The Generalized Cross Entropy Method, with Applications to Probability Density Estimation

Zdravko I. Botev and Dirk P. Kroese*

The University of Queensland

Department of Mathematics
The University of Queensland
Brisbane 4072
AUSTRALIA
E-mail: {botev,kroese}@maths.uq.edu.au

Abstract: The fundamental problem in statistical learning is to determine the simplest model that explains a given set of empirical data and which uses as few assumptions as possible. Many classical approaches to the statistical learning problem impose extra (artificial) assumptions in order to provide a unique and well-behaved solution of the problem. In this paper we describe a simple and general framework for statistical modelling which unifies many recent advances in Monte Carlo statistical methods. The approach combines information-theoretic ideas based on generalized cross-entropy principles with constrained functional optimization fundamentals. The effectiveness of the approach is demonstrated through an application to density estimation and data modeling.

AMS 2000 subject classifications: Primary 94A17, 60K35; secondary 68Q32, 93E14.
Keywords and phrases: Cross entropy, information theory, Monte Carlo simulation, statistical modeling, kernel smoothing, functional optimization, bandwidth selection, calculus of variations.

1. Introduction

The main problem in *statistical learning* is to find or estimate the sparsest probability model for a given collection of empirical data with the introduction of as little extraneous information as possible. Representing the model via a probability distribution, the question is how to efficiently sample from this unknown *target* distribution. A similar problem (but simpler) occurs in *Monte Carlo simulation*, where the objective is to efficiently sample from

*Supported by the Australian Research Council,under grant number DP0558957

some target density f whose functional form is specified (usually) up to an unknown constant.

Despite its apparent simplicity, many important applications can be formulated in this framework. Examples are:

1. *Monte Carlo integration*, where the problem is to estimate integrals of the form $\int_{\mathscr{X}} H(\mathbf{x})\, d\mathbf{x}$, for an arbitrary function H and set $\mathscr{X}$. These problems can be efficiently solved by sampling from the target $f(\mathbf{x}) = c\,|H(\mathbf{x})|$, where c is an unknown normalizing constant.
2. *Rare-event simulation*, where a small probability $\ell = \mathbb{P}_h(S(\mathbf{X}) \geqslant \gamma)$ needs to be estimated, for some real-valued function S of a random variable $\mathbf{X}$ with probability density h. This problem is solved efficiently by sampling from the minimum variance importance sampling density [31] $f(\mathbf{x}) = c\, I_{\{S(\mathbf{x}) \geqslant \gamma\}}\, h(\mathbf{x})$, where I denotes the indicator function.
3. *Global maximization*, where some non-smooth or discrete multidimensional multimodal function S needs to be maximized on a set $\mathscr{X}$. Here a sequence of targets could be the uniform distributions on the level sets $\{\mathbf{x} : S(\mathbf{x}) \geqslant \gamma_t\}$, for increasing levels $\gamma_1, \gamma_2, \ldots$.

We stress again the important fact that all of the above problems can be solved efficiently provided one can estimate an optimal (in some well-defined sense) probability density from a given set of empirical data. Thus density estimation is not only an important tool for data analysis in the toolkit of the applied statistician, but is also a crucial for the performance of many Population-based Monte Carlo simulation techniques attacking the problems of Rare-event simulation, Global Nonlinear or Non-smooth Optimization and Multidimensional Integration. Statistical learning, such as probability density estimation, is typically an *ill-posed* problem, in the sense that extraneous assumptions need to be introduced for unique stable and well-behaved solutions to exist. Consider for example the continuous data $\mathcal{X}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ in Figure 1, represented by plusses. The objective is to find an optimal in some sense model for $\mathcal{X}_n$ using as few assumptions as possible, or, more specifically, to best estimate the corresponding probability density from the finite number of observations and specifications. Which is the best possible probabilistic model for the data, the uni-modal probability density function or the multi-modal one?

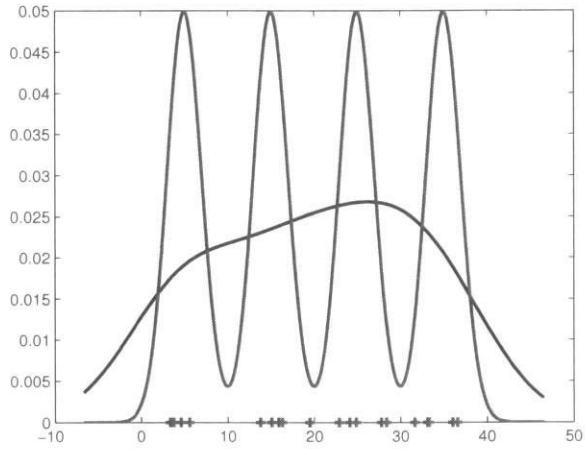

FIG 1. *Which density represents the data the best?*

There are reasons to prefer the simpler and sparser model, as the data seems not numerous enough to justify multiple modes. However, the data was in fact generated from the multimodal density. Yet the unimodal curve, representing the current state-of-the-art in density estimation [16], is not even multi-modal. This is partly what makes the problem ill-posed. We may have reasons to prefer the uni-modal curve but the multi-modal curve is also a reasonable model for the data. The data simply does not provide enough information to give a unique or well-defined solution to the density estimation problem, and extra information and assumptions need to be introduced in the model in order to obtain unique and stable solutions.

The parametric approach to statistical learning, advocated by Fisher, is to specify the model up to a small number of parameters and to estimate these optimally via the likelihood principle. A major problem with this classical paradigm is that one has to specify the probability density function *subjectively*. Moreover, it is hard to verify the validity of the parametric model assumptions. For example, [33] argues that with large samples, goodness-of-fit tests almost always reject quite reasonable models. Bayesian statistics is, in essence, also a parametric approach, because a functional form for the model is assumed.

The non-parametric approach to statistical learning, initiated by Pearson, takes a more direct path, by trying to estimate the entire probability density, rather than a few parameters of a subjectively specified function. Currently the most popular non-parametric approach to density estimation is the *kernel approach* (for a general introduction see [33], [42], [36])