Neural Networks 98 (2018) 263–270

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Necessary and sufficient conditions of proper estimators based on self density ratio for unnormalized statistical models

Kazuyuki Hiraoka [a,*], Toshihiko Hamada [a], Gen Hori [b,c]

[a] National Institute of Technology, Wakayama College, Wakayama, 644-0023, Japan
[b] Faculty of Business Administration, Asia University, Tokyo, 180-8629, Japan
[c] Brain Science Institute, RIKEN, Saitama, 351-0198, Japan

ARTICLE INFO

Article history:
Received 28 January 2016
Received in revised form 12 April 2017
Accepted 28 November 2017
Available online 11 December 2017

Keywords:
Unnormalized statistical model
Consistency
Score matching
Non-local scoring rules
Self density ratio

ABSTRACT

The largest family of density-ratio based estimators is obtained for unnormalized statistical models under the assumption of properness. They do not require normalization of the probability density function (PDF) because they are based on the density ratio of the same PDF at different points; therefore, the multiplicative normalization constant cancels out. In contrast with most existing work, a single necessary and sufficient condition is given here, rather than merely sufficient conditions for proper criteria for estimation. The condition implies that an extended Bregman divergence framework with data-dependent noise (Gutmann & Hirayama, 2011) gives the largest family of proper criteria in the present case. This properness yields consistent estimation as long as some mild conditions are satisfied. The present study shows that the above-mentioned framework gives an "upper bound" for attempts to extend Hyvärinen's score matching and therefore provides a perspective for studies in this direction.

1. Introduction

The intractability of normalization is a major computational bottleneck in machine learning with complex statistical models. Consequently, the probability distribution is known only up to a multiplicative normalization constant in many cases. A typical example is deep learning, where contrastive divergence (CD; Hinton, 2002) is used to avoid this difficulty (Hinton, Osindero, & Teh, 2006; Roux & Bengio, 2008). Similar problems appear in various other applications (Gutmann & Hyvärinen, 2013a), including graphical models (Koller, Friedman, Getoor, & Taskar, 2007), unsupervised feature learning (Bengio, Courville, & Vincent, 2012), computational neuroscience (Gutmann & Hyvärinen, 2013b; Köster & Hyvärinen, 2010), modeling of images (Li, 2009; Rangarajan & Chellappa, 1995), natural language processing (Bengio, Ducharme, Vincent, & Jauvin, 2003), and social networks (Robins, Pattison, Kalish, & Lusher, 2007).

Approaches to the estimation of unnormalized statistical models can be classified as follows. Typical strategies used to avoid normalization are elimination, estimation, and approximation. The normalization constant is eliminated by criteria based on $(\log f)'$ or $f(v)/f(u)$ for a probability density function (PDF) $f(x)$ in the elimination strategy. In the estimation strategy, the normalization constant is regarded as an additional unknown parameter and is estimated together with the original unknown parameters in the statistical model. In the approximation strategy, f is replaced by an approximation. The algorithms can also be divided into deterministic methods and stochastic methods (here, Monte Carlo methods). Pseudorandom numbers are used in the latter methods, and so the result changes for every run. Existing methods are classified in Table 1 (Gutmann & Hyvärinen, 2013a; Sohl-Dickstein, Battaglino, & DeWeese, 2011).

Score matching (SM; Hyvärinen, 2005) is a deterministic method that avoids normalization. It uses only the ratios f'/f and f''/f in its criterion so that it sidesteps the normalization problem. Ratio matching (RM; Hyvärinen, 2007b) and generalized score matching (GSM; Lyu, 2009) provide extensions for discrete data. GSM also gives a rich family of estimators for continuous data: in this case $\mathcal{L}f/f$ is used instead of f'/f in SM for an arbitrary linear operator $\mathcal{L}$. Another wide extension of SM can be found in the generic forms of local estimators investigated in the theory of local scoring rules (LSRs; Ehm & Gneiting, 2012; Parry, Dawid, & Lauritzen, 2012).

In noise-contrastive estimation (NCE; Gutmann & Hyvärinen, 2010, 2012), the normalization constant is estimated through discrimination between the observed data and some artificially generated noise. Its extension (Pihlaja, Gutmann, & Hyvärinen, 2010) is also classified to the same category: the parameters and the normalization constant are obtained simultaneously by minimization of an objective functional. Their further extension is

* Corresponding author.
E-mail addresses: hiraoka@wakayama-nct.ac.jp (K. Hiraoka), hamada@wakayama-nct.ac.jp (T. Hamada), hori@brain.riken.jp (G. Hori).

https://doi.org/10.1016/j.neunet.2017.11.018
0893-6080/

Table 1
Approaches to estimation of unnormalized models. From left to right: (Generalized) score matching, ratio matching, local scoring rules, Bregman divergence framework with data-dependent noise, noise-contrastive estimation, mean field theory, variational Bayes techniques, pseudolikelihood, minimum probability flow learning, Monte Carlo maximum likelihood estimation, and contrastive divergence. BDF-DDN appears twice because it can be used in multiple ways.

	Elimination	Estimation	Approximation
Deterministic	(G)SM, RM,LSR, BDF-DDN		MF, VB, PL, MPF
Monte Carlo	BDF-DDN	NCE	MCML, CD

given by the Bregman divergence framework (BDF; Gutmann & Hirayama, 2011). BDF is flexible enough to be used across several of the strategies in Table 1. In particular, the normalization constant is eliminated in BDF with data-dependent noise (BDF-DDN; Gutmann & Hirayama, 2011), which gives yet another extension of SM. Though BDF-DDN is used in a Monte Carlo style in Gutmann and Hirayama (2011), it can be viewed as a sum of randomly selected deterministic criteria. Since each of these criteria yields a stand-alone deterministic estimator, BDF-DDN also appears in the "deterministic" row in Table 1.

The normalization constant is directly approximated in Monte Carlo maximum likelihood (MCML) estimation (Delman, 2002; Geyer, 1994), which can lead to an estimate with large variance unless the random sampling is designed carefully (Gutmann & Hyvärinen, 2013a; Pihlaja et al., 2010). CD approximates the gradient of the log-likelihood by the Markov chain Monte Carlo method instead of solving an intractable integration problem. The Markov chain in CD is replaced with a deterministic flow on the set of probability distributions in minimum probability flow learning (MPF; Sohl-Dickstein et al., 2011). Pseudolikelihood (PL; Besag, 1975) approximates the joint probability distribution as a computationally tractable product of conditional distributions. Mean field (MF) theory and variational Bayes (VB) techniques (Attias, 2000; Kappen & Rodríguez, 1997; Tanaka, 1998) are also in the same category. A variety of methods based on sampling and numerical integration have also been studied (Haykin, 2008).

Several conditions have been proposed for being a 'good' estimator in statistics. One basic condition is consistency (Lehmann & Casella, 1998), which is the requirement that the estimate should converge to the true value as the sample size increases. For example, maximum likelihood estimation (MLE) is well known to be consistent in general and can be used for a wide variety of applications in regular models (Lehmann & Casella, 1998). Construction of a consistent estimator is not trivial for unnormalized statistical models, however. MLE is intractable because the calculation of the normalization constant is not feasible in this case. Among the existing methods classified in Table 1, SM, NCE, LSR, GSM, and BDF-DDN are known to realize consistent estimation for unnormalized statistical models. However, theoretical analysis of CD is difficult in spite of its practical usefulness (Carreira-Perpigñán & Hinton, 2005).

In contrast to most existing work, this paper gives a single necessary and sufficient condition rather than merely sufficient conditions for proper criteria for estimation. The concept of properness is defined in Section 2.3; it yields consistent estimation as long as some mild conditions are satisfied. Though properness is a minimum requirement for usefulness, it logically implies an extended BDF-DDN from the above necessary and sufficient condition. In other words, this extended BDF-DDN gives the largest family of proper criteria under some assumptions.

Estimators based on density ratios are attracting a great deal of attention (Fishman, 1996; Sugiyama, Kawanabe, & Chui, 2010; Sugiyama, Suzuki, & Kanamori, 2012) for various statistical data processing tasks (Bickel, Bogojeska, Lengauer, & Scheffer, 2008; Hido, Tsuboi, Kashima, Sugiyama, & Kanamori, 2008; Shimodaira, 2000; Storkey & Sugiyama, 2007; Sugiyama, Krauledat, & Müller, 2007; Sugiyama, Takeuchi, Suzuki, Kanamori, & Hachiya, 2009; Suzuki & Sugiyama, 2009a, b; Suzuki, Sugiyama, Kanamori, & Sese, 2009; Suzuki, Sugiyama, Sese, & Kanamori, 2008; Zadrozny, 2004). So we focus on the deterministic elimination approach for continuous data, and we investigate estimators based on the density ratio of the same PDF $f(x)$ at different points. A typical example of such estimators is provided by minimization of the sample average over

$$M(x) = \log\left(1 + \frac{f(x+a)}{f(x)}\right) + \log\left(1 + \frac{f(x-a)}{f(x)}\right) \tag{1}$$

with a positive constant $a > 0$ (Gutmann & Hirayama, 2011).

This paper is organized as follows. Section 2 presents the definition of the unnormalized statistical model, the motivation for our approach, and the framework for estimation based on self density ratios. Section 3 gives the largest family of proper criteria for estimation in two equivalent styles: a generic form for proper criteria and the Bregman divergence framework. Section 4 concludes the paper.

The contents in Sections 2.1, 2.3, and 3.2 have been reported briefly in Hiraoka, Hamada, and Hori (2014) without mathematical proof. In this paper, we give the mathematical proofs along with a detailed discussion of our approach.

2. Background and overview

2.1. Unnormalized statistical model

We consider the estimation of an unnormalized statistical model $\mathcal{G} = (g, \Theta)$. The model $\mathcal{G}$ represents a family of PDFs for a random vector X,

$$f(x;\theta) = \frac{1}{C(\theta)} g(x;\theta), \qquad x = (x_1, \ldots, x_N) \in \mathcal{X} \subset \mathbb{R}^N$$

with a given function $g > 0$ and a vector of unknown parameters

$$\theta = (\theta_1, \ldots, \theta_K) \in \Theta \subset \mathbb{R}^K,$$

where $\mathbb{R}$ is the set of real numbers. The calculation of

$$C(\theta) = \int_{\mathcal{X}} g(x;\theta)\, dx > 0$$

is assumed to be intractable throughout this paper. Let $\theta^* \in \Theta$ be the true vector of parameters, which we aim to estimate from i.i.d. samples $x(1), \ldots, x(n) \sim f(\ \cdot\ ; \theta^*)$, without using $C(\theta)$ explicitly. In particular, we consider estimators $\hat{\theta}$ of the form

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{t} M(x(t);\theta)$$

for some penalty criterion M, which is called a scoring rule (Ehm & Gneiting, 2012; Parry et al., 2012).

Smoothness of functions is supposed implicitly and the term "almost surely (all)", which means "except for probability (measure) zero", is omitted throughout the discussion below.

2.2. Limitation of local estimators

Before describing our approach, we point out a limitation of local estimators to motivate the use of non-local estimators. We consider the one-dimensional case $N = K = 1$ for simplicity in this subsection. SM is local in the sense that its criterion

$$M^{\mathrm{SM}}(x;\theta) = \frac{\frac{\partial^2}{\partial x^2} g(x;\theta)}{g(x;\theta)} - \frac{1}{2}\left(\frac{\frac{\partial}{\partial x} g(x;\theta)}{g(x;\theta)}\right)^2$$

depends on only g and its derivatives at the observed sample x; it never uses $g(x+3)$, for example. Though this design looks natural