Estimating Unnormalised Models by Score Matching

Michael Gutmann

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, University of Edinburgh

Spring Semester 2019

Program

1. Basics of score matching
2. Practical objective function for score matching

Program

1. Basics of score matching
   - Basic ideas of score matching
   - Objective function that captures the basic ideas but cannot be computed
2. Practical objective function for score matching

Problem formulation

- We want to estimate the parameters $\boldsymbol{\theta}$ of a parametric statistical model for a random vector $\mathbf{x} \in \mathbb{R}^d$.
- Given: iid data $\mathbf{x}_i, \ldots, \mathbf{x}_n$ that are assumed to be observations of $\mathbf{x}$ that has pdf p_*
- Further notation: $p(\boldsymbol{\xi}; \boldsymbol{\theta})$ is the model pdf; $\boldsymbol{\xi} \in \mathbb{R}^d$ is a dummy variable.
- Assumptions:
  - Model $p(\boldsymbol{\xi}; \boldsymbol{\theta})$ is known only up the partition function

$$p(\boldsymbol{\xi}; \boldsymbol{\theta}) = \frac{\tilde{p}(\boldsymbol{\xi}; \boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \qquad Z(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi}} \tilde{p}(\boldsymbol{\xi}; \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\xi}$$

  - Functional form of $\tilde{p}$ is known (can be easily computed)
  - Partition function $Z(\boldsymbol{\theta})$ cannot be computed analytically in closed form and numerical approximation is expensive.
- Goal: Estimate the model without approximating the partition function $\mathbf{Z}(\boldsymbol{\theta})$.

Basic ideas of score matching

- Maximum likelihood estimation can be considered to find parameter values $\hat{\boldsymbol{\theta}}$ so that

$$p(\boldsymbol{\xi}; \hat{\boldsymbol{\theta}}) \approx p_*(\boldsymbol{\xi}) \qquad \text{or} \qquad \log p(\boldsymbol{\xi}; \hat{\boldsymbol{\theta}}) \approx \log p_*(\boldsymbol{\xi})$$

  (as measured by Kullback-Leibler divergence, see Barber 8.7)

- Instead of estimating the parameters $\boldsymbol{\theta}$ by matching (log) densities, score matching identifies parameter values $\hat{\boldsymbol{\theta}}$ for which the derivatives (slopes) of the log densities match

$$\nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi}; \hat{\boldsymbol{\theta}}) \approx \nabla_{\boldsymbol{\xi}} \log p_*(\boldsymbol{\xi})$$

- $\nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi}; \boldsymbol{\theta})$ does not depend on the partition function:

$$\nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \left[\log \tilde{p}(\boldsymbol{\xi}; \boldsymbol{\theta}) - \log Z(\boldsymbol{\theta})\right] = \nabla_{\boldsymbol{\xi}} \log \tilde{p}(\boldsymbol{\xi}; \boldsymbol{\theta})$$

The score function (in the context of score matching)

- Define the model score function $\mathbb{R}^d \to \mathbb{R}^d$ as

$$\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \log p(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_1} \\ \vdots \\ \frac{\partial \log p(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_d} \end{pmatrix} = \nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi};\boldsymbol{\theta})$$

  While defined in terms of $p(\boldsymbol{\xi};\boldsymbol{\theta})$, we also have

$$\boldsymbol{\psi}(\boldsymbol{\xi};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \log \tilde{p}(\boldsymbol{\xi};\boldsymbol{\theta})$$

- Similarly, define the data score function as

$$\boldsymbol{\psi}_*(\boldsymbol{\xi}) = \nabla_{\boldsymbol{\xi}} \log p_*(\boldsymbol{\xi})$$

Definition of the SM objective function

- Estimate $\boldsymbol{\theta}$ by minimising a distance between model score function $\psi(\boldsymbol{\xi};\boldsymbol{\theta})$ and score function of observed data $\psi_*(\boldsymbol{\xi})$

$$
\begin{aligned}
J_{\mathrm{sm}}(\boldsymbol{\theta}) &= \frac{1}{2}\int_{\boldsymbol{\xi}\in\mathbb{R}^m} p_*(\boldsymbol{\xi})\|\psi(\boldsymbol{\xi};\boldsymbol{\theta}) - \psi_*(\boldsymbol{\xi})\|^2 d\boldsymbol{\xi} \\
&= \frac{1}{2}\mathbb{E}_*\|\psi(\mathbf{x};\boldsymbol{\theta}) - \psi_*(\mathbf{x})\|^2 \qquad (\mathbf{x}\sim p_*)
\end{aligned}
$$

  where $\mathbb{E}_*$ denotes the expectation $\mathbb{E}_{p_*}$ with respect to p_*
- Since $\psi(\boldsymbol{\xi};\boldsymbol{\theta}) = \nabla_{\boldsymbol{\xi}} \log \tilde{p}(\boldsymbol{\xi};\boldsymbol{\theta})$ does not depend on $Z(\boldsymbol{\theta})$ there is no need to compute the partition function.
- Knowing the unnormalised model $\tilde{p}(\boldsymbol{\xi};\boldsymbol{\theta})$ is enough.
- Expectation $\mathbb{E}_*$ with respect to p_* can be approximated as sample average over the observed data, but what about ψ_*?

Program

1. Basics of score matching
   - Basic ideas of score matching
   - Objective function that captures the basic ideas but cannot be computed

2. Practical objective function for score matching

Program

1. Basics of score matching

2. Practical objective function for score matching
   - Integration by parts to obtain a computable objective function
   - Simple example

Reformulation of the SM objective function

- In the objective function we have the score function of the data distribution $\boldsymbol{\psi}_*$. How to compute it?
- In fact, no need to compute it because the score matching objective function J_{sm} can be expressed as

$$J_{\mathrm{sm}}(\boldsymbol{\theta}) = \mathbb{E}_* \sum_{j=1}^{d} \left[\partial_j \psi_j(\mathbf{x}; \boldsymbol{\theta}) + \frac{1}{2} \psi_j^2(\mathbf{x}; \boldsymbol{\theta}) \right] + \text{const.}$$

where the constant does not depend on $\boldsymbol{\theta}$, and

$$\psi_j(\boldsymbol{\xi}; \boldsymbol{\theta}) = \frac{\partial \log \tilde{p}(\boldsymbol{\xi}; \boldsymbol{\theta})}{\partial \xi_j} \qquad \partial_j \psi_j(\boldsymbol{\xi}; \boldsymbol{\theta}) = \frac{\partial^2 \log \tilde{p}(\boldsymbol{\xi}; \boldsymbol{\theta})}{\partial \xi_j^2}$$

Proof (general idea)

- Use Euclidean distance and expand the objective function J_{sm}

$$
\begin{aligned}
J_{\mathrm{sm}}(\boldsymbol{\theta}) &= \frac{1}{2}\mathbb{E}_* \|\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta}) - \boldsymbol{\psi}_*(\mathbf{x})\|^2 \\
&= \frac{1}{2}\mathbb{E}_* \|\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})\|^2 - \mathbb{E}_* \left[\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})^\top \boldsymbol{\psi}_*(\mathbf{x})\right] + \frac{1}{2}\mathbb{E}_* \|\boldsymbol{\psi}_*(\mathbf{x})\|^2 \\
&= \frac{1}{2}\mathbb{E}_* \|\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})\|^2 - \sum_{j=1}^{d} \mathbb{E}_* \left[\psi_j(\mathbf{x};\boldsymbol{\theta})\psi_{*,j}(\mathbf{x})\right] + \mathrm{const}
\end{aligned}
$$

- First term does not depend on $\boldsymbol{\psi}_*$. The ψ_j and $\psi_{*,j}$ are the j-th elements of the vectors $\boldsymbol{\psi}$ and $\boldsymbol{\psi}_*$, respectively. Constant does not depend on $\boldsymbol{\theta}$.
- The trick is to use integration by parts for the second term to get an objective function which does not involve $\boldsymbol{\psi}_*$.

Proof (not examinable)

$$
\begin{aligned}
\mathbb{E}_* \left[\psi_j(\mathbf{x};\boldsymbol{\theta})\psi_{*,j}(\mathbf{x})\right] &= \int_{\boldsymbol{\xi}} p_*(\boldsymbol{\xi})\psi_{*,j}(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})d\boldsymbol{\xi} \\
&= \int_{\boldsymbol{\xi}} p_*(\boldsymbol{\xi})\frac{\partial \log p_*(\boldsymbol{\xi})}{\partial \xi_j}\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})d\boldsymbol{\xi} \\
&= \prod_{k\neq j}\int_{\xi_k}\left(\int_{\xi_j} p_*(\boldsymbol{\xi})\frac{\partial \log p_*(\boldsymbol{\xi})}{\partial \xi_j}\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})d\xi_j\right)d\xi_k \\
&= \prod_{k\neq j}\int_{\xi_k}\left(\int_{\xi_j} \frac{\partial p_*(\boldsymbol{\xi})}{\partial \xi_j}\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})d\xi_j\right)d\xi_k
\end{aligned}
$$

Use integration by parts

$$
\begin{aligned}
\int_{\xi_j} \frac{\partial p_*(\boldsymbol{\xi})}{\partial \xi_j}\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})d\xi_j &= \left[p_*(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})\right]_{a_j}^{b_j} - \int_{\xi_j} p_*(\boldsymbol{\xi})\frac{\partial \psi_j(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_j}d\xi_j \\
&= -\int_{\xi_j} p_*(\boldsymbol{\xi})\frac{\partial \psi_j(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial \xi_j}d\xi_j,
\end{aligned}
$$

where the a_j and b_j specify the boundaries of the data pdf p_* along dimension j and where we assume that $\left[p_*(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})\right]_{a_j}^{b_j} = 0$.

Proof (not examinable)

If $[p_*(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})]_{a_j}^{b_j} = 0$

$$\begin{aligned}
\mathbb{E}_* \left[\psi_j(\mathbf{x};\boldsymbol{\theta})\psi_{*,j}(\mathbf{x})\right] &= -\prod_{k\neq j}\int_{\xi_k}\left(\int_{\xi_j} p_*(\boldsymbol{\xi})\frac{\partial\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial\xi_j}d\xi_j\right)d\xi_k \\
&= -\int_{\boldsymbol{\xi}} p_*(\boldsymbol{\xi})\frac{\partial\psi_j(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial\xi_j}d\boldsymbol{\xi} \\
&= -\mathbb{E}_*\left[\partial_j\psi_j(\mathbf{x};\boldsymbol{\theta})\right]
\end{aligned}$$

so that

$$\begin{aligned}
J_{\mathrm{sm}}(\boldsymbol{\theta}) &= \frac{1}{2}\mathbb{E}_*\|\boldsymbol{\psi}(\mathbf{x};\boldsymbol{\theta})\|^2 - \sum_{j=1}^{d} -\mathbb{E}_*\left[\partial_j\psi_j(\mathbf{x};\boldsymbol{\theta})\right] + \text{const} \\
&= \mathbb{E}_*\sum_{j=1}^{d}\left[\partial_j\psi_j(\mathbf{x};\boldsymbol{\theta}) + \frac{1}{2}\psi_j^2(\mathbf{x};\boldsymbol{\theta})\right] + \text{const}
\end{aligned}$$

Replacing the expectation / integration over the data density p_* by a sample average over the observed data gives a computable objective function for score matching.

Final method of score matching

- Given iid data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the score matching estimate is

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \left[\partial_j \psi_j(\mathbf{x}_i; \boldsymbol{\theta}) + \frac{1}{2} \psi_j(\mathbf{x}_i; \boldsymbol{\theta})^2 \right]$$

  ψ_j is the partial derivative of the log unnormalised model $\log \tilde{p}$ with respect to the j-th coordinate (slope) and $\partial_j \psi_j$ its second partial derivative (curvature).
- Parameter estimation with intractable partition functions without approximating the partition function.

Requirements

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \left[\partial_j \psi_j(\mathbf{x}_i; \boldsymbol{\theta}) + \frac{1}{2} \psi_j(\mathbf{x}_i; \boldsymbol{\theta})^2 \right]$$

Requirements:

- technical (from proof): $[p_*(\boldsymbol{\xi})\psi_j(\boldsymbol{\xi}; \boldsymbol{\theta})]_{a_j}^{b_j} = 0$, where the a_j and b_j specify the boundaries of the data pdf p_* along dimension j
- smoothness: second derivatives of $\log \tilde{p}(\boldsymbol{\xi}; \boldsymbol{\theta})$ with respect to the ξ_j need to exist, and should be smooth with respect to $\boldsymbol{\theta}$ so that $J(\boldsymbol{\theta})$ can be optimised with gradient-based methods.

Simple example

- $\tilde{p}(\xi;\theta) = \exp(-\theta\xi^2/2)$, parameter $\theta > 0$ is the precision.
- The slope and curvature of the log unnormalised model are

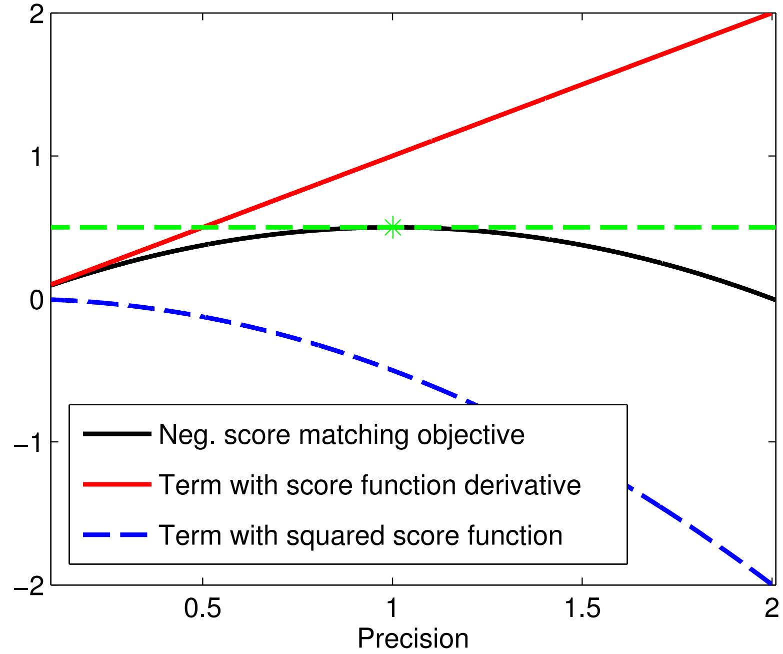
$$\psi(\xi;\theta) = \partial_\xi \log \tilde{p}(\xi;\theta) = -\theta\xi, \qquad \partial_\xi \psi(\xi;\theta) = -\theta.$$

- If p_* is Gaussian, $\lim_{\xi\to\pm\infty} p_*(\xi)\psi(\xi;\theta) = 0$ for all θ.

- Score matching objective

$$J(\theta) = -\theta + \frac{1}{2}\theta^2 \frac{1}{n}\sum_{i=1}^n x_i^2$$

$$\Rightarrow \hat{\theta} = \left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)^{-1}$$



- For Gaussians, same as the MLE.

Extensions

- Score matching as presented here only works for $\mathbf{x} \in \mathbb{R}^d$
- There are extensions for discrete and non-negative random variables (not examinable)

  https://www.cs.helsinki.fi/u/ahyvarin/papers/CSDA07.pdf
- Can be shown to be part of a general framework to estimate unnormalised models (not examinable)

  https://michaelgutmann.github.io/assets/papers/Gutmann2011b.pdf
- Overall message: in some situations, other learning criteria than likelihood are preferable.

Program recap

1. Basics of score matching
   - Basic ideas of score matching
   - Objective function that captures the basic ideas but cannot be computed
2. Practical objective function for score matching
   - Integration by parts to obtain a computable objective function
   - Simple example