Learning Equivariant Energy Based Models with Equivariant Stein Variational Gradient Descent

35th Conference on Neural Information Processing Systems (NeurIPS 2021)

Priyank Jaini*, Lars Holdijk*, Max Welling
University of Amsterdam

Reporter: Fanmeng Wang
March 24,2022

Outline

- Introduction
- Background
- Equivariant Stein Variational Gradient Descent(E-SVGD)
- Equivariant Joint Energy Model
- Experiments

- **Introduction**
- Background
- Equivariant Stein Variational Gradient Descent(E-SVGD)
- Equivariant Joint Energy Model
- Experiments

Incorporate symmetries in probabilistic models

- Many real-world observations comprise symmetries and admit probabilistic models that are invariant to such symmetry transformations.

- In this paper, we focus on the problem of efficient sampling and learning of equivariant probability densities by incorporating symmetries in probabilistic models.

- We propose ***Equivariant Stein Variational Gradient Descent*** algorithm for sampling from invariant densities and define ***Equivariant Energy Based Models*** to model invariant densities that are learned using contrastive divergence.

- Introduction
- **Background**
- Equivariant Stein Variational Gradient Descent(E-SVGD)
- Equivariant Joint Energy Model
- Experiments

Key definitions and notations

- Let G be a **group** acting on R^d through a representation R : $G \to GL(d)$ where $GL(d)$ is the general linear group on R^d, such that $\forall g \in G$, $g \to R_g$.
- A **density** π is G-*invariant* if $\forall g \in G$ and x $\in R^d$, $\pi(R_g x) = \pi(x)$.
- A **function** $f(\cdot)$ is G-*equivariant* if $\forall g \in G$ and x $\in R^d$, $f(R_g x) = R_g f(x)$.
- We denote with $O(x)$ the **orbit** of an element $x \in X$ defined as $O(x) := \{x' : x' = R_g x, \forall g \in G\}$
- We call $\pi_{|G}$ the **factorized density** of a G-*invariant* density π where $\pi_{|G}$ has support on the set $X_{|G}$ where the elements of $X_{|G}$ are indexing the orbits i.e. if x, $\tilde{x} \in X_{|G}$, then $x \neq R_g \tilde{x}$, $\forall g \in G$.

Equivariant Learning

Give access to an i.i.d. samples $\{x_1, \dots, x_n\} \sim \pi$ from a G-invariant density π, we want to approximate π.

- **Method:**
  Learning an equivariant normalizing flow that transforms **a simple latent G-invariant density** q_0 to **the target density** π through a series of G-equivariant diffeomorphic transformations $\mathbf{T} = (\boldsymbol{T}_1, \boldsymbol{T}_2, \dots, \boldsymbol{T}_k)$ i.e. $\pi := \boldsymbol{T}_{\#q_0}$.

- **Drawback:**
  It requires **T** to not only be a F-equivariant diffeomorphism(F is a proper sub-group of G), but computation of the inverse and Jacobian must be cheap as well.

- **Solution:**
  Using continuous normalizing flows that define a dynamical system through a time-dependent Lipschitz velocity field Ψ: $R^d \times R_+ \to R^d$ with the following system of ordinary differential equations(ODEs):

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \Psi(\boldsymbol{x}(t), t), \qquad \boldsymbol{x}(0) = \boldsymbol{z} \tag{1}$$

Stein Variational Gradient Descent (SVGD)

A particle optimization variational inference method that combines the paradigms of sampling and variational inference for Bayesian inference problems.

- This is achieved in a series of T discrete steps that transform the set of particles $\{x_i^0\}_{i=1}^n \sim q_0(x)$ sampled from a base distribution q_0 (e.g. Gaussian) at t = 0 using the map:

$$x^t = \boldsymbol{T}(x) \coloneqq x^{t-1} + \varepsilon \cdot \Psi(x^{t-1})$$

where ε is the step size and $\Psi(\cdot)$ is a vector field. $\Psi(\cdot)$ is chosen such that it maximally decreases the KL divergence between the push-forward density $q_t(x) = \boldsymbol{T}_{\#q_{t-1}(x)}$ and the target π(x).

Stein Variational Gradient Descent (SVGD)

- If Ψ is restricted to the unit ball of an RKHS H_k^d with positive definite kernel k: $R^d \times R^d \to R$: the direction of steepest descent that maximizes the negative gradient of the KL divergence is given by:

$$\Psi^*_{q,\pi}(x) := \arg\max_{\Psi \in \mathcal{H}_k^d} -\nabla_\varepsilon \mathrm{KL}(q||\pi)|_{\varepsilon \to 0} = \mathbb{E}_{x \sim q}[\mathrm{trace}(\mathcal{A}_\pi \Psi(x))], \qquad (2)$$

- Replacing the expectation in the update with a Monte Carlo sum over the current set of particles that represent q we get:

$$x_i^{t+1} \leftarrow x_i^t + \varepsilon \tilde{\Psi}^*(x_i^t), \text{ where, } \tilde{\Psi}^*(x_i^t) := \frac{1}{n} \sum_{j=1}^{n} \Big(\underbrace{\nabla_{x_j^t} k(x_j^t, x_i)}_{\text{repulsive force}} - \underbrace{k(x_j^t, x_i) \cdot \nabla_{x_j^t} \mathrm{E}(x_j^t)}_{\text{attractive force}} \Big) \quad (3)$$

A combination of repulsive force and attractive force in Equation(3) can be used to avoid falling into local optimality.

- Furthermore, geometric information using pre-conditioning matrices can be incorporated in Equation (3) by using **matrix valued kernels** leading to the following **generalized form** of SVGD :

$$x_i^{t+1} \leftarrow x_i^t + \frac{\varepsilon}{n} \sum_{j=1}^{n} \left(\nabla_{x_j^t} \boldsymbol{K}(x_j^t, x_i) - \boldsymbol{K}(x_j^t, x_i) \cdot \nabla_{x_j^t} \mathrm{E}(x_j^t) \right), \qquad (4)$$

- Introduction
- Background
- **Equivariant Stein Variational Gradient Descent(E-SVGD)**
- Equivariant Joint Energy Model
- Experiments

Equivariant Stein Variational Gradient Descent (E-SVGD)

We call the updates in Equations (3) & (4) ***equivariant Stein variational gradient descent*** when the kernel k(;) (and K(;) respectively) is invariant (equivariant) and the initial set of particles $\{x_1^0, \dots, x_n^0\}$ are sampled from an invariant density q_0.

- **Proof:**

$$x_i^{t+1} \leftarrow x_i^t + \varepsilon \tilde{\Psi}^*(x_i^t), \text{ where, } \tilde{\Psi}^*(x_i^t) := \frac{1}{n}\sum_{j=1}^{n} \Big(\underbrace{\nabla_{x_j^t} k(x_j^t, x_i)}_{\text{repulsive force}} - \underbrace{k(x_j^t, x_i) \cdot \nabla_{x_j^t} \mathrm{E}(x_j^t)}_{\text{attractive force}} \Big) \quad (3)$$

  Since the initial distribution q_0 is F-invariant, the update in Equation (3) is G-equivariant if Ψ is G-equivariant.

  If k(·,·) is G-invariant then $\nabla_x k(\cdot, x)$ is G-equivariant. Since π = exp(-E(x)) is G-invariant, ∇_xE(x) is also G-equivariant.

  Thus, both the terms for Ψ are G-equivariant if k(·,·) is G-equivariant making the update in Equation (3) G-equivariant.

  The result follows similarly for Equation (4) when K(·,·) is G-equivariant.

- Therefore, all that is required to sample from a G-invariant density using equivariant SVGD is to construct a **positive definite kernel** that is G-equivariant.

Equivariant Stein Variational Gradient Descent (E-SVGD)

- **Advantages:**

Equivariant SVGD, is able to model long-range interactions among particles due to the use of equivariant kernel. Intuitively, any point x is able to exert these forces on any other point x' in equivariant SVGD if x' is in the neighborhood of any point in the orbit O(x) of x.

This ability to capture long-range interactions by equivariant Stein variational gradient descent subsequently makes it more efficient in sample complexity and running time with better sample quality, and makes it more robust to different initial configurations of the particles compared to SVGD.

- Introduction
- Background
- Equivariant Stein Variational Gradient Descent(E-SVGD)
- **Equivariant Joint Energy Model**
- Experiments

Energy Based Models (EBMs)

- Given a set of samples $\{x_1, \ldots, x_n\}$, **energy-based models (EBMs)** learn an energy function $E_\theta(x)$: $R^d \rightarrow R$ that defines a probability distribution

$$\tilde{\pi}_\theta(\mathrm{x}) = \exp(-E_\theta(x)) / Z_\theta$$

where $Z_\theta = \int \exp(-E_\theta(x))\, dx$ is the partition function.

- **EBMs** are usually trained by maximizing the log-likelihood of the data under the given model:

$$\theta^* := \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{ML}}(\theta) = \mathbb{E}_{\boldsymbol{x}\sim\pi}\left[-\log \tilde{\pi}_{\boldsymbol{\theta}}(x)\right] \qquad (5)$$

- Contrastive divergence provides a paradigm to learn **EBMs** using maximum likelihood estimation without needing to compute Z_θ by approximating the gradient of $\nabla_\theta L_{ML}(\theta)$ in Equation (5) as follows:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{ML}}(\theta) \approx \mathbb{E}_{\boldsymbol{x}^+\sim\pi}\left[\nabla_{\boldsymbol{\theta}}\mathrm{E}_{\boldsymbol{\theta}}(x^+)\right] - \mathbb{E}_{\boldsymbol{x}^-\sim\tilde{\pi}_{\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{\theta}}\mathrm{E}_{\boldsymbol{\theta}}(x^-)\right] \qquad (6)$$

Intuitively, the gradient in Equation (6) drives the model such that it assigns higher energy to the negative samples x^- sampled from the current model and decreases the energy of the positive samples x^+ which are the data-points from the target distribution.

Equivariant Energy Based Models

Equivariant energy based models that are trained contrastively using our proposed equivariant Stein variational gradient descent algorithm to learn invariant (unnormalized) densities given access to i.i.d. samples $\{x_1, \ldots, x_n\} \sim \pi$.

- Since, training an **EBMs** using MLE requires sampling from the current model $\tilde{\pi}_\theta(\mathrm{x})$, successful training of **EBMs** relies heavily on sampling strategies that lead to faster mixing.

- Fortuitously, since $E_\theta(.)$ in our present setting is G-equivariant, we propose to use our equivariant Stein variational gradient sampler for more efficient training of the **equivariant energy based model**.

Algorithm 1: Equivariant EBM training

Input: $\{x_1^+, x_2^+, \cdots, x_m^+\} \sim \pi(x)$
while *not converged* **do**
  ▷ *Generate samples from current eqNN model* $\mathrm{E}_{\boldsymbol{\theta}}$
  $\{x_1^-, x_2^-, \cdots, x_m^-\} = \mathsf{EquivariantSVGD}(\mathrm{E}_{\boldsymbol{\theta}})$;
  ▷ *Optimize objective* $\mathcal{L}_{\mathsf{ML}}(\theta)$:
  $\Delta\boldsymbol{\theta} \leftarrow \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{\theta}}(x_i^+) - \nabla_{\boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{\theta}}(x_i^-)$;
  ▷ *Update* $\boldsymbol{\theta}$ *using* $\Delta\boldsymbol{\theta}$ *and Adam optimizer*
end

Equivariant Joint Energy Model

We can extend equivariant energy based models to equivariant joint energy models.

- Let $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ be a set of samples with observations x_i and labels y_i. Given a parametric function $f_\theta : R^d \to R^k$, a classifier uses the conditional distribution $\tilde{\pi}_\theta(y \mid x) \propto \exp(f_\theta(x)[y])$ where $f_\theta(x)[y]$ is the logit corresponding to the y^{th} class label.

- These logits can be used to define the joint density $\tilde{\pi}_\theta(x , y)$ and marginal density $\tilde{\pi}_\theta(x)$ as follows:

$$\tilde{\pi}_{\boldsymbol{\theta}}(x, y) = \frac{\exp\left(f_{\boldsymbol{\theta}}(x)[y]\right)}{Z_{\boldsymbol{\theta}}}, \quad \text{and,} \quad \tilde{\pi}_{\boldsymbol{\theta}}(x) = \frac{\sum_y \exp\left(f_{\boldsymbol{\theta}}(x)[y]\right)}{Z_{\boldsymbol{\theta}}} \tag{7}$$

- We can train this model by maximizing the log-likelihood of the joint distribution as follows:

$$\mathcal{L}(\theta) := \mathcal{L}_{\mathsf{ML}}(\theta) + \mathcal{L}_{\mathsf{SL}}(\theta) = \log \tilde{\pi}_{\boldsymbol{\theta}}(x) + \log \tilde{\pi}_{\boldsymbol{\theta}}(y|x) \tag{8}$$

where $L_{SL}(\theta)$ is the supervised loss which is the cross-entropy loss in the case of classification

- Introduction
- Background
- Equivariant Stein Variational Gradient Descent(E-SVGD)
- Equivariant Joint Energy Model
- **Experiments**

Experiments

- Reconstruct potential function describing a many-body particle system (DW-4) trained using limited number of meta-stable States.
  - In this many-body particles system, a double-well potential describes the configuration of four particles that is invariant to rotations, translations and, permutation of the particles.
  - In our experiment, we show that given access to only a single example of each metastable state configuration, an **equivariant EBMs** trained with E-SVGD can recover other states with similar energy as those of in the training set.

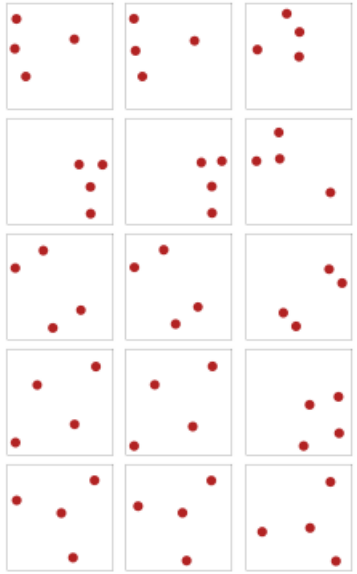

Figure 7: *Col. 1*: Samples from true potential energy. *Col. 2*: Samples from EBM trained with SVGD. *Col. 3*: Samples from equivariant EBM trained with E-SVGD.

Experiments

- Hybrid (generative & discriminative) model invariant to rotations for FashionMNIST trained using dataset with no rotations.

  - In this experiment, we take the FashionMNIST dataset with training set consisting regular images whereas the test set is processed to contain images that are randomly rotated using the C4-symmetry group.

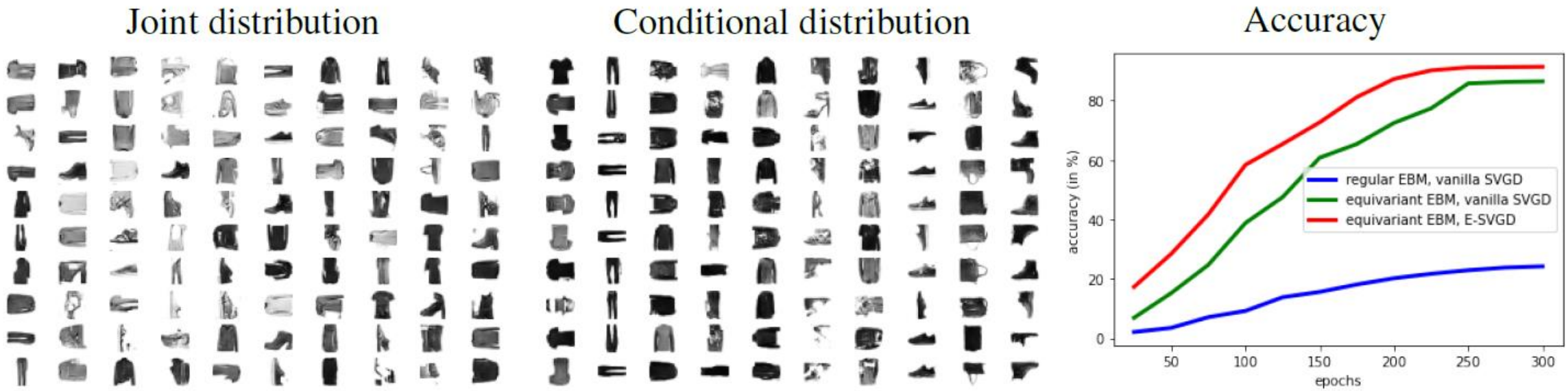

Figure 6: *Left & Center:* Samples generated from joint and class-conditional distribution using equivariant EBM. *Right:* Plot of classification accuracy vs. training iterations for equivariant and regular EBMs trained using vanilla SVGD and E-SVGD.

Thanks!