Probabilistic Modelling and Reasoning Solutions for Tutorial 7 — Additional Practice

Spring 2019
Michael Gutmann

Exercise 1. *Score matching for the exponential family*

In the lecture, we have derived the objective function $J(\boldsymbol{\theta})$ for score matching,

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

where ψ_j is the partial derivative of the log model-pdf $\log p(\mathbf{x};\boldsymbol{\theta})$ with respect to the j-th coordinate (slope) and $\partial_j\psi_j$ its second partial derivative (curvature). The observed data are denoted by $\mathbf{x}_1,\ldots,\mathbf{x}_n$ and $\mathbf{x}\in\mathbb{R}^m$.

The goal of this exercise is to show that for statistical models of the form

$$\log p(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K}\theta_k F_k(\mathbf{x}) - \log Z(\boldsymbol{\theta}), \qquad \mathbf{x}\in\mathbb{R}^m, \tag{2}$$

the score matching objective function becomes a quadratic form, which can be optimised efficiently (see e.g. Barber Appendix A.5.3).

The set of models above are called the (continuous) exponential family, or also log-linear models because the models are linear in the parameters θ_k. Since the exponential family generally includes probability mass functions as well, the qualifier "continuous" may be used to highlight that we are here considering continuous random variables only. The functions $F_k(\mathbf{x})$ are assumed to be known; they are the sufficient statistics (see e.g. Barber Section 8.5).

(a) Denote by $\mathbf{K}(\mathbf{x})$ the matrix with elements $K_{kj}(\mathbf{x})$,

$$K_{kj}(\mathbf{x}) = \frac{\partial F_k(\mathbf{x})}{\partial x_j}, \qquad k=1\ldots K, \quad j=1\ldots m, \tag{3}$$

and by $\mathbf{H}(\mathbf{x})$ the matrix with elements $H_{kj}(\mathbf{x})$,

$$H_{kj}(\mathbf{x}) = \frac{\partial^2 F_k(\mathbf{x})}{\partial x_j^2}, \qquad k=1\ldots K, \quad j=1\ldots m. \tag{4}$$

Furthermore, let $\mathbf{h}_j(\mathbf{x}) = (H_{1j}(\mathbf{x}),\ldots,H_{Kj}(\mathbf{x}))^\top$ be the j–th column vector of $\mathbf{H}(\mathbf{x})$.
Show that for the continuous exponential family, the score matching objective in Equation (1) becomes

$$J(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top\mathbf{r} + \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{M}\boldsymbol{\theta}, \tag{5}$$

where

$$\mathbf{r} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{h}_j(\mathbf{x}_i), \qquad \mathbf{M} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top. \tag{6}$$

Solution. For

$$\log p(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{K}\theta_k F_k(\mathbf{x}) - \log Z(\boldsymbol{\theta}) \tag{S.1}$$

the first derivative with respect to x_j, the j-th element of $\mathbf{x}$, is

$$\psi_j(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial x_j} \tag{S.2}$$

$$= \sum_{k=1}^{K} \theta_k \frac{\partial F_k(\mathbf{x})}{\partial x_j} \tag{S.3}$$

$$= \sum_{k=1}^{K} \theta_k K_{kj}(\mathbf{x}). \tag{S.4}$$

The second derivative is

$$\partial_j \psi_j(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial^2 \log p(\mathbf{x};\boldsymbol{\theta})}{\partial x_j^2} \tag{S.5}$$

$$= \sum_{k=1}^{K} \theta_k \frac{\partial^2 F_k(\mathbf{x})}{\partial x_j^2} \tag{S.6}$$

$$= \sum_{k=1}^{K} \theta_k H_{kj}(\mathbf{x}), \tag{S.7}$$

which we can write more compactly as

$$\partial_j \psi_j(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{h}_j(\mathbf{x}). \tag{S.8}$$

The score matching objective in Equation (1) features the sum $\sum_j \psi_j(\mathbf{x};\boldsymbol{\theta})^2$. The term $\psi_j(\mathbf{x};\boldsymbol{\theta})^2$ equals

$$\psi_j(\mathbf{x};\boldsymbol{\theta})^2 = \left[\sum_{k=1}^{K} \theta_k K_{kj}(\mathbf{x})\right]^2 \tag{S.9}$$

$$= \sum_{k=1}^{K} \sum_{k'=1}^{K} K_{kj}(\mathbf{x}) K_{k'j}(\mathbf{x}) \theta_k \theta_{k'}, \tag{S.10}$$

so that

$$\sum_{j=1}^{m} \psi_j(\mathbf{x};\boldsymbol{\theta})^2 = \sum_{j=1}^{m} \sum_{k=1}^{K} \sum_{k'=1}^{K} K_{kj}(\mathbf{x}) K_{k'j}(\mathbf{x}) \theta_k \theta_{k'} \tag{S.11}$$

$$= \sum_{k=1}^{K} \sum_{k'=1}^{K} \theta_k \theta_{k'} \left[\sum_{j=1}^{m} K_{kj}(\mathbf{x}) K_{k'j}(\mathbf{x})\right], \tag{S.12}$$

which can be more compactly expressed using matrix notation. Noting that

$$\sum_{j=1}^{m} K_{kj}(\mathbf{x}_i) K_{k'j}(\mathbf{x}_i)$$

equals the (k, k') element of the matrix-matrix product $\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top$,

$$\sum_{j=1}^{m} K_{kj}(\mathbf{x}_i) K_{k'j}(\mathbf{x}_i) = \left[\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\right]_{k,k'}, \tag{S.13}$$

we can write

$$\sum_{j=1}^{m} \psi_j(\mathbf{x};\boldsymbol{\theta})^2 = \sum_{k=1}^{K}\sum_{k'=1}^{K} \theta_k\theta_{k'} \left[\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\right]_{k,k'} \tag{S.14}$$

$$= \boldsymbol{\theta}^\top \mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top \boldsymbol{\theta} \tag{S.15}$$

where we have used that for some matrix $\mathbf{A}$

$$\boldsymbol{\theta}^\top \mathbf{A}\boldsymbol{\theta} = \sum_{k,k'} \theta_k\theta_{k'}[\mathbf{A}]_{k,k'} \tag{S.16}$$

where $[\mathbf{A}]_{k,k'}$ is the (k,k') element of the matrix $\mathbf{A}$.
Inserting the expressions into Equation (1) gives

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\partial_j\psi_j(\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{2}\psi_j(\mathbf{x}_i;\boldsymbol{\theta})^2\right] \tag{S.17}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\partial_j\psi_j(\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\psi_j(\mathbf{x}_i;\boldsymbol{\theta})^2 \tag{S.18}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{\theta}^\top\mathbf{h}_j(\mathbf{x}_i) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}^\top\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\boldsymbol{\theta} \tag{S.19}$$

$$= \boldsymbol{\theta}^\top\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{h}_j(\mathbf{x}_i)\right] + \frac{1}{2}\boldsymbol{\theta}^\top\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{K}(\mathbf{x}_i)\mathbf{K}(\mathbf{x}_i)^\top\right]\boldsymbol{\theta} \tag{S.20}$$

$$= \boldsymbol{\theta}^\top\mathbf{r} + \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{M}\boldsymbol{\theta}, \tag{S.21}$$

which is the desired result.

(b) *The pdf of a zero mean Gaussian parametrised by the variance σ^2 is*

$$p(x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}. \tag{7}$$

The (multivariate) Gaussian is a member of the exponential family. By comparison with Equation (2), we can re-parametrise the statistical model $\{p(x;\sigma^2)\}_{\sigma^2}$ and work with

$$p(x;\theta) = \frac{1}{Z(\theta)}\exp\left(\theta x^2\right), \qquad \theta < 0, \qquad x \in \mathbb{R}, \tag{8}$$

instead. The two parametrisations are related by $\theta = -1/(2\sigma^2)$. Using the previous result on the (continuous) exponential family, determine the score matching estimate $\hat{\theta}$, and show that the corresponding $\hat{\sigma}^2$ is the same as the maximum likelihood estimate. This result is noteworthy because unlike in maximum likelihood estimation, score matching does not need the partition function $Z(\theta)$ for the estimation.

Solution. By comparison with Equation (2), the sufficient statistics $F(x)$ is x^2.

We first determine the score matching objective function. For that, we need to determine the quantities $\mathbf{r}$ and $\mathbf{M}$ in Equation (6). Here, both $\mathbf{r}$ and $\mathbf{M}$ are scalars, and so are the matrices **K** and **H** that define **r** and **M**. By their definitions, we obtain

$$K(x) = \frac{\partial F(x)}{\partial x} = 2x \tag{S.22}$$

$$H(x) = \frac{\partial^2 F(x)}{\partial x^2} = 2 \tag{S.23}$$

$$r = 2 \tag{S.24}$$

$$M = \frac{1}{n}\sum_{i=1}^{n} K(x_i)^2 \tag{S.25}$$

$$= 4m_2 \tag{S.26}$$

where m_2 denotes the second empirical moment,

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2. \tag{S.27}$$

With Equation (1), the score matching objective thus is

$$J(\theta) = 2\theta + \frac{1}{2}4m_2\theta^2 \tag{S.28}$$

$$= 2\theta + 2m_2\theta^2 \tag{S.29}$$

A necessary condition for the minimiser to satisfy is

$$\frac{\partial J(\theta)}{\partial \theta} = 2 + 4\theta m_2 \tag{S.30}$$

$$= 0 \tag{S.31}$$

The only parameter value that satisfies the condition is

$$\hat{\theta} = -\frac{1}{2m_2}. \tag{S.32}$$

The second derivative of $J(\theta)$ is

$$\frac{\partial^2 J(\theta)}{\theta^2} = m_2, \tag{S.33}$$

which is positive (as long as all data points are non-zero). Hence $\hat{\theta}$ is a minimiser.

From the relation $\theta = -1/(2\sigma^2)$, we obtain that the score matching estimate of the variance σ^2 is

$$\hat{\sigma}^2 = -\frac{1}{2\hat{\theta}} = m_2. \tag{S.34}$$

We can obtain the score matching estimate $\hat{\sigma}^2$ from $\hat{\theta}$ in this manner for the same reason that we were able to work with transformed parameters in maximum likelihood estimation.

For zero mean Gaussians, the second moment m_2 is the maximum likelihood estimate of the variance, which shows that the score matching and maximum likelihood estimate are here the same. While the two methods generally yield different estimates, the result also holds for multivariate Gaussians where the score matching estimates also equal the maximum likelihood estimates (see the original article on score matching http://jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf).