- diffusion equations.
- Fokker-Planck equation
- SDE.
- §1.

SDE are used to model the evolution of random diffusion processes.

Time-evolving probability density p_t of these diffusion processes is governed by Fokker-Planck.

unnormalized posterior sampling via a discretization of the Langevin diffusion.

- [32] showed that the Fokker-Planck equation is equivalent to following the Gradient flow of an entropy functional in Wasserstein space. JKO scheme.
- one way: use a fixed discretization of the domain and apply standard numerical integration methods. limited to small dimensions since the discretization of spaces grow exponentially.
- alternative to domain discretization is particle simulation. the particles are approximately distributed according to the stationary Dist, but no density estimate is readily available.
- Another way to avoid discritization is to parameterize the density of p_t. [48] variational methods to minimize the divergence between the predictive and the true density.

- [24] introduced a parametric method to compute JKO steps via entropy-regularized optimal transport. The method yields unnormalized probability density without direct sample access.
- Recent works : ICNN : $\nabla\psi_\theta : \mathbb{R}^D \to \mathbb{R}^D$ too computationally expensive to be applied directly to JKO steps.
- This paper : [6]
  - scalable parametric method to approximate Wasserstein gradient flows via JKO stepping via ICNN.
    - leverage Brenier's Thm to bypass the costly computation of the Wasserstein distance, and parametrize the OT map as the gradient of an ICNN.
    - Given sample access to the initial measure ρ_0, use SGD to sequentially learn time-discretized JKO dynamics of ρ_t.
    - The trained model can sample from a continuous approximation of ρ_t and compute its density $\frac{d\rho_t}{dx}(x)$.
  - Compute gradient flows for the Fokker-Plank free energy functional F_{FP} given by (5). can be generalized to other cases.

(5) : $F_{FP}(\rho) = \mathcal{U}(\rho) - \beta^{-1}\mathcal{E}(\rho)$.

$\mathcal{U}(\rho) = \int_{\mathbb{R}^D} \mathcal{V}(x)\, d\rho(x)$ is the potential energy ; $\mathcal{V} : \mathbb{R}^D \to \mathbb{R}$ is the potential function.

$\beta > 0$ is the magnitude; $\mathcal{E}(\rho) = -\int_{\mathbb{R}^D} \log \frac{d\rho}{dx}(x)\, d\rho(x)$ is the entropy.

• §2. Wasserstein Gradient Flow.

W-2 Distance

• Def: [Wasserstein-2 distance]. The (squared) Wasserstein-2 metric W_2 between $\mu, \nu \in P_2(\mathbb{R}^D)$ is

$$W_2^2(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^D \times \mathbb{R}^D} \|x-y\|_2^2 \, d\pi(x,y).$$

where the minimum is over measures π on $\mathbb{R}^D \times \mathbb{R}^D$ with marginals μ and ν respectively [68].

• By Brenier's Thm [13], exist μ-unique map $\nabla\psi^*: \mathbb{R}^D \to \mathbb{R}^D$ which is gradient of a convex function $\psi^*: \mathbb{R}^D \to \mathbb{R} \cup \{+\infty\}$ satisfying $\nabla\psi^* \# \mu = \nu$ [46].

$$\Rightarrow W_2^2(\mu,\nu) = \int_{\mathbb{R}^D} \|x - \nabla\psi^*(x)\|_2^2 \, d\mu(x).$$

Wasserstein Gradient Flows.

• In Euclidean case, gradient flows along a function $f: \mathbb{R} \to \mathbb{R}$ follow the steepest descent direction and are defined through the ODE $\frac{dx_t}{dt} = -\nabla f(x_t)$.

• Discretization of this flow leads to the gradient descent minimization algorithm. When the functionals are defined over the space of measures equipped with W-2 metric, it is called Wasserstein-gradient flow.

• [4] for exposition of gradient flows in metric space or [59 Ch8] for intro.

• A curve of measures $\{\rho_t\}_{t\in\mathbb{R}_+}$ following the Wasserstein gradient flow of a functional F solves the continuity equation

$$\frac{\partial \rho_t}{\partial t} = \operatorname{div}\left(\rho_t \nabla_x F'(\rho_t)\right), \quad \text{s.t. } \rho_0 = \rho^0.$$

where $F'(\cdot)$ is the first variation of F [4, Thm 8.3.1]. The term on the right can be understood as the gradient of F in Wasserstein space, a vector field perturbatively rearranging the mass in ρ_t to yield the steepest possible local change of F.

• Many applications come from the connection between Wasserstein gradient flows and SDEs. Consider an $\mathbb{R}^D$-valued stochastic process $\{X_t\}_{t\in\mathbb{R}_+}$ governed by the following Itô SDE.

$$dX_t = -\nabla \Xi(X_t)\,dt + \sqrt{2\beta^{-1}}\,dW_t, \quad \text{s.t. } X_0 \sim \rho^0. \qquad \cdots\cdots (3)$$

where $\Xi : \mathbb{R}^D \to \mathbb{R}$ is the potential function, W_t is the standard Wiener process, and $\beta > 0$ is the magnitude. The solution of (3) is called an advection-diffusion process. The marginal measure ρ_t of X_t at each time satisfies the Fokker-Planck equation with fixed diffusion coefficient:

$$\frac{\partial \rho_t}{\partial t} = \operatorname{div}(\nabla \Xi(x)\,\rho_t) + \beta^{-1}\Delta \rho_t, \quad \text{s.t. } \rho_0 = \rho^0. \qquad \cdots\cdots (4).$$

(4) is the Wasserstein gradient flow (2) for F given by the Fokker-Plank free energy functional [32].

$$F_{FB}(\rho) = \mathcal{U}(\rho) - \beta^{-1}\mathcal{E}(\rho). \qquad \cdots\cdots (5).$$

$U(\rho) = \int_{\mathbb{R}^n} \Psi(x)\, d\rho(x)$ is the potential energy and $\mathcal{E}(\rho) = -\int_{\mathbb{R}^n} \log \frac{d\rho}{dx}(x)\, d\rho(x)$ is the entropy.

As a result, to solve the SDE (3), one may compute the Wasserstein gradient flow of the Fokker-Planck equation with the free-energy functional F_{FP} given by (5).

JKO Scheme

- Computing Wasserstein gradient flows is challenging. The closed form solution is typically unknown, necessitating numerical approximation techniques. JKO - approximation of the dynamics of ρ_t in (2) [32]. It consists of a time-discretization update of the continuous flow given by:

$$\rho^{(k)} \leftarrow \underset{\rho \in P_2(\mathbb{R}^n)}{\arg\min} \left[F(\rho) + \frac{1}{2h} \cdot W_2^2(\rho^{(k-1)}, \rho) \right].$$

where $\rho^{(0)} = \rho^0$ is the initial condition and $h > 0$ is the time-discretization step size. This discrete time gradient flow converges to the continuous ones as $h \to 0$. i.e. $\rho^{(k)} \approx \rho_{t=h}$. Method was further developed in [4, 60]. But performing JKO iterations remains challenging : the minimization w.r.t W_2.

- Common approach to perform JKO steps is to discretize the spatial domain. For support size $\lesssim 10^6$, (6) can be solve by standard OT algorithms [51]. In Dimensions $D \geq 3$, discrete supports can hardly approximate continuous Dist and hence the dynamics of gradient flows.

To tackle this issue, [24] propose a stochastic parametric method to approximate the density of ρ_ε.

Their method uses entropy-regularized OT, which is biased.

§5 Computing Wasserstein Gradient Flows with ICNN.

JKO Reformulation via Optimal Push-forwards Maps

• Key idea: replace the optimization (6) over probability measures by an optimization over convex function, an idea inspired by [11]. set $\rho = \nabla\psi \# \rho^{(k-1)}$. Convex function ψ.

$$\psi^{(k)} \leftarrow \underset{\text{Convex } \psi}{\arg\min} \left[F(\nabla\psi \# \rho^{(k-1)}) + \frac{1}{2h} W_2^2(\rho^{(k-1)}, \nabla\psi \# \rho^{(k-1)}) \right]. \quad \cdots\cdots (7).$$

To proceed to the next step of JKO scheme, we define $\rho^{(k)} = \nabla\psi^{(k)} \# \rho^{(k-1)}$.

Since ρ is the pushforward of $\rho^{(k-1)}$ by the gradient of a convex function $\nabla\psi$, the W_2^2 term in (7) can be evaluated explicitly, simplifying the W-2 distance term in (7):

$$\psi^{(k)} \leftarrow \underset{\text{convex } \psi}{\arg\min} \left[F(\nabla\psi \# \rho^{(k-1)}) + \frac{1}{2h} \int_{\mathbb{R}^0} \| x - \nabla\psi(x) \|_2^2 \, d\rho^{(k-1)}(x) \right]. \quad \cdots\cdots (8)$$

This formulation avoids the difficulty of computing W-2 distances. An additional advantage is that we can sample from $\rho^{(k)}$. Since $\rho^{(k)} = [\nabla\psi^{(k)} \circ \cdots \circ \nabla\psi^{(1)}] \# \rho^0$, one may sample $x_0 \sim \rho^{(0)}$, and then $\nabla\psi^{(k)} \circ \cdots \circ \nabla\psi^{(1)}(x_0)$ gives a sample from $\rho^{(k)}$.

• Moreover, if functions $\psi^{(\cdot)}$ are strictly convex, then gradients $\nabla\psi^{(\cdot)}$ are invertible.

In this case, the density $\frac{d\rho^{(k)}}{dx}$ of $\rho^{(k)} = \left[\nabla\psi^{(k)} \circ \cdots \circ \nabla\psi^{(1)}\right] \# \rho^0$ is computable by the change of variables formula (assuming $\psi^{(\cdot)}$ are twice differentiable).

density: $$\frac{d\rho^{(k)}}{dx}(x_k) = \left[\det \nabla^2\psi^{(k)}(x_{k-1})\right]^{-1} \cdots \left[\det \nabla^2\psi^{(1)}(x_0)\right]^{-1} \cdot \frac{d\rho^{(0)}}{dx}(x_0) \quad \cdots (9).$$

where $x_i = \nabla\psi^{(i)}(x_{i-1})$ for $i = 1, \ldots, k$ and $\frac{d\rho^{(0)}}{dx}$ is the density of $\rho^{(0)}$.

Stochastic Optimization for JKO via ICNNs.

In general, the solution $\psi^{(k)}$ of (8) is intractable since it requires optimization over all convex fun.

To tackle this issue, [11] discretizes the space of convex function. The approach also requires discretization of measures $\rho^{(k)}$ limiting this method to small Dims.

• ICNNs [6] satisfying a universal approximating property among convex fun [20].

• ICNNs are parametric models of the form $\psi_\theta : \mathbb{R}^D \to \mathbb{R}$ with ψ_θ convex w.r.t its input.

• The parameters are optimized via deep learning optimization techniques such as SGD.

The JKO step then becomes finding the optimal parameters θ^* for ψ_θ:

$$\theta^* \leftarrow \arg\min_\theta \left[F(\nabla\psi_\theta \# \rho^{(k-1)}) + \frac{1}{2h}\int_{\mathbb{R}^D} \|x - \nabla\psi_\theta(x)\|_2^2 \, d\rho^{(k-1)}(x) \right] \quad \cdots (10).$$

If the functional F can be estimated stochastically using random batches from $\rho^{(k-1)}$, then SGD can be used to optimize θ. F_{FP} given by (5) is an example of such a functional:

Thm 1 [Estimator of F_{FP}]. Let $\rho \in P_{2,ac}(\mathbb{R}^D)$ and $T : \mathbb{R}^D \to \mathbb{R}^D$ be a diffeomorphism. For a random batch $x_1, \dots, x_N \sim \rho$, the expression $\left[\hat{\mathcal{U}}_T(x_1, \dots, x_N) - \beta^{-1} \Delta \hat{\mathcal{E}}_T(x_1, \dots, x_N) \right]$, where

$$\hat{\mathcal{U}}_T(x_1, \dots, x_N) := \frac{1}{N} \sum_{n=1}^{N} \Phi(T(x_n)) \text{ and}$$

$$\Delta \hat{\mathcal{E}}_T(x_1, \dots, x_N) := \frac{1}{N} \sum_{n=1}^{N} \log \left| \det \nabla T(x_n) \right|,$$

is an estimator of $F_{FP}(T\#\rho)$ up to a constant (w.r.t T) shift given by $\beta^{-1}\mathcal{E}(\rho)$.

proof: $\hat{\mathcal{U}}_T$ is a straightforward unbiased estimator for $\mathcal{U}(T\#\rho)$.

Let p and p_T be the densities of ρ and $T\#\rho$. Since T is a diffeomorphism, we have

$p_T(y) = p(x) \cdot \left| \det \nabla T(x) \right|^{-1}$, where $x = T^{-1}(y)$. Using the change of variables formula, we have.

$$\mathcal{E}(\rho) = -\int_{\mathbb{R}^D} \log \frac{d\rho}{dx}(x) \, d\rho(x)$$

$$\mathcal{E}(T\#\rho) = -\int_{\mathbb{R}^D} p_T(y) \cdot \log p_T(y) \, dy$$

$$= -\int_{\mathbb{R}^D} p(x) \cdot \left| \det \nabla T(x) \right|^{-1} \cdot \log \left[p(x) \cdot \left| \det \nabla T(x) \right|^{-1} \right] \cdot \left| \det \nabla T(x) \right| dx$$

$$= -\int_{\mathbb{R}^D} p(x) \log p(x) \, dx + \int_{\mathbb{R}^D} p(x) \cdot \log \left| \det \nabla T(x) \right| dx$$

$$= \mathcal{E}(\rho) + \int_{\mathbb{R}^D} p(x) \log \left| \det \nabla T(x) \right| dx$$

$$\Rightarrow \Delta \mathcal{E}_T(\rho) := \mathcal{E}(T\#\rho) - \mathcal{E}(\rho) = \int_{\mathbb{R}^D} p(x) \log \left| \det \nabla T(x) \right| dx = \int_{\mathbb{R}^D} \log \left| \det \nabla T(x) \right| d\rho(x).$$

which explains that $\Delta\hat{\mathcal{E}}_T$ is an unbiased estimator of $\Delta\mathcal{E}_T(\rho)$.

Thus, $\hat{\mathcal{U}}_T - \beta^{-1}\Delta\hat{\mathcal{E}}_T$ is an estimator for $F_{FP}(\rho) = \mathcal{U}(T\#\rho) - \beta^{-1}\mathcal{E}(T\#\rho)$ up to a shift to $\beta^{-1}\mathcal{E}(\rho)$.

To apply Thm 1, take $T \leftarrow \nabla\psi_\theta$, $\rho \leftarrow \rho^{(k-1)}$. to obtain a stochastic estimator for $F_{FP}(\nabla\psi_\theta \# \rho^{(k-1)})$ in (10). Here $\beta^{-1}\mathcal{E}(\rho^{(k-1)})$ is θ-independent and constant since $\rho^{(k-1)}$ is fixed, so the offset of the estimator plays no role in the optimization w.r.t θ.

Alg 1: Stochastic JKO method for F_{FP}. The training is done solely based on random samples from the initial measure ρ^0: its density is not needed.

Alg 1: Fokker-Planck JKO via ICNNs

- Inputs:
  - Initial measure ρ^0 accessible by samples;
  - JKO discretization step $h > 0$, number of JKO steps $K > 0$.
  - target potential $\mathcal{D}(x)$, diffusion process temperature β^{-1}.
  - batch size N.
- Output. trained ICNN models $\{\psi^{(k)}\}_{k=1}^{K}$ representing JKO steps.

for $k = 1, 2, \ldots, K$ do:

  $\psi_\theta \leftarrow$ basic ICNN model;

  for $i = 1, 2, \ldots$ do:

    Sample batch $Z \sim \rho^0$ of size N;

    $X \leftarrow \nabla\psi^{(k-1)} \circ \cdots \circ \nabla\psi^{(1)}(Z)$

    $\hat{W}_2^2 \leftarrow \frac{1}{N} \sum_{x \in X} \|\nabla\psi_\theta(x) - x\|_2^2$

    $\hat{\mathcal{U}} \leftarrow \frac{1}{N} \sum_{x \in X} \mathcal{U}(\nabla\psi_\theta(x))$.

    $\hat{\Delta\mathcal{E}} \leftarrow \frac{1}{N} \sum_{x \in X} \log\det \nabla^2\psi_\theta(x)$

    $\hat{\mathcal{L}} \leftarrow \frac{1}{2h} \hat{W}_2^2 + \hat{\mathcal{U}} - \beta^{-1}\hat{\Delta\mathcal{E}}$.

    Perform a gradient step over θ by using $\frac{\partial \hat{\mathcal{L}}}{\partial \theta}$

  $\psi^{(k)} \leftarrow \psi_\theta$

This alg assumes F is the Fokker-Planck diffusion energy functional. can generalize to any F.

Computing the Density of Diffusion Process.

Our alg provides a computable density for $\rho^{(k)}$. sample from $\rho^{(k)}$ while simultaneously computing the density of the samples. evaluate $\frac{d\rho^{(k)}}{dx}(x_k)$ for any arbitrary $x_k \in \mathbb{R}^D$.

§ Potential function.

The Langevin Eqn: $\lambda \frac{dX_t}{dt} = -\frac{\partial V(x)}{\partial x} + \eta(t)$. ····· (1).

X_t is the position of a particle in a potential $V(x)$ and noise term $\eta(t)$.

(1) can be written as SDE

drift term diffusion term.

$$dX_t = -\nabla V(x)\,dt + \sqrt{2}\,dB_t \quad \cdots\cdots (2).$$

(2) has a unique steady-state distribution that does not change along the trajectory (X_t) of the particle. This means that if X_0 is distributed according to some probability density function p_∞, then X_t is also distributed according to p_∞ $\forall t \geq 0$. If we set potential V in (2) s.t. $p_\infty = \pi$.

- The steady-state distribution: choosing the potential.

The Fokker-Plank equation is a PDE that describes the evolution of a probability DisE over time under the effect of drift forces and random (or noise) fores.:

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}\left[\frac{\partial V(x)}{\partial x} p(x,t)\right] + \frac{\partial^2 p(x,t)}{\partial x^2} \quad \cdots\cdots (3).$$

The steady-state solution of the Fokker-Plank equation is given by $\frac{\partial p(x,t)}{\partial t} = 0$. If p_∞ is the steady-state Distribution, we have

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}\left[\frac{\partial V(x)}{\partial x} p_\infty(x) + \frac{\partial p_\infty(x)}{\partial x}\right] = \frac{\partial}{\partial x} J(x) = 0. \quad \cdots\cdots (4).$$

where $J(x)$ denotes the probability "flux". (4) implies that $J(x)$ must be a constant.

$P_\infty(x)$ and $\frac{\partial P_\infty(x)}{\partial x}$ must also satisfy certain boundary conditions.

boundary conditions: $\begin{cases} J(x) = \text{constant} \\ \frac{\partial P_\infty(x)}{\partial x} = 0 \text{ at } x = \pm\infty \end{cases}$. $P_\infty(x) = 0$ at $x = \pm\infty \Rightarrow J(x) = 0$ at $x = \pm\infty$.

$$\Rightarrow J(x) = \frac{\partial V(x)}{\partial x} P_\infty(x) + \frac{\partial P_\infty(x)}{\partial x} = 0. \quad \cdots (5).$$

$$\Rightarrow P_\infty(x) \propto \exp(-V(x)). \quad \cdots (6)$$

(6) represents a Gibbs distribution. We can sample from energy-based model of form.

$\pi(x) = \frac{\exp(-E(x))}{Z}$, let $E(x) = V(x)$. we can set $V(x) = -\log \pi(x)$.