A Framework for Probability Density Estimation

John Shawe-Taylor
Centre for Computational Statistics
and Machine Learning,
University College London,
Gower Street, London WC1E 6BT
jst@cs.ucl.ac.uk

Alex Dolia
Southampton Statistical Sciences
Research Institute,
University of Southampton,
Southampton SO17 1BJ
od@soton.ac.uk

Abstract

The paper introduces a new framework for learning probability density functions. A theoretical analysis suggests that we can tailor a distribution for a class of tasks by training it to fit a small subsample. Experimental evidence is given to support the theoretical analysis.

1 Introduction

The question of probability density estimation lies at the core of data modelling and machine learning. It is regarded as the hardest task since good estimation of the probability density can be used to solve other problems such as regression and classification. Furthermore, recent results show that L_1 density approximation of a discrete distribution requires sample sizes supra-polynomial in the cardinality of the support [2].

Vapnik [9] has argued that it is frequently better to learn the quantity you are interested in rather than go indirectly through a harder problem. This has certainly proved a good strategy for problems such as classification. This paper is concerned with an intermediate option where we may not have a single well-defined task to solve, but at the same time wish to avoid trying to accurately model the full probability density function in either an L_1 or KL divergence sense. We therefore consider a family of tasks (formalised in a so-called Touchstone Class) and ask that the learned density should be accurate on tasks drawn from this class.

Our main result is that constraining the learning with just a small sample from the Touchstone Class ensures good expected performance across the whole class with high probability. Hence, we can diversify the applicability of our learned density at a relatively low extra cost.

We further present experimental results that verify that the effect predicted by the theory can indeed be observed in practical experiments. Before launching into the details we give two potential applications as motivation for the approach.

In many cases probabilistic inference involves two phases, the learning of a distribution and subsequently inference of probabilities of certain configurations within the learned model. A great deal of emphasis and work has been devoted to the inference phase, but relatively little work has been done ensuring that the model accurately represents the information required by the inference. This paper aims to address this question by developing a general framework within which we can control what applications of a density function we wish to model.

The question of learning a density with limited resources in such a way that the result will be useful for a range of potential application scenarios can also play an important role in sensor networks. Here many devices cooperate with limited computing and bandwidth to assemble a range of information about the environment without prior knowledge of precisely what information may be required by users of the network. The analysis we have developed could help to guide the density learning to ensure that the range of anticipated queries can be accurately answered.

The rest of this section aims to place our work in the context of earlier approaches to density estimation.

Mukherjee and Vapnik [10, 6] provide the main inspiration for the approach taken here. They show how a one class SVM can be augmented by constraints that constrain the cumulative density up to each of the training points to be well approximated by their empirical estimate. They added constraints corresponding to all of the training examples. Our analysis suggests that adding only a small proportion of these constraints will give almost as good a fit. This theoretical analysis is borne out by the results of our first experiment shown in Figure 1 in which we plot the total misfit as a function of the number of constraints added. None corresponds the the one class SVM and all to the result of Mukherjee and Vapnik. As predicted the loss falls very quickly with the addition of just a few constraints before levelling out. The analysis undertaken by Mukherjee and Vapnik [6] bounds the KL divergence between the target density and a sequence of estimators and so does not address the weaker notion that we consider here.

We should also distinguish our work from level set estimation [4, 11]. This is the problem of finding the regions in which the density function exceeds a certain value. This problem has no obvious formulation within the framework that we consider and we leave open the potential connections between the theories concerning the accuracy of algorithms for this task developed in [4, 11] and the analysis developed here.

Finally, the paper [5] presents work that is closely related to our framework. They consider choosing a distribution by maximising the relative entropy subject to fitting the marginals for all of a finite set of features or in our terminology Touchstone functions. The result of such a constrained optimisation is known to be a Gibbs distribution from the exponential family. They do not, however, consider the possibility of generalising over a Touchstone class, but rather bound the log loss of the estimated maximum entropy distribution with that of other Gibbs distributions in the class.

2 Learning model

As indicated above learning a probability density function (pdf) can be viewed as learning an oracle that can answer a variety of questions. In order to fully specify the learning task we propose that the set of questions that can be asked of the oracle be specified. The following definition makes this notion precise.

Definition 1 *A* touchstone class *for learning a probability density function (pdf) on a measurable space $\mathcal{X}$ is a class of measurable real-valued functions $\mathcal{F}$ on $\mathcal{X}$ with a distribution $P_{\mathcal{F}}$ defined over $\mathcal{F}$. Given an unknown pdf function p, the* error $\mathrm{err}(\hat{p})$ *of an approximate pdf function $\hat{p}$ is defined as*

$$\mathrm{err}(\hat{p}) = \mathbb{E}_{f\sim P_{\mathcal{F}}}\left[\ell(\mathbb{E}_p[f], \mathbb{E}_{\hat{p}}[f])\right],$$

where ℓ is an appropriate loss function such as the absolute value, its square or an epsilon-insensitive version of either.

Note that taking an ϵ-insensitive binary valued loss would make the error measure equal to the probability of the estimate being out by more than ϵ.

We begin by giving examples of touchstone classes that can motivate the definition.

Example 2 *If we take $\mathcal{F}_{\mathcal{S}}$ to be the set of indicator functions $I_{S(\mathbf{a})}$ for sets of the form*

$$\mathcal{S} = \{S(\mathbf{a})\colon \mathbf{a} \in \mathbb{R}^n\}$$

where

$$S(\mathbf{a}) = \{\mathbf{x} \in \mathbb{R}^n\colon \mathbf{x}_i \leq \mathbf{a}_i, 1 \leq i \leq n\} \subset \mathbb{R}^n$$

the error measure assesses the accuracy of the cumulative density function defined by $\hat{p}$, since

$$\mathrm{err}(\hat{p}) = \mathbb{E}_{S(\mathbf{a})\sim P_S}\left[|P(S(\mathbf{a})) - \hat{P}(S(\mathbf{a}))|\right]$$

where P_S is a distribution over the sets $S(\mathbf{a})$ and P ($\hat{P}$) is the probability distribution for the density p ($\hat{p}$). This is the approach taken by Mukherjee and Vapnik [10, 6] where they attempt to match the values $P(S(\mathbf{x}_i))$ and $\hat{P}(S(\mathbf{x}_i))$ for all training examples $\mathbf{x}_i, i = 1, \dots, m$. Note that in their case P_S is implicitly chosen to be equal to the input distribution, while a touchstone class allows for any distribution for $\mathcal{P}_{\mathcal{F}}$. Mukherjee and Vapnik form part of the inspiration for the framework proposed here.

Example 3 *A further example of a touchstone class is given by extending to a general class of indicator functions $\mathcal{T}$ to obtain the touchstone class $\mathcal{F}_{\mathcal{T}}$. We would now require that $\hat{P}(S)$ be a good estimate of the probability $P(S)$ for a randomly drawn set from the class $\mathcal{T}$. This corresponds to the measure between distributions introduced by Ben-David* et al. *[3]. Their results are an important example of the power of the proposed approach.*

Example 4 *For a third example consider a distribution over $\{0,1\}^n$. The touchstone class $\mathcal{F}_{\mathcal{I}}$ is taken as a set of 'projection' functions $\pi_{\mathbf{i},\mathbf{v}}$ onto subsets $\mathbf{i} = \{i_1, \dots, i_{|\mathbf{i}|}\} \in \mathcal{I}$ of variables drawn from a set $\mathcal{I} \subseteq 2^{\{1,\dots,n\}}$ with prescribed values $\mathbf{v} \in \{0,1\}^{|\mathbf{i}|}$*

$$\mathcal{F}_{\mathcal{I}} = \left\{\pi_{\mathbf{i},\mathbf{v}}\colon \mathbf{i} \in \mathcal{I}, \mathbf{v} \in \{0,1\}^{|\mathbf{i}|}\right\},$$

where

$$\pi_{\mathbf{i},\mathbf{v}}(\mathbf{x}) = \begin{cases} 1 & \textit{if } \mathbf{x}_{i_j} = \mathbf{v}_j, \textit{ for } j = 1, \dots, |\mathbf{i}|, \\ 0; & \textit{otherwise.} \end{cases}$$

For this case the expectation $\mathbb{E}_p[\pi_{\mathbf{i},\mathbf{v}}]$ is the marginal for the variables indexed by $\mathbf{i}$ set to the values $\mathbf{v}$. Hence, the framework includes the computation of marginal distributions over prescribed subsets of binary variables.

We expect that learning will proceed by choosing a particular $\hat{p}$ from a class of modelling densities $\mathcal{P}$ for which the evaluation of $\mathbb{E}_{\hat{p}}[f]$ can be computed exactly. We introduce the definition of approximation that we will use to guide the learning.

Definition 5 *We say that $\hat{p} \in \mathcal{P}$ is an ϵ-approximation of the true density p with respect to the Touchstone Class $\mathcal{F}$ and loss function ℓ, if* $\mathrm{err}(\hat{p}) \leq \epsilon$.

Definition 6 *We say that a class of densities $\mathcal{P}$ is learnable with respect to the Touchstone Class $\mathcal{F}$ and loss function ℓ if there is an algorithm $\mathcal{A}$ such that given any $p \in \mathcal{P}$, $\epsilon > 0$ and $\delta > 0$, $\mathcal{A}$ given as input a sample of m points according to p where m is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, returns an estimate $\hat{p} \in \mathcal{P}$ that with probability $1 - \delta$ over the choice of random sample is an ϵ-approximation of p wrt $\mathcal{F}$ and ℓ.*

3 Analysis Framework

In this section we indicate the style of analysis that we propose for the model described above. The results given here are applicable to all of the possible scenarios discussed. In the next section we will consider the application of this approach to implementations of the cases described in Examples 2 and 3.

The aim of the theoretical analysis is to derive bounds on the $\mathrm{err}(\hat{p})$ for an estimate $\hat{p}$ of the pdf p in terms of quantities that can then be optimised in algorithms designed to approximate the pdf p.

There are two phases to the estimation, first we must estimate the accuracy of $\hat{p}$ for a particular function $f \in \mathcal{F}$, and secondly we need to consider the expectation of this quantity over a random choice of f according to $P_{\mathcal{F}}$.

Ignoring for the moment the first phase, the second phase can be viewed as learning a function q that maps $\mathcal{F}$ to the reals:

$$q : f \in \mathcal{F} \longmapsto q(f) = \mathbb{E}_q(f) \in \mathbb{R},$$

where we have deliberately overloaded the notation of q. This is a supervised learning problem with the target function given by $f \mapsto p(f) = \mathbb{E}_p(f)$, that is a standard regression problem modulo the fact that we do not have exact evaluations of $\mathbb{E}_p(f)$ for our training sample.

This brings us to the problem covered by the first phase, namely estimating $\mathbb{E}_p(f)$ for a given $f \in \mathcal{F}$. Since the expectation of the function $f(x)$ is $\mathbb{E}_p(f)$, the empirical estimate $\hat{\mathbb{E}}(f) = (1/m)\sum_i f(x_i)$ of the expected value of this f should give a good estimate of its true value.

The rest of this section is concerned with results that ensure both of these phases give good approximations.

Again we first consider the second phase. Now we consider the Rademacher complexity of our distribution class $\mathcal{P}$ from which $\hat{p}$ is chosen. We first give the definitions and main result. Note that we have removed the standard absolute value from the definition of Rademacher complexity as the main result holds in the stronger form given here (see for example [1]).

Definition 7 *For a sample $S = \{x_1, \cdots, x_m\}$ generated by a distribution D on a set X and a real-valued function class $\mathcal{F}$ with a domain X, the empirical Rademacher complexity of $\mathcal{F}$ is the random variable*

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \middle| x_1, \cdots, x_m \right] \quad (1)$$

where $\sigma = \{\sigma_1, \cdots, \sigma_m\}$ are independent uniform $\{\pm 1\}$-valued Rademacher random variables. The Rademacher complexity of $\mathcal{F}$ is

$$R_m(\mathcal{F}) = \mathbb{E}_S\left[\hat{R}_m(\mathcal{F})\right] = \mathbb{E}_{S\sigma}\left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i)\right] \quad (2)$$

Theorem 8 *Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be a class of functions mapping from S to $[0,1]$. Let $(x_i)_{i=1}^m$ be drawn independently according to a probability distribution D. Then with probability at least $1 - \delta$ over random draws of samples of size m, every $f \in \mathcal{F}$ satisfies*

$$\begin{aligned} \mathbb{E}_D[f(x)] &\leq \hat{\mathbb{E}}[f(x)] + R_m(\mathcal{F}) + 3\sqrt{\tfrac{\ln(2/\delta)}{2m}} \\ &\leq \hat{\mathbb{E}}[f(x)] + \hat{R}_m(\mathcal{F}) + 3\sqrt{\tfrac{\ln(2/\delta)}{2m}} \end{aligned} \quad (3)$$

Before beginning the analysis we quote Hoeffding's inequality.

Theorem 9 (Hoeffding's inequality) *If $X_1, \ldots, X_n$ are independent random variables satisfying $X_i \in [a_i, b_i]$, and if we define the random variable $S_n = \sum_{i=1}^n X_i$, then it follows that*

$$P\{|S_n - \mathbb{E}[S_n]| \geq \varepsilon\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Definition 10 *For a class $\mathcal{P}$ of distributions and a Touchstone Class $\mathcal{F}$ of functions we define the $\mathcal{F}$-derived class of functions to be*

$$\mathcal{P}_{\mathcal{F}} = \{f \in \mathcal{F} \mapsto \mathbb{E}_p[f] : p \in \mathcal{P}\}.$$

Furthermore we define the empirical ℓ-loss of a density $q \in \mathcal{P}_{\mathcal{F}}$ with respect to finite sets $S_f \subseteq \mathcal{F}$ and $S_x \subseteq \mathcal{X}$, as

$$\hat{\mathbb{E}}_f[\ell(\hat{\mathbb{E}}_x[f], \mathbb{E}_q[f])],$$

where $\hat{\mathbb{E}}_f$ refers to the empirical expectation using the sample S_f and $\hat{\mathbb{E}}_x$ to the empirical expectation using S_x.

We can now state our first result.

Theorem 11 *Let $\mathcal{F}$ be a Touchstone Class and $\mathcal{P}$ a class of distributions such that there exists a polynomial Q with the property that for $m \geq Q(1/\epsilon)$,*

$$R_m(\mathcal{P}_{\mathcal{F}}) \leq \epsilon,$$

where the associated symmetric loss function ℓ has range $[0,1]$, satisfies the triangle inequality and is Lipschitz continuous with constant L. Then an algorithm that can select a function from $\mathcal{P}_{\mathcal{F}}$ that minimises the empirical ℓ loss can learn $\mathcal{P}$ with respect to the function class $\mathcal{F}$.

Proof: Given $\epsilon > 0$ and $\delta > 0$, choose

$$m_f = \max\left\{Q(4/\epsilon), \frac{72}{\epsilon^2}\ln\frac{4}{\delta}\right\}. \qquad (4)$$

Sample m_f functions S_f from $\mathcal{F}$ according to $P_{\mathcal{F}}$. Now sample m_x input points S_x according to p where

$$m_x = \frac{8L^2}{\epsilon^2}\ln\frac{4m_f}{\delta}. \qquad (5)$$

Now let $\hat{p}$ be the density approximation returned by the algorithm that minimises

$$\hat{\mathbb{E}}_f[\ell(\mathbb{E}_{\hat{p}}[f], \hat{\mathbb{E}}_x[f])].$$

Since, the algorithm minimises the empirical ℓ loss we have

$$\hat{\mathbb{E}}_f[\ell(\mathbb{E}_{\hat{p}}[f], \hat{\mathbb{E}}_x[f])] \leq \hat{\mathbb{E}}_f[\ell(\mathbb{E}_p[f], \hat{\mathbb{E}}_x[f])]. \qquad (6)$$

An application of Hoeffding's inequality shows that the choice of m_x ensures that for a fixed function f, $|\hat{\mathbb{E}}_x[f] - \mathbb{E}_p[f]| \geq \epsilon/(4L)$ with probability at most

$$2\exp\left(-\frac{m_x\epsilon^2}{2L^2}\right) \leq \frac{\delta}{2m_f}$$

so that with probability $1 - \delta/2$

$$\sup_{f\in S_f} |\hat{\mathbb{E}}_x[f] - \mathbb{E}_p[f]| \leq \epsilon/(4L).$$

Together with equation (6) and the triangle and Lipschitz property of the loss ℓ this implies that

$$\begin{aligned}\hat{\mathbb{E}}_f[\ell(\mathbb{E}_p[f], \mathbb{E}_{\hat{p}}[f])] &\leq \hat{\mathbb{E}}_f[\ell(\mathbb{E}_p[f], \hat{\mathbb{E}}_x[f])] \\ &\quad + \hat{\mathbb{E}}_f[\ell(\mathbb{E}_{\hat{p}}[f], \hat{\mathbb{E}}_x[f])] \\ &\leq 2\hat{\mathbb{E}}_f[\ell(\mathbb{E}_p[f], \hat{\mathbb{E}}_x[f])] \\ &\leq 2\hat{\mathbb{E}}_f\left[L|\hat{\mathbb{E}}_x[f] - \mathbb{E}_p[f]|\right] \\ &\leq \epsilon/2. \qquad (7)\end{aligned}$$

Equation (7) bounds the empirical estimate in the application of the Rademacher Theorem 8 to the function class $\mathcal{P}_{\mathcal{F}}$ with probability at least $1 - \delta/2$:

$$\begin{aligned}\mathrm{err}(\hat{p}) &= \mathbb{E}_{f\sim P_{\mathcal{F}}}\left[\ell(\mathbb{E}_p[f], \mathbb{E}_{\hat{p}}[f])\right] \\ &\leq \hat{\mathbb{E}}_f\left[\ell(\mathbb{E}_p[f], \mathbb{E}_{\hat{p}}[f])\right] + R_{m_f}(\mathcal{P}_{\mathcal{F}}) + 3\sqrt{\frac{\ln(4/\delta)}{2m_f}} \quad (8)\end{aligned}$$

The choice of m_f ensures that the last two terms sum to $\epsilon/2$. Hence, with probability at least $1 - \delta$ we have the required total bound of $\mathrm{err}(\hat{p}) \leq \epsilon$. □

The result is couched in the slightly traditional framework of prescribing a given accuracy and confidence, but nonetheless we believe illustrates some of the constraints implicit in the framework.

The main points to highlight are as follows.

- The required number of function samples m_f depends principally on the Rademacher complexity of the class $P_{\mathcal{F}}$, that is the class of densities that are being used, and only indirectly (as inputs) on the Touchstone class $\mathcal{F}$ itself. We will see an example of this dependency in the next section.
- The sample complexity m_x is very benign for small L as for example when using an L_1 norm, since its main dependence is on $\ln m_f$.
- The main insight that the analysis provides is that we can expect to get good approximation across the Touchstone class by choosing a density that gives good performance on a small random sample of these functions.

Section 5 will present experiments to illustrate these points, particularly the last item. First, however, in the next section we introduce the specific function classes that will be used and derive bounds on their performance.

4 Support Vector Density Estimation

We now define a specific class of density functions inspired again by Mukherjee and Vapnik [10]. The starting point is the one class SVM [7] but with a kernel κ normalised so that for all $\mathbf{z} \in \mathcal{X}$,

$$\int_{\mathcal{X}} \kappa(\mathbf{x}, \mathbf{z})d\mathbf{x} = 1.$$

The standard choice for κ is a normalised Gaussian

$$\kappa(\mathbf{x}, \mathbf{z}) = \frac{1}{\left(\sqrt{2\pi\sigma}\right)^d}\exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right)$$

where d is the dimension of the input space. In general we assume that there is a finite constant C_κ such that

$$C_\kappa := \sup_{\mathbf{z},\mathbf{z}'}\sqrt{\kappa(\mathbf{z}, \mathbf{z}')} = \sqrt{\kappa(\mathbf{x}, \mathbf{x})},$$

for all $\mathbf{x} \in \mathcal{X}$ with $\kappa(\mathbf{z}, \mathbf{x}) \geq 0$ for all $\mathbf{x}$, $\mathbf{z}$. If we now consider learning a density function in a dual representation

$$q(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}),$$

the constraint $\sum_{i=1}^{m} \alpha_i = 1$ ensures that the density is correctly normalised, that is q satisfies

$$q(\mathcal{X}) = \int_{\mathcal{X}} q(\mathbf{x}) d\mathbf{x} = 1.$$

We therefore define a sequence of spaces $\mathcal{P}(B)$ parametrised by $B \in \mathbb{R}^+$ to be

$$\mathcal{P}(B) = \left\{ q_{\mathbf{w}} : \mathbf{x} \mapsto \langle \mathbf{w}, \phi(\mathbf{x}) \rangle \middle| \, \|\mathbf{w}\| \leq B, q_{\mathbf{w}}(\mathcal{X}) = 1 \right\},$$

where ϕ is the feature mapping corresponding to κ. The corresponding space $\mathcal{P}_{\mathcal{F}}(B)$ is given by

$$\mathcal{P}_{\mathcal{F}}(B) = \left\{ q_{\mathbf{w}} : f \mapsto \mathbb{E}_{q_{\mathbf{w}}}[f] \middle| \, \|\mathbf{w}\| \leq B, q_{\mathbf{w}}(\mathcal{X}) = 1 \right\}. \tag{9}$$

We can evaluate $\mathbb{E}_{q_{\mathbf{w}}}[f]$ as follows

$$\begin{aligned} \mathbb{E}_{q_{\mathbf{w}}}[f] &= \int_{\mathcal{X}} q_{\mathbf{w}}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{X}} \langle \mathbf{w}, \phi(\mathbf{x}) \rangle f(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{X}} \langle \mathbf{w}, f(\mathbf{x}) \phi(\mathbf{x}) \rangle d\mathbf{x} \\ &= \left\langle \mathbf{w}, \int_{\mathcal{X}} f(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x} \right\rangle, \end{aligned}$$

implying that we are working in a linear space defined by the feature map

$$\phi_{\mathcal{F}} : f \longmapsto \int_{\mathcal{X}} f(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x}.$$

The corresponding inner product or kernel function $\kappa_{\mathcal{F}}$ is given by

$$\kappa_{\mathcal{F}}(f, g) = \langle \phi_{\mathcal{F}}(f), \phi_{\mathcal{F}}(g) \rangle = \int_{\mathcal{X}^2} f(\mathbf{x}) g(\mathbf{z}) \kappa(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z}.$$

We quote a standard result for the Rademacher complexity of linear function spaces.

Theorem 12 *If $\kappa : X \times X \to \mathbb{R}$ is a kernel, and $S = \{x_1, \cdots, x_m\}$ is a sample of point from X, then the empirical Rademacher complexity of the class $\mathcal{F}_B$ of linear functions in the kernel defined feature space with norm bounded by B satisfies*

$$\hat{R}_m(\mathcal{F}) \leq \frac{2B}{m} \sqrt{\sum_{i=1}^{m} \kappa(x_i, x_i)} = \frac{2B}{m} \sqrt{\operatorname{tr}(K)}, \tag{10}$$

where K is the kernel matrix defined on the set S and tr *is the trace of a matrix.*

We have a lemma bounding the empirical Rademacher complexity of the space $\mathcal{P}_{\mathcal{F}}(B)$.

Lemma 13 *Let $\mathcal{P}_{\mathcal{F}}$ be defined by equation (9) with respect to the kernel κ and the function space $\mathcal{F}$. Then the empirical Rademacher complexity of $\mathcal{P}_{\mathcal{F}}(B)$ on the sample $\{f_1, \ldots, f_{m_f}\}$ is bounded by*

$$\hat{R}_{m_f}(\mathcal{P}_{\mathcal{F}}(B)) \leq \frac{2B}{m_f} \sqrt{\sum_{i=1}^{m_f} \min\left(C_\kappa^2 \|f_i\|_{L_1}^2, \|f_i\|_{L_1} \|f_i\|_{L_\infty} \right)}.$$

Proof: In order to apply Theorem 12, we must compute the trace of the kernel matrix K corresponding to the sample $\{f_1, \ldots, f_{m_f}\}$. Consider the i entry

$$\begin{aligned} \kappa_{\mathcal{F}}(f_i, f_i) &= \int_{\mathcal{X}^2} f_i(\mathbf{x}) f_i(\mathbf{z}) \kappa(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \\ &\leq C_\kappa^2 \int_{\mathcal{X}^2} |f_i(\mathbf{x})| |f_i(\mathbf{z})| d\mathbf{x} d\mathbf{z} \\ &= \left(\int_{\mathcal{X}} |f_i(\mathbf{x})| d\mathbf{x} \right)^2 \\ &= C_\kappa^2 \|f_i\|_{L_1}^2, \end{aligned}$$

for the first term of the minimum. For the second term

$$\begin{aligned} \kappa_{\mathcal{F}}(f_i, f_i) &= \int_{\mathcal{X}^2} f_i(\mathbf{x}) f_i(\mathbf{z}) \kappa(\mathbf{x}, \mathbf{z}) d\mathbf{x} d\mathbf{z} \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} f_i(\mathbf{z}) \kappa(\mathbf{x}, \mathbf{z}) d\mathbf{z} f_i(\mathbf{x}) d\mathbf{x} \\ &\leq \|f_i\|_{L_\infty} \int_{\mathcal{X}} \int_{\mathcal{X}} \kappa(\mathbf{x}, \mathbf{z}) d\mathbf{z} |f_i(\mathbf{x})| d\mathbf{x} \\ &= \|f_i\|_{L_\infty} \int_{\mathcal{X}} |f_i(\mathbf{x})| d\mathbf{x} \\ &= \|f_i\|_{L_\infty} \|f_i\|_{L_1} \end{aligned}$$

as required. □

This in turn provides a bound on the Rademacher complexity for a function space with bounded L_1 norm.

Corollary 14 *Let $\mathcal{P}_{\mathcal{F}}$ be defined by equation (9) with respect to the kernel κ and the function space $\mathcal{F}$ satisfying $\|f\|_{L_1} \leq C$ for $f \in \mathcal{F}$. Then the Rademacher complexity of $\mathcal{P}_{\mathcal{F}}(B)$ is bounded by*

$$R_{m_f}(\mathcal{P}_{\mathcal{F}}(B)) \leq \frac{2BCC_\kappa}{\sqrt{m_f}}.$$

Note that the corollary implies that $\mathcal{P}_{\mathcal{F}}(B)$ satisfies the Rademacher condition of Theorem 11 for the polynomial

$$Q(1/\epsilon) = \frac{4B^2C^2C_\kappa^2}{\epsilon^2}.$$

As indicated above the application of Theorem 11 is slightly unnatural as in practice we are typically not able to specify the size m_x of the sample of inputs. We therefore now present a bound on the error of a pdf function returned by an algorithm in terms of the sample sizes and complexities.

Theorem 15 *Suppose that we learn a pdf function $\hat{p}$ in the space $\mathcal{P}_{\mathcal{F}}(B)$ defined in equation (9) based on a sample of m_x inputs and m_f sample functions from the space $\mathcal{F}$. Then with probability at least $1-\delta$ over the generation of the two samples we can bound the error of $\hat{p}$ by*

$$\begin{aligned}\mathrm{err}(\hat{p}) \quad \leq \quad & L\sqrt{\frac{2}{m_x}\ln\frac{4m_f}{\delta}} + \hat{\mathbb{E}}_f[\ell(\mathbb{E}_{\hat{p}}[f], \hat{\mathbb{E}}_x[f])] \\ & + \frac{2BC_\kappa}{m_f}\sqrt{\sum_{i=1}^{m_f}\|f_i\|_{L_1}^2} + \sqrt{\frac{9}{2m_f}\ln\frac{4}{\delta}}\end{aligned} \tag{11}$$

Proof: The bound is derived from the empirical Rademacher version of the general Rademacher bound of Theorem 8 to the sampling over functions so that it holds with probability at least $1-\delta/2$. The first two terms come from applying the triangle inequality to the empirical error term

$$\begin{aligned}\hat{\mathbb{E}}_f[\ell(\mathbb{E}_p[f], \mathbb{E}_{\hat{p}}[f])] \quad \leq \quad & \hat{\mathbb{E}}_f[\ell(\mathbb{E}_p[f], \hat{\mathbb{E}}_x[f])] \\ & + \hat{\mathbb{E}}_f[\ell(\mathbb{E}_{\hat{p}}[f], \hat{\mathbb{E}}_x[f])]\end{aligned}$$

and bounding the first term using Hoeffding's inequality applied to ensure the inequality holds with probability $1-\delta/2$. The third term is the bound on the empirical Rademacher complexity given in Lemma 13. Hence the result follows. □

The form of the bound in Theorem 15 motivates the optimisations implemented in our algorithmic strategy. We see that the bound on the norm of the weight vector in the feature space appears and so this is introduced into the objective. Furthermore the second term corresponds to the amount by which the constraints fail to be satisfied. This term is realised by introducing slack variables that measure the slack in the constraints and the sum of the slack variables is incorporated into the objective controlled by a regularisation parameter D. The final form of the optimisation that minimises the bound of Theorem 15 is as follows:

$$\begin{array}{ll}\min_{\alpha,\xi} & \sum_{i,j=1}^{m_x}\alpha_i\alpha_j\kappa(\mathbf{x}_i,\mathbf{x}_j) + D\sum_{j=1}^{m_f}\xi_j \\ \text{subj to} & \sum_{i=1}^{m_x}\alpha_i = 1 \\ \multicolumn{2}{l}{\ell\left(\sum_{i=1}^{m_x}\alpha_i\int_{\mathcal{X}}\kappa(\mathbf{x}_i,\mathbf{x})f_j(\mathbf{x})d\mathbf{x}, \frac{1}{m_x}\sum_{i=1}^{m_x}f_j(x_i)\right) \leq \xi_j} \\ & \quad \text{and } \xi_j \geq 0 \text{ for } j = 1,\dots,m_f, \\ & \alpha_i \geq 0 \text{ for } i = 1,\dots,m_x.\end{array} \tag{12}$$

Note that if the number of constraints $m_f = 0$ we arrive at the smallest enclosing hypershere or one class SVM problem (see [8] for details). In the experiments described below we will vary m_f to interpolate between this case and the more stringent fitting implied by larger m_f. The one class SVM can also be regularised by using the box constraint on the dual variables by placing an upper bound $\alpha_i \leq 1/(\nu m_x)$. Further regularisation can be obtained by using an ϵ-insensitive loss function for ℓ, where ϵ provides a further trade-off between under and overfitting.

5 Experimental Results

We present two experiments that aim to verify the main insights highlighted at the end of the previous section.

The first interpolates between the framework introduced by Mukherjee and Vapnik [10, 6] and the one class Support Vector Machine [7]. We can view the latter as performing novelty detection, but if used with a normalised Gaussian kernel it also determines a pdf. The former takes this approach but adds constraints corresponding to minimising the empirical L_1-loss for the full Touchstone class $\mathcal{F}_{\mathcal{S}}$ defined in Example 2. In other words it does not pick a sample of functions from $\mathcal{F}_{\mathcal{S}}$ but rather includes all of the functions. We perform a series of experiments plotting the average error as we take smaller samples from $\mathcal{F}_{\mathcal{S}}$ finishing with the one class SVM that corresponds to adding no constraints.

The 2-dimensional data used in the experiment is generated artificially as a mixture of two Gaussians $p(x) = 0.5N(\nu_1, \sigma_1^2) + 0.5N(\nu_2, \sigma_2^2)$.

Figure 1 shows the plot of the L_2 error as a function of the number of constraints included in the optimisation. Observe that as predicted by the theory the error falls very fast as the first few constraints are included and then levels out to decrease only relatively slowly as larger numbers of constraints are added. We used the version without slack variables for the experiments reported here, but included an ϵ-insensitive loss function:

$$\begin{array}{ll}\min_\alpha & \sum_{i,j=1}^{m_x}\alpha_i\alpha_j\kappa(\mathbf{x}_i,\mathbf{x}_j) \\ \text{subject to} & \\ & |\mathbb{E}_{q_\alpha}[f_j] - \hat{\mathbb{E}}_x[f_j]| \leq \epsilon \text{ for } j = 1,\dots,m_f, \\ & \sum_{i=1}^{m_x}\alpha_i = 1,\ \alpha_i \geq 0 \text{ for } i = 1,\dots,m_x,\end{array} \tag{13}$$

where

$$\begin{aligned}\mathbb{E}_{q_\alpha}[f_j] \quad &= \quad \sum_{i=1}^{m_x}\alpha_i\int_{\mathcal{X}}\kappa(\mathbf{x}_i,\mathbf{x})f_j(\mathbf{x})d\mathbf{x} \\ &= \quad \sum_{i=1}^{m_x}\alpha_i\prod_{k=1}^{n}\left(1 - Q\left(\frac{x_{k,j}-x_{k,i}}{\sigma}\right)\right), \\ Q(x) \quad &= \quad \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp(-t^2)dt,\end{aligned} \tag{14}$$

and

$$\hat{\mathbb{E}}_x[f_j] = \frac{1}{m_x}\sum_{i=1}^{m_x} f_j(x_i) = \frac{1}{m_x}\sum_{i=1}^{m_x}\prod_{k=1}^{n}\theta(x_{k,j}-x_{k,i}), \tag{15}$$

with $\theta(x) = 1$ if $x > 0$ and $\theta(x) = 0$ if $x \leq 0$.

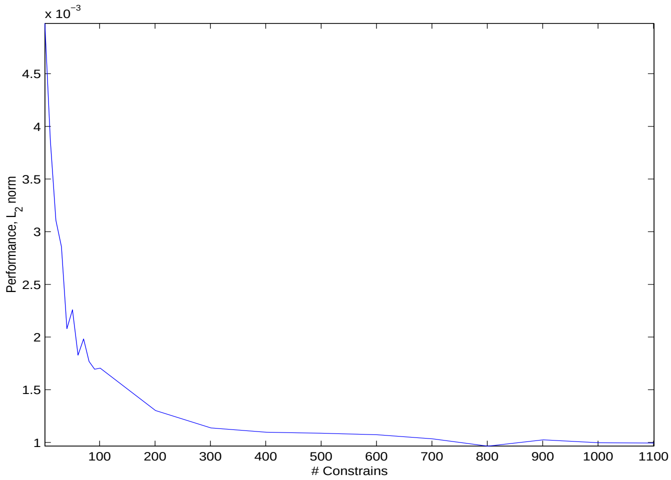

Figure 1: Plot of L_2 error over the Touchstone class of Example 2 as a function of the size of m_f

The first experiment verifies that the approach works for the simple case that inspired the analysis. We now consider a higher (10) dimensional example with an uncountably infinite Touchstone class in order to verify how the approach scales to more complex cases. Using orthants to the left of data points does not make sense in high dimensional spaces as these become an exponentially small proportion of the space, so the Touchstone class is now chosen to be the set of all halfspaces with their closest point to the origin generated by an isotropic Gaussian centred at the origin. Hence, in the second example

$$\begin{aligned}\mathbb{E}_{q_\alpha}[f_j] &= \sum_{i=1}^{m_x}\alpha_i\int_{\mathcal{X}}\kappa(\mathbf{x}_i,\mathbf{x})f_j(\mathbf{x})d\mathbf{x} \\ &= \sum_{i=1}^{m_x}\alpha_i\left(1 - Q\left(\frac{\mathbf{x}_i'\mathbf{w}_j/\|\mathbf{w}_j\| - \|\mathbf{w}_j\|}{\sigma}\right)\right),\end{aligned} \tag{16}$$

and

$$\begin{aligned}\hat{\mathbb{E}}_s[f_j] &= \frac{1}{m_x}\sum_{i=1}^{m_x} f_j(x_i) \\ &= \frac{1}{m_x}\sum_{i=1}^{m_x}\theta\left(\frac{\mathbf{x}_i'\mathbf{w}_j/\|\mathbf{w}_j\| - \|\mathbf{w}_j\|}{\sigma}\right).\end{aligned} \tag{17}$$

The data distribution is generated by a 10 dimensional mixture of two Gaussian with centers $(0,\ldots,0)$ and $(1,2,0,\ldots,0)$ and $\sigma = 1$.

Figure 2 shows the training (blue unbroken) and test (red dashed) L_2 error as a function of the number of introduced constraints. Again $x = 0$ corresponds to the one class SVM and as predicted by the theoretical analysis, very rapid falls are observed in the average loss as a relatively small number of constraints are added. The test error tracks the training error very tightly indicating that overfitting has not occurred for this sample size $m_x = 100$. Figure 3 shows the same two plots for a smaller sample size $m_x = 50$. Here we can observe a divergence between training and test error for small numbers of constraints indicating that

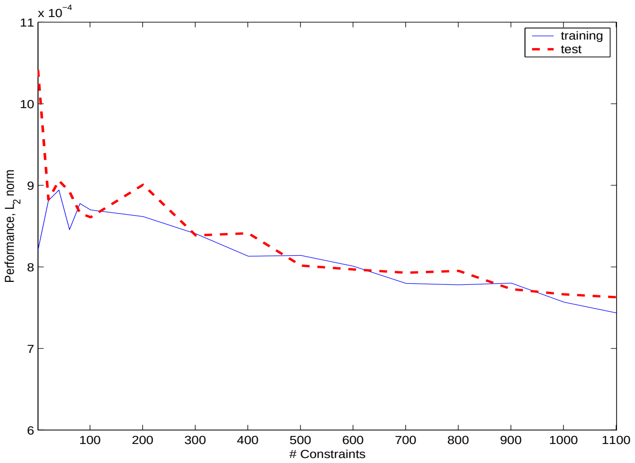

Figure 2: The average training (blue unbroken) and test (red dashed) L_2 error as a function of the number of constraints (size of the sample m_f) – sample size $m_x = 100$

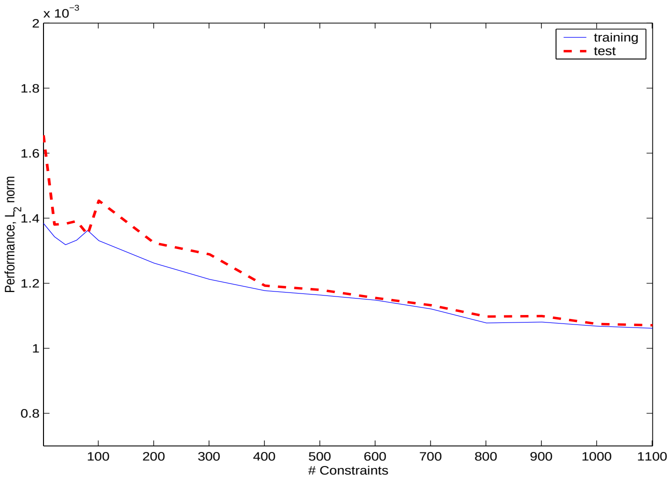

Figure 3: The average training (blue unbroken) and test (red dashed) L_2 error as a function of the number of constraints (size of the sample m_f) – sample size $m_x = 50$

some overfitting has occurred in this regime.

Finally, in Figure 4 we see a similar plot with $m_x = 500$. Here the learning effect of just a small number of constraints is particularly evident.

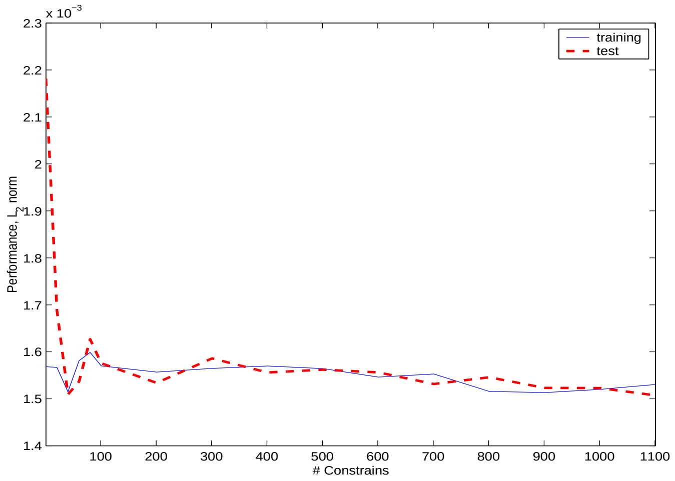

Figure 4: The average training (blue unbroken) and test (red dashed) L_2 error as a function of the number of constraints (size of the sample m_f) – sample size $m_x = 500$

6 Conclusions

The preliminary experiments verify the main thrust of the analysis, namely that a relatively small number of constraints allows the pdf to 'learn' the best fit. This is in line with the predictions made by the theoretical framework developed in the earlier sections.

We believe that the approach has many potential applications. For example the example used for the second experiment could use unlabelled data to get a good estimate of the probabilities between two parallel hyperplanes. This could then be used to guide a semi-supervised classification algorithm that attempted to separate the small amount of labelled data with a wide slab with low probability, this being the equivalent of a large margin for a fully labelled dataset.

The more ambitious aim is to use the approach to integrate into a single analysis the learning of a density and inference of probabilities from that density. By learning the density with a suitably parameterised model with constraints ensuring a good fit for the marginals of interest, we can expect algorithms for approximate inference that can be implemented efficiently over the model to give good estimates with appropriate bounds on their accuracy.

Acknowledgement This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

[1] Amiran Ambroladze, Emilio Parrado-Hernández, and John Shawe-Taylor. Complexity of pattern classes and Lipschitz property. *Theoretical Computer Science, special issue arising from ALT 2005*, to appear, 2006.

[2] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *Foundations of Computer Science (FOCS)*, 2000.

[3] S. Ben-David, J. Gehrke, and R. Schuller. A theoretical framework for learning from a pool of disparate data sources. In *Proceedings of KDD'02*, 2002.

[4] S. Ben-David and M. Lindenbaum. Learning distributions by their density levels – a paradigm for learning without a teacher. *Journal of Computer and Systems Sciences*, 55(1):171–182, 1997.

[5] Miroslav Dudik, Steve J. Phillips, and Robert E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Proceedings of the 19th Annual Conference on Learning Theory*, pages 123–138, 2006.

[6] Sayan Mukherjee and Vladimir Vapnik. Multivariate density estimation: an SVM approach. Technical Report AIM-1653, 1999.

[7] B. Schölkopf, J.C. Platt, J.S. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[8] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

[9] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[10] Vladimir Vapnik and Sayan Mukherjee. Support vector method for multivariate density estimation. In *Proceedings of the Neural Information Processing Conference (NIPS)*, pages 659–665, 1999.

[11] Régis Vert. *Theoretical Insights on Density Level Set Estimation with Applications to Anomaly Detection*. PhD Thesis, Ecole Normale Superieure, Paris, 2006.