Score Matching

Kaibo Tang

July 2, 2024

1 Introduction

Estimation by score matching[1] was introduced by Hyvärinen in 2005. It offers an alternative to traditional Markov Chain Monte Carlo methods for estimating models where the probability density function (pdf) is known only up to a multiplicative normalization constant.

Suppose $\mathbf{x} \in \mathbb{R}^n$ whose pdf is given by $p_{\mathbf{x}}(\cdot)$. We want want to estimate $\boldsymbol{\theta}$ from $\mathbf{x}$ such that the estimate $\hat{\boldsymbol{\theta}}$ allows us to approximate $p_{\mathbf{x}}(\cdot)$ by $p(\cdot; \hat{\boldsymbol{\theta}})$. Assume that we can only compute the pdf up to a multiplicative normalization constant given by $Z(\boldsymbol{\theta})$:

$$p(\boldsymbol{\xi}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} q(\boldsymbol{\xi}; \boldsymbol{\theta}). \tag{1}$$

2 Score Matching Estimator

The author first introduces the concept of score functions. The score function of the model pdf $\boldsymbol{\psi}$ is given by

$$\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \ln p(\boldsymbol{\xi}; \boldsymbol{\theta}), \tag{2}$$

and the score function of the underlying data pdf $\boldsymbol{\psi}_{\mathbf{x}}$ is given by

$$\boldsymbol{\psi}_{\mathbf{x}}(\cdot) = \nabla_{\boldsymbol{\xi}} \ln p_{\mathbf{x}}(\cdot). \tag{3}$$

[1] jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf

The score matching estimator of $\boldsymbol{\theta}$ is then given by the one that *minimizes the expected squared distance between* $\boldsymbol{\psi}(\cdot, \boldsymbol{\theta})$ *and* $\boldsymbol{\psi}_{\mathbf{x}}(\cdot)$, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}), \tag{4}$$

where

$$J(\boldsymbol{\theta}) = \frac{1}{2} \int_{\boldsymbol{\xi} \in \mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \|\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) - \boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})\|_2^2 d\boldsymbol{\xi}. \tag{5}$$

In the paper, the author proves two theorems that are central to the understanding of the method. Here, I will provide the theorems and brief proofs.

Theorem 2.1. *Assume the model score function* $\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta})$ *and the data pdf* $p_{\mathbf{x}}(\boldsymbol{\xi})$ *are differentiable. In addition, assume*

1. *the expectations* $E_{\mathbf{x}}(\|\boldsymbol{\psi}(\mathbf{x}; \boldsymbol{\theta})\|_2^2)$ *and* $E_{\mathbf{x}}(\|\boldsymbol{\psi}_{\mathbf{x}}(\mathbf{x})\|_2^2)$ *are finite* $\forall \boldsymbol{\theta}$*, and*
2. $\forall \boldsymbol{\theta}$*,* $p_{\mathbf{x}}(\boldsymbol{\xi}) \boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) \to 0$ *as* $\|\boldsymbol{\xi}\|_2 \to \infty$*.*

Then, the objective function in (5) can be written as

$$J(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi} \in \mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \sum_{i=1}^{n} \left[\partial_i \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) + \frac{1}{2} \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta})^2 \right] d\boldsymbol{\xi} + \text{const}, \tag{6}$$

where $\psi_i(\cdot; \boldsymbol{\theta})$ *is the* i*-th component of the model score function, and* $\partial_i \psi_i(\cdot; \boldsymbol{\theta})$ *is the partial of* $\psi_i(\cdot; \boldsymbol{\theta})$ *with respect to* ξ_i*.*

Proof. Starting from (5), we have

$$J(\boldsymbol{\theta}) = \int_{\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \left[\frac{1}{2} (\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}))^2 + \frac{1}{2} (\boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi}))^2 - \boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})^T \boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) \right] d\boldsymbol{\xi}. \tag{7}$$

Notice that the first term in the bracket in (7) becomes the second term under the summation in (6), and the second term in the bracket in (7) does not depend on $\boldsymbol{\theta}$ and can thus be discarded. Hence, we only need to show

$$-\int_{\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})^T \boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) d\boldsymbol{\xi} = \int_{\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \sum_{i=1}^{n} \partial_i \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) d\boldsymbol{\xi}. \tag{8}$$

Notice that by the definition of $\boldsymbol{\psi}_{\mathbf{x}}$ and chain rule, we have

$$\psi_{\mathbf{x},i}(\boldsymbol{\xi}) = \frac{\partial \ln p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_i} = \frac{1}{p_{\mathbf{x}}(\boldsymbol{\xi})} \frac{\partial p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_i}. \tag{9}$$

Using (9), we can rewrite the LHS of (8) as following:

$$\text{LHS} = -\sum_{i=1}^{n} \int_{\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \psi_{\mathbf{x},i}(\boldsymbol{\xi}) \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) d\boldsymbol{\xi} \tag{10}$$

$$= -\sum_{i=1}^{n} \int_{\mathbb{R}^n} \frac{\partial p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_i} \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) d\boldsymbol{\xi}. \tag{11}$$

By integration by parts, we have

$$\int_{\mathbb{R}} \frac{\partial p_{\mathbf{x}}(\boldsymbol{\xi})}{\partial \xi_i} \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) d\xi_i + \int_{\mathbb{R}} \frac{\partial \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta})}{\partial \xi_i} p_{\mathbf{x}}(\boldsymbol{\xi}) d\xi_i = p_{\mathbf{x}}(\boldsymbol{\xi}) \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) |_{\xi_i = -\infty}^{\xi_i = \infty}. \tag{12}$$

Notice that the RHS of (12) vanishes by assumption 2 since $\|\boldsymbol{\xi}\|_2 \to \infty$ as $|\xi_i| \to \infty$. Lastly, recalling Fubini's theorem, using (12) and following from (11), we have

$$\text{LHS} = \sum_{i=1}^{n} \int_{\mathbb{R}^n} \frac{\partial \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta})}{\partial \xi_i} p_{\mathbf{x}}(\boldsymbol{\xi}) d\boldsymbol{\xi} = \int_{\mathbb{R}^n} p_{\mathbf{x}}(\boldsymbol{\xi}) \sum_{i=1}^{n} \partial_i \psi_i(\boldsymbol{\xi}; \boldsymbol{\theta}) d\boldsymbol{\xi} = \text{RHS}. \tag{13}$$

□

Remark 2.1.1. *In the proof given by the author,*[2] *the "multivariate version of such partial integration" that the author claim to use is in fact just the usual integration by parts in the single variable case. But notice how, in the proof provided here, Fubini's theorem is used to apply this trick for single variable case to the multivariate case.*

Remark 2.1.2. *In practice, if we sample sufficient observations of* $\mathbf{x}$ *denoted by* $\mathbf{x}(1), \cdots, \mathbf{x}(T)$*. Then the objective in (6) can be approximated by*

$$J_T(\boldsymbol{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} \left[\partial_i \psi_i(\mathbf{x}(t); \boldsymbol{\theta}) + \frac{1}{2} \psi_i(\mathbf{x}(t); \boldsymbol{\theta})^2 \right] + \text{const.} \tag{14}$$

This is true due to the law of large numbers, which, in this case, states that, for a fixed $\boldsymbol{\theta}$*, the sample average* $J_T(\boldsymbol{\theta})$ *becomes arbitrarily close to the expected value* $J(\boldsymbol{\theta})$ *for* T *sufficiently large. As we shall see later in Theorem 2.4, under additional assumptions, this convergence becomes uniform in* $\boldsymbol{\theta}$*.*

[2]See Appendix A, pp.706-708.

Remark 2.1.3 (Motivation of Score Matching)**.** *One of the major motivations of score matching is to drop the normalization constant $Z(\boldsymbol{\theta})$. Indeed,*

$$\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \ln p(\boldsymbol{\xi};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \ln q(\boldsymbol{\xi};\boldsymbol{\theta}) + \nabla_{\boldsymbol{\xi}} \ln \frac{1}{Z(\boldsymbol{\theta})} = \nabla_{\boldsymbol{\xi}} \ln q(\boldsymbol{\xi};\boldsymbol{\theta}). \quad (15)$$

Hence, we have

$$\psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) = \frac{\partial \ln q(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_i} \text{ and } \partial_i \psi_i(\boldsymbol{\xi};\boldsymbol{\theta}) = \frac{\partial \psi_i(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_i} = \frac{\partial^2 \ln q(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_i^2}. \quad (16)$$

Notice that $Z(\boldsymbol{\theta})$ does not appear at all in the objective function in (6). Nevertheless, as is shown in Theorem 2.2, without having to explicitly compute $Z(\boldsymbol{\theta})$, the estimator in (4) is well-defined.

Theorem 2.2 (Well-definedness)**.** *Assume that there exists a unique $\boldsymbol{\theta}^*$ such that the pdf of $\mathbf{x}$ follows the model parameterized by $\boldsymbol{\theta}^*$, i.e., $p_{\mathbf{x}}(\cdot) = p(\cdot;\boldsymbol{\theta}^*)$ almost everywhere (a.e.).*[3] *Here, we say the solution is unique if*

$$p(\cdot;\tilde{\boldsymbol{\theta}}^*) = p(\cdot;\boldsymbol{\theta}^*) \text{ a.e. } \Longrightarrow \tilde{\boldsymbol{\theta}}^* = \boldsymbol{\theta}^*. \quad (17)$$

In addition, assume that $q(\boldsymbol{\xi};\boldsymbol{\theta}) > 0 \; \forall \boldsymbol{\xi}, \boldsymbol{\theta}$. Then

$$J(\boldsymbol{\theta}) = 0 \iff \boldsymbol{\theta} = \boldsymbol{\theta}^*. \quad (18)$$

Proof. [$\Longrightarrow$] Assume $J(\boldsymbol{\theta}) = 0$. Since $Z(\boldsymbol{\theta}^*) > 0$ and $q(\boldsymbol{\xi};\boldsymbol{\theta}^*) > 0 \; \forall \boldsymbol{\xi}$, we have

$$p_{\mathbf{x}}(\boldsymbol{\xi}) = p(\boldsymbol{\xi},\boldsymbol{\theta}^*) = \frac{1}{Z(\boldsymbol{\theta}^*)} q(\boldsymbol{\xi};\boldsymbol{\theta}^*) > 0 \text{ a.e.} \quad (19)$$

Since $J(\boldsymbol{\theta}) = 0$, we have $\boldsymbol{\psi}_{\mathbf{x}}(\cdot) = \boldsymbol{\psi}(\cdot;\boldsymbol{\theta})$ a.e., which implies that $\ln p_{\mathbf{x}}(\cdot)$ and $\ln p(\cdot;\boldsymbol{\theta})$ differ only by an additive constant. But since both $p_{\mathbf{x}}(\cdot)$ and $p(\cdot;\boldsymbol{\theta})$ are pdf's, the constant has to be 0. Hence, $p(\cdot;\boldsymbol{\theta}) = p_{\mathbf{x}}(\cdot) = p(\cdot;\boldsymbol{\theta}^*)$ a.e. Lastly, by uniqueness, we have $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

[$\Longleftarrow$] The converse is trivial. □

Lastly, in this note, a detailed proof of Corollary 3 in the article[4] is provided, along with some additional mild regularity assumptions that guarantees the consistency of the score matching estimator obtained by minimization of J_T in (14).

[3] With respect to the Lebesgue measure.

[4] See p.698.

Lemma 2.3. *Let $(\mathbf{x}_n)$ be a bounded sequence in $\mathbb{R}^n$. Then, $(\mathbf{x}_n)$ is convergent $\iff$ every convergent subsequence of $(\mathbf{x}_n)$ converges to the same limit.*

Proof. [$\Longleftarrow$] We prove the contrapositive. Suppose $\mathbf{x}_n$ does not converge to $\mathbf{x}$, i.e., $\exists \epsilon > 0$ s.t. $|\mathbf{x}_n - \mathbf{x}| \geq \epsilon \ \forall n \in \mathbb{N}$. By Bolzano-Weierstrass theorem, we can construct a convergent subsequence of $(\mathbf{x}_n)$ say $(\mathbf{x}_{n_k})$ such that $\mathbf{x}_{n_k}$ does not converge to $\mathbf{x}$.

[$\Longrightarrow$] The converse is trivial. □

Theorem 2.4 (Corollary 3)**.** *Assume $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^m$ compact. Let $f : \mathbb{R}^n \times \boldsymbol{\Theta} \to \mathbb{R}$ be given by*

$$f(\boldsymbol{\xi}, \boldsymbol{\theta}) := \frac{1}{2}\|\boldsymbol{\psi}(\boldsymbol{\xi}; \boldsymbol{\theta}) - \boldsymbol{\psi}_{\mathbf{x}}(\boldsymbol{\xi})\|_2^2. \tag{20}$$

In addition to the assumptions made in Theorem 2.1 and 2.2, assume the following:

1. *$\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, $f(\boldsymbol{\xi}, \boldsymbol{\theta})$ is Lebesgue measurable and continuous for almost all $\boldsymbol{\xi} \in \mathbb{R}^n$, and*

2. *$\exists$ dominating function $d : \mathbb{R}^n \to \mathbb{R}$ s.t. $E_{\boldsymbol{\xi}}[d(\boldsymbol{\xi})] < \infty$ and*

$$\|f(\boldsymbol{\xi}, \boldsymbol{\theta})\| \leq d(\boldsymbol{\xi}) \ \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{21}$$

Then, the score matching estimator obtained by minimization of J_T in (14) of Remark 2.1.2 is consistent, i.e.,

$$\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}^* \text{ as } T \to \infty \tag{22}$$

where $\hat{\boldsymbol{\theta}}_T$ is the estimator obtained by minimization of J_T:

$$\hat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}} J_T(\boldsymbol{\theta}), \tag{23}$$

and we recall $\boldsymbol{\theta}^*$ is the unique parameter such that $p_{\mathbf{x}}(\cdot) = p(\cdot; \boldsymbol{\theta}^*)$ a.e. as in Theorem 2.2.

Proof. From assumptions 1 and 2, along with the compactness of $\boldsymbol{\Theta}$, by the uniform law of large numbers,[5] we have $J_T \to J$ uniformly in $\boldsymbol{\theta}$.

[5] en.wikipedia.org/wiki/Law_of_large_numbers#Uniform_laws_of_large_numbers

Consider an arbitrary subsequence $(\hat{\boldsymbol{\theta}}_{T_k})$ of the sequence $(\hat{\boldsymbol{\theta}}_T)$ s.t.

$$\hat{\boldsymbol{\theta}}_{T_k} \to \hat{\boldsymbol{\theta}}' \text{ as } k \to \infty. \tag{24}$$

By (23), we have

$$J_{T_k}(\hat{\boldsymbol{\theta}}_{T_k}) \leq J_{T_k}(\boldsymbol{\theta}) \; \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}. \tag{25}$$

Since $J_{T_k} \to J$ uniformly, we have $J_{T_k}(\boldsymbol{\theta}) \to J(\boldsymbol{\theta})$ point-wise. Additionally, by the uniform convergence of J_{T_k} and (24), we have

$$J_{T_k}(\hat{\boldsymbol{\theta}}_{T_k}) \to J(\hat{\boldsymbol{\theta}}'). \tag{26}$$

Hence, from (25) and (26), we have

$$J(\hat{\boldsymbol{\theta}}') \leq J(\boldsymbol{\theta}) \; \forall \boldsymbol{\theta} \in \boldsymbol{\Theta}, \tag{27}$$

i.e.,

$$\hat{\boldsymbol{\theta}}' = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \implies J(\hat{\boldsymbol{\theta}}') = 0 \iff \hat{\boldsymbol{\theta}}' = \boldsymbol{\theta}^*. \tag{28}$$

Therefore, we have

$$\hat{\boldsymbol{\theta}}_{T_k} \to \boldsymbol{\theta}^* \text{ as } k \to \infty. \tag{29}$$

Lastly, by Lemma 2.3, we have

$$\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}^* \text{ as } T \to \infty, \tag{30}$$

which finishes our proof. □

Remark 2.4.1. *Lastly, I would like to remind the reader what we have managed to show so far in Theorem 2.4. We demonstrate the practical significance of score matching, i.e., that the estimator obtained by minimization of J_T becomes arbitrarily close the true underlying parameter value for sample size T sufficiently large.*

3 Examples

The focus of this note is on the theoretical aspect of score matching estimator. Hence, simulations are beyond the scope of this note and are left to the readers as exercise :)