 2017; Rotskoff et al. 2019). Compared with the commonly used fixed weight strategy, these dynamical weight adjustment schemes usually lead to less approximation error, especially when the number of particles is limited (Zhang et al. 2022).

Contribution: In this paper, we propose the first ParVI methods which possess both accelerated position update and dynamical weight adjustment simultaneously. Specif-

*Corresponding author.

Methods \ Features	Accelerated position update	Dynamic weight adjustment	Dissimilarity and associated smoothing approach	Underlying probability space
SVGD (Liu and Wang 2016)	✗	✗	KL-RKHS	Wasserstein
BLOB (Craig and Bertozzi 2016)	✗	✗	KL-BLOB	Wasserstein
KSDD (Korba et al. 2021)	✗	✗	KSD-KSDD	Wasserstein
ACCEL (Taghvaei and Mehta 2019)	✓	✗	KL-GFSD	Wasserstein
WNES, WAG (Liu et al. 2019)	✓	✗	General	Wasserstein
AIG (Wang and Li 2022)	✓	✗	KL-GFSD	Information (General)
DPVI (Zhang et al. 2022)	✗	✓	General	WFR
GAD-PVI (Ours)	✓	✓	General	IFR (General)

Table 1: Feature-by-Feature comparison of different ParVIs.

ically, we first construct a novel Information-Fisher-Rao (IFR) probability space, which augment the original information space with an orthogonal Fisher-Rao structure. Then, we originate a novel Semi-Hamiltonian IFR (SHIFR) flow in this space, which simplifies the influence of the kinetic energy on the velocity field in the Hamiltonian IFR flow[1]. By discretizing the SHIFR flow, a practical General Accelerated Dynamic-weight Particle-based Variational Inference (GAD-PVI) framework is proposed. The main contribution of our paper are listed as follows:

- We investigate the convergence property of the SHIFR flow and show that the target distribution π is the stationary distribution of the proposed semi-Hamiltonian flow for proper dissimilarity functional $\mathcal{D}(\cdot|\pi)$. Moreover, our theoretical result also shows that the augmented Fisher-Rao structure yields an additional decrease on the local functional dissipation, compared to the Hamiltonian flow in the vanilla information space.
- We derive an effective finite-particle approximation to the SHIFR flow, which directly evolves the position, weight, and velocity of the particles via a set of ordinary differential equations. The finite particle system is compatible with different dissimilarity and associated smoothing approaches. We prove that the mean-field limit of the proposed particle system converges to the exact SHIFR flow under mild condition.
- By adopting explicit Euler discretization to the finite-particle system, we architect the General Accelerated Dynamic-weight Particle-based Variational Inference (GAD-PVI) framework, which update positions in an acceleration manner and dynamically adjust weights. We derive nine GAD-PVI instances by using three different dissimilarity functionals and associated smoothing approaches(KL-BLOB, KL-GFSD and KSD-KSDD) on the Wasserstein/Kalman-Wasserstein/Stein IFR space, respectively.

We evaluate our algorithms on various synthetic and real-world tasks. The empirical results demonstrate the superiority of our GAD-PVI methods.

[1]Though the Hamiltonian IFR flow seems a natural choice, it is generally infeasible to obtain practical algorithm by discretizing this flow. Please check the Appendix A.2 for a detailed discussion of Hamiltonian IFR flow. Our appendix can be downloaded at https://github.com/zituitui/GAD-PVI_paper

Notation. Given a probability measure μ on $\mathbb{R}^d$, we denote $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ if its second moment is finite. For a given functional $\mathcal{F}(\cdot) : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, $\frac{\delta \mathcal{F}(\tilde{\mu})}{\delta \mu}(\cdot) : \mathbb{R}^d \to \mathbb{R}$ denote its first variation at $\mu = \tilde{\mu}$. We use $C(\mathbb{R}^n)$ to denote continuous functions from $\mathbb{R}^n$ to $\mathbb{R}$. We denote $\mathbf{x}^i \in \mathbb{R}^d$ as the i-th particle, for $i \in \{1...M\}$. We denote the Dirac delta distribution with point mass located at $\mathbf{x}^i$ as $\delta_{\mathbf{x}^i}$, and use $f * g : \mathbb{R}^d \to \mathbb{R}$ to denote the convolution operation between $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$. Besides, we use ∇ and $\nabla \cdot ()$ to denote the gradient and the divergence operator, respectively. We denote a general information probability space as $(\mathcal{P}(\mathbb{R}^n), G(\mu))$, where $G(\mu)[\cdot]$ denotes the one-to-one information metric tensor mapping elements in the tangent space $T_\mu \mathcal{P}(\mathbb{R}^n) \subset C(\mathbb{R}^n)$ to the cotangent space $T^*_\mu \mathcal{P}(\mathbb{R}^n) \subset C(\mathbb{R}^n)$. The inverse map of $G(\mu)[\cdot]$ is denoted as $G^{-1}(\mu)[\cdot] : T^*_\mu \mathcal{P}(\mathbb{R}^n) \to T_\mu \mathcal{P}(\mathbb{R}^n)$.

Preliminaries

When dealing with Bayesian inference tasks, variational inference methods approximate the target posterior π with an easy-to-sample distribution μ, and recast the inference task as an optimization problem over $\mathcal{P}_2(\mathbb{R}^d)$ (Ranganath, Gerrish, and Blei 2014):

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^n)} \mathcal{F}(\mu) := \mathcal{D}(\mu|\pi). \tag{1}$$

To solve this optimization problem, Particle-based Variational Inference (ParVI) methods generally simulate the gradient flow of $\mathcal{F}(\mu)$ in certain probability space with a finite particle system, which transport the initial empirical distribution towards the target distribution π iteratively. Given an information metric tensor $G(\mu)[\cdot]$, the gradient flow in the information probability space $(\mathcal{P}(\mathbb{R}^n), G(\mu))$ takes the following form(Ambrosio, Gigli, and Savaré 2008):

$$\partial_t \mu_t = -G(\mu_t)^{-1} \left[\frac{\delta \mathcal{F}(\mu_t)}{\delta \mu} \right]. \tag{2}$$

Wasserstein Gradient Flow and Classical ParVIs

Since the seminal work Stein Variational Gradient Descent (SVGD)(Liu and Wang 2016), many ParVI methods focus on flows in the Wasserstein space, where the inverse of the Wasserstein metric tensor writes

$$G^W(\mu)^{-1}[\Phi] = -\nabla \cdot (\mu \nabla \Phi), \Phi \in T^*_\mu \mathcal{P}(\mathbb{R}^n), \tag{3}$$

and the Wasserstein gradient flow is defined as

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla \frac{\delta \mathcal{F}(\mu_t)}{\delta \mu}). \tag{4}$$

Based on the probability flow (4) on the density, existing ParVIs maintain a set of particles $\mathbf{x}_t^i$ and directly modify the particle position according to the following ordinary differential equation

$$\mathrm{d}\mathbf{x}_t^i = \nabla \frac{\delta \mathcal{F}(\tilde{\mu}_t)}{\delta \mu}(\mathbf{x}_t^i)\mathrm{d}t, \tag{5}$$

where $\tilde{\mu}_t = \sum_{i=1}^M w_t^i \delta_{\mathbf{x}_t^i}$ denotes the empirical distribution. Since the first total variation $\frac{\delta \mathcal{F}(\tilde{\mu}_t)}{\delta \mu}$ of $\mathcal{F}$ might be not well-defined for the discrete empirical distribution, various ParVI methods have proposed by choosing different dissimilarity $\mathcal{F}$ and associated smoothing approaches for $\frac{\delta \mathcal{F}(\tilde{\mu}_t)}{\delta \mu}$, e.g., KL-BLOB(Chen et al. 2018a), KL-GFSD(Liu et al. 2019), and KSD-KSDD(Korba et al. 2021).

Hamiltonian Gradient Flows and Accelerated ParVIs

The following Hamiltonian gradient flow in the general information probability space has recently been utilized to derive more efficient ParVI methods

$$\begin{cases} \partial_t \mu_t = \dfrac{\delta}{\delta \Phi} \mathcal{H}(\mu_t, \Phi_t), \\ \partial_t \Phi_t = -\gamma_t \Phi_t - \dfrac{\delta}{\delta \mu} \mathcal{H}(\mu_t, \Phi_t), \end{cases} \tag{6}$$

where Φ_t denote the Hamiltonian velocity and $\mathcal{H}(\mu_t, \Phi_t) = \frac{1}{2}\int \Phi_t G(\mu_t)^{-1}[\Phi_t]\, dx + \mathcal{F}(\mu_t)$ denotes the Hamiltonian potential. Note that the Hamiltonian flow (6) can be regarded as the second-order accelerated version of the information gradient flow (2), and usually converges faster to the equilibrium of the target distribution under mild condition(Carrillo, Choi, and Tse 2019; Taghvaei and Mehta 2019; Wang and Li 2022). Though the form of the Hamiltonian flow (6) seems complicated, it induces a simple augmented particle system $(\mathbf{x}_t^i, \mathbf{v}_t^i)$, which evolves the position $\mathbf{x}_t^i$ and velocity $\mathbf{v}_t^i$ of particles simultaneously. As the position update rule of $\mathbf{x}_t^i$ also uses the extra velocity information, the induced system is said to have an accelerated position update. By discretizing the continuous particle system, several accelerated ParVI methods have been proposed, which converge faster to the target distribution in numerous real-world Bayesian inference tasks (Taghvaei and Mehta 2019; Wang and Li 2022).

Wasserstein-Fisher-Rao Flow and Dynamic-Weight ParVIs

Recently, the Wasserstein-Fisher-Rao (WFR) Flow has been used to derive effective dynamic weight adjustment approaches to mitigate the fixed-weight restriction of ParVIs(Zhang et al. 2022). The inverse of WFR metric tensor is

$$G^{WFR}(\mu)^{-1}[\Phi] = -\nabla \cdot (\mu \nabla \Phi) + (\Phi - \int \Phi)\mu, \tag{7}$$

where $\Phi \in T_\mu^* \mathcal{P}(\mathbb{R}^n)$, and the WFR gradient flow writes:

$$\partial_t \mu_t = \underbrace{\nabla \cdot (\mu_t \nabla \frac{\delta \mathcal{F}(\mu_t)}{\delta \mu})}_{\text{Wasserstein transport}} - \underbrace{(\frac{\delta \mathcal{F}(\mu_t)}{\delta \mu} - \int \frac{\delta \mathcal{F}(\mu_t)}{\delta \mu}\mathrm{d}\mu_t)\mu_t}_{\text{Fisher-Rao variational distortion}}. \tag{8}$$

Since the WFR space can be regarded as orthogonal sum of the Wasserstein space and the Fisher-Rao space, Zhang et al. (2022) mimic a JKO splitting scheme for the WFR flow, which deal with the position and the weight with the Wasserstein transport and the Fisher-Rao variational distortion, respectively. Given a set of particles with position $\mathbf{x}_t^i$ and weight $\mathbf{w}_t^i$, the Fisher-Rao distortion can be approximated by the following ode

$$\frac{\mathrm{d}}{\mathrm{d}t} w_t^i = -\left(\frac{\delta \mathcal{F}(\tilde{\mu}_t)}{\delta \mu}(\mathbf{x}_t^i) - \sum_{i=1}^M w_t^i \frac{\delta \mathcal{F}(\tilde{\mu}_t)}{\delta \mu}(\mathbf{x}_t^i)\right) w_t^i. \tag{9}$$

According to the ode (9), Zhang et al. (2022) derive two dynamical weight-adjustment scheme and propose the Dynamic-Weight Particle-Based Variational Inference (DPVI) framework, which is compatible with several dissimilarity functionals and associated smoothing approaches.

Methodology

In this section, we present our General Accelerated Dynamic-weight Particle-based Variational Inference (GAD-PVI) framework, detailted in Algorithm 1. We first introduce a novel augmented Information-Fisher-Rao space, and originate the Semi-Hamiltonian-Information-Fisher-Rao (SHIFR) flow in the space. The theoretical analysis on SHIFR shows that it usually possesses an additional decrease on the local functional dissipation compared to the Hamiltonian flow in the original information space. Then, effective finite-particle systems, which directly evolve the position, weight, and velocity of the particles via a set of ordinary differential equations, are constructed based on SHIFR flows in several IFR spaces with different underlying information metric tensors. We demonstrate that the mean-field limit of the constructed particle system exactly converges to the SHIFR flow in the corresponding probability space. Next, we develop the GAD-PVI framework by discretizing these continuous-time finite-particles formulations, which enables simultaneous accelerated updates of particles' positions and dynamic adjustment of particles' weights. We present nine effective GAD-PVI algorithms that use different underlying information metric tensors, dissimilarity functionals and the associated finite-particle smoothing approaches.

Information-Fisher-Rao Space and Semi-Hamiltonian-Information-Fisher-Rao Flows

To define the augmented Information-Fisher-Rao probability space, we introduce the Information-Fisher-Rao metric tensor $G^{IFR}(\mu)$, whose inverse is defined as follows.

$$G^{IFR}(\mu)^{-1}[\Phi] = G^I(\mu)^{-1}[\Phi] + (\Phi - \int \Phi \mathrm{d}\mu)\mu, \tag{10}$$

where $\Phi \in T_\mu^* \mathcal{P}(\mathbb{R}^n)$ and $G^I(\mu)$ denotes certain underlying information metric tensor. Note that $G^{IFR}(\mu)$ is formed by the inf-convolution of $G^I(\mu)$ and Fisher-Rao metric tensor.

Based on $G^{IFR}(\mu)$, we introduce the following novel semi-Hamiltonian flow of $\mathcal{F}$ on the Information-Fisher-Rao space $(\mathcal{P}(\mathbb{R}^n), G^{IFR}(\mu))$

$$\begin{cases} \partial_t \mu_t = \dfrac{\delta}{\delta \Phi} \mathcal{H}^{IFR}(\mu_t, \Phi_t), \\ \partial_t \Phi_t = -\gamma_t \Phi_t - \dfrac{1}{2} \dfrac{\delta}{\delta \mu} \left(\displaystyle\int \Phi_t G^I(\mu_t)^{-1} \left[\Phi_t\right] dx \right) - \dfrac{\delta \mathcal{F}(\mu_t)}{\delta \mu}. \end{cases} \tag{11}$$

where Φ_t denote the Hamiltonian velocity and

$$\begin{aligned} \mathcal{H}^{IFR}(\mu_t, \Phi_t) = & \underbrace{\frac{1}{2} \int \Phi_t G^I(\mu_t)^{-1} \left[\Phi_t\right] dx}_{\text{Information kinetic energy}} \\ & + \underbrace{\frac{1}{2} \int \Phi_t (\Phi_t - \int \Phi d\mu_t) d\mu_t}_{\text{Fisher-Rao kinetic energy}} + \underbrace{\frac{\delta \mathcal{F}(\mu_t)}{\delta \mu}}_{\text{potential energy}}, \end{aligned} \tag{12}$$

denotes the Hamiltonian potential in the IFR space. Compared to the full Hamiltonian flow of $\mathcal{F}$ in the IFR space, the SHIFR flow (11) ignores the influence of the Fisher-Rao kinetic energy on the Hamiltonian field Φ_t. Later, we will show that SHIFR can be directly transformed into a particle system consisting of odes on the positions, velocities and weights of particles for proper underlying information metric tensor, while it is generally infeasible to obtain such a direct particle system according to the corresponding full Hamiltonian flow because it is difficult to handle the Fisher-Rao kinetic energy. As the kinetic energy term vanishes when near the equilibrium of the flow, therefore it is acceptable for the SHIFR flow to neglect this intractable term and still has the target distribution π as its stationary distribution. Moreover, this semi-Hamiltonian flow would converge faster compare to the Hamiltonian flow in the original information space on account of extra local descending property. Due to the limit of space, we defer the discussion of the stationary analysis and functional dissipation quantitative analysis of the SHIFR flow to Appendix A.4. Please refer to Proposition 2 and Proposition 3 for details.

With different underlying information metric tensor $G^I(\mu)$ in $\mathcal{H}^{IFR}(\mu_t, \Phi_t)$, we can obtain different SHIFR flows. Suitable $G^I(\mu)$ includes the Wasserstein metric tensor,the Kalman-Wasserstein metric tensor (KW-metric) and the Stein metric tensor (S-metric). For instance, the SHIFR flow with Wasserstein metric (Wasserstein-SHIFR flow) writes:

$$\begin{cases} \partial_t \mu_t = -\nabla \cdot (\mu_t \nabla \Phi_t) - \left(\dfrac{\delta \mathcal{F}(\mu_t)}{\delta \mu} - \displaystyle\int \dfrac{\delta \mathcal{F}(\mu_t)}{\delta \mu} \mathrm{d}\mu_t \right) \mu_t, \\ \partial_t \Phi_t = -\gamma_t \Phi_t - \|\nabla \Phi_t\|^2 - \dfrac{\delta \mathcal{F}(\mu_t)}{\delta \mu}. \end{cases} \tag{13}$$

Note that in the subsequent section, we focus on the Wasserstein-SHIFR flow, and defer the detailed formulations with respect to KW-SHIFR and S-SHIFR to the Appendix B.1 and B.2 due to limited space.

Algorithm 1: General Accelerated Dynamic-weight Particle-based Variational Inference (GAD-PVI) framework

Input: Initial distribution $\tilde{\mu}_0 = \sum_{i=1}^M w_0^i \delta_{\mathbf{x}_0^i}$, position adjusting step-size η_{pos}, weight adjusting step-size η_{wei}, velocity field adjusting step-size η_{vel}, velocity damping parameter γ.

1: Choose a suitable functional $\mathcal{F}$ and its smoothing strategy $U_{\tilde{\mu}}$ from KL-BLOB/KL-GFSD/KSD-KSDD
2: **for** $k = 0, 1, ..., T-1$ **do**
3: **for** $i = 1, 2, ..., M$ **do**
4: Update positions $\mathbf{x}_{k+1}^i$'s according to (18).
5: **end for**
6: **for** $i = 1, 2, ..., M$ **do**
7: Adjust velocity field $\mathbf{v}_{k+1}^i$'s according to (19).
8: **end for**
9: **for** $i = 1, 2, ..., M$ **do**
10: Adjust weights w_{k+1}^i's according to (20).
11: **end for**
12: **end for**
13: **Output**:$\tilde{\mu}_T = \sum_{i=1}^M w_T^i \delta_{\mathbf{x}_T^i}$.

Finite-Particles Formulations to SHIFR Flows

Now, we derive the finite-particle approximation to the SHIFR flow, which directly evolves the position $\mathbf{x}_t^i$, weight w_t^i, and velocity $\mathbf{v}_t^i$ of the particles. Specifically, we construct the following ordinary differential equation system to simulate the Wasserstein-SHIFR flow (13)

$$\begin{cases} \mathrm{d}\mathbf{x}_t^i = \mathbf{v}_t^i \mathrm{d}t, \\ \mathrm{d}\mathbf{v}_t^i = (-\gamma \mathbf{v}_t^i - \nabla \dfrac{\delta \mathcal{F}(\tilde{\mu}_t)}{\delta \mu}(\mathbf{x}_t^i)) \mathrm{d}t, \\ \mathrm{d}w_t^i = -\left(\frac{\delta \mathcal{F}(\tilde{\mu}_t)}{\delta \mu}(\mathbf{x}_t^i) - \sum_{i=1}^M w_t^i \frac{\delta \mathcal{F}(\tilde{\mu}_t)}{\delta \mu}(\mathbf{x}_t^i) \right) w_t^i \mathrm{d}t, \\ \tilde{\mu}_t = \sum_{i=1}^M w_t^i \delta_{\mathbf{x}_t^i}. \end{cases} \tag{14}$$

The following proposition demonstrates that the mean-field limit of the particle system (14) corresponds precisely to the Wasserstein-SHIFR flow in (13).

Proposition 1. *Suppose the empirical distribution $\tilde{\mu}_0^M$ of M weighted particles weakly converges to a distribution μ_0 when $M \to \infty$. Then, the path of* (14) *starting from $\tilde{\mu}_0^M$ and Φ_0 with initial velocity* $\mathbf{0}$ *weakly converges to a solution of the Wasserstein-SHIFR gradient flow* (13) *starting from $\mu_t|_{t=0} = \mu_0$ = and $\Phi_t|_{t=0} = \mathbf{0}$ as $M \to \infty$:*

GAD-PVI Framework

Generally, it is impossible to obtain an analytic solution of the continuous finite-particles formulations (14), thus a numerical integration method is required to derive an approximate solution. Note that any numerical solver, such as the implicit Euler method (Platen and Bruti-Liberati 2010) and higher-order Runge-Kutta method (Butcher 1964) can be used. Here, we follow the tradition of ParVIs to adopt the first-order explicit Euler discretization (Süli and Mayers 2003) since it is efficient and easy-to-implement (Zhang et al. 2022), and propose our GAD-PVI framework, as listed in Algorithm 1.

Dissimilarity Functionals and Smoothing Approaches To develop practical GAD-PVI methods, we must first select a dissimilarity functional $\mathcal{F}$. The commonly used underlying functionals are KL-divergence (Liu and Wang 2016; Liu et al. 2019; Wang and Li 2022) and Kernel-Stein-Discrepancy (Korba et al. 2020). Once a dissimilarity functional $\mathcal{F}$ has been chosen, we need to select a smoothing approach to approximate the first variation of the empirical approximation, as the value of $\frac{\delta\mathcal{F}(\cdot)}{\mu}$ at an empirical distribution $\tilde{\mu} = \sum_{i=1}^{M} w^i \delta_{\mathbf{x}^i}$ is generally not well-defined. Smoothing strategies allow us to approximate the first variation value at the discrete empirical distribution. Generally, a smoothed approximation to the first total variation is denoted as $U_{\tilde{\mu}}(\cdot) \approx \frac{\delta\mathcal{F}(\tilde{\mu})}{\delta\mu}(\cdot)$. The commonly used smoothing approaches in the ParVI area, namely BLOB (with KL-divergence as $\mathcal{F}$) (Craig and Bertozzi 2016), GFSD (with KL-divergence as $\mathcal{F}$) (Liu et al. 2019), and KSDD (with Kernel Stein Discrepancy as $\mathcal{F}$) (Korba et al. 2021), are all compatible with our GAD-PVI framework.

Here, we describe the dissimilarity functional KL-divergence and the associated BLOB smoothing approach as an example. The first total variation of the KL is

$$\frac{\delta\mathcal{F}(\mu)}{\delta\mu}(\cdot) := \frac{\delta KL(\mu|\pi)}{\delta\mu}(\cdot) = -\log\pi(\cdot) + \log\mu(\cdot). \quad (15)$$

As $\log\mu(\mathbf{x})$ is ill-defined for the discrete empirical distribution $\tilde{\mu}_k$, BLOB smoothing approach reformulate the intractable term $\log\mu$ as $\frac{\delta}{\delta\mu}\mathbb{E}_\mu[\log\mu]$ and smooth the density with a kernel function K, resulting in the approximation

$$\begin{aligned}\log\tilde{\mu} &\approx \frac{\delta}{\delta\tilde{\mu}}\mathbb{E}_{\tilde{\mu}}[\log(\tilde{\mu} * K)] \quad (16)\\ &:= \log\textstyle\sum_{i=1}^{M} w^i K(\cdot,\mathbf{x}^i) + \sum_{i=1}^{M} \frac{w^i K(\cdot,\mathbf{x}^i)}{\sum_{j=1}^{M} w^j K(\mathbf{x}^i,\mathbf{x}^j)}.\end{aligned}$$

for a discrete density $\tilde{\mu} = \sum_{i=1}^{M} w^i\mathbf{x}^i$. This leads to the following approximation results:

$$\begin{aligned}U_{\tilde{\mu}_k}(\mathbf{x}) = &-\log\pi(\mathbf{x}) + \log\textstyle\sum_{i=1}^{M} w_k^i K(\mathbf{x},\mathbf{x}_k^i)\\ &+ \sum_{i=1}^{M} \frac{w_k^i K(\mathbf{x},\mathbf{x}_k^i)}{\sum_{j=1}^{M} w_k^j K(\mathbf{x}_k^i,\mathbf{x}_k^j)}.\end{aligned} \quad (17)$$

Details regarding other dissimilarity functionals and smoothing approaches are included in the Appendix B.3.

Updating Rules Once the functional $\mathcal{F}$ and its empirical approximation of the first variation $U_{\tilde{\mu}} \approx \frac{\delta\mathcal{F}(\tilde{\mu})}{\delta\mu}$ is decided, we adopt a Jacobi-type strategy to update the position $\mathbf{x}_k^i$, velocity field $\mathbf{v}_k^i$ and the weight w_k^i, i.e., the calculations in the $k+1$-th iteration are totally based on the variables obtained in the k-th iteration. Therefore, starting from M weighted particles located at $\{\mathbf{x}_0^i\}_{i=1}^M$ with weights $\{w_0^i\}_{i=0}^M$ and $\{\mathbf{v}_0^i = 0\}_{i=0}^M$, GAD-PVI w.r.t. the Wasserstein-SHIFR flow first updates the positions of particles according to the following rule:

$$\mathbf{x}_{k+1}^i = \mathbf{x}_k^i + \eta_{pos}\mathbf{v}_k^i. \quad (18)$$

Then, it adjusts the velocity field as

$$\mathbf{v}_{k+1}^i = (1-\gamma\eta_{vel})\mathbf{v}_k^i - \eta_{vel}\nabla U_{\tilde{\mu}_k}(\mathbf{x}_k^i), \quad (19)$$

and particles' weights as following:

$$w_{k+1}^i = w_k^i - \eta_{wei}(U_{\tilde{\mu}_k}(\mathbf{x}_k^i) - \textstyle\sum_{j=1}^{M} w_k^j U_{\tilde{\mu}_k}(\mathbf{x}_k^j)). \quad (20)$$

Here $\tilde{\mu}_k = \sum_{i=1}^{M} w_k^i \delta_{\mathbf{x}_k^i}$ denotes the empirical distribution, and $\eta_{pos}/\eta_{vel}/\eta_{wei}$ are the discretization stepsizes. It can be verified that the total mass of $\tilde{\mu}_k$ is conserved and $\tilde{\mu}_k$ remains a valid probability distribution during the whole procedure of GAD-PVI, i.e. $\sum_i w_k^i = 1$ for all k. The detailed updating rules of GAD-PVI w.r.t. the KW-SHIFR and S-SHIFR can be found in Appendix B.3.

Notice that, compared to the classical ParVIs, the position acceleration scheme and dynamic-weight scheme only bring *little* extra computational cost, because the number of time-complexity-bottleneck operation, i.e. calculation of $U_{\tilde{\mu}}$ and $\nabla U_{\tilde{\mu}}$, remains the same.

An Alternative Weight Adjusting Approach Except for Continuous Adjusting (CA) strategy, the Duplicate/Kill (DK) strategy, which is a probabilistic discretization strategy to the Fisher-Rao part of (13), can also be adopted in GAD-PVI. This strategy duplicates/kills particle $\mathbf{x}_{k+1}^i$ according to an exponential clock with instantaneous rate:

$$R_{k+1}^i = -\eta_{wei}(\frac{\delta\mathcal{F}(\tilde{\mu}_k)}{\delta\mu}(\mathbf{x}_k^i) - \sum_{j=1}^{M} w_k^j \frac{\delta\mathcal{F}(\tilde{\mu}_k)}{\delta\mu}(\mathbf{x}_k^j)). \quad (21)$$

Specifically, if $R_{k+1}^i > 0$, duplicate the particle $\mathbf{x}_{k+1}^i$ with probability $1-\exp(-R_{k+1}^i)$, and kill another one with uniform probability to conserve the total mass; if $R_{k+1}^i < 0$, kill the particle $\mathbf{x}_{k+1}^i$ with probability $1-\exp(R_{k+1}^i)$, and duplicate another one with uniform probability. By replacing the CA strategy (20) in the GAD-PVI framework, we could obtain the DK variants of GAD-PVI methods.

GAD-PVI Instances With different underlying information metric tensors (W-metric, KW-metic and S-metric), weight adjustment approaches(CA and DK) and dissimilarity functionals/associated smoothing approaches(KL-BLOB, KL-GFSD and KSD-KSDD), we can derive 18 different intances of GAD-PVI, named as WGAD/KWGAD/SGAD-CA/DK-BLOB/GFSD/KSDD.

Experiments

In this section, we conduct empirical studies with our GAD-PVI algorithms. Here, we focus on the instances of GAD-PVI w.r.t. the W-SHIFR flows, i.e.,WGAD-CA/DK-BLOB/GFSD. The experimental results on methods w.r.t. the KW-SHIFR and S-SHIFR flows are provided in the Appendix C. Note that we do not include GAD-PVI methods with the KSDD smoothing approaches, as they are more computationally expensive and have been widely reported to be less stable(Korba et al. 2020; Zhang et al. 2022). We include four classes of methods as our baseline: classical ParVI algorithms (SVGD, GFSD and BLOB), the Nesterov accelerated ParVI algorithms (WNES-BLOB/GFSD),

Algorithm	Smoothing Strategy	
	BLOB	GFSD
ParVI	1.570e-01±2.210e-04	2.143e-01±7.424e-04
WAIG	1.572e-01±2.070e-04	2.142e-01±7.048e-04
WNES	1.571e-01±3.011e-04	2.138e-01±7.771e-04
DPVI-DK	1.568e-01±1.496e-03	2.142e-01±2.712e-03
DPVI-CA	1.285e-01±2.960e-04	1.638e-01±4.332e-04
WGAD-DK	1.561e-01±1.155e-03	2.142e-01±1.501e-03
WGAD-CA	**1.274e-01±2.964e-04**	**1.626e-01±4.842e-04**

Table 2: Averaged W_2 for the GP task with dataset LIDAR.

the Hamiltonian accelerated ParVI algorithms (WAIG-BLOB/GFSD) and the Dynamic-weight ParVI algorithms (DPVI-CA/DK-BLOB/GFSD).

We compare the performance of these algorithms on two simulations, i.e., a 10-D Single-mode Gaussian model (SG) and a Gaussian mixture model (GMM), and two real-world applications, i.e. Gaussian Process (GP) regression and Bayesian neural network (BNN). For all the algorithms, the particles' weights are initialized to be equal. In the first three experiments, we tune the parameters to achieve the best W_2 distance. In the BNN task, we split $1/5$ of the training set as our validation set to tune the parameters. Note that, the position step-size are tuned via grid search for the fixed-weight ParVI algorithms, then used in the corresponding dynamic-weight algorithms. The acceleration parameters and weight adjustment parameters are tuned via grid search for each specific algorithm. We repeat all the experiments 10 times and report the average results. Due to limited space, only parts of the results are reported in this section. We refer readers to the Appendix C for the results on SG and additional results for GMM, GP and BNN.

Gaussian Mixture Model

We consider approximating a 10-D Gaussian mixture model with two components, weighted by $1/3$ and $2/3$ respectively. We run all algorithms with particle number $M \in \{32, 64, 128, 256, 512\}$.

In Figure 1, we report the W_2 distance between the empirical distribution generated by each algorithm and the target distribution w.r.t. iterations of different ParVI methods. We generate 5,000 samples from the target distribution π as reference to evaluate the W_2 distance by using the POT library (Flamary et al. 2021). The results demonstrate that our GAD-PVI algorithms consistently outperform their counterpart with only one (or none) of the accelerated position update strategy and dynamic weight adjustment approach. Besides, the CA weight-adjustment approach usually result a lower W_2 compared to the DK scheme, and WGAD-CA-BLOB/GFSD usually have the fastest convergence and the lowest final W_2 distance to the target.

Gaussian Process Regression

The Gaussian Process (GP) model is widely adopted for the uncertainty quantification in regression problems (Rasmussen 2003). We follow the experiment setting in (Chen et al. 2018b), and use the dataset LIDAR which consists of

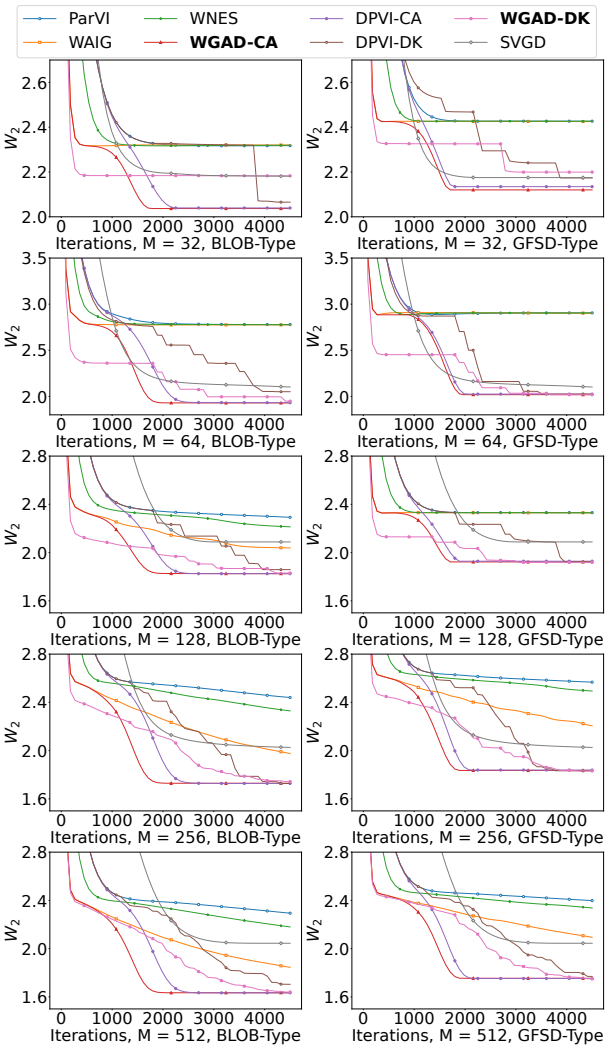

Figure 1: W_2 to the target w.r.t. iterations in the GMM task.

221 observations. In this task, we set the particle number to $M = 128$ for all the algorithms.

We report the W_2 distance between the empirical distribution after 10000 iterations and the target distribution in Table 2. The target distribution is approximated by 10000 reference particles generated by the HMC method(Brooks et al. 2011). It can be observed that both the accelerated position update and the dynamic weight adjustment result in a decreased W_2 and GAD-PVI algorithms consistently achieve lowest W_2 to the target. Besides, the results also show that the CA variants usually outperforms their DK counterpart, as CA is able to adjust the weight continuously on $[0, 1]$ while DK set the weight either to 0 or $1/M$.

In Figure 2, we plot the contour lines of the log posterior and the particles generated by four representative algorithms, namely BLOB, WAIG-BLOB, DPVI-CA-BLOB, and WGAD-CA-BLOB, at different iterations (0, 100, 500, 2000, 10000). The results indicate that the particles in WAIG-BLOB and WGAD-CA-BLOB exhibit a faster con-

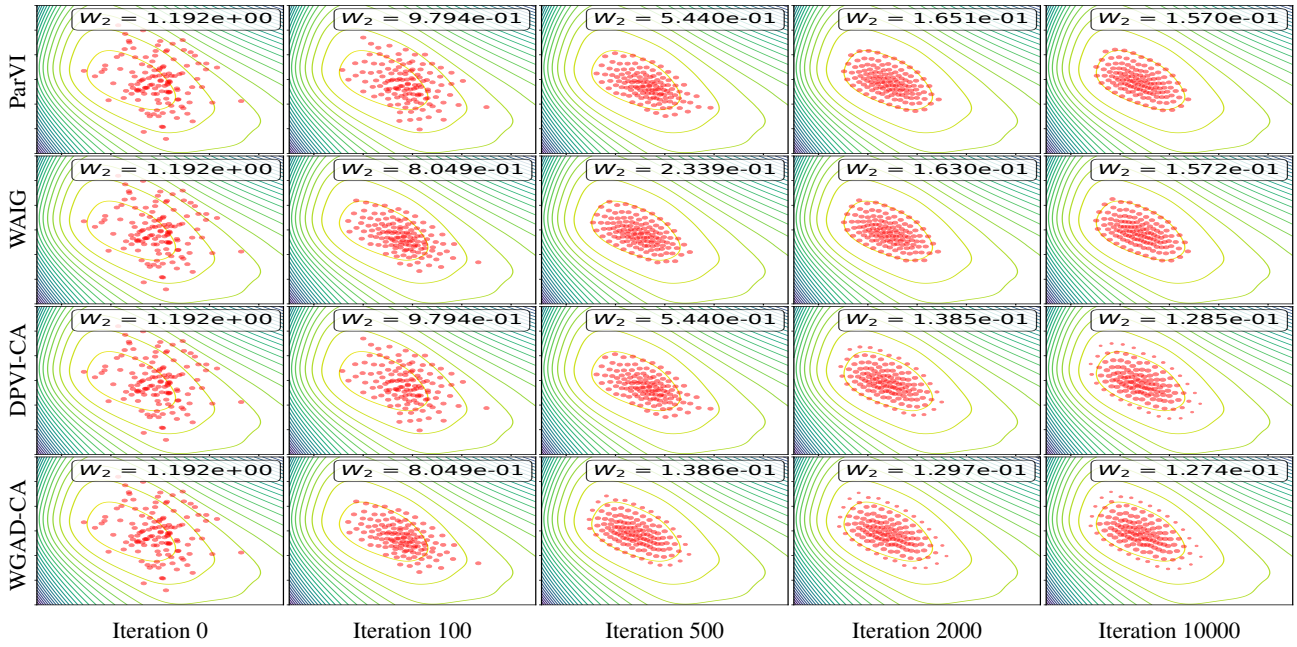

Figure 2: The contour lines of the log posterior in the Gaussian Process task (all variants with BLOB strategy).

vergence to the high probability area of the target due to their accelerated position updating strategy, and the DPVI-CA and WGAD-CA algorithms finally offer a broader final coverage, as the CA dynamic weight adjustment strategy enables the particles to represent the region with arbitrary local density mass instead of a fixed $1/M$ mass.

Bayesian Neural Network

In this experiment, we study a Bayesian regression task with Bayesian neural network on 4 datasets from UCI and LIBSVM. We follow the experiment setting from (Liu and Wang 2016; Zhang et al. 2022), which models the output as a Gaussian distribution and uses a Gamma$(1, 0.1)$ prior for the inverse covariance. We use a one-hidden-layer neural network with 50 hidden units and maintain 128 particles. For all the datasets, we set the batchsize as 128.

We present the Root Mean Squared Error (RMSE) of various ParVI algorithms in Table 3. The results demonstrate that the combination of the accelerated position updating strategy and the dynamically weighted adjustment leads to a lower RMSE. Notably, WGAD-CA type algorithms outperform other methods in the majority of cases.

algorithms	Datasets		
	Concrete	kin8nm	RedWine
ParVI-SVGD	6.323e+00	8.020e-02	6.330e-01
ParVI-BLOB	6.313e+00	7.891e-02	6.318e-01
WAIG-BLOB	6.063e+00	7.791e-02	6.267e-01
WNES-BLOB	6.112e+00	7.690e-02	6.264e-01
DPVI-DK-BLOB	6.285e+00	7.889e-02	6.294e-01
DPVI-CA-BLOB	6.292e+00	7.789e-02	6.298e-01
WGAD-DK-BLOB	6.058e+00	7.688e-02	6.267e-01
WGAD-CA-BLOB	**6.047e+00**	**7.629e-02**	**6.263e-01**
ParVI-GFSD	6.314e+00	7.891e-02	6.317e-01
WAIG-GFSD	6.105e+00	7.794e-02	6.265e-01
WNES-GFSD	6.123e+00	7.756e-02	6.263e-01
DPVI-DK-GFSD	6.291e+00	7.882e-02	6.277e-01
DPVI-CA-GFSD	6.290e+00	7.791e-02	6.298e-01
WGAD-DK-GFSD	6.099e+00	7.726e-02	6.265e-01
WGAD-CA-GFSD	**6.088e+00**	**7.634e-02**	**6.260e-01**

Table 3: Averaged Test $RMSE$ in the BNN task.

Conclusion

In this paper, we propose the General Accelerated Dynamic-Weight Particle-based Variational Inference (GAD-PVI) framework, which adopts an accelerated position update scheme and dynamic weight adjustment approach simultaneously. Our GAD-PVI framework is developed by discretizing the Semi-Hamiltonian Information Fisher-Rao (SHIFR) flow on the novel Information-Fisher-Rao space. The theoretical analysis demonstrate that the SHIFR flow yields additional decrease on the local functional dissipation compared to the Hamiltonian flow in the vanilla information space. We propose effective particle system which evolve the position, weight, velocity of particles via a set of odes for the SHIFR flows with different underlying information metrics. By directly discretizing the proposed particle system, we obtain our GAD-PVI framework. Several effective instances of the GAD-PVI framework have been provided by employing three distinct dissimilarity functionals and associated smoothing approaches under the Wasserstein/Kalman-Wasserstein/Stein metric. Empirical studies demonstrate the faster convergence and reduced approximation error of GAD-PVI methods over the SOTAs.

Acknowledgments

This work is supported by National Key Research and Development Program of China under Grant 2020AAA0107400 and National Natural Science Foundation of China (Grant No: 62206248).

References

Ambrosio, L.; Gigli, N.; and Savaré, G. 2008. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Brooks, S.; Gelman, A.; Jones, G.; and Meng, X.-L. 2011. *Handbook of markov chain monte carlo*. CRC press.

Butcher, J. C. 1964. Implicit runge-kutta processes. *Mathematics of Computation*, 18(85): 50–64.

Carrillo, J. A.; Choi, Y.-P.; and Tse, O. 2019. Convergence to equilibrium in Wasserstein distance for damped Euler equations with interaction forces. *Communications in Mathematical Physics*, 365: 329–361.

Chen, C.; Zhang, R.; Wang, W.; Li, B.; and Chen, L. 2018a. A unified particle-optimization framework for scalable Bayesian sampling. *arXiv preprint arXiv:1805.11659*.

Chen, W. Y.; Mackey, L.; Gorham, J.; Briol, F.-X.; and Oates, C. 2018b. Stein points. In *ICML*, 844–853. PMLR.

Craig, K.; and Bertozzi, A. 2016. A blob method for the aggregation equation. *Mathematics of computation*, 85(300): 1681–1717.

Dong, J.; Cong, Y.; Sun, G.; Fang, Z.; and Ding, Z. 2021. Where and How to Transfer: Knowledge Aggregation-Induced Transferability Perception for Unsupervised Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2): 1–18.

Dong, J.; Cong, Y.; Sun, G.; Zhong, B.; and Xu, X. 2020. What Can Be Transferred: Unsupervised Domain Adaptation for Endoscopic Lesions Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4022–4031.

Flamary, R.; Courty, N.; Gramfort, A.; Alaya, M. Z.; Boisbunon, A.; Chambon, S.; Chapel, L.; Corenflos, A.; Fatras, K.; Fournier, N.; Gautheron, L.; Gayraud, N. T.; Janati, H.; Rakotomamonjy, A.; Redko, I.; Rolet, A.; Schutz, A.; Seguy, V.; Sutherland, D. J.; Tavenard, R.; Tong, A.; and Vayer, T. 2021. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78): 1–8.

Gallouët, T. O.; and Monsaingeon, L. 2017. A JKO Splitting Scheme for Kantorovich–Fisher–Rao Gradient Flows. *SIAM Journal on Mathematical Analysis*, 49(2): 1100–1130.

Korba, A.; Aubin-Frankowski, P.-C.; Majewski, S.; and Ablin, P. 2021. Kernel Stein Discrepancy Descent. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 5719–5730. PMLR.

Korba, A.; Salim, A.; Arbel, M.; Luise, G.; and Gretton, A. 2020. A non-asymptotic analysis for Stein variational gradient descent. *NeurIPS*, 33.

Lafferty, J. D. 1988. The Density Manifold and Configuration Space Quantization. *Transactions of the American Mathematical Society*, 305(2): 699–741.

Li, L.; qiang liu; Korba, A.; Yurochkin, M.; and Solomon, J. 2023. Sampling with Mollified Interaction Energy Descent. In *The Eleventh International Conference on Learning Representations*.

Liu, C.; Zhuo, J.; Cheng, P.; Zhang, R.; and Zhu, J. 2019. Understanding and accelerating particle-based variational inference. In *ICML*, 4082–4092.

Liu, Q.; and Wang, D. 2016. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*.

Liu, Y.; Shang, F.; Cheng, J.; Cheng, H.; and Jiao, L. 2017. Accelerated first-order methods for geodesically convex optimization on Riemannian manifolds. *Advances in Neural Information Processing Systems*, 30.

Platen, E.; and Bruti-Liberati, N. 2010. *Numerical solution of stochastic differential equations with jumps in finance*, volume 64. Springer Science & Business Media.

Ranganath, R.; Gerrish, S.; and Blei, D. 2014. Black box variational inference. In *Artificial intelligence and statistics*, 814–822. PMLR.

Rasmussen, C. E. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*, 63–71. Springer.

Rotskoff, G.; Jelassi, S.; Bruna, J.; and Vanden-Eijnden, E. 2019. Global convergence of neuron birth-death dynamics. *arXiv preprint arXiv:1902.01843*.

Shen, Z.; Heinonen, M.; and Kaski, S. 2021. Derandomizing MCMC dynamics with the diffusion Stein operator. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 17507–17517. Curran Associates, Inc.

Süli, E.; and Mayers, D. F. 2003. *An introduction to numerical analysis*. Cambridge university press.

Taghvaei, A.; and Mehta, P. 2019. Accelerated flow for probability distributions. In *International Conference on Machine Learning*, 6076–6085. PMLR.

Wang, Y.; and Li, W. 2022. Accelerated Information Gradient Flow. *Journal of Scientific Computing*, 90: 11.

Zhang, C.; Li, Z.; Du, X.; and Qian, H. 2022. DPVI: A Dynamic-Weight Particle-Based Variational Inference Framework. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 4900–4906. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zhang, H.; and Sra, S. 2018. An estimate sequence for geodesically convex optimization. In *Conference On Learning Theory*, 1703–1723. PMLR.

Zhu, M.; Liu, C.; and Zhu, J. 2020. Variance Reduction and Quasi-Newton for Particle-Based Variational Inference. In *ICML*, 11576–11587. PMLR.