# Paper Review: Estimation of Non-Normalized Statistical Models by Score Matching

Beomsu Kim*

June 29, 2021

**Paper Information.**

- Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. Journal of Machine Learning Research, 6(24):695–709, 2005.

## 1 Introduction

- Assume we observe a random vector $\mathbf{x} \in \mathbb{R}^n$ which has a pdf denoted by $p_{\mathbf{x}}(\cdot)$.
- We have a parametrized density model $p(\cdot;\theta)$, where $\theta$ is an $m$-dimensional vector of parameters.
- We want to approximate $p_{\mathbf{x}}(\cdot)$ by $p(\cdot;\hat{\theta})$ for the estimated parameter value $\hat{\theta}$.
- We only are able to compute the pdf given by the model up to a multiplicative constant $Z(\boldsymbol{\theta})$.

$$p(\xi;\theta) = \frac{1}{Z(\boldsymbol{\theta})} q(\xi;\boldsymbol{\theta}), \qquad Z(\boldsymbol{\theta}) = \int_{\xi\in\mathbb{R}^n} q(\xi;\boldsymbol{\theta})\, d\xi$$

  where the calculation of $Z(\boldsymbol{\theta})$ is intractable, even by numerical methods.

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

## 2 Estimation by Score Matching

- Define

$$\Psi(\xi;\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial \log p(\xi;\boldsymbol{\theta})}{\partial \xi_1} \\ \vdots \\ \frac{\partial \log p(\xi;\boldsymbol{\theta})}{\partial \xi_n} \end{bmatrix} = \begin{bmatrix} \Psi_1(\xi;\boldsymbol{\theta}) \\ \vdots \\ \Psi_n(\xi;\boldsymbol{\theta}) \end{bmatrix} = \nabla_\xi \log p(\xi;\boldsymbol{\theta}).$$

  We call this the score function, although according to the conventional definition, it is actually the gradient with respect the parameters.

- Observe that

$$\Psi(\xi;\boldsymbol{\theta}) = \nabla_\xi \log p(\xi;\boldsymbol{\theta}) = \nabla_\xi \log q(x;\boldsymbol{\theta}) - \nabla_\xi \log Z(\boldsymbol{\theta}) = \nabla_\xi \log q(\xi;\boldsymbol{\theta}). \tag{1}$$

- Likewise, we denote by

$$\Psi_{\mathbf{x}}(\cdot) = \nabla_\xi \log p_{\mathbf{x}}(\cdot)$$

  the score function of the distribution of observed data $\mathbf{x}$.

- We now propose that the model is estimated by minimizing the expected squared distance between the model score function $\Psi(\cdot;\boldsymbol{\theta})$ and the data score function $\Psi_{\mathbf{x}}(\cdot)$.

$$J(\boldsymbol{\theta}) = \frac{1}{2} \int_{\xi \in \mathbb{R}^n} p_{\mathbf{x}}(\xi) \|\Psi(\xi;\boldsymbol{\theta}) - \Psi_{\mathbf{x}}(\xi)\|^2 \, d\xi. \tag{2}$$

- Our *score matching* estimator of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}).$$

- The motivation for this estimator is that the score function can be directly computed from $q$ as in (1), and we do not need to compute $Z$.
- However, this may still seem to be a very difficult way of estimation $\boldsymbol{\theta}$, since we might have to compute an estimator of the data score function $\Psi_{\mathbf{x}}$ from the observed sample, which is basically a non-parametric estimation problem.
- No such non-parametric estimation is needed, because we can use a simple trick of partial integration to compute the objective function very easily.

**Theorem 1** *Assume that*

1. *the model score function* $\Psi(\xi;\boldsymbol{\theta})$ *is differentiable,*
2. *the data pdf* $p_{\mathbf{x}}(\xi)$ *is differentiable,*
3. *the expectations* $\mathbb{E}_{\mathbf{x}}[\|\Psi(\mathbf{x};\boldsymbol{\theta})\|^2]$ *and* $\mathbb{E}_{\mathbf{x}}[\|\Psi_{\mathbf{x}}(\mathbf{x})\|^2]$ *are finite for any* $\boldsymbol{\theta}$*,*
4. $p_{\mathbf{x}}(\xi)\Psi(\xi;\boldsymbol{\theta}) \to 0$ *as* $\|\xi\| \to \infty$*.*

*Then, the objective function in* (2) *can be expressed as*

$$J(\boldsymbol{\theta}) = \int_{\xi\in\mathbb{R}^n} p_{\mathbf{x}}(\xi) \sum_{i=1}^{n} \left[\partial_i \Psi_i(\xi;\boldsymbol{\theta}) + \frac{1}{2}\Psi_i(\xi;\boldsymbol{\theta})^2\right] d\xi + C \tag{3}$$

*where* $C$ *is a constant independent of* $\boldsymbol{\theta}$*, and*

$$\partial_i \Psi_i(\xi;\boldsymbol{\theta}) = \frac{\partial \Psi_i(\xi;\boldsymbol{\theta})}{\partial \xi_i} = \frac{\partial^2 \log q(\xi;\boldsymbol{\theta})}{\partial \xi_i^2}.$$

*Proof.* Definition (2) gives

$$\begin{aligned} J(\boldsymbol{\theta}) &= \int p_{\mathbf{x}}(\xi) \left[\frac{1}{2}\|\Psi_{\mathbf{x}}(\xi)\|^2 + \frac{1}{2}\|\Psi(\xi;\boldsymbol{\theta})\|^2 - \Psi_{\mathbf{x}}(\xi)^\top \Psi(\xi;\boldsymbol{\theta})\right] d\xi \\ &= \int p_{\mathbf{x}}(\xi) \left[\frac{1}{2}\|\Psi_{\mathbf{x}}(\xi)\|^2\right] d\xi + \int p_{\mathbf{x}}(\xi) \left[\frac{1}{2}\|\Psi(\xi;\boldsymbol{\theta})\|^2\right] d\xi + \int p_{\mathbf{x}}(\xi) \left[-\Psi_{\mathbf{x}}(\xi)^\top \Psi(\xi;\boldsymbol{\theta})\right] d\xi. \end{aligned}$$

This factorization uses assumption 3. The first term is independent of $\boldsymbol{\theta}$, so we set it as a constant $C$. Also,

$$\int p_{\mathbf{x}}(\xi) \left[\frac{1}{2}\|\Psi(\xi;\boldsymbol{\theta})\|^2\right] d\xi = \int p_{\mathbf{x}}(\xi) \sum_{i=1}^{n} \left[\frac{1}{2}\Psi_i(\xi;\boldsymbol{\theta})^2\right] d\xi$$

Finally, by assumptions 1, 2, and 4,

$$\begin{aligned} &\int p_{\mathbf{x}}(\xi) \left[-\Psi_{\mathbf{x}}(\xi)^\top \Psi(\xi;\boldsymbol{\theta})\right] d\xi \\ &= \int p_{\mathbf{x}}(\xi) \sum_{i=1}^{n} \left[-\frac{\partial \log p_{\mathbf{x}}(\xi)}{\partial \xi_i} \cdot \Psi_i(\xi;\boldsymbol{\theta})\right] d\xi \\ &= \int \sum_{i=1}^{n} \left[-\frac{\partial p_{\mathbf{x}}(\xi)}{\partial \xi_i} \cdot \Psi_i(\xi;\boldsymbol{\theta})\right] d\xi \\ &= -\sum_{i=1}^{n} \left[\lim_{\xi_i\to\infty} p_{\mathbf{x}}(\xi)\cdot\Psi_i(\xi;\boldsymbol{\theta}) - \lim_{\xi_i\to-\infty} p_{\mathbf{x}}(\xi)\cdot\Psi_i(\xi;\boldsymbol{\theta})\right] + \int \sum_{i=1}^{n} p_{\mathbf{x}}(\xi)\cdot\partial_i\Psi_i(\xi;\boldsymbol{\theta})\, d\xi \\ &= \int p_{\mathbf{x}}(\xi) \sum_{i=1}^{n} \partial_i \Psi_i(\xi;\boldsymbol{\theta})\, d\xi \end{aligned}$$

using integration by parts. This concludes the proof. □

- In practice, we have $T$ observations of the random vector $\mathbf{x}$, denoted by $\mathbf{x}(1),\ldots,\mathbf{x}(T)$. The sample version of $J$ is obvious obtained from (3) as

$$\tilde{J}(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n} \left[\partial_i \Psi_i(\mathbf{x}(t);\boldsymbol{\theta}) + \frac{1}{2}\Psi_i(\mathbf{x}(t);\boldsymbol{\theta})^2\right] + C \tag{4}$$

  which is asymptotically equivalent to $J$ due to the law of large numbers.

**Theorem 2** *Assume that*

1. $p_{\mathbf{x}}(\cdot) = p(\cdot; \boldsymbol{\theta}^*)$ *for some* $\boldsymbol{\theta}^*$,
2. *no other parameter value gives a pdf that is equal to* $p(\cdot; \boldsymbol{\theta}^*)$,
3. $q(\xi; \boldsymbol{\theta}) > 0$ *for all* $\xi$, $\boldsymbol{\theta}$.

*We then have*

$$J(\boldsymbol{\theta}) = 0 \iff \boldsymbol{\theta} = \boldsymbol{\theta}^*.$$

*Proof.* Assume $J(\boldsymbol{\theta}) = 0$. Then, the assumption $q > 0$ implies $p_{\mathbf{x}}(\xi) > 0$ for all $\xi$, which implies that $\Psi_{\mathbf{x}}(\cdot)$ and $\Psi(\cdot; \boldsymbol{\theta})$ are equal. This implies $\log p_{\mathbf{x}}(\cdot) = \log p(\cdot; \boldsymbol{\theta}) + c$ for some constant $c$. But $c$ is necessarily 0 because both $p_{\mathbf{x}}$ and $p(\cdot; \boldsymbol{\theta})$ are pdfs. Thus, $p_{\mathbf{x}} = p(\cdot; \boldsymbol{\theta})$. By assumption, only $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ fulfills this equality, so necessarily $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, and we have proven the implication from left to right. The converse is trivial. □

**Corollary 3** *Under the assumptions of the preceding Theorems, the score matching estimator obtained by minimization of* $\tilde{J}$ *is consistent, i.e., it converges in probability towards the true value of* $\boldsymbol{\theta}$ *when the sample size approaches infinity, assuming that the optimization algorithm is able to find the global minimum.*

*Proof.* As sample size approaches infinity, $\tilde{J}$ converges to $J$ in probability by the law of large numbers. Thus, the estimator converges to a point where $J$ is globally minimized. By Theorem 2, the global minimum is unique and found at the true parameter value (obviously, $J$ cannot be negative). □

- The result of consistency assumes that the global minimum of $\tilde{J}$ is found by the optimization algorithm used in the estimation. In practice, this may not be true, in particular because there may be several local minima. Then, the consistency is of local nature, i.e., the estimator is consistent if the optimization iteration is started sufficiently close to the true value.
- Note that consistency implies asymptotic unbiasedness.