

Paper Review: A Connection Between Score Matching and Denoising Autoencoders

Beomsu Kim*

July 6, 2021

Paper Information.

- Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. Neural Computation, 23(7):1661–1674, 2011.

1 Introduction

- $q(x)$: unknown true pdf, $x \in \mathbb{R}^d$.
- $D_n = \{x^{(1)}, \dots, x^{(n)}\}$: training set of n i.i.d. samples from $q(x)$.
- $q_0(x) = \frac{1}{n} \sum_{i=1}^n \delta(\|x - x^{(i)}\|)$: empirical pdf associated with D_n .
- $q_\sigma(\tilde{x} \mid x) = \frac{1}{(2\pi)^{d/2}\sigma^d} e^{-\frac{1}{2\sigma^2}\|\tilde{x}-x\|^2}$: smoothing kernel or noise model.
- $q_\sigma(\tilde{x}, x) = q_\sigma(\tilde{x} \mid x)q_0(x)$: joint pdf.
- $q_\sigma(\tilde{x}) = \frac{1}{n} \sum_{i=1}^n q_\sigma(\tilde{x}, x^{(i)})$: Parzen density estimate based on D_n , obtainable by marginalizing $q_\sigma(\tilde{x}, x)$.
- $p(x; \theta)$: density model with parameters θ .
- $J_1 \cong J_2$: means $J_1(\theta)$ and $J_2(\theta)$ have the same set of minimizers.
- $\mathbb{E}_{q(x)}[g(x)] = \int_x q(x)g(x) dx$: expectation with respect to distribution $q(x)$.
- $\text{softplus}(x) = \log(1 + e^x)$: will be applied elementwise to vectors.
- $\text{sigmoid}(x) = \frac{1}{1+e^{-x}} = \text{softplus}'(x)$: will be applied elementwise to vectors.
- I : identity matrix.
- W^\top : transpose of matrix W .
- W_i : vector for i th row of W .
- $W_{\cdot,j}$: vector for j th column of W .

*Department of Mathematical Sciences, KAIST. Email beomsu.kim@kaist.ac.kr

2 Denoising Autoencoders

- A training input $x \in D_n$ is first corrupted by additive Gaussian noise of covariance $\sigma^2 I$ yielding corrupted input $\tilde{x} = x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. This corresponds to conditional density

$$q_\sigma(\tilde{x} | x) = \frac{1}{(2\pi)^{d/2} \sigma^d} e^{-\frac{1}{2\sigma^2} \|\tilde{x} - x\|^2}.$$

- The corrupted version \tilde{x} is encoded into a hidden representation h through an affine mapping followed by a nonlinearity.

$$h = \text{encode}(\tilde{x}) = \text{sigmoid}(W\tilde{x} + b)$$

- The hidden representation h is decoded into reconstruction x^r through affine mapping.

$$x^r = \text{decode}(h) = W^\top h + c$$

- The parameters $\theta = \{W, b, c\}$ are optimized so that the expected squared reconstruction error $\|x^r - x\|^2$ is minimized, i.e., the objective function being minimized by such a denoising autoencoder (DAE) is

$$\begin{aligned} J_{DAE_\sigma}(\theta) &= \mathbb{E}_{q_\sigma(\tilde{x}, x)} [\|\text{decode}(\text{encode}(\tilde{x})) - x\|^2] \\ &= \mathbb{E}_{q_\sigma(\tilde{x}, x)} [\|W^\top \text{sigmoid}(W\tilde{x} + b) + c - x\|^2]. \end{aligned}$$

3 Score Matching

3.1 Explicit Score Matching (ESM)

- Define the energy-based model

$$p(x; \theta) = \frac{1}{Z(\theta)} \exp(-E(x; \theta)), \quad Z(\theta) = \int_x \exp(-E(x; \theta)) dx.$$

- Define the score function

$$\Psi(x; \theta) = \frac{\partial \log p(x; \theta)}{\partial x}.$$

- Explicit score matching minimizes

$$J_{ESM_q}(\theta) = \mathbb{E}_{q(x)} \left[\frac{1}{2} \left\| \Psi(x; \theta) - \frac{\partial \log q(x)}{\partial x} \right\|^2 \right].$$

3.2 Implicit Score Matching (ISM)

- Define

$$J_{ISM_q}(\theta) = \mathbb{E}_{q(x)} \left[\frac{1}{2} \|\Psi(x; \theta)\|^2 + \sum_{i=1}^d \frac{\partial \Psi_i(x; \theta)}{\partial x_i} \right].$$

- Hyvärinen in *Estimation of Non-normalized Statistical Models by Score Matching* shows that

$$J_{ESM_q} \cong J_{ISM_q}.$$

4 Linking Score Matching to the DAE Objective

4.1 Matching the Score of a Non-Parametric Estimator

- Matching $\Psi(x; \theta)$ with the score of Parzen windows density estimator $q_\sigma(\tilde{x})$ yields

$$J_{ESM_\sigma}(\theta) = \mathbb{E}_{q_\sigma(\tilde{x})} \left[\frac{1}{2} \left\| \Psi(\tilde{x}; \theta) - \frac{\partial \log q_\sigma(\tilde{x})}{\partial \tilde{x}} \right\|^2 \right].$$

- $q_\sigma(\tilde{x})$ satisfies the regularities conditions for implicit score matching, so

$$J_{ESM_\sigma} \cong J_{ISM_\sigma}.$$

- This equivalence breaks in the limit $\sigma \rightarrow 0$, because q_σ no longer satisfies the regularity conditions, and J_{ESM_σ} can no longer be computed (whereas J_{ISM_σ} remains well-behaved).

4.2 Denoising Score Matching (DSM)

- We define the following denoising score matching (DSM) objective

$$J_{DSM_{q_\sigma}}(\theta) = \mathbb{E}_{q_\sigma(\tilde{x}, x)} \left[\frac{1}{2} \left\| \Psi(\tilde{x}; \theta) - \frac{\partial q_\sigma(\tilde{x} | x)}{\partial \tilde{x}} \right\|^2 \right].$$

- The underlying intuition is that following the gradient Ψ of the log-density at some corrupted point \tilde{x} should ideally move us towards the clean sample x .
- In other words, the model $p(x; \theta)$ should assign higher density to x than \tilde{x} .
- Indeed, with the considered Gaussian kernel we have

$$\frac{\partial \log q_\sigma(\tilde{x} | x)}{\partial \tilde{x}} = \frac{1}{\sigma^2}(x - \tilde{x})$$

and this direction corresponds to moving from \tilde{x} back towards clean sample x .

- We observe that

$$J_{ESM_\sigma} \cong J_{DSM_{q_\sigma}}.$$

- Also, for an appropriate choice of $p(x; \theta)$,

$$J_{DSM_{q_\sigma}} \cong J_{DAE_\sigma}.$$

Proposition 1.

$$J_{ESM_\sigma} \cong J_{DSM_{q_\sigma}}$$

Proof. Observe that

$$\begin{aligned} J_{ESM_\sigma}(\theta) &= \mathbb{E}_{q_\sigma(\tilde{x})} \left[\frac{1}{2} \left\| \Psi(\tilde{x}; \theta) - \frac{\partial \log q_\sigma(\tilde{x})}{\partial \tilde{x}} \right\|^2 \right] \\ &= \mathbb{E}_{q_\sigma(\tilde{x})} \left[\frac{1}{2} \|\Psi(\tilde{x}; \theta)\|^2 \right] - \mathbb{E}_{q_\sigma(\tilde{x})} \left[\left\langle \Psi(\tilde{x}; \theta), \frac{\partial \log q_\sigma(\tilde{x})}{\partial \tilde{x}} \right\rangle \right] + C_1 \end{aligned}$$

for a constant C_1 independent of θ . We also have

$$\begin{aligned} \mathbb{E}_{q_\sigma(\tilde{x})} \left[\left\langle \Psi(\tilde{x}; \theta), \frac{\partial \log q_\sigma(\tilde{x})}{\partial \tilde{x}} \right\rangle \right] &= \int_{\tilde{x}} q_\sigma(\tilde{x}) \left\langle \Psi(\tilde{x}; \theta), \frac{\partial \log q_\sigma(\tilde{x})}{\partial \tilde{x}} \right\rangle d\tilde{x} \\ &= \int_{\tilde{x}} \left\langle \Psi(\tilde{x}; \theta), \frac{\partial}{\partial \tilde{x}} q_\sigma(\tilde{x}) \right\rangle d\tilde{x} \\ &= \int_{\tilde{x}} \left\langle \Psi(\tilde{x}; \theta), \frac{\partial}{\partial \tilde{x}} \int_x q_0(x) q_\sigma(\tilde{x} | x) dx \right\rangle d\tilde{x} \\ &= \int_{\tilde{x}} \left\langle \Psi(\tilde{x}; \theta), \int_x q_0(x) \frac{\partial q_\sigma(\tilde{x} | x)}{\partial \tilde{x}} dx \right\rangle d\tilde{x} \\ &= \int_{\tilde{x}} \left\langle \Psi(\tilde{x}; \theta), \int_x q_0(x) q_\sigma(\tilde{x} | x) \frac{\partial \log q_\sigma(\tilde{x} | x)}{\partial \tilde{x}} dx \right\rangle d\tilde{x} \\ &= \int_{\tilde{x}} \int_x q_0(x) q_\sigma(\tilde{x} | x) \left\langle \Psi(\tilde{x}; \theta), \frac{\partial \log q_\sigma(\tilde{x} | x)}{\partial \tilde{x}} \right\rangle dx d\tilde{x} \\ &= \int_{\tilde{x}} \int_x q_\sigma(\tilde{x}, x) \left\langle \Psi(\tilde{x}; \theta), \frac{\partial \log q_\sigma(\tilde{x} | x)}{\partial \tilde{x}} \right\rangle dx d\tilde{x} \\ &= \mathbb{E}_{q_\sigma(\tilde{x}, x)} \left[\left\langle \Psi(\tilde{x}; \theta), \frac{\partial \log q_\sigma(\tilde{x} | x)}{\partial \tilde{x}} \right\rangle \right]. \end{aligned}$$

This shows that

$$\begin{aligned} J_{ESM_\sigma}(\theta) &= \mathbb{E}_{q_\sigma(\tilde{x})} \left[\frac{1}{2} \|\Psi(\tilde{x}; \theta)\|^2 \right] - \mathbb{E}_{q_\sigma(\tilde{x})} \left[\left\langle \Psi(\tilde{x}; \theta), \frac{\partial \log q_\sigma(\tilde{x})}{\partial \tilde{x}} \right\rangle \right] + C_1 \\ &= \mathbb{E}_{q_\sigma(\tilde{x}, x)} \left[\frac{1}{2} \|\Psi(\tilde{x}; \theta)\|^2 \right] - \mathbb{E}_{q_\sigma(\tilde{x}, x)} \left[\left\langle \Psi(\tilde{x}; \theta), \frac{\partial \log q_\sigma(\tilde{x} | x)}{\partial \tilde{x}} \right\rangle \right] + C_1 \\ &= \mathbb{E}_{q_\sigma(\tilde{x}, x)} \left[\frac{1}{2} \left\| \Psi(\tilde{x}; \theta) - \frac{\partial q_\sigma(\tilde{x} | x)}{\partial x} \right\|^2 \right] - \mathbb{E}_{q_\sigma(\tilde{x}, x)} \left[\frac{1}{2} \left\| \frac{\partial q_\sigma(\tilde{x} | x)}{\partial x} \right\|^2 \right] + C_1 \\ &= J_{DSM_{q_\sigma}}(\theta) + C_1 - C_2 \end{aligned}$$

where C_2 is another constant independent of θ . This proves

$$J_{ESM_\sigma} \cong J_{DSM_{q_\sigma}}.$$

□

Proposition 2.

$$J_{DSM_{q\sigma}} \cong J_{DAE_\sigma}$$

Proof. We choose

$$p(x; \theta) = \frac{1}{Z(\theta)} \exp(-E(x; \theta))$$

where for $\theta = \{W, b, c\}$,

$$E(x; \theta) = -\frac{\langle c, x \rangle - \frac{1}{2}\|x\|^2 + \sum_i \text{softplus}(\langle W_i, x \rangle + b_i)}{\sigma^2}.$$

We then have

$$\begin{aligned} \Psi_j(x; \theta) &= \frac{\partial \log p(x; \theta)}{\partial x_j} \\ &= -\frac{\partial E(x; \theta)}{\partial x_j} \\ &= \frac{1}{\sigma^2} \left(c_j - x_j + \sum_i \text{softplus}'(\langle W_i, x \rangle + b_i) \frac{\partial(\langle W_i, x \rangle + b_i)}{\partial x_j} \right) \\ &= \frac{1}{\sigma^2} \left(c_j - x_j + \sum_i \text{sigmoid}(\langle W_i, x \rangle + b_i) W_{ij} \right) \\ &= \frac{1}{\sigma^2} (c_j - x_j + \langle W_{:,j}, \text{sigmoid}(Wx + b) \rangle) \end{aligned}$$

which we can write as the single equation

$$\Psi(x; \theta) = \frac{1}{\sigma^2} (W^\top \text{sigmoid}(Wx + b) + c - x).$$

It follows that

$$\begin{aligned} J_{DSM_{q\sigma}}(\theta) &= \mathbb{E}_{q_\sigma(\tilde{x}, x)} \left[\frac{1}{2} \left\| \Psi(\tilde{x}; \theta) - \frac{\partial q_\sigma(\tilde{x} | x)}{\partial x} \right\|^2 \right] \\ &= \mathbb{E}_{q_\sigma(\tilde{x}, x)} \left[\frac{1}{2} \left\| \frac{1}{\sigma^2} (W^\top \text{sigmoid}(Wx + b) + c - x) - \frac{1}{\sigma^2} (x - \tilde{x}) \right\|^2 \right] \\ &= \frac{1}{2\sigma^4} \mathbb{E}_{q_\sigma(\tilde{x}, x)} [\|W^\top \text{sigmoid}(W\tilde{x} + b) + c - x\|^2] \\ &= \frac{1}{2\sigma^4} J_{DAE_\sigma}(\theta) \end{aligned}$$

and so

$$J_{DSM_{q\sigma}} \cong J_{DAE_\sigma}.$$

□