# Introduction to Score-matching

Justin T Chiu

July 5, 2023

## Goals

1. What is an energy-based model and why are they hard to train?
2. What is score-matching, and how can it be used to train an EBM?
3. How does score-matching relate to diffusion models?

# Energy-Based Models (EBM)

# Problem setup: Density estimation

- Observations from true model $x \sim p^*(x)$
- Goal: Learn a model $p(x)$ that's close to $p^*(x)$
  - Capture uncertainty / variability over $x$
- Participation: Give examples of an $x$ we model, and how $p(x)$ is parameterized
  - Ex: Language modeling uses Transformers for $p(x) = \prod_t p(x_t | x_{<t})$

# Running example: Image generation

- "Solved": Finite-class density estimation
  - Softmax assigns a score to each $E(x)$ then normalizes

$$softmax(x) = \frac{\exp(E(x))}{\sum_x \exp(E(x))}$$

- Image generation
  - Every change in a single pixel is a new class
  - Size: 1024 x 1024, each pixel has 256 * 3 values

# Image generation models

- Autoregressive: Break down generation from left-to-right

$$p(x) = \prod_t p(x_{ij} | x_{<i,j}, x_{\bullet,<j})$$

- Latent variable model: Specify break down more flexibly

$$p(x) = \sum_z p(x|z)p(z)$$

- Energy-based model: Don't force breakdown of decision process

# EBM drawing

# What is an EBM?

- Globally normalized over images $x$

$$p(x) = \frac{\exp(E(x))}{Z}$$
$$Z = \int_x \exp(E(x))$$

- Computation of the partition function $Z$ is hard
  - Integrate $E(x)$ over all possible images
- Goal of training: maximize likelihood (minimize KL div)
  - Need to compute $p(x)$ and therefore $Z$
  - Next: How to avoid computing partition function $Z$

# Training an EBM

# KL divergence to Fisher divergence

- Standard: Minimize KL divergence

$$E_{p^*(x)} \log \frac{p^*(x)}{p(x)} = E_{p^*(x)} \log p^*(x) - E_{p^*(x)} \log p(x)$$

- Instead: Minimize Fisher divergence

$$E_{p^*(x)} \left\| \nabla \log \frac{p^*(x)}{p(x)} \right\|_2^2$$

- Avoid computing partition function $Z$

# Fisher divergence