RC23697 (W0508-139) August 24, 2005
Mathematics

IBM Research Report

Surrogate Cost Techniques in Classification and Regression Analysis

S. L. Hantler
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Surrogate Cost Techniques in Classification and Regression Analysis

S. L. Hantler
TJ Watson Research Center
Yorktown Heights, NY 10598

Abstract

We study the problems of classification and regression analysis when the set of classes or parameters is a σ-compact metric space, by means of surrogate cost minimization. We give a natural sufficient condition for the optimal classifier or estimator to be of the form $\mathcal{T}f$ when the function f minimizes a cost function which is a surrogate for the actual loss defined on pairs of classes. Sequences of functions whose expectations converge to the infimum of the expectations of all such functions can then be found by minimizing the sample averages of training sets. This result extends and sharpens previous results.

1 Introduction

Our interest is to classify or estimate a parameter of objects by minimizing a surrogate cost of functions defined on those objects rather than by minimizing the distance, usually called actual loss, between the actual class or parameter of the object and its estimated class or parameter. In this section, we describe the problem informally and in section 2 we establish the notation and assumptions that we will use throughout the paper. In section 3 we state the problem precisely and give the main mathematical results of this paper which are applied to the problems of classification and regression analysis in sections 5 and 6.

We are given a set, Ω, of objects and a set, $\mathcal{Y}$, of classes (in classification) or parameters (in regression analysis). There is a probability distribution, P, on the objects and, corresponding to each object, $\omega \in \Omega$, a probability distribution, $\pi(\omega)$, on the classes or parameters, corresponding to the probability distribution of classes or parameters given that object. These combine in an intuitively clear and standard way to determine an overall probability, $\mathbf{P}$, on the product space, $\Omega \times \mathcal{Y}$. For each object $\omega \in \Omega$, we seek to find a good estimate, $\Lambda(\omega) \in \mathcal{Y}$.

The quality of the estimate is measured by a loss, $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where $L(y_1, y_2)$ is the loss incurred by incorrectly estimating y_2 by y_1. The goal is to find a function, $\Lambda : \Omega \to \mathcal{Y}$ so that the expected loss, $E[L(\Lambda(\omega), y)]$ is as small as possible. In the two

problems we are studying, symmetric loss classification and regression analysis, $(\mathcal{Y}, L)$ is a metric space.

Empirical risk minimization is a technique for finding Λ, given a finite sample of pairs $\{\omega_i, y_i\}$. In empirical risk minimization Λ is selected from a prespecified set, $\mathcal{F}$, of candidate functions $\Omega \to \mathcal{Y}$, such that Λ nearly minimizes the actual loss on the sample. That is, $\Lambda \in \mathcal{F}$ is chosen so that

$$\sum_i L(\Lambda(\omega_i), y_i) - \inf_{f \in \mathcal{F}} \sum_i L(f(\omega_i), y_i) \tag{1}$$

is small. For empirical risk minimization to be effective, minimizing the sample loss over $\mathcal{F}$ must be nearly as good as minimizing the actual expected loss, with high probability, over samples drawn from $\Omega \times \mathcal{Y}$ according to $\mathbf{P}$. This requires that $\mathcal{F}$ contain a good estimator, that minimizing sample loss be effective in minimizing expected loss, and that the associated minimization problem be tractable. Empirical risk minimization is well studied in statistical machine learning and in statistics and we will not discuss it further in this paper.

Our goal is to study another technique for finding Λ, (empirical) surrogate risk minimization. In surrogate risk minimization the optimal estimator Λ is the composition of two functions. There is a prespecified set of functions, $\mathfrak{F}$ mapping $\Omega \to \mathcal{S}$, a subset of the real valued functions on $\mathcal{Y}$, and a prespecified function $\mathcal{T} : \mathcal{S} \to \mathcal{Y}$. In surrogate risk minimization $f(\omega)(y)$ serves as a surrogate loss estimate of the actual loss of using $\mathcal{T} \cdot f(\omega)$ to classify or estimate y. The aim is find $\widehat{f}$ whose expected surrogate loss is small and choose $\Lambda = \mathcal{T} \cdot \widehat{f}$. As in empirical risk minimization, this is done on a sample and $\widehat{f} \in \mathfrak{F}$ is chosen so that

$$\sum_i \widehat{f}(\omega_i)(y_i) - \inf_{f \in \mathfrak{F}} \sum_i f(\omega_i)(y_i), \tag{2}$$

is small. A way to think about it is that we have replaced the problem of minimizing the expected actual loss, $E[L(\Lambda(\omega), y)]$ over $\mathcal{Y}$ valued functions on Ω by the problem of minimizing the expected surrogate cost, $E[\widehat{f}(\omega)(y)]$ over S valued functions on Ω. The method will be successful when the probability that $\mathcal{T} \cdot \widehat{f}(\omega)$ is near the optimal estimator of ω is nearly 1 when $E[\widehat{f}(\omega)(y)]$ is small. The purpose of this paper is to give conditions under which that happens.

As the success of empirical risk minimization depends on the choice of $\mathcal{F}$, in surrogate risk minimization, the choice of $\mathfrak{F}$ and of $\mathcal{T}$ must be made to ensure that minimizing the sample surrogate cost is effective in minimizing the expectation of the surrogate cost, that the associated minimization problem be tractable and that the estimator obtained from a function with small surrogate cost has small actual loss. An advantage of surrogate risk minimization is that one may be able to choose a family of surrogate cost functions

for which the minimization problem (2) is easier than (1), the minimization problem for the actual loss. For example, when $\mathcal{S}$ is convex the minimization problem (2) is convex with its attendant computational advantages.

We give sufficient conditions on the family $\mathfrak{F}$ and the function $\mathcal{T}$ in countable classification and in regression analysis on compact, convex subsets of Euclidean space that ensure that minimizing the sample surrogate cost also minimizes the expected actual loss. The practitioner can then choose any class of functions dense in $\mathfrak{F}$ and be assured that the procedure converges to an optimal solution of the problem, or he may choose a class of functions well approximated by functions in $\mathfrak{F}$ and be assured that the surrogate risk minimization procedure converges to a function which approximates the optimal estimator within a known tolerance. These are matters best understood in the context of specific problems and will not be treated here.

Zhang [3] gave a sufficient condition for $\mathcal{S} \subset \mathbb{R}^k$ and an associated function, $\mathcal{T} = \text{argmin} : \mathbb{R}^k \to \{1, 2, \ldots, k\}$ to be suitable for surrogate risk minimization for zero-one loss classification of a set of objects into k classes. We are interested in extensions to countable classification and to regression analysis in Euclidean spaces. The method employed in [3], relying on the fact that the set of classes is finite and that the loss function is zero-one loss, does not seem to generalize to the problems of interest here. We present a method which applies to countable sets (extending the result in [3]) as well as to regression analysis on compact, convex subsets of $\mathbb{R}^k$. In addition to applying to a larger class of problems, we feel that our analysis is more straightforward and better explains what is happening.

To give a unified treatment of our results, we present them in a general setting, but the reader should have in mind the simple case covered in [3], that of a finite set $\mathcal{Y}$, of k elements, with $L(y_1, y_2) = 0$ or 1 as $y_1 = y_2$ or not, often referred to as zero-one loss. In that case, $\mathcal{S} \subset \mathbb{R}^k$, a separable metrizable topological vector space, and the Borel probabilities on $\mathcal{Y}$ are the nonnegative elements of $\mathcal{S} \subset \mathbb{R}^k$ whose l^1 norms are one. The reduction of the results in section 3 to this case will be carried out in section 5.

2 Notation and Assumptions

This section establishes notation and conventions which will be used throughout the remainder of this paper.

We are given a probability space, $\{\Omega, \Sigma, P\}$, a topological vector space $\mathcal{X}$, a subset $\mathcal{S} \subset \mathcal{X}$ and the topological dual $\mathcal{X}^*$ of $\mathcal{X}$ with the strong topology. We also have

$\mathcal{D} \subset \mathcal{S} \times \mathcal{X}^*$,

$$\mathcal{M} \subset \{\mu \in \mathcal{X}^* : \exists s \in \mathcal{S} \ni \langle s, \mu \rangle < \inf_{\{s:(s,\mu)\in\mathcal{D}\}} \langle s, \mu \rangle\}$$

and

$$\{\mu \in \mathcal{M} : (s, \mu) \in \mathcal{D}\}$$

is open for each $s \in \mathcal{S}$. We denote by $\mathfrak{F}$ the space of measurable mappings $\Omega \to \mathcal{S}$, and by $\mathfrak{M}$ the space of measurable mappings $\Omega \to \mathcal{M}$.

Section 3 is devoted to proving Theorems 1 and 2, the main results of this paper.

3 Main Results

We will show now that under suitable conditions on $\mathcal{S}$ when the expectations of a certain sequence of functions converges in mean to the infimum of the expectations of such functions, then the function values must avoid a certain prohibited set with probability converging to 1. Several technical details obscure the view. First, we must show that an auxiliary function is Borel measurable and then we must approximate the means of the functions by restricting the integration to 'nice subsets' of Ω.

Theorem 1. *If $\mathcal{S}$ is separable, $\pi \in \mathfrak{M}, f_j \in \mathfrak{F}$ and $E[\langle f_j, \pi \rangle] \to \inf_{f\in\mathfrak{F}} E[\langle f, \pi \rangle] > -\infty$ then $P[(f_j, \pi) \in \mathcal{D}] \to 0$.*

Proof. Since $\mathcal{S}$ is separable it has a countable dense subset, $\{s_i : i \in \mathbb{N}\}$. For $\alpha \geq 0, i \in \mathbb{N}$ let

$$\mathcal{M}_{i,\alpha} = \left\{\mu \in \mathcal{M} : \inf_{\{s:(s,\mu)\in\mathcal{D}\}} \langle s, \mu \rangle - \langle s_i, \mu \rangle > \alpha\right\}.$$

If $\emptyset \neq \mathcal{M}_\alpha = \bigcup_i \mathcal{M}_{i,\alpha}$, define $\sigma_\alpha : \mathcal{M}_\alpha \to \{s_i : i \in \mathbb{N}\}$ by

$$\sigma_\alpha(\mu) = s_{\min\{j:\mu\in\mathcal{M}_{j,\alpha}\}}$$

and notice that

$$\langle s - \sigma_\alpha(\mu), \mu \rangle > \alpha$$

whenever $\mu \in \mathcal{M}_\alpha$ and $(s, \mu) \in \mathcal{D}$.

We claim that, when $\mathcal{M}_\alpha \neq \emptyset$ the function $\sigma_\alpha : \mathcal{M}_\alpha \to \mathcal{S}$ is Borel measurable. To see this, since $\sigma_\alpha^{-1}(\{s_j\}) = \mathcal{M}_{j,\alpha} \bigcap \left(\bigcup_{i=0}^{j-1} \mathcal{M}_{i,\alpha}\right)^c$, it suffices to show that $\mathcal{M}_{i,\alpha}$ is a Borel set for every i and α. For each $s \in \mathcal{S}$, let $F_s : \mathcal{M} \to \mathcal{S}$ be defined by

$$F_s(\mu) = \begin{cases} \langle s, \mu \rangle & (s, \mu) \in \mathcal{D} \\ \infty & (s, \mu) \notin \mathcal{D}. \end{cases}$$

Since $\{\mu \in \mathcal{M} : (s,\mu) \in \mathcal{D}\}$ is open in the relative topology on $\mathcal{M}$ for each $s \in \mathcal{S}$ and $\mu \to \langle s,\mu\rangle$ is continuous, F_s is upper semicontinuous for every $s \in \mathcal{S}$. Since $\inf_{\{s:(s,\mu)\in\mathcal{D}\}}\langle s,\mu\rangle = \inf_{s\in\mathcal{S}} F_s(\mu)$, it follows that $\inf_{\{s:(s,\mu)\in\mathcal{D}\}}\langle s,\mu\rangle - \langle s_i,\mu\rangle$ is upper semicontinuous and $\mathcal{M}_{i,\alpha}$ is an F_σ set.

Suppose $\epsilon > 0$ is given. Since $\{s_i : i \in \mathbb{N}\}$ is dense in $\mathcal{S}$ and $\mathcal{M} \subset \mathcal{X}^*, \mathcal{M}_\alpha \nearrow \mathcal{M}$ as $\alpha \searrow 0$, so we may choose $\alpha > 0$ so that $P[\pi^{-1}(\mathcal{M}_\alpha)] \geq 1 - \frac{\epsilon}{2}$ and define

$$f_j^*(\omega) = \begin{cases} \sigma_\alpha \cdot \pi(\omega) & \omega \in (f_j,\pi)^{-1}(\mathcal{D}) \bigcap \pi^{-1}(\mathcal{M}_\alpha) \\ f_j(\omega) & \omega \notin (f_j,\pi)^{-1}(\mathcal{D}) \bigcap \pi^{-1}(\mathcal{M}_\alpha) \end{cases}$$

Then $f_j^* \in \mathfrak{F}$ and since $\inf_{f\in\mathfrak{F}} E[\langle f,\pi\rangle] > -\infty$, we may choose J so large that $j > J$ implies that $\frac{\epsilon\alpha}{2} > E[\langle f_j,\pi\rangle] - \inf_{f\in\mathfrak{F}} E[\langle f,\pi\rangle]$. It follows that for $j > J$

$$\begin{aligned} \frac{\epsilon\alpha}{2} > E[\langle f_j,\pi\rangle] - \inf_{f\in\mathfrak{F}} E[\langle f,\pi\rangle] &\geq E[\langle f_j,\pi\rangle - \langle f_j^*,\pi\rangle] \\ &= \int_{(f_j,\pi)^{-1}(\mathcal{D})\bigcap\pi^{-1}(\mathcal{M}_\alpha)} \langle f_j(\omega) - \sigma_\alpha\cdot\pi(\omega),\pi(\omega)\rangle dP \\ &\geq \alpha \cdot P\left[(f_j,\pi)^{-1}(\mathcal{D})\bigcap\pi^{-1}(\mathcal{M}_\alpha)\right] \\ &\geq \alpha\left(P\left[(f_j,\pi)^{-1}(\mathcal{D})\right] - \frac{\epsilon}{2}\right). \end{aligned}$$

We conclude that $\epsilon > P\left[(f_j,\pi)^{-1}(\mathcal{D})\right]$ completing the proof. □

Now, if $\mathcal{X}$ is a separable metrizable topological vector space, e.g. a separable normed linear space, then every one of its subsets is separable, proving Theorem 2.

Theorem 2. *If $\mathcal{X}$ is separable and metrizable, $\pi \in \mathfrak{M}$, and $f_j \in \mathfrak{F}, E[\langle f_j,\pi\rangle] \to \inf_{f\in\mathfrak{F}} E[\langle f,\pi\rangle] > -\infty$ then $P[(f_j,\pi) \in \mathcal{D}] \to 0$.*

As discussed in Section 1, important applications of Theorems 1 and 2 are to the problems of classification and regression analysis. In the next section, we show how these theorems apply to these problems.

4 Applications

Before proceeding to classification and regression analysis, we specialize Theorem 2 to a form suitable to those applications.

For the remainder of this paper, we assume that $(\mathcal{Y}, L)$ is a σ-compact metric space and that $\mathcal{X} = C_0(\mathcal{Y})$, the separable Banach space of continuous real valued functions on

$\mathcal{Y}$ tending to 0 at ∞, with sup norm. Then $\mathcal{X}^*$ is the Banach space of bounded Borel measures on $\mathcal{Y}$ with total variation norm.

Further, we assume we are given $\mathcal{T} : \mathcal{S} \to \mathcal{Y}$ Borel measurable, and $\mathcal{T}^* : C_0(\mathcal{Y})^* \to 2^{\mathcal{Y}}$ so that for each $y \in \mathcal{Y}$ the set $\{x^* \in \mathcal{X}^* : L(y, \mathcal{T}^* x^*) > \epsilon\}$ is open in $\mathcal{X}^*$.

Finally, we assume $\mathcal{M}$ is a subset of the Borel probabilities on $\mathcal{Y}$ i.e., nonnegative Borel measures on $\mathcal{Y}$ which assign measure 1 to $\mathcal{Y}$, and that $\forall \mu \in \mathcal{M}$ and $\forall \epsilon > 0, \exists s \in S \ni$

$$\int_{\mathcal{Y}} s d\mu < \inf_{\{s: L(\mathcal{T}s, \mathcal{T}^*\mu) > \epsilon\}} \int_{\mathcal{Y}} s d\mu.$$

The elements of $\mathfrak{M}$ are transition probabilities on the pair of measurable spaces, (Ω, Σ) and $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. The pair P and $\pi \in \mathfrak{M}$ determines a unique probability, $\mathbf{P}$ on $\{\Omega \times \mathcal{Y}, \Sigma \otimes \mathcal{B}(\mathcal{Y})\}$ as described, for example, in [2].

Corollary 1 is simply a restatement, in this context, of Theorem 2.

Corollary 1. *If $\pi \in \mathfrak{M}$, $f_j \in \mathfrak{F}$ and*

$$\int_{\Omega \times \mathcal{Y}} f_j(\omega)(y) d\mathbf{P}(\omega, y) \to \inf_{f \in \mathfrak{F}} \int_{\Omega \times \mathcal{Y}} f(\omega)(y) d\mathbf{P}(\omega, y) > -\infty$$

then $\mathcal{T} f_j \to \mathcal{T}^* \pi$ in probability.

We next identify the appropriate spaces $\mathcal{Y}$ and $\mathcal{S}$ and the appropriate functions $\mathcal{T}$ and $\mathcal{T}^*$ in the cases of interest. In particular, $\mathcal{T}^*$ will be chosen so that $\mathcal{T}^*\mu$ is the set of optimal classifiers or estimators of μ. We begin, in section 5, with classification.

5 Classification

Suppose that $\mathcal{Y}$ is a countable set. Without loss of generality, we assume $\mathcal{Y} \subset \mathbb{N}$, identifying classes with natural numbers. In this case, $\mathcal{C}_0(\mathcal{Y})$ can be identified with the set of real valued sequences converging to zero and $\mathcal{C}_0(\mathcal{Y})^*$ is the set of absolutely summable sequences. Choose $\mathcal{S} \subset \{x \in C_0(\mathcal{Y}) : x(y) \geq K \ \forall y \in \mathcal{Y}\}$ for some K. Then $\mathcal{M}$ is the set of nonnegative elements of $\mathcal{C}_0(\mathcal{Y})^*$ whose sum is 1, and $\langle s, \mu \rangle = \sum_{i \in \mathcal{Y}} s(i)\mu(i)$.

Now, let $\mathcal{T} : \mathcal{S} \to \mathcal{Y}$ be defined by $\mathcal{T}(s) = \min\{i : s_i \geq s_k, \forall k \in \mathcal{Y}\}$ and $\mathcal{T}^* x^*$ be the set of indices of the maximal elements of x^*. Recall that a measurable function $\Lambda : \Omega \to \mathcal{Y}$ is an optimal classifier of ω whenever $\Lambda(\omega) \in \mathcal{T}^*\pi(\omega)$ almost everywhere.

First notice that for each $y \in \mathcal{Y}$ the set $\{x^* \in \mathcal{X}^* : L(y, \mathcal{T}^* x^*) > \epsilon\}$ is open in $\mathcal{X}^*$. If $\forall \mu \in \mathcal{M}, \exists s \in \mathcal{S}$ so that

$$\sum_{i \in \mathcal{Y}} s(i)\mu(i) < \inf_{\{s : L(\mathcal{T}s, \mathcal{T}^*\mu) > \epsilon\}} \sum_{i \in \mathcal{Y}} s(i)\mu(i) \tag{3}$$

then by Corollary 1, if $\{f_j\} : \Omega \to \mathcal{S}$ is chosen so that

$$\int_{\Omega \times \mathcal{Y}} f_j(\omega)(i) d\mathbf{P}(\omega, i) \to \inf_{f \in \mathfrak{F}} \int_{\Omega \times \mathcal{Y}} f(\omega)(i) d\mathbf{P}(\omega, i) \geq K$$

then $\mathcal{T} f_j(\omega)$ converges in probability to an optimal estimator of $\pi(\omega)$ in probability.

When $\mathcal{Y}$ is finite, we recover the result of [3], without the unnecessary hypothesis that $\mathcal{S}$ be the range of a continuous function.

Example: Let $\mathcal{S}$ be the set of sequences with exactly one nonzero entry, that entry being negative. It is easy to verify that $\mathcal{S}$ satisfies (3) in this case. Notice that the cardinality of $\mathcal{S}$ is the same as that of $\mathcal{Y}$ which is obviously best possible, so $\mathcal{S}$ is, in some sense, optimal for countable classification problems. We note that the convex hull of $\mathcal{S}$ is $-\mathcal{M}$ when $\mathcal{Y}$ is finite. □

We now apply Corollary 1 to study regression analysis on compact, convex subsets of $\mathbb{R}^n$.

6 Regression Analysis in $\mathbb{R}^n$

Suppose that $\mathcal{Y}$ is a compact, convex subset of $\mathbb{R}^n$, and the actual loss is given by $L(y_1, y_2) = \|y_1 - y_2\|^2$, often called mean square loss.

Let $\mathcal{S} \subset \mathcal{C}_0(\mathcal{Y})$ and $\mathcal{M} \subset \{\mu \in \mathcal{C}_0(\mathcal{Y})^* : \mu \geq 0, \int_{\mathcal{Y}} d\mu(y) = 1\}$, the Borel probabilities on $\mathcal{Y}$. Then $\langle s, \mu \rangle = \int_{\mathcal{Y}} s(y) d\mu(y)$.

For $s \in \mathcal{S}$ define $\mathcal{T}(s) = \int_{\mathcal{Y}} y s(y) d\lambda(y)$ and for $\mu \in \mathcal{C}_0(\mathcal{Y})^*$ define $\mathcal{T}^*\mu$ to be the singleton set $\{\int_{\mathcal{Y}} y d\mu(y)\}$. Recall that a measurable function $\Lambda : \Omega \to \mathcal{Y}$ is an optimal mean square error estimator of $\pi(\omega)$ whenever $\Lambda(\omega) \in \mathcal{T}^*\pi(\omega)$ almost everywhere.

Of course, for each $y \in \mathcal{Y}$ the set $\{x^* \in \mathcal{X}^* : L(y, \mathcal{T}^* x^*) > \epsilon\}$ is open in $\mathcal{X}^*$. If $\forall \mu \in \mathcal{M}, \exists s \in \mathcal{S}$ so that

$$\int_{\mathcal{Y}} s(y) d\mu(y) < \inf_{\{s : \|\mathcal{T}(s) - \mathcal{T}^*\mu\| > \epsilon\}} \int_{\mathcal{Y}} s(y) d\mu(y) \tag{4}$$

then by Corollary 1, if $\{f_j\} : \Omega \to \mathcal{S}$ is chosen so that

$$\int_{\Omega\times\mathcal{Y}} f_j(\omega)(y)d\mathbf{P}(\omega,i) \to \inf_{f\in\mathfrak{F}} \int_{\Omega\times\mathcal{Y}} f(\omega)(y)d\mathbf{P}(\omega,i) \geq 0,$$

then $\mathcal{T}(f_j)(\omega)$ converges in probability P to the optimal estimator of $\pi(\omega)$ in probability.

Example: Assume now that the Lebesgue measure of $\mathcal{Y}$ is positive and normalize it, for notational convenience, to be 1, denoting it by λ. We now assume that $\mathcal{M}$ consists of the Borel probabilities, μ, on $\mathcal{Y}$ with Radon-Nikodym derivative $\widehat{\mu} \in \mathcal{L}^2(\mathcal{Y}, d\lambda)$ and $\|\widehat{\mu}\|_2 \leq K$.

Let $\mathcal{S} = \{x \in C_0(\mathcal{Y}) : \|x\|_2 \leq K\}$ and for $\epsilon > 0$

$$\mathcal{D}_\epsilon = \{(s,\mu) \in \mathcal{S}\times\mathcal{M} : \|\langle s,\mu\rangle s - \widehat{\mu}\|_2^2 > \epsilon\}.$$

An easy calculation shows that $\inf_{\{s:(s,\mu)\in\mathcal{D}_\epsilon\}}\langle s,\mu\rangle > -M\|\widehat{\mu}\|_2 \geq -MK$. But, since $\mathcal{S}$ is dense in $\{\widehat{\mu} \in \mathcal{M} : \|\widehat{\mu}\|_2 \leq K\}$ and $\langle \frac{-M\widehat{\mu}}{\|\widehat{\mu}\|_2}, \mu\rangle = -M\|\widehat{\mu}\|_2$, we have, for every $\mu \in \mathcal{M}, \exists s \in \mathcal{S} \ni$

$$\int_{\mathcal{Y}} s(y)d\mu(y) < \inf_{\{s:(s,\mu)\in\mathcal{D}_\epsilon\}} \int_{\mathcal{Y}} s(y)d\mu(y).$$

It follows from Corollary 1 that if $\{f_j\} : \Omega \to \mathcal{S}$ is chosen so that

$$\int_{\Omega\times\mathcal{Y}} f_j(\omega)(y)d\mathbf{P}(\omega,y) \to \inf_{f\in\mathfrak{F}} \int_{\Omega\times\mathcal{Y}} f(\omega)(y)d\mathbf{P}(\omega,y) \geq -MK$$

then

$$P\left\{\omega : \|\langle f_j(\omega), \pi(\omega)\rangle f_j(\omega) - \pi(\omega)\|_2^2 > \epsilon\right\} \to 0,$$

so that

$$P\left\{\omega : \left\|\frac{f_j(\omega)}{\|f_j\|_1} - \pi(\omega)\right\|_2^2 > \epsilon\right\} \to 0,$$

and

$$\frac{1}{\|f_j\|_1}\int_{\mathcal{Y}} y f_j(\omega)(y)d\lambda(y)$$

converges to the optimal estimator of π in probability. □

7 Conclusion

We have given sufficient conditions for a set of continuous functions on a σ-compact metric space together with a map from that set to the metric space to be sufficient

for empirical surrogate risk minimization to be effective for classification and regression analysis. We have also given examples for classification among countable classes and regression analysis in Euclidean spaces. Although we do not pursue this avenue, it seems that the regression analysis result applies more generally to regression analysis in any compact, convex subset of a locally compact topological group.

Acknowledgments

Thanks are due to Al Taylor for his valuable insights, to Mark Brodie for patient explanations of machine learning, and to Richard Gail and Jon Lenchner for several helpful conversations.

References

[1] Edwards, R E., "Functional Analysis: Theory and Applications," Holt, Rinehart and Winston, Inc., NY, **(1965)**.

[2] Neveu, J., "Mathematical Foundations of the Calculus of Probability," Holden-Day, Inc., San Francisco, **(1965)**.

[3] Zhang, T., "Statistical Analysis of Some Multi-Category Large Margin Classification Methods," *Journal of Machine Learning Research*, **5 (2004)**, 1225–1251.