# Energy-Based Models

Instructor: John Thickstun

Discussion Board: Available on Ed

Zoom Link: Available on Canvas

Instructor Contact: thickstn@cs.washington.edu

Course Webpage: https://courses.cs.washington.edu/courses/cse599i/20au/

# Big Picture

- Why is generative modeling hard?
- We need to assign a probability (density) to every point $x \in \mathcal{X}$.
- Why does this seem to be harder than classification?
- Classification: assign a class label to every point $x \in \mathcal{X}$.
- There is a global constraint on densities: $\int_{\mathcal{X}} p(x)\, dx = 1.$

# Energy-Based Models

- What if we just forget about the global constraint?
- Learn an unconstrained energy functional $E_\theta : \mathcal{X} \to \mathbb{R}$.
- The energy functional implicitly defines a probability density $p_\theta$.
- E.g. for any energy $E_\theta$ we can define an associated Gibbs distribution

$$p_\theta(x) = \frac{1}{Z_\theta} e^{-E_\theta(x)}, \text{ where } Z_\theta = \int_{\mathcal{X}} e^{-E_\theta(y)}\, dy.$$

# How Do We Use an EBM?

- Learn an unconstrained energy functional $E_\theta : \mathcal{X} \to \mathbb{R}$.

- For any energy $E_\theta$ we can define an associated Gibbs distribution

$$p_\theta(x) = \frac{1}{Z_\theta} e^{-E_\theta(x)}, \text{ where } Z_\theta = \int_{\mathcal{X}} e^{-E_\theta(y)}\, dy.$$

- How do we sample $x \sim p_\theta$ given an energy functional $E_\theta$?

- How do we train $E_\theta$ so that $p_\theta \approx p$?

# How Do We Use an EBM?

- Learn an unconstrained energy functional $E_\theta : \mathcal{X} \to \mathbb{R}$.

- For any energy $E_\theta$ we can define an associated Gibbs distribution

$$p_\theta(x) = \frac{1}{Z_\theta} e^{-E_\theta(x)}, \text{ where } Z_\theta = \int_{\mathcal{X}} e^{-E_\theta(y)}\, dy.$$

- **How do we sample $x \sim p_\theta$ given an energy functional $E_\theta$?**

- How do we train $E_\theta$ so that $p_\theta \approx p$?

# Sampling From an EBM

- Generate samples $x \sim p_\theta$ given energy $E_\theta : \mathcal{X} \to \mathbb{R}$ where

$$p_\theta(x) = \frac{1}{Z_\theta} e^{-E_\theta(x)}, \text{ where } Z_\theta = \int_{\mathcal{X}} e^{-E_\theta(y)}\, dy.$$

- Classical statistics to the rescue!
- Markov-Chain Monte Carlo (MCMC).
- Construct a Markov-Chain with stationary distribution $p_\theta$.

# Langevin Dynamics

- If $\mathcal{X} = \mathbb{R}^d$, Langevin dynamics are defined by a stochastic differential equation

$$\frac{\partial x}{\partial t} = \nabla_x \log p_\theta(x)\, dt + \sqrt{2}\, dW_t.$$

- The term $dW_t$ is white noise: the derivative of Brownian motion $W_t$.
- Can make sense of this derivative with the machinery of Ito integration.
- Fokker-Planck equation: stationary distribution is $p_\theta(x)$.
- More precisely, $D(x_t \parallel p_\theta) \to 0 \text{ as } t \to \infty$.

# Discretized Langevin Dynamics

- Continuous Langevin dynamics process:

$$\frac{\partial x}{\partial t} = \nabla_x \log p_\theta(x)\, dt + \sqrt{2}\, dW_t.$$

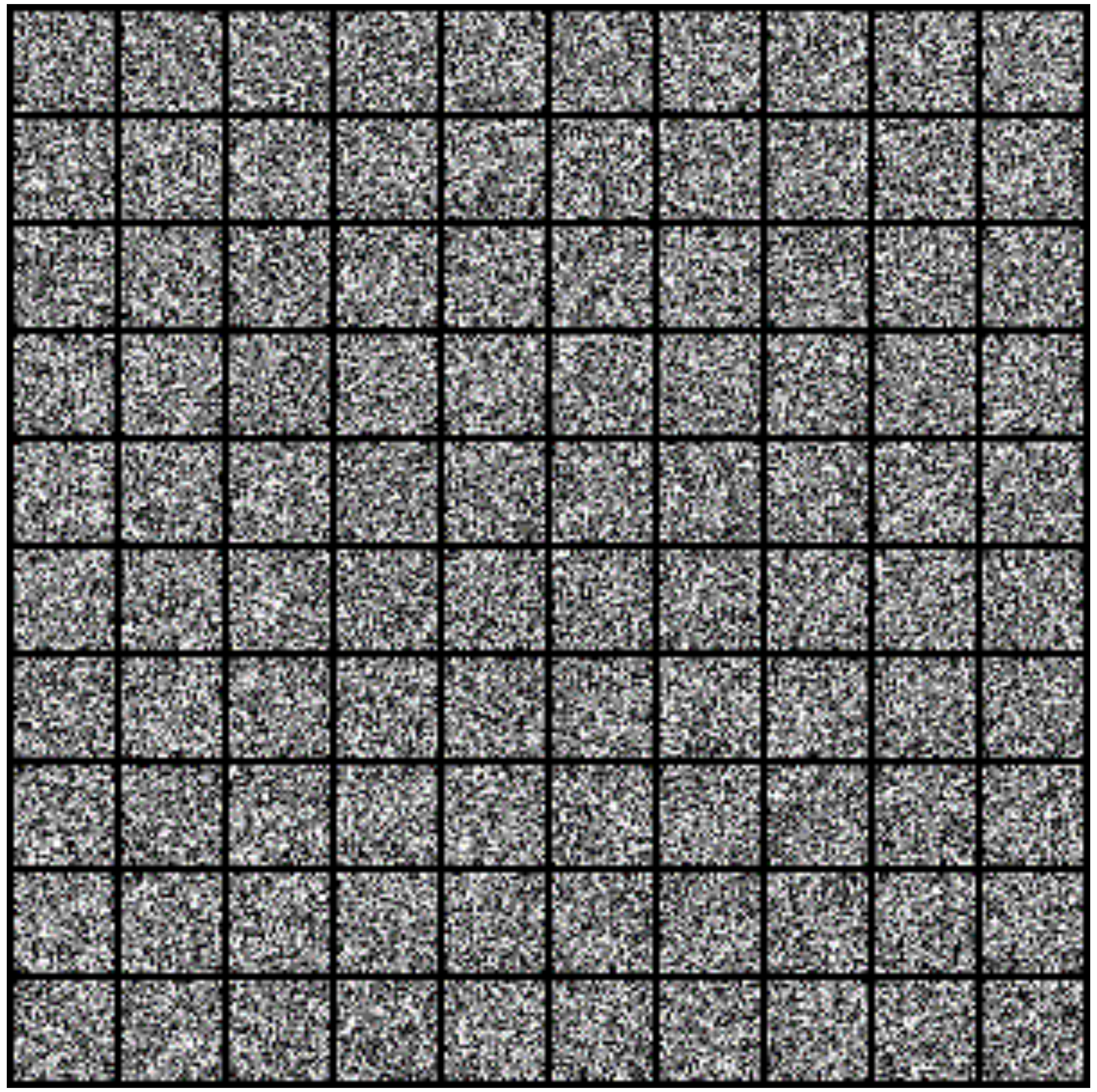
- Can't construct a diffusion. Discretize and construct a Markov chain:

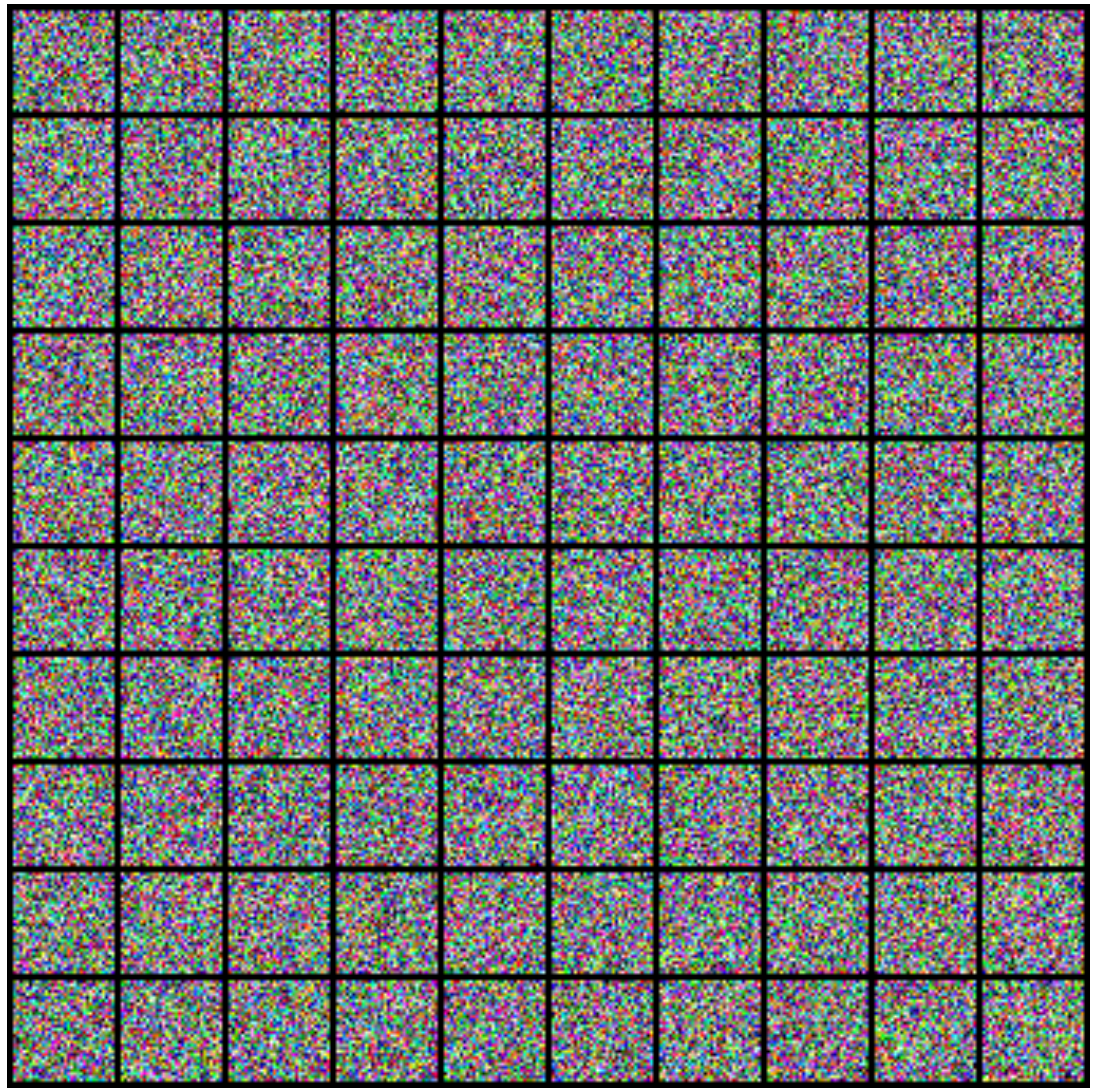
$$x_{t+1} = x_t - \eta \nabla_x \log p_\theta(x_t) + \sqrt{2\eta}\varepsilon_t.$$

- Where the noise terms are sampled i.i.d. $\varepsilon_t \sim \mathcal{N}(0, I)$.

- Analogous to Euler discretization of a (deterministic) differential equation.

# Langevin Dynamics in Practice





Sampling from MNIST (left) and CIFAR-10 (right) EBMs Using Langevin Dynamics

Song and Ermon, Neurips 2019

# 5-Minute Break

# How Do We Use an EBM?

- Learn an unconstrained energy functional $E_\theta : \mathcal{X} \to \mathbb{R}$.

- For any energy $E_\theta$ we can define an associated Gibbs distribution

$$p_\theta(x) = \frac{1}{Z_\theta} e^{-E_\theta(x)}, \text{ where } Z_\theta = \int_{\mathcal{X}} e^{-E_\theta(y)}\, dy.$$

- How do we sample $x \sim p_\theta$ given an energy functional $E_\theta$?

- How do we train $E_\theta$ so that $p_\theta \approx p$?

# How Do We Use an EBM?

- Learn an unconstrained energy functional $E_\theta : \mathcal{X} \to \mathbb{R}$.

- For any energy $E_\theta$ we can define an associated Gibbs distribution

$$p_\theta(x) = \frac{1}{Z_\theta} e^{-E_\theta(x)}, \text{ where } Z_\theta = \int_{\mathcal{X}} e^{-E_\theta(y)}\, dy.$$

- How do we sample $x \sim p_\theta$ given an energy functional $E_\theta$? Langevin dynamics:

$$\nabla_x \log p_\theta(x) = -\nabla_x E_\theta(x) - \nabla_x \log Z_\theta = -\nabla_x E_\theta(x).$$

- **How do we train $E_\theta$ so that $p_\theta \approx p$?**

# Estimating the Gradient Field

- Do we even need an energy functional?

- Langevin dynamics just needs gradients:

$$
\begin{aligned}
x_{t+1} &= x_t - \eta \nabla_x \log p_\theta(x_t) + \sqrt{2\eta}\varepsilon_t \\
&= x_t + \eta \nabla_x E_\theta(x_t) + \sqrt{2\eta}\varepsilon_t .
\end{aligned}
$$

- Just learn $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ to approximate gradients $s(x) = \nabla_x \log p(x)$.

- The function $s : \mathbb{R}^d \to \mathbb{R}^d$ is called the score function; we want $s_\theta \approx s$.

# Score Matching

- Want to learn $s_\theta : \mathbb{R}^d \to \mathbb{R}^d$ such that $s_\theta(x) \approx s(x) = \nabla_x \log p(x)$.

- What is a good way to quantify $s_\theta \approx s$? How about MSE?

$$\mathop{\mathbb{E}}_{x \sim p}\left[\frac{1}{2}\|s_\theta(x) - \nabla_x \log p(x)\|_2^2\right].$$

- Minimize the MSE using the following identity:

$$\arg\min_\theta \mathop{\mathbb{E}}_{x \sim p}\left[\frac{1}{2}\|s_\theta(x) - \nabla_x \log p(x)\|_2^2\right] = \arg\min_\theta \mathop{\mathbb{E}}_{x \sim p}\left[\operatorname{tr}\left(\nabla_x s_\theta(x)\right) + \frac{1}{2}\|s_\theta(x)\|_2^2\right].$$

# Implicit Score Matching

Proposition [Hyvärinen, 2005]:

$$\arg\min_{\theta} \mathop{\mathbb{E}}_{x\sim p}\left[\frac{1}{2}\|s_\theta(x) - \nabla_x \log p(x)\|_2^2\right] = \arg\min_{\theta} \mathop{\mathbb{E}}_{x\sim p}\left[\mathrm{tr}\left(\nabla_x s_\theta(x)\right) + \frac{1}{2}\|s_\theta(x)\|_2^2\right].$$

Proof. Step 1 (expand the quadratic):

$$\arg\min_{\theta} \mathop{\mathbb{E}}_{x\sim p}\left[\frac{1}{2}\|s_\theta(x) - \nabla_x \log p(x)\|_2^2\right] = \arg\min_{\theta} \mathop{\mathbb{E}}_{x\sim p}\left[\frac{1}{2}\|s_\theta(x)\|^2 - s_\theta(x)^T \nabla_x \log p(x)\right].$$

Step 2 (integration by parts):

$$\mathop{\mathbb{E}}_{x\sim p}\left[s_\theta(x)^T \nabla_x \log p(x)\right] = \sum_{i=1}^{d} \int_{\mathcal{X}} s_\theta(x)_i \frac{\partial \log p(x)}{\partial x_i} p(x)\,dx = \sum_{i=1}^{d} \int_{\mathcal{X}} s_\theta(x)_i \frac{\partial p(x)}{\partial x_i}\,dx$$

$$= -\sum_{i=1}^{d} \int_{\mathcal{X}} \frac{s_\theta(x)_i}{\partial x_i} p(x)\,dx = -\int_{\mathcal{X}} \mathrm{tr}\left(\nabla_x s_\theta(x)\right) p(x)\,dx = -\mathop{\mathbb{E}}_{x\sim p}\left[\mathrm{tr}\left(\nabla_x s_\theta(x)\right)\right].$$