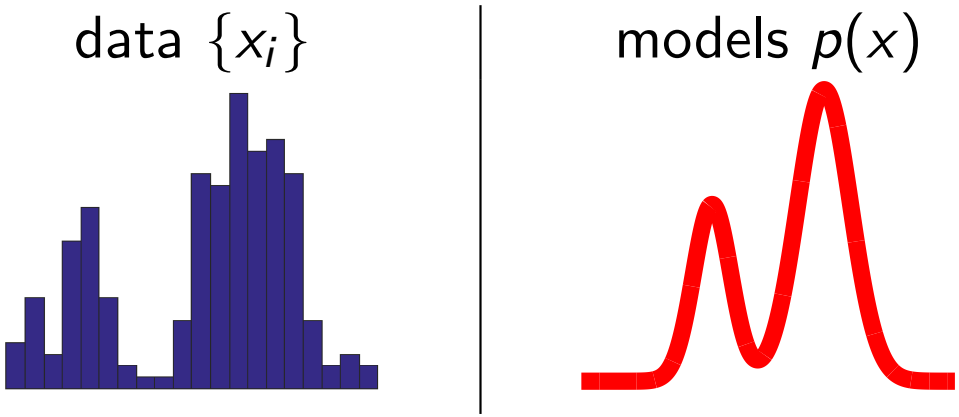
Probabilistic Learning and Inference Using Stein Discrepancy

Qiang Liu
Dartmouth College

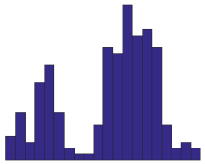
December 24, 2016

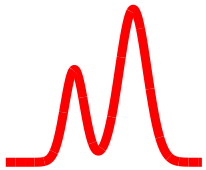
Machine Learning and Statistics



Data-Model Discrepancy

$$\mathbb{D}\Big(\underbrace{\quad}_{\text{data } \{x_i\}_{i=1}^n} \;,\; \underbrace{\quad}_{\text{model } p} \Big)$$





- **Learning:** Given $\{x_i\}$, find an optimal p:

$$\min_{p} \mathbb{D}(\{x_i\}, \quad p).$$

- **Sampling (or numerical quadrature)**: Given p, find optimal $\{x_i\}$:

$$\min_{\{x_i\}} \mathbb{D}(\{x_i\}, \quad p).$$

- **Model checking (e.g., goodness of fit test)**: Given both p and $\{x_i\}$, tell if they are consistent:

$$\mathbb{D}(\{x_i\}, \quad p) \stackrel{?}{=} 0.$$

Unnormalized Distributions

- In practice, many distributions are unnormalized densities:

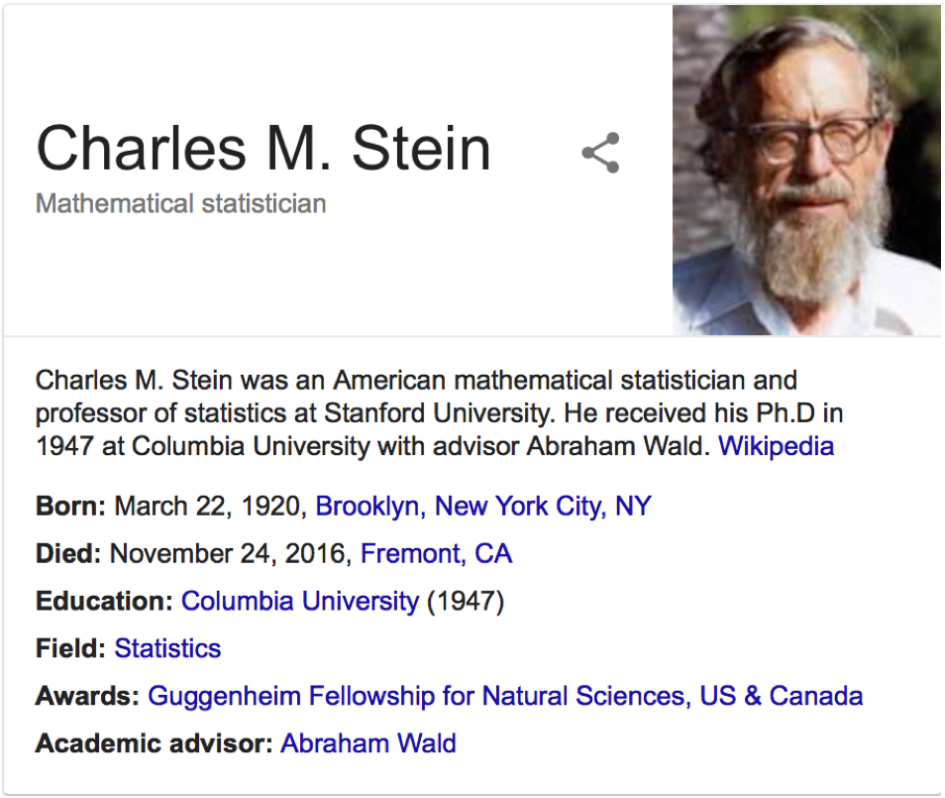
$$p(x) = \frac{1}{Z}\bar{p}(x), \qquad Z = \int \bar{p}(x)dx.$$

  Z: often critically difficult to calculate.

- Widely appears in Bayesian inference, (deep) probabilistic graphical models, energy-based models, etc.
- Highly difficult to learn and sample and evaluate.
  - Traditional methods: KL divergence + MCMC / variational inference, etc. Many drawbacks.

Stein's Method [Stein, 1972]

- A set of theoretical technique for proving approximation and limit theorems in probability theory.
  - central limit theorem, Berry-Esseen bounds, concentration inequalities, etc.
- Often remarkably powerful. A large body of theoretical work.



More ...

Stein's Method

- The key idea (that we will exploit): Characterizing a distribution p with a Stein operator $\mathcal{T}_p$, such that

$$p = q \quad \Longleftrightarrow \quad \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] = 0.$$

- For continuous distributions with differentiable density $p(x)$,

$$\mathcal{T}_p \phi(x) \stackrel{def}{=} \langle \nabla_x \log p(x), \quad \phi(x) \rangle + \nabla_x \cdot \phi(x).^1$$

[1] $\nabla_x \cdot \phi = \sum_i \partial_{x_i} \phi$

Stein's Method

- The key idea (that we will exploit): Characterizing a distribution p with a Stein operator $\mathcal{T}_p$, such that

$$p = q \quad \Longleftrightarrow \quad \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] = 0.$$

- For continuous distributions with differentiable density $p(x)$,

$$\mathcal{T}_p \phi(x) \stackrel{def}{=} \langle \nabla_x \log p(x), \quad \phi(x) \rangle + \nabla_x \cdot \phi(x).^1$$

[1] $\nabla_x \cdot \phi = \sum_i \partial_{x_i} \phi$

Stein's Method

- The key idea (that we will exploit): Characterizing a distribution p with a Stein operator $\mathcal{T}_p$, such that

$$p = q \quad \Longleftrightarrow \quad \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] = 0.$$

- For continuous distributions with differentiable density $p(x)$,

$$\mathcal{T}_p \phi(x) \overset{def}{=} \langle \underbrace{\nabla_x \log p(x)}_{\text{score function}}, \quad \phi(x) \rangle + \nabla_x \cdot \phi(x).^1$$

  - Score function $\boldsymbol{s}_p(x) = \nabla_x \log p(x) = \frac{\nabla_x p(x)}{p(x)}$, independent of normalization constant Z!
  - General methods for constructing Stein operators: *the generator method, density method*, etc.

[1] $\nabla_x \cdot \phi = \sum_i \partial_{x_i} \phi$

Stein's Method

$p = q$, Stein's Identity : $\mathbb{E}_{x\sim p}[\langle \nabla_x \log p(x), \quad \phi(x)\rangle + \nabla_x \cdot \phi(x)] = 0$:

Why?

- Use integration by parts, assuming zero boundary conditions.

$$\int p(x)\nabla_x \phi(x) + \phi(x)\nabla_x p(x) dx = p(x)\phi(x)\Big|_{-\infty}^{+\infty} = 0.$$

Stein's Method

$p = q$, Stein's Identity : $\mathbb{E}_{x\sim p}[\langle \nabla_x \log p(x), \quad \phi(x)\rangle + \nabla_x \cdot \phi(x)] = 0$:

Stein's identity: an infinite number of identities (moment equations), indexed by testing function ϕ. Lots of applications:

- Learning probabilistic models from data
  - Score matching [Hyvärinen, 2005, Lyu, 2009, Sriperumbudur et al., 2013]
  - Spectrum methods [Sedghi and Anandkumar, 2014]
- Variance reduction [Oates et al., 2014, 2016, 2017]
- Feature learning [Janzamin et al., 2014]
- Optimization [Erdogdu, 2015]
- and many more ...

Stein's Method

$p \neq q \quad \Rightarrow \quad \exists$ some ϕ, $\mathbb{E}_{x\sim q}[\mathcal{T}_p\phi(x)] \neq 0$:

Why (method I)?

- We can show (denote $s_p(x) = \nabla_x \log p(x)$):

$$\mathbb{E}_{x\sim q}[\mathcal{T}_p\phi(x)] = \mathbb{E}_{x\sim q}[\mathcal{T}_p\phi(x)] - \mathbb{E}_{x\sim p}[\mathcal{T}_p\phi(x)]$$

Stein's Method

$p \neq q \quad \Rightarrow \quad \exists$ some ϕ, $\mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] \neq 0$:

Why (method I)?

- We can show (denote $\boldsymbol{s}_p(x) = \nabla_x \log p(x)$):

$$\begin{aligned}\mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] &= \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] - \mathbb{E}_{x\sim q}[\mathcal{T}_q \phi(x)] \\ &= \mathbb{E}_{x\sim q}[\langle \boldsymbol{s}_p(x) - \boldsymbol{s}_q(x), \quad \phi(x)\rangle]\end{aligned}$$

  - Stein operator: essentially the inner product with the difference of score functions $\boldsymbol{s}_p - \boldsymbol{s}_q$.

  - Unless $\nabla_x \log p(x) \equiv \nabla_x \log q(x)$, we can always find a $\phi(x)$ to get non-zero.

Stein's Method

$p \neq q \quad \Rightarrow \quad \exists$ some ϕ, $\mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] \neq 0$:

Why (method II)?

- Let $x \sim q$ and $q_{[\epsilon\phi]}$ the density of $x' = x + \epsilon\phi(x)$, then

$$\frac{\partial}{\partial \epsilon} \mathrm{KL}(q_{[\epsilon\phi]} \mid\mid p)\big|_{\epsilon=0} = -\mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)].$$

Equals zero only at the stationary points of KL divergence (i.e., $p = q$).

Stein Discrepancy

$$p \neq q \quad \Longrightarrow \quad \exists \phi, \text{ such that } \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] \neq 0$$

- Define Stein discrepancy between p and q:

$$\mathbb{D}(q, \ p) = \max_{\phi \in \mathcal{F}} \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)]$$

  $\mathcal{F}$: a rich enough set of functions.

- It gives a functional optimization. Traditional Stein's method takes $\mathcal{F}$ to be sets of functions with bounded Lipschitz norm; computationally difficult for practical use.
- Gorham and Mackey [2015]: Derived a computable Stein discrepancy by enforcing Lipschitz constraints on a finite number of points, solved by linear programming.

Stein Discrepancy

$$p \neq q \quad \Longrightarrow \quad \exists \phi, \text{ such that } \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] \neq 0$$

- Define Stein discrepancy between p and q:

$$\mathbb{D}(q, \ p) = \max_{\phi \in \mathcal{F}} \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)]$$

  $\mathcal{F}$: a rich enough set of functions.

- It gives a functional optimization. Traditional Stein's method takes $\mathcal{F}$ to be sets of functions with bounded Lipschitz norm; computationally difficult for practical use.
- Gorham and Mackey [2015]: Derived a computable Stein discrepancy by enforcing Lipschitz constraints on a finite number of points, solved by linear programming.

Kernelized Stein discrepancy [Liu, Lee, Jordan. 2016]

- Let $k(x, x')$ be a positive definite kernel, and $\mathcal{H}_0$ its related reproducing kernel Hilbert space (RKHS). $\mathcal{H} = \mathcal{H}_0 \times \cdots \times \mathcal{H}_0$.
- Kernelized Stein discrepancy (KSD): take $\mathcal{F}$ to be the unit ball of RKHS.

$$\mathbb{D}(q,\ p) = \max_{\phi \in \mathcal{F}} \mathbb{E}_{x \sim q}[\mathcal{T}_p \phi(x)], \qquad \mathcal{F} = \{\phi \in \mathcal{H} \colon ||\phi||_{\mathcal{H}} \leq 1\}$$

Kernelized Stein discrepancy [Liu, Lee, Jordan. 2016]

- Let $k(x, x')$ be a positive definite kernel, and $\mathcal{H}_0$ its related reproducing kernel Hilbert space (RKHS). $\mathcal{H} = \mathcal{H}_0 \times \cdots \times \mathcal{H}_0$.
- Kernelized Stein discrepancy (KSD): take $\mathcal{F}$ to be the unit ball of RKHS.

$$\mathbb{D}(q,\ p) = \max_{\phi \in \mathcal{F}} \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)], \qquad \mathcal{F} = \{\phi \in \mathcal{H} \colon ||\phi||_{\mathcal{H}} \leq 1\}$$

then it has a closed form solution

$$\mathbb{D}(q,\ p)^2 = \mathbb{E}_{x,x'\sim q}[\kappa_p(x, x')]$$

where $\kappa_p(x, x') = \mathcal{T}_p^x(\mathcal{T}_p^{x'} \otimes k(x, x'))$

$$= s_p(x)^\top k(x, x') s_p(x') + s_p(x)^\top \nabla_{x'} k(x, x') + \nabla_x k(x, x')^\top s_p(x') + \Delta k(x, x')$$

where $\mathcal{T}_p^x$ is Stein operator w.r.t. x.

Kernelized Stein discrepancy [Liu, Lee, Jordan. 2016]

- Let $k(x, x')$ be a positive definite kernel, and $\mathcal{H}_0$ its related reproducing kernel Hilbert space (RKHS). $\mathcal{H} = \mathcal{H}_0 \times \cdots \times \mathcal{H}_0$.
- Kernelized Stein discrepancy (KSD): take $\mathcal{F}$ to be the unit ball of RKHS.

$$\mathbb{D}(q,\ p) = \max_{\phi \in \mathcal{F}} \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)], \qquad \mathcal{F} = \{\phi \in \mathcal{H} \colon ||\phi||_{\mathcal{H}} \leq 1\}$$

then it has a closed form solution

$$\mathbb{D}(q,\ p)^2 = \mathbb{E}_{x,x'\sim q}[\kappa_p(x, x')]$$

where $\kappa_p(x, x') = \mathcal{T}_p^x(\mathcal{T}_p^{x'} \otimes k(x, x'))$

$$= s_p(x)^\top k(x, x') s_p(x') + s_p(x)^\top \nabla_{x'} k(x, x') + \nabla_x k(x, x')^\top s_p(x') + \Delta k(x, x')$$

where $\mathcal{T}_p^x$ is Stein operator w.r.t. x. **Key: Stein operator is linear.**

Kernelized Stein discrepancy [Liu, Lee, Jordan. 2016]

- Let $k(x, x')$ be a positive definite kernel, and $\mathcal{H}_0$ its related reproducing kernel Hilbert space (RKHS). $\mathcal{H} = \mathcal{H}_0 \times \cdots \times \mathcal{H}_0$.
- Kernelized Stein discrepancy (KSD): take $\mathcal{F}$ to be the unit ball of RKHS.

$$\mathbb{D}(q,\ p) = \max_{\phi \in \mathcal{F}} \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)], \qquad \mathcal{F} = \{\phi \in \mathcal{H} \colon ||\phi||_{\mathcal{H}} \leq 1\}$$

then it has a closed form solution

$$\mathbb{D}(q,\ p)^2 = \mathbb{E}_{x,x'\sim q}[\kappa_p(x, x')] \underbrace{\approx \frac{1}{\mathbf{n(n-1)}} \sum_{\mathbf{i \neq j}} \kappa_{\mathbf{p}}(\mathbf{x_i}, \mathbf{x_j})}_{\text{empirical estimation (U-statistic)}}$$

where $\kappa_p(x, x') = \mathcal{T}_p^x(\mathcal{T}_p^{x'} \otimes k(x, x'))$

$$= s_p(x)^\top k(x, x') s_p(x') + s_p(x)^\top \nabla_{x'} k(x, x') + \nabla_x k(x, x')^\top s_p(x') + \Delta k(x, x')$$

where $\mathcal{T}_p^x$ is Stein operator w.r.t. x.

Empirical Kernelized Stein Discrepancy

- Given $\{x_i\}$ drawn from (unknown) $q(x)$, the U-statistic provides unbiased estimator of $\mathbb{D}(q,\ p)^2$:

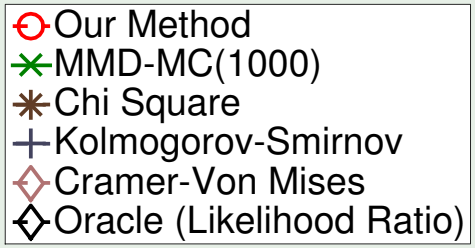
$$\mathbb{D}^2(\{x_i\},\ p) \stackrel{def}{=} \frac{1}{n(n-1)} \sum_{i \neq j} \kappa_p(x_i, x_j).$$

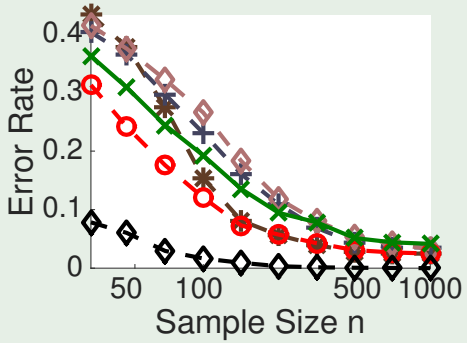
- Asymptotic distribution is well understood:
  - If $p \neq q$, $\mathbb{D}^2(\{x_i\},\ p) = \mathbb{D}^2(q,\ p) + O_p(1/\sqrt{n})$ (asymptotic normal)
  - If $p = q$, $\mathbb{D}^2(\{x_i\},\ p) = O_p(1/n)$. (infinite sum of χ^2 distributions)
- Goodness-of-fit test: test if $\{x_i\}$ is drawn from p.
  - Reject the null if $\mathbb{D}^2(\{x_i\},\ p) > \gamma$.
  - Threshold γ decided using a generalized bootstrap procedure by Arcones and Gine [1992], Huskova and Janssen [1993].

Goodness-of-fit Test

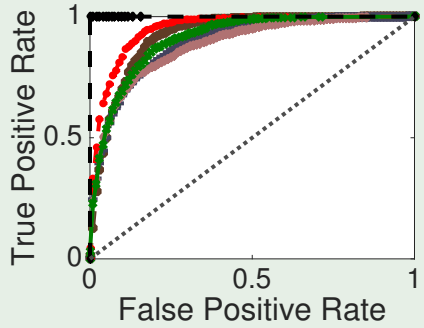
1D Gaussian mixture model (GMM)

- Simulate samples either from the true, or the perturbed model with equal probabilities. Use GOF tests to tell if the sample is drawn from the true model.





(significance $\alpha = 0.05$)

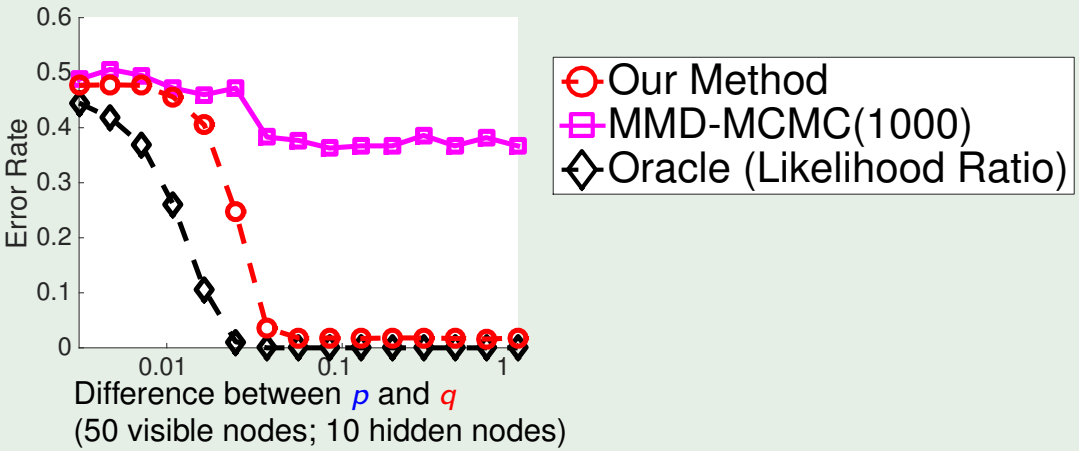


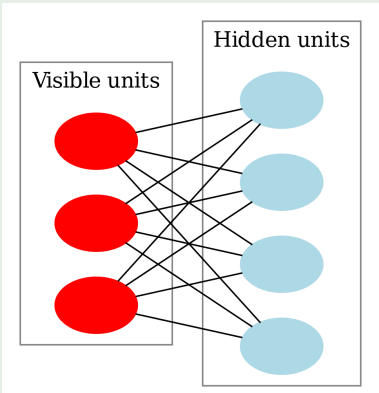
(Fixed $n = 100$)

Goodness-of-fit Test

Gaussian-Bernoulli Restricted Boltzmann Machine

- Gaussian visible nodes + binary hidden nodes; effectively a Gaussian mixture with exponential number of mixture components.





Connection with Other Discrepancy Measures

- Maximum mean discrepancy (MMD):

$$\mathbb{M}(q,\ p) = \max_{f \in \mathcal{H}_0} \{\mathbb{E}_p f - \mathbb{E}_q f \qquad s.t. \qquad ||f||_{\mathcal{H}_0} \leq 1\}.$$

$\mathcal{H}_0$ is the RKHS related to $k(x, x')$.

- KSD can be treated as a MMD using the "Steinalized" kernel $\kappa_p(x, x') = \mathcal{T}_p^x(\mathcal{T}_p^{x'} \otimes k(x, x'))$, which depends on p (KSD is asymmetric):

$$\mathbb{D}(q,\ p) = \max_{f \in \mathcal{H}_p} \{\mathbb{E}_p f - \mathbb{E}_q f \qquad s.t. \qquad ||f||_{\mathcal{H}_p} \leq 1\}$$

  - $\mathcal{H}_p$ is the RKHS of $k_p(x, x')$.
  - $\mathcal{H}_p$ is the image of Stein operator $\mathcal{T}_p\phi$: $\mathcal{H}_p = \{f = \mathcal{T}_p\phi \colon \phi \in \mathcal{H}\}$.
  - $\mathcal{H}_p$ is the "tangent space" of p: $\mathbb{E}_p[f] = 0,\ \forall f \in \mathcal{H}_p$ (will discuss more).

Connection with Other Discrepancy Measures

Fisher divergence

- Fisher divergence: $\mathbb{F}(q, p) = \mathbb{E}_{x \sim q}[||\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)||_2^2]$.
- Used as a learning objective in score matching.
- KSD is a smoothed version of Fisher divergence; we can show

$$\mathbb{D}(q, p) = \mathbb{E}_{x \sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x, x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))].$$

Connection with Other Discrepancy Measures

Fisher divergence

- Fisher divergence: $\mathbb{F}(q, p) = \mathbb{E}_{x\sim q}[||\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)||_2^2]$.
- Used as a learning objective in score matching.
- KSD is a smoothed version of Fisher divergence; we can show

$$\mathbb{D}(q, p) = \mathbb{E}_{x\sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x, x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))].$$

KL Divergence

- Fisher divergence = derivative of KL when variables are perturbed by i.i.d. Gaussian (debruijn identity).
- KSD = derivative of KL when variables are perturbed by smooth functions in RKHS.

Related Work on Stein + RKHS

- Chwialkowski et al. [2016]: Independent work on quite the same idea.
- Oates et al. [2014, 2016, 2017]: Combined Stein's identity with RKHS; used for deriving a super-efficient variance reduction method.

- **Numerical Quadrature**: Given p, find points $\{x_i\}$ to "fool" the goodness-of-fit test:

$$\min_{\{x_i\}} \sum_{ij} \kappa_p(x_i, x_j).$$

- **Numerical Quadrature**: Given p, find points $\{x_i\}$ to "fool" the goodness-of-fit test:

$$\min_{\{x_i\}} \sum_{ij} \kappa_p(x_i, x_j).$$

- Unfortunately, does not work well in practice: Difficult non-convex optimization.

Consider a simpler problem:

- Given $\{x_i\}$ generated arbitrarily (e.g., by MCMC or bootstrap).
- Find weights $\{w_i\}$ so that $\{w_i,\ x_i\}$ approximates p in that $\sum_i w_i h(x_i) \approx \mathbb{E}_{x\sim p}[h(x)]$.

Consider a simpler problem:

- Given $\{x_i\}$ generated arbitrarily (e.g., by MCMC or bootstrap).
- Find weights $\{w_i\}$ so that $\{w_i,\ x_i\}$ approximates p in that $\sum_i w_i h(x_i) \approx \mathbb{E}_{x\sim p}[h(x)]$.
- Minimizing the empirical kernelized Stein discrepancy:

$$\min_{\{w_i\}} \Big\{\mathbb{D}(\{w_i, x_i\};\ p) \equiv \sum_{ij} w_i w_j \kappa_p(x_i, x_j) \quad s.t. \quad \sum_i w_i = 1,\ \ w_i \geq 0\Big\}.$$

- This is easy to solve: convex quadratic programming.

Consider a simpler problem:

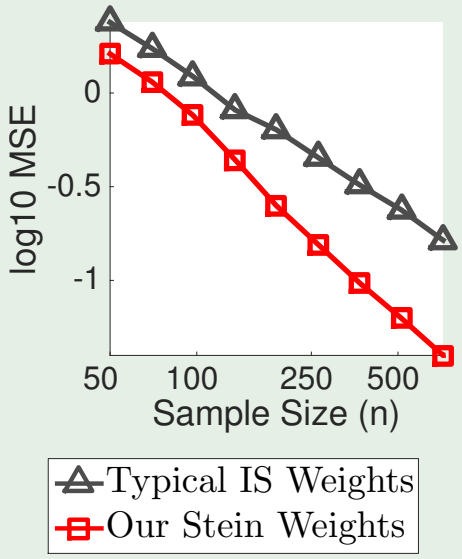
- Given $\{x_i\}$ generated arbitrarily (e.g., by MCMC or bootstrap).
- Find weights $\{w_i\}$ so that $\{w_i,\ x_i\}$ approximates p in that $\sum_i w_i h(x_i) \approx \mathbb{E}_{x\sim p}[h(x)]$.
- Minimizing the empirical kernelized Stein discrepancy:

$$\min_{\{w_i\}} \Big\{ \mathbb{D}(\{w_i, x_i\};\ p) \equiv \sum_{ij} w_i w_j \kappa_p(x_i, x_j) \quad s.t. \quad \sum_i w_i = 1,\ \ w_i \geq 0 \Big\}.$$

- Better convergence rate than the typical $O(n^{-1/2})$ Monte Carlo rate:

$$\sum_i w_i h(x_i) - \mathbb{E}_p h = O(n^{-\alpha/2}),\ \ 1 < \alpha \leq 2.$$

if $\{x_i\}$ is i.i.d. drawn from some unknown q and $h \in \mathcal{H}_p$.



Related: control variates and Bayesian MC [e.g., Briol et al., 2015, Bach, 2015].

But how to find a set of good point $\{x_i\}$ to approximate p?

Stein Variational Gradient Descent

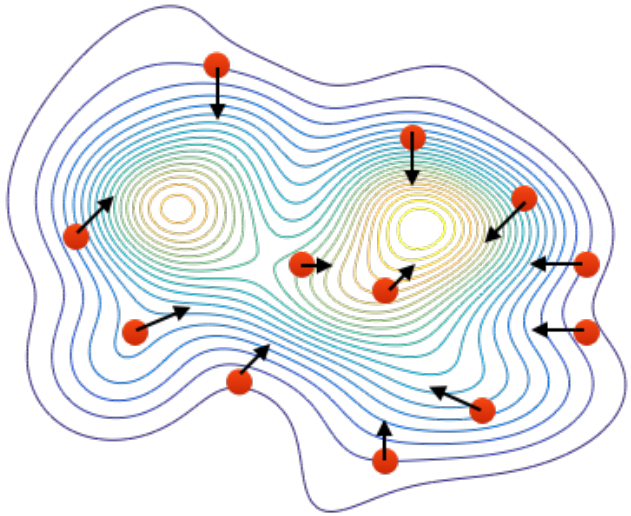
- Directly minimize $\mathrm{KL}(\{x_i\} \,||\, p)$.
- Idea: Iteratively move $\{x_i\}_{i=1}^n$ towards the target p by updates of form

$$x_i' \leftarrow x_i + \epsilon\phi(x_i),$$

where ϕ is a perturbation direction chosen to maximumly decrease the KL divergence with p, that is,

$$\phi = \arg\max_{\phi \in \mathcal{F}} \left\{ -\frac{\partial}{\partial \epsilon} \mathrm{KL}(q_{[\epsilon\phi]} \,||\, p)\big|_{\epsilon=0} \right\},$$

where $q_{[\epsilon\phi]}$ is the density of $x' = x + \epsilon\phi(x)$ when the density of x is q.



Stein Variational Gradient Descent

Closely relates to Stein operator:

$$-\frac{\partial}{\partial \epsilon}\mathrm{KL}(q_{[\epsilon\phi]} \mid\mid p)\big|_{\epsilon=0} = \mathbb{E}_{x\sim q}[\mathcal{T}_p\phi(x)].$$

where $q_{[\epsilon\phi]}$ is the density of $x' = x + \epsilon\phi(x)$ when the density of x is q.

Gives another interpretation of Stein discrepancy:

$$\mathbb{D}(q,\ p) = \max_{\phi\in\mathcal{F}}\Big\{-\frac{\partial}{\partial \epsilon}\mathrm{KL}(q_{[\epsilon\phi]} \mid\mid p)\big|_{\epsilon=0}\Big\}$$

The optimal direction has a closed form when $\mathcal{F}$ is the unit ball of RKHS $\mathcal{H}$:

$$\begin{aligned}\phi^*(\cdot) &= \mathbb{E}_{x\sim q}[\mathcal{T}_p k(x,\cdot)] \\ &= \mathbb{E}_{x\sim q}[\nabla_x log p(x) k(x,\cdot) + \nabla_x k(x,\cdot)].\end{aligned}$$

Stein Variational Gradient Descent

Closely relates to Stein operator:

$$-\frac{\partial}{\partial \epsilon}\mathrm{KL}(q_{[\epsilon\phi]} \mid\mid p)\big|_{\epsilon=0} = \mathbb{E}_{x\sim q}[\mathcal{T}_p\phi(x)].$$

where $q_{[\epsilon\phi]}$ is the density of $x' = x + \epsilon\phi(x)$ when the density of x is q.

Gives another interpretation of Stein discrepancy:

$$\mathbb{D}(q,\ p) = \max_{\phi\in\mathcal{F}}\Big\{ -\frac{\partial}{\partial \epsilon}\mathrm{KL}(q_{[\epsilon\phi]} \mid\mid p)\big|_{\epsilon=0}\Big\}$$

The optimal direction has a closed form when $\mathcal{F}$ is the unit ball of RKHS $\mathcal{H}$:

$$\begin{aligned}\phi^*(\cdot) &= \mathbb{E}_{x\sim q}[\mathcal{T}_p k(x,\cdot)] \\ &= \mathbb{E}_{x\sim q}[\nabla_x log p(x) k(x,\cdot) + \nabla_x k(x,\cdot)].\end{aligned}$$

Stein Variational Gradient Descent

Closely relates to Stein operator:

$$-\frac{\partial}{\partial \epsilon}\mathrm{KL}(q_{[\epsilon\phi]} \mid\mid p)\big|_{\epsilon=0} = \mathbb{E}_{x\sim q}[\mathcal{T}_p\phi(x)].$$

where $q_{[\epsilon\phi]}$ is the density of $x' = x + \epsilon\phi(x)$ when the density of x is q.

Gives another interpretation of Stein discrepancy:

$$\mathbb{D}(q,\ p) = \max_{\phi\in\mathcal{F}}\Big\{ -\frac{\partial}{\partial \epsilon}\mathrm{KL}(q_{[\epsilon\phi]} \mid\mid p)\big|_{\epsilon=0}\Big\}$$

The optimal direction has a closed form when $\mathcal{F}$ is the unit ball of RKHS $\mathcal{H}$:

$$\begin{aligned}\phi^*(\cdot) &= \mathbb{E}_{x\sim q}[\mathcal{T}_p k(x,\cdot)] \\ &= \mathbb{E}_{x\sim q}[\nabla_x log p(x) k(x,\cdot) + \nabla_x k(x,\cdot)].\end{aligned}$$

Stein Variational Gradient Descent

Approximating $\mathbb{E}_{x\sim q}[\cdot]$ with empirical averaging over the current points gives:

$$x_i \leftarrow x_i + \epsilon \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[\,\nabla_x logp(x)\, k(x, x_i) + \nabla_x k(x, x_i)\,], \quad \forall i = 1, \ldots, n.$$

- Deterministically transport probability mass from initialize q_0 to target p.

Stein Variational Gradient Descent

Approximating $\mathbb{E}_{x\sim q}[\cdot]$ with empirical averaging over the current points gives:

$$x_i \leftarrow x_i + \epsilon \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[\underbrace{\nabla_x logp(x)}_{gradient}\, k(x, x_i) + \underbrace{\nabla_x k(x, x_i)}_{\text{repulsive force}}], \quad \forall i = 1, \ldots, n.$$

Two terms:

- $\nabla_x logp(x)$: moves the particles $\{x_i\}$ towards high probability regions of $p(x)$.
- $\nabla_x k(x, x')$: enforce diversity in $\{x_i\}$ (otherwise all x_i collapse to modes of $p(x)$).

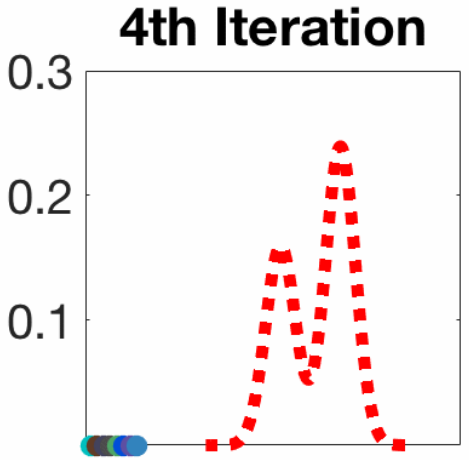
Stein Variational Gradient Descent

Approximating $\mathbb{E}_{x\sim q}[\cdot]$ with empirical averaging over the current points gives:

$$x_i \leftarrow x_i + \epsilon \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[\underbrace{\nabla_x logp(x)}_{\textit{gradient}}\, k(x,x_i) + \underbrace{\nabla_x k(x,x_i)}_{\text{repulsive force}}], \quad \forall i = 1,\ldots,n.$$

Two terms:

- $\nabla_x logp(x)$: moves the particles $\{x_i\}$ towards high probability regions of $p(x)$.
- $\nabla_x k(x,x')$: enforce diversity in $\{x_i\}$ (otherwise all x_i collapse to modes of $p(x)$).



Movie viewable in Adobe Acrobat Reader

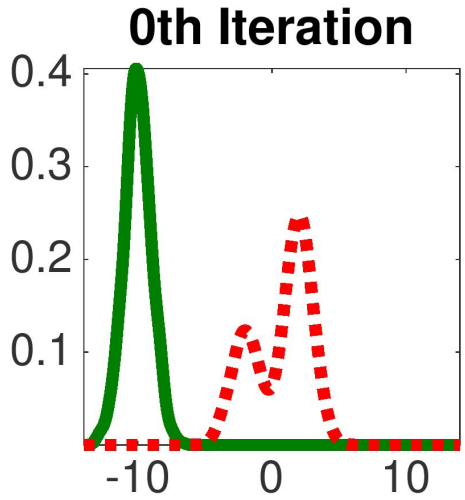
Stein Variational Gradient Descent

Approximating $\mathbb{E}_{x\sim q}[\cdot]$ with empirical averaging over the current points gives:

$$x_i \leftarrow x_i + \epsilon \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[\underbrace{\nabla_x logp(x)}_{\textit{gradient}}\, k(x,x_i) + \underbrace{\nabla_x k(x,x_i)}_{\text{repulsive force}}], \quad \forall i = 1,\ldots,n.$$

Two terms:

- $\nabla_x logp(x)$: moves the particles $\{x_i\}$ towards high probability regions of $p(x)$.
- $\nabla_x k(x,x')$: enforce diversity in $\{x_i\}$ (otherwise all x_i collapse to modes of $p(x)$).



Movie viewable in Adobe Acrobat Reader

Stein Variational Gradient Descent

Approximating $\mathbb{E}_{x\sim q}$ with empirical averaging over the current points gives:

$$x_i \leftarrow x_i + \epsilon \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[\underbrace{\nabla_x logp(x)}_{\textit{gradient}}\, k(x, x_i) + \underbrace{\nabla_x k(x, x_i)}_{\text{repulsive force}}], \quad \forall i = 1, \ldots, n.$$

- When using a single particle ($n = 1$), it reduces to standard gradient ascent for $\max_x \log p(x)$ (i.e., maximum a posteriori (MAP)):

$$x \leftarrow x + \epsilon \nabla_x \log p(x).$$

- Typical Monte Carlo / MCMC: perform worse when $n = 1$.

Stein variational gradient descent is a $\cdots$

- nonparametric variational inference.
- deterministic sampling.
- gradient-based quadrature method.

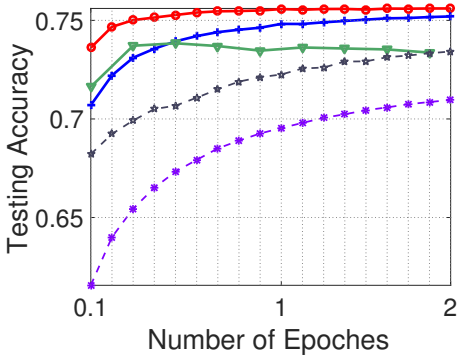
As $n \to \infty$ and $\epsilon \to 0$, the evolution of the density of the particles is governed by a gradient flow

$$\frac{\partial}{\partial t} q_t(x) = -\tilde{\nabla} \mathrm{KL}(q_t \mid\mid p),$$

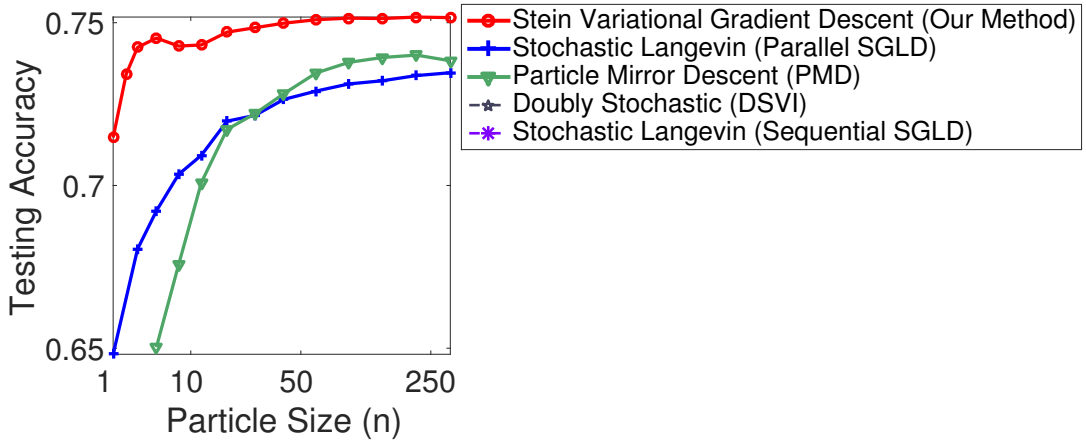
which decreases KL divergence monotonically

$$\frac{\partial}{\partial t} \mathrm{KL}(q_t \mid\mid p) = -\mathbb{D}(q_t, \ p)^2.$$

Bayesian Logistic Regression



(a) Particle size $n = 100$



(b) Results at 3000 iteration (≈ 0.32 epoches)

Bayesian Neural Network

- Test Bayesian neural nets on UCI dataset (with 20 particles).
- Compare with probabilistic back propagation (PBP) [Hernández-Lobato and Adams, 2015].

	Avg. Test RMSE		Avg. Test LL		Avg. Time (Secs)	
Dataset	**PBP**	**Our Method**	**PBP**	**Our Method**	**PBP**	**Ours**
Boston	2.977 ± 0.093	$\mathbf{2.957 \pm 0.099}$	-2.579 ± 0.052	$\mathbf{-2.504 \pm 0.029}$	18	**16**
Concrete	5.506 ± 0.103	$\mathbf{5.324 \pm 0.104}$	-3.137 ± 0.021	$\mathbf{-3.082 \pm 0.018}$	33	**24**
Energy	1.734 ± 0.051	$\mathbf{1.374 \pm 0.045}$	-1.981 ± 0.028	$\mathbf{-1.767 \pm 0.024}$	25	**21**
Kin8nm	0.098 ± 0.001	$\mathbf{0.090 \pm 0.001}$	0.901 ± 0.010	$\mathbf{0.984 \pm 0.008}$	118	**41**
Naval	0.006 ± 0.000	$\mathbf{0.004 \pm 0.000}$	3.735 ± 0.004	$\mathbf{4.089 \pm 0.012}$	173	**49**
Combined	4.052 ± 0.031	$\mathbf{4.033 \pm 0.033}$	-2.819 ± 0.008	$\mathbf{-2.815 \pm 0.008}$	136	**51**
Protein	4.623 ± 0.009	$\mathbf{4.606 \pm 0.013}$	-2.950 ± 0.002	$\mathbf{-2.947 \pm 0.003}$	682	**68**
Wine	0.614 ± 0.008	$\mathbf{0.609 \pm 0.010}$	-0.931 ± 0.014	$\mathbf{-0.925 \pm 0.014}$	26	**22**
Yacht	$\mathbf{0.778 \pm 0.042}$	0.864 ± 0.052	$\mathbf{-1.211 \pm 0.044}$	-1.225 ± 0.042	25	25
Year	$8.733 \pm \mathrm{NA}$	$\mathbf{8.684 \pm NA}$	$-3.586 \pm \mathrm{NA}$	$\mathbf{-3.580 \pm NA}$	7777	**684**

Learning model from data: Given observed data $\{x_{obs,i}\}_{i=1}^n$ drawn from

$$p(x \mid \theta) = \frac{1}{Z} \exp(-\psi(x;\ \theta)), \qquad Z = \int \exp(-\psi(x;\ \theta))dx.$$

We want to estimate parameter θ.

- Deep energy model: $\psi(x;\ \theta)$ is some deep convolutional neural network.

Learning model from data: Given observed data $\{x_{obs,i}\}_{i=1}^n$ drawn from

$$p(x \mid \theta) = \frac{1}{Z} \exp(-\psi(x;\ \theta)), \qquad Z = \int \exp(-\psi(x;\ \theta))dx.$$

We want to estimate parameter θ.

- Deep energy model: $\psi(x;\ \theta)$ is some deep convolutional neural network.
- Maximum likelihood estimator:

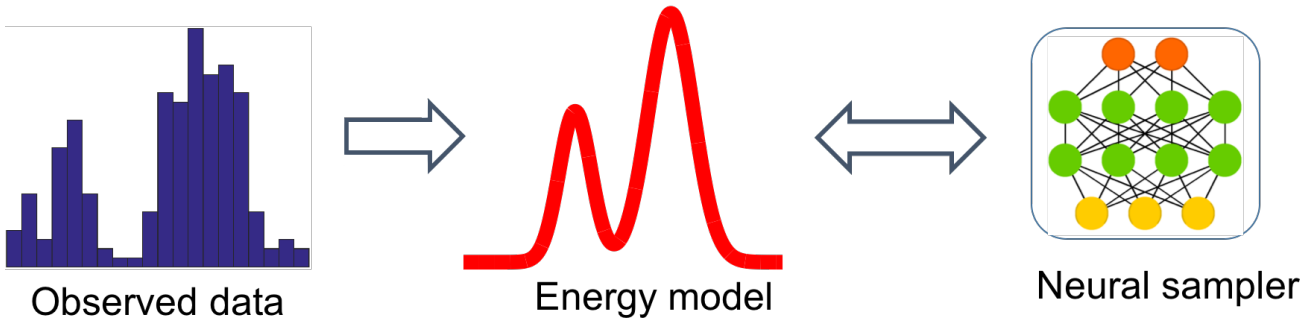
$$\max_\theta \left\{ L(\theta) \equiv \sum_{i=1}^n \log p(x_{obs,i} \mid \theta) \right\}$$

Gradient: $$\nabla_\theta L(\theta) = -\underbrace{\hat{\mathbb{E}}_{obs}[\partial_\theta \psi(x;\ \theta)]}_{\text{Data averaging}} + \underbrace{\mathbb{E}_\theta[\partial_\theta \psi(x;\ \theta)]}_{\text{Model averaging}}.$$

- Difficulty: requires to sample from $p(x|\theta)$ to estimate the model averaging at every gradient iteration.

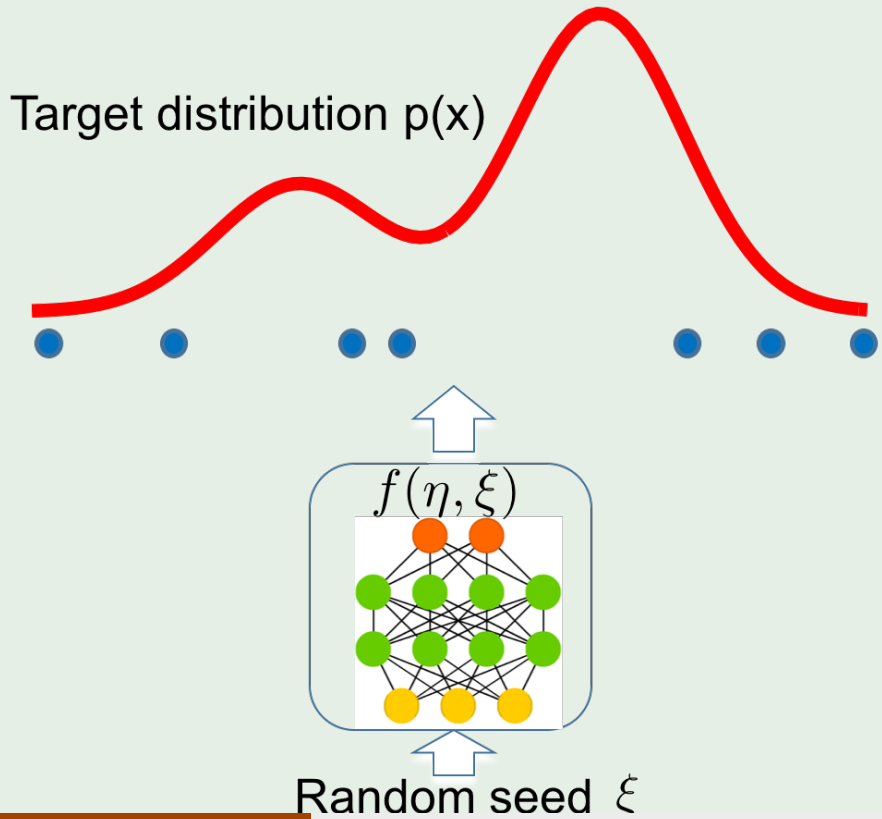
Amortized Inference

- Here, we have to solve many similar inference problems (e.g., sample from $p(x|\theta)$ at each iteration).
- We should not solve each problem from scratch.
- "Amortized inference": train a neural network to "learn to draw samples" from $p(x|\theta)$ and adaptively adjust network parameters as θ updates.



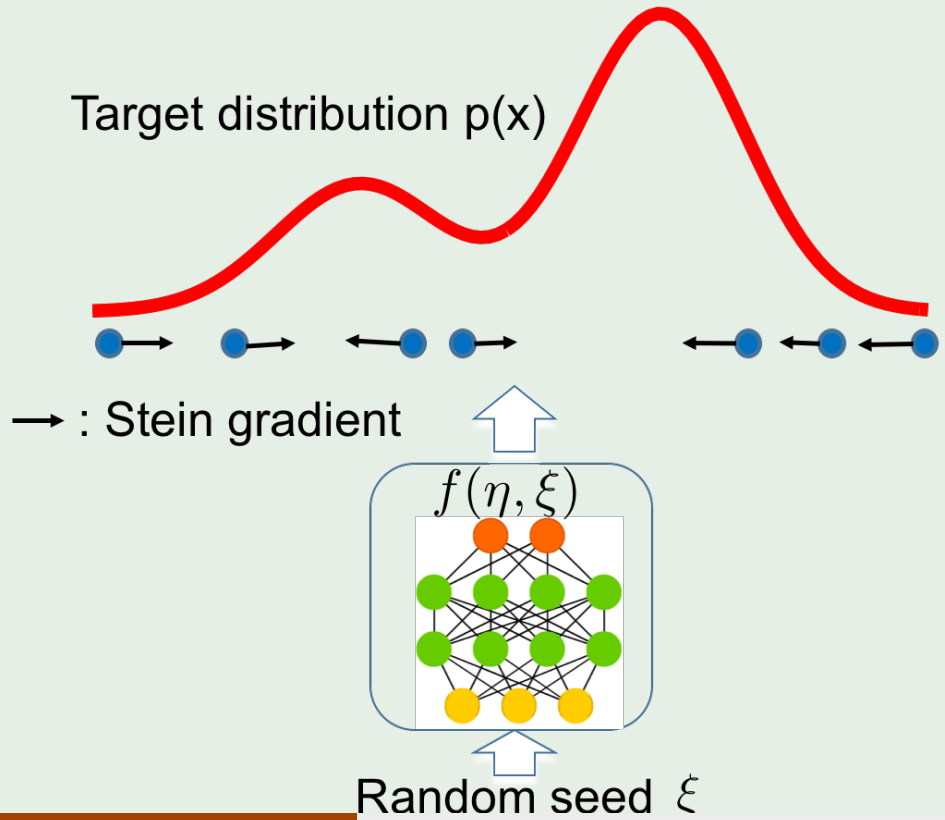
"Learning to Sample"

- Given $p(x)$ and a neural network $f(\eta,\ \xi)$ with parameter η and random input ξ.
- Find η to match the density of random output $x = f(\eta,\ \xi)$ with $p(x)$.
- Idea: Iteratively adjust η to make the output move along the Stein variational gradient direction.



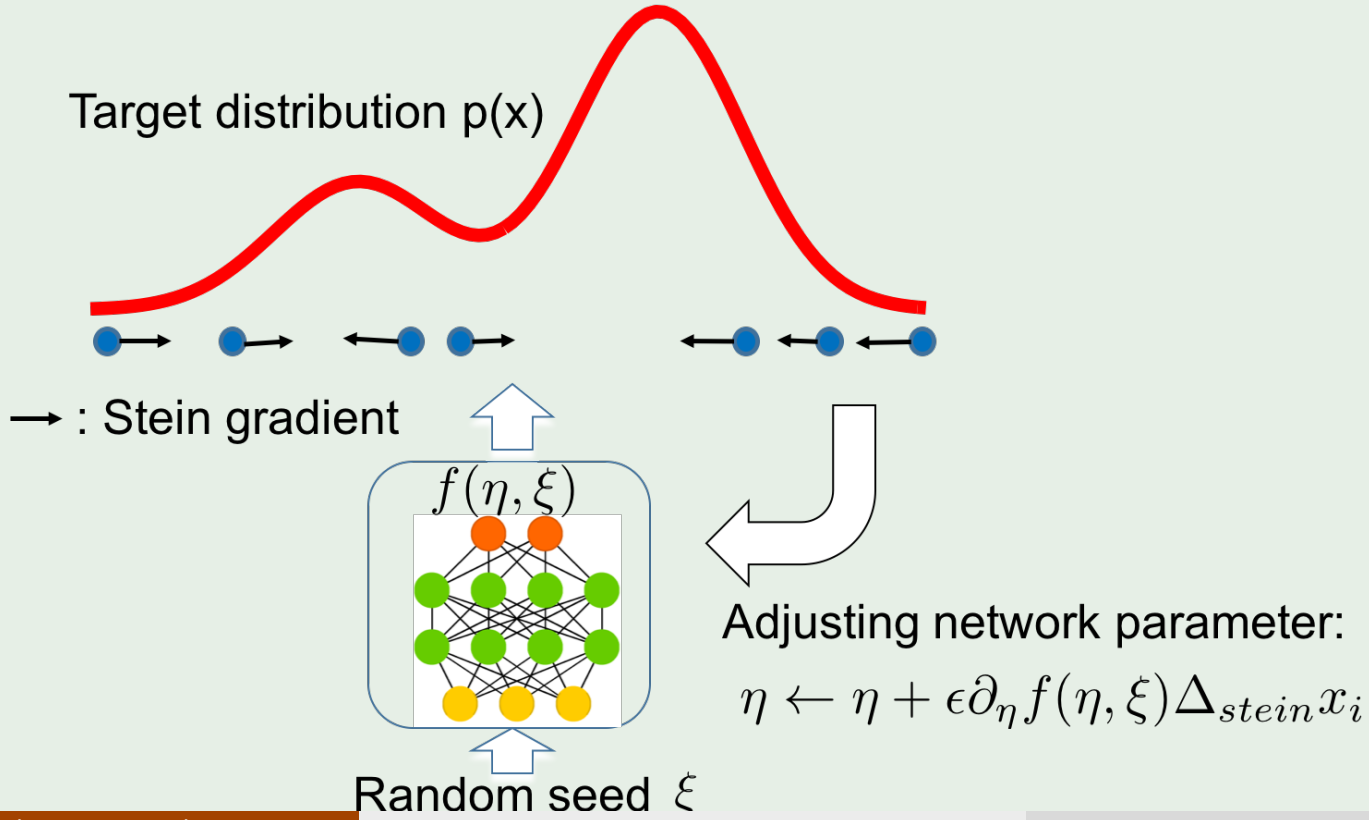
"Learning to Sample"

- Given $p(x)$ and a neural network $f(\eta,\ \xi)$ with parameter η and random input ξ.
- Find η to match the density of random output $x = f(\eta,\ \xi)$ with $p(x)$.
- Idea: Iteratively adjust η to make the output move along the Stein variational gradient direction.



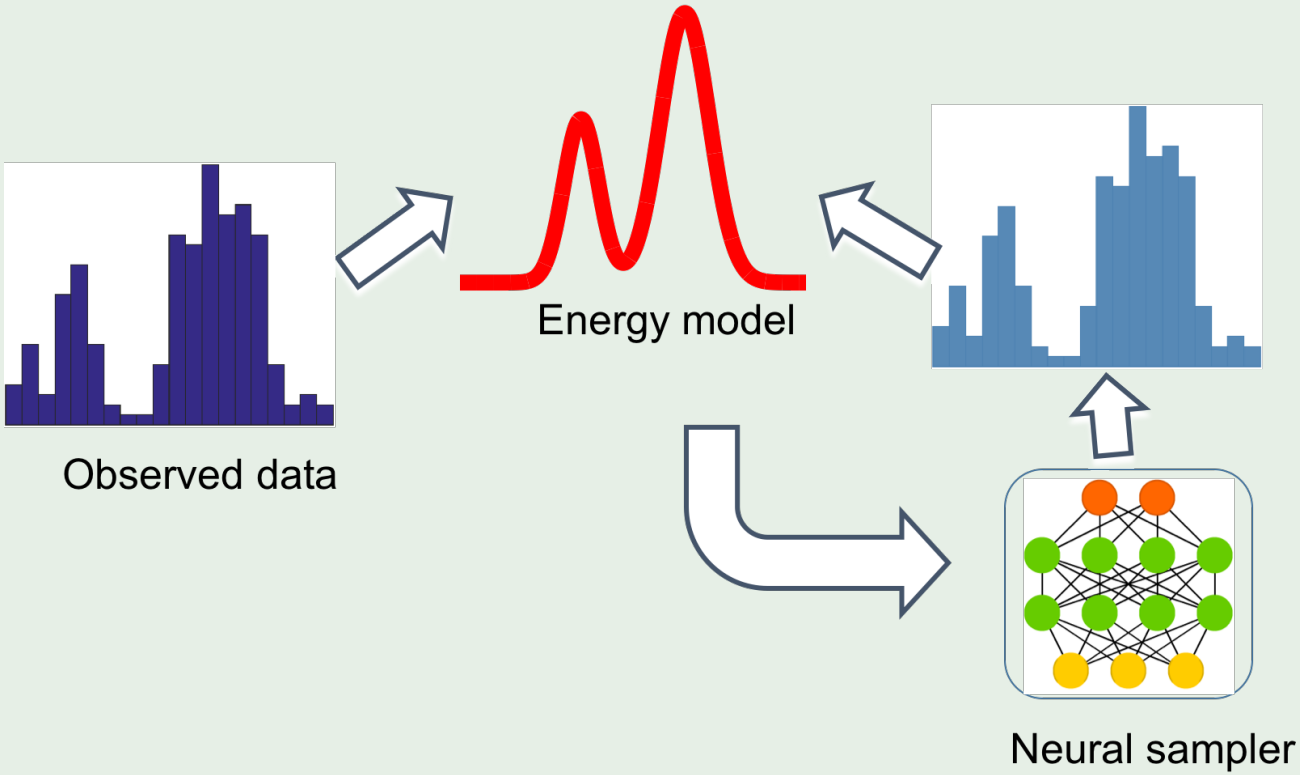
"Learning to Sample"

- Given $p(x)$ and a neural network $f(\eta,\ \xi)$ with parameter η and random input ξ.
- Find η to match the density of random output $x = f(\eta,\ \xi)$ with $p(x)$.
- Idea: Iteratively adjust η to make the output move along the Stein variational gradient direction.



MLE Learning as an Adversarial Game

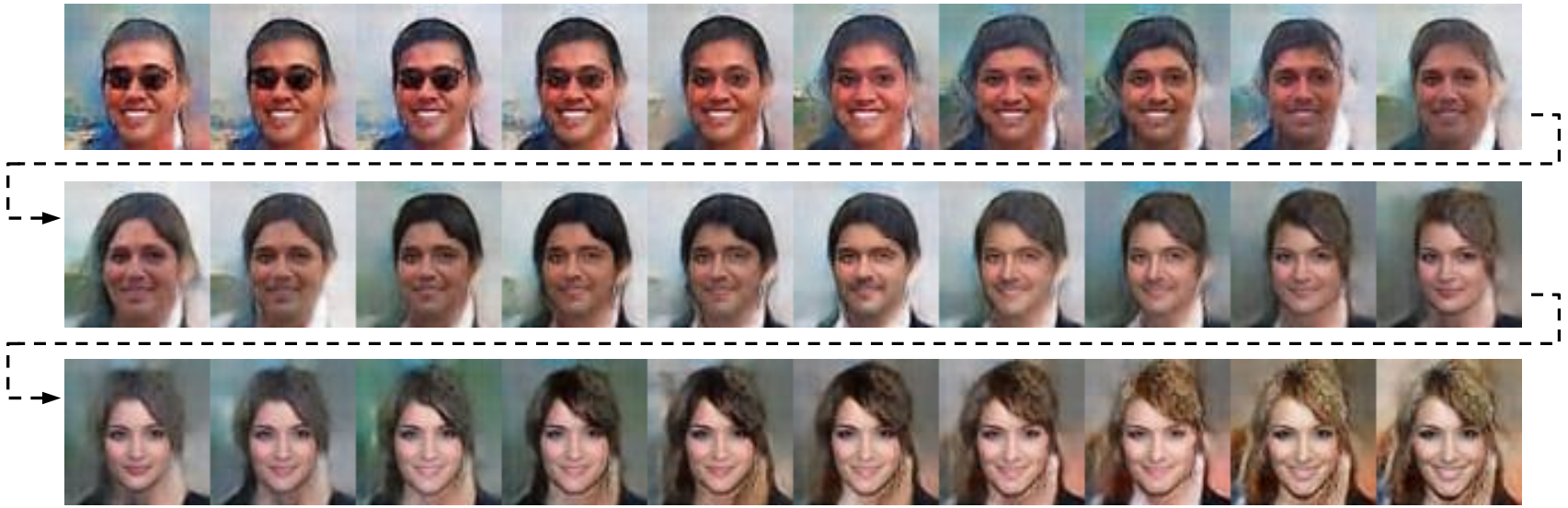
- *Can be treated as an adversarial process between the energy model and the neural sampler.*
- *Similar to generative adversarial networks (GAN) [Goodfellow et al., 2014].*



Real images

Generated by neural sampler

- It learns to "linearize" the semantics of the data distribution.
- Changing the random input ξ smoothly.



Real images

Generated by Neural Sampler

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

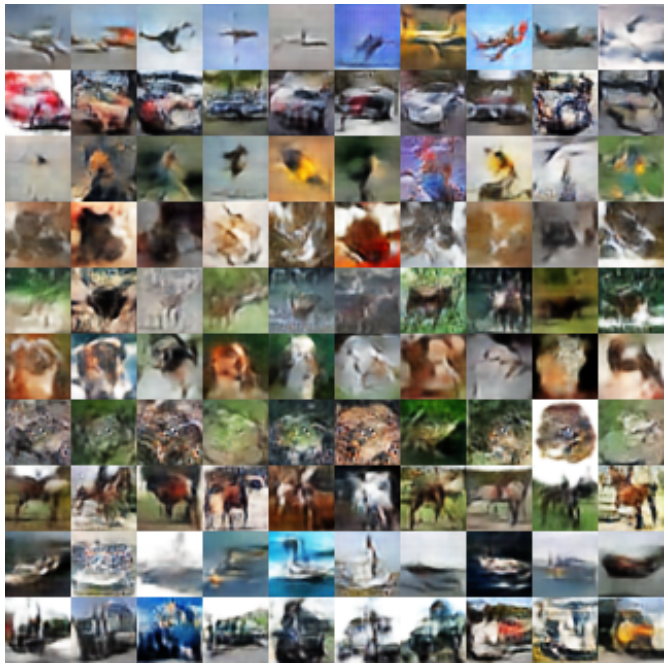
DCGAN [Radford et al., 2015]

Stein

Inception Score				
	Real Training Set	500 Duplicate	DCGAN	Stein
Model Trained on ImageNet	11.237	11.100	6.581	6.351
Model Trained on CIFAR-10	9.848	9.807	7.368	7.428

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

DCGAN [Radford et al., 2015]



Stein

Testing Accuracy			
Real Training Set	500 Duplicate	DCGAN	Stein
92.58 %	44.96 %	44.78 %	63.89 %

Conclusion

- Stein discrepancy Combined with RKHS, variational inference, Monte Carlo etc.
- Provides new tools for many perspectives of probabilistic inference & learning.
- More ideas from Stein's method can be potentially useful for practical machine learning.
- More applications and theories!

Conclusion

- Stein discrepancy Combined with RKHS, variational inference, Monte Carlo etc.
- Provides new tools for many perspectives of probabilistic inference & learning.
- More ideas from Stein's method can be potentially useful for practical machine learning.
- More applications and theories!

Thank You

- Liu, Lee, Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation `https://arxiv.org/abs/1602.03253`
- Liu, Lee. Black-box Importance Sampling `https://arxiv.org/abs/1610.05247`
- Liu, Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm `https://arxiv.org/abs/1608.04471`
- Wang, Liu. Learning to Draw Samples: With Application to Amortized MLE for Generative Adversarial Learning `https://arxiv.org/abs/1611.01722`

References I

M. A. Arcones and E. Gine. On the bootstrap of U and V statistics. The Annals of Statistics, pages 655–674, 1992.

F. Bach. On the equivalence between quadrature rules and random features. arXiv preprint arXiv:1502.06800, 2015.

F.-X. Briol, C. Oates, M. Girolami, M. A. Osborne, D. Sejdinovic, et al. Probabilistic integration: A role for statisticians in numerical analysis? arXiv preprint http://arxiv.org/abs/1512.00933, 2015.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In ICML, 2016.

M. A. Erdogdu. Newton-stein method: An optimization method for glms via stein's lemma. arXiv preprint arXiv:1511.08895, 2015.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014.

J. Gorham and L. Mackey. Measuring sample quality with stein's method. In NIPS, pages 226–234, 2015.

J. M. Hernández-Lobato and R. P. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In ICML, 2015.

M. Huskova and P. Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. The Annals of Statistics, pages 1811–1823, 1993.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. In Journal of Machine Learning Research, pages 695–709, 2005.

M. Janzamin, H. Sedghi, and A. Anandkumar. Score function features for discriminative learning: Matrix and tensor framework. arXiv preprint arXiv:1412.2863, 2014.

S. Lyu. Interpretation and generalization of score matching. In UAI, pages 359–366, 2009.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for monte carlo integration. arXiv preprint arXiv:1410.2392, 2014.

C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami. Convergence rates for a class of estimators based on stein's identity. arXiv preprint arXiv:1603.03220, 2016.

C. J. Oates, M. Girolami, and N. Chopin. Control functionals for monte carlo integration. Journal of the Royal Statistical Society, Series B, 2017.

References II

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

H. Sedghi and A. Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. arXiv preprint arXiv:1412.3046, 2014.

B. Sriperumbudur, K. Fukumizu, R. Kumar, A. Gretton, and A. Hyvärinen. Density estimation in infinite dimensional exponential families. arXiv preprint arXiv:1312.3516, 2013.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory, pages 583–602, 1972.

Kernelized Stein Discrepancy (KSD)

- Is KSD a valid discrepancy: $p = q \iff \mathbb{D}(q, p) = 0$?
- We can show

$$S(q, p) = \mathbb{E}_{x,x' \sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x, x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))]$$

(Recall that $\mathbb{E}_q[\mathcal{T}_p \phi(x)] = \mathbb{E}_q[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))\phi(x)^\top]$)

- We just need $k(x, x')$ to be integrally strictly positive definite:

$$\int g(x)k(x, x')g(x')dx > 0 \quad \forall g \in L_2 \setminus \{0\}$$

Kernelized Stein Discrepancy (KSD)

- Is KSD a valid discrepancy: $p = q \iff \mathbb{D}(q, \ p) = 0$?
- We can show

$$S(q, \ p) = \mathbb{E}_{x,x' \sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x, x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))]$$

(Recall that $\mathbb{E}_q[\mathcal{T}_p \phi(x)] = \mathbb{E}_q[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))\phi(x)^\top]$)

- We just need $k(x, x')$ to be integrally strictly positive definite:

$$\int g(x)k(x, x')g(x')dx > 0 \quad \forall g \in L_2 \setminus \{0\}$$

Kernelized Stein Discrepancy (KSD)

- Is KSD a valid discrepancy: $p = q \iff \mathbb{D}(q, \ p) = 0$?
- We can show

$$S(q, \ p) = \mathbb{E}_{x,x' \sim q}[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))^\top k(x, x')(\boldsymbol{s}_p(x') - \boldsymbol{s}_q(x'))]$$

(Recall that $\mathbb{E}_q[\mathcal{T}_p \phi(x)] = \mathbb{E}_q[(\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x))\phi(x)^\top]$)

- We just need $k(x, x')$ to be integrally strictly positive definite:

$$\int g(x)k(x, x')g(x')dx > 0 \quad \forall g \in L_2 \setminus \{0\}$$

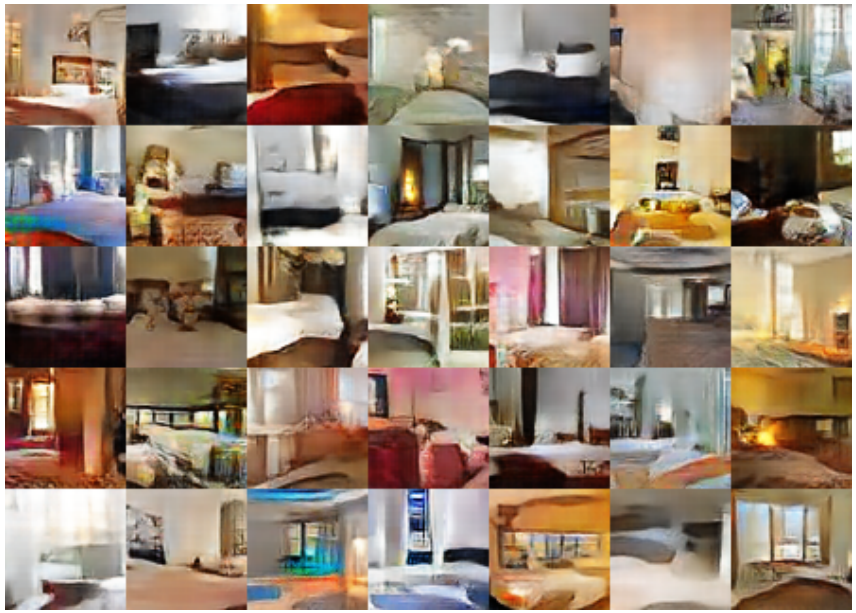
More face images generated by our neural sampler on CelebA.

DCGAN

Stein

MNIST images generated by DCGAN [Radford et al., 2015] and our neural sampler.

DCGAN



Stein

Images generated by DCGAN [Radford et al., 2015] and our neural sampler.