Estimating the Moments of a Random Vector with Applications

John Shawe-Taylor[1],Nello Cristianini[2]

[1]Department of Computer Science,
Royal Holloway, University of London
England

[2]Department of Statistics,
University of California at Davis,
USA

jst@cs.rhul.ac.uk
nello@wald.ucdavis.edu

Abstract − A general result about the quality of approximation of the mean of a distribution by its empirical estimate is proven that does not involve the dimension of the feature space. Using the kernel trick this gives also bounds the quality of approximation of higher order moments. A number of applications are derived of interest in learning theory including a new novelty detection algorithm and rigorous bounds on the Robust Minimax Classification algorithm.

1 Introduction

Many statistical analyses rely on estimating the moments of a distribution from a sample. The core result of this paper is a measure of the quality of the estimation of the mean of a distribution through a bound on the norm of the error between the true mean and the empirical one. The bound has the flavour of learning theory bounds in that it does not involve the dimension of the feature space but rather a bound on the radius of the ball containing the support of the distribution. The result is therefore applicable in feature spaces defined by a kernel, the type of representation ubiquitous in kernel-based learning methods.

The second key ingredient of the paper is the observation that higher moments are just means in the feature spaces defined by polynomial kernels of the appropriate degree. This implies that the basic result gives as corollaries bounds on the errors made in estimating higher order moments.

Many learning algorithms make explicit or implicit use of means and covariances of the input distribution. Our basic theorem can therefore be used to derive a number of interesting learning theoretic results. These include new novelty detection algorithms based on the empirical mean and covariance matrix.

A more explicit use of the mean and covariance of a distribution was made in an algorithm recently proposed by Lanckriet *et al.* [1]. Their algorithm known as Robust Minimax Classification optimises the probability of misclassification subject to the assumption that the empirical and true means and covariances coincide. Our bounds on the differences between the true and empirical estimates of these quantities mean that we can provide rigorous bounds on the generalisation error of their algorithm.

The paper is organised as follows. Section 2 presents the core theoretical result. Section 3 applies it to a simple novelty detection algorithm and introduces the link between higher order moments and polynomial kernels. This gives a bound on the accuracy of estimates of higher order moments. Section 4 applies the results to give bounds on the Robust Minimax Classification algorithm of Lanckriet *et al.* [1]. We finish with some conclusions.

2 Base Result

The first question we will consider is that of the stability of a fixed function of a finite dataset. In other words how different will the value of this same function be on another dataset generated by the same source? The key property that we will require of the relevant quantity or random variable is known as *concentration*. A random variable that is concentrated is very likely to assume values close to its expectation as values become exponentially unlikely as a function of their distance from the mean. For a concentrated quantity we will therefore be confident that it will assume very similar values on new datasets generated from the same source. This is the case, for example, for the function 'average height of the female individuals' used above. There are many results that assert the concentration of a random variable provided it exhibits certain properties. These results are often referred to as concentration inqualities. Here we quote one of the best known theorems that is usually attributed to McDiarmid.

Theorem 1 *(McDiarmid [2]) Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,x_i' \in A} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)|$$

$$\leq \quad c_i,\ 1 \leq i \leq n.$$

Then for all $\epsilon > 0$,

$$P\{f(X_1,\ldots,X_n) - \mathbb{E}f(X_1,\ldots,X_n) \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) \tag{1}$$

Another well-used inequality that bounds the deviation from the mean for the special case of sums of random variables is Hoeffding's Inequality. We quote it here where it will be seen to be a simple special case of McDiarmid's Inequality for the case where

$$f(X_1,\ldots,X_n) = \sum_{i=1}^n X_i. \tag{2}$$

Theorem 2 ***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*

$$P\{|S_n - \mathbb{E}[S_n]| \geq \varepsilon\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{3}$$

As an example consider the average of a set of ℓ instances $r_1, r_2, \ldots, r_\ell$ of a random variable R given by a probability distribution P on the interval $[a, b]$. Taking $X_i = r_i/\ell$ it follows in the notation of Hoeffding's Inequality that

$$S_\ell = \frac{1}{\ell}\sum_{i=1}^\ell r_i = \hat{\mathbb{E}}[R], \tag{4}$$

where $\hat{\mathbb{E}}[R]$ denotes the sample average of the random variable R. Furthermore

$$\mathbb{E}[S_n] = \mathbb{E}\left[\frac{1}{\ell}\sum_{i=1}^\ell r_i\right] = \frac{1}{\ell}\sum_{i=1}^\ell \mathbb{E}\left[r_i\right] = \mathbb{E}[R], \tag{5}$$

so that an application of Hoeffding's Inequality gives

$$P\{|\hat{\mathbb{E}}[R] - \mathbb{E}[R]| \geq \varepsilon\} \leq 2\exp\left(-\frac{2\ell\varepsilon^2}{(b-a)^2}\right), \tag{6}$$

indicating an exponential decay of probability with the difference between observed sample average and the true average. Notice that the probability also decays exponentially with the size of the sample.

The example of the average of a random variable raises the question of how reliably we can estimate the average of a random vector $\phi(\mathbf{x})$, where ϕ is a mapping from the input space X into a feature space $\mathcal{F}$ corresponding to a kernel $k(\cdot,\cdot)$. This is equivalent to asking how close the centre of mass of the projections of a training sample

$$S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\} \tag{7}$$

will be to the true expectation

$$\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x})] = \int_X \phi(\mathbf{x}) dP(\mathbf{x}), \tag{8}$$

where $P(\mathbf{x})$ is the probability distribution generating the data with support

$$\text{supp}(P) = \{\mathbf{x}\colon P(\mathbf{x}) > 0\}.$$

We denote the centre of mass of the training sample with

$$\bar{\phi}_S = \frac{1}{\ell}\sum_{i=1}^\ell \phi(\mathbf{x}_i). \tag{9}$$

We introduce the following real valued function of the sample S as our measure of the accuracy of the estimate

$$g(S) = \left\|\bar{\phi}_S - \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x})]\right\|. \tag{10}$$

We can apply McDiarmid's theorem to the random variable $g(S)$ by bounding the change in this quantity when $\mathbf{x}_i$ is replaced by $\mathbf{x}_i'$ to give S'

$$\begin{aligned} |g(S) - g(S')| &= \left|\left\|\bar{\phi}_S - \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x})]\right\| - \left\|\bar{\phi}_{S'} - \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x})]\right\|\right| \\ &\leq \left\|\bar{\phi}_S - \bar{\phi}_{S'}\right\| = \frac{1}{\ell}\left\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_i')\right\| \leq \frac{2R}{\ell}, \end{aligned}$$

where $R = \sup_{\mathbf{x}\in\text{supp}(P)}\|\phi(\mathbf{x})\|$. Hence, applying McDiarmid with $c_i = 2R/\ell$, we obtain

$$P\left\{g(S) - \mathbb{E}_S[g(S)] \geq \epsilon\right\} \leq \exp\left(-\frac{2\ell\epsilon^2}{4R^2}\right). \tag{11}$$

We are now at the equivalent point after the application of Hoeffding's inequality in the one dimensional case. But in higher dimensions we no longer have a simple expression for $\mathbb{E}_S[g(S)]$. We need therefore to consider the more involved argument (see explanation below)

$$\begin{aligned} \mathbb{E}_S[g(S)] &= \mathbb{E}_S\left[\left\|\bar{\phi}_S - \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x})]\right\|\right] = \mathbb{E}_S\left[\left\|\bar{\phi}_S - \mathbb{E}_{\tilde{S}}[\bar{\phi}_{\tilde{S}}]\right\|\right] \\ &= \mathbb{E}_S\left[\left\|\mathbb{E}_{\tilde{S}}[\bar{\phi}_S - \bar{\phi}_{\tilde{S}}]\right\|\right] \leq \mathbb{E}_{S\tilde{S}}\left[\left\|\bar{\phi}_S - \bar{\phi}_{\tilde{S}}\right\|\right] \\ &= \mathbb{E}_{\sigma S\tilde{S}}\left[\frac{1}{\ell}\left\|\sum_{i=1}^\ell \sigma_i\left(\phi(\mathbf{x}_i) - \phi(\tilde{\mathbf{x}}_i)\right)\right\|\right] \\ &= \mathbb{E}_{\sigma S\tilde{S}}\left[\frac{1}{\ell}\left\|\sum_{i=1}^\ell \sigma_i\phi(\mathbf{x}_i) - \sum_{i=1}^\ell \sigma_i\phi(\tilde{\mathbf{x}}_i)\right\|\right] \qquad (12) \\ &\leq 2\mathbb{E}_{S\sigma}\left[\frac{1}{\ell}\left\|\sum_{i=1}^\ell \sigma_i\phi(\mathbf{x}_i)\right\|\right] \qquad (13) \\ &\leq \frac{2}{\ell}\mathbb{E}_{S\sigma}\left[\left(\left\langle\sum_{i=1}^\ell \sigma_i\phi(\mathbf{x}_i), \sum_{j=1}^\ell \sigma_j\phi(\mathbf{x}_j)\right\rangle\right)^{1/2}\right] \\ &\leq \frac{2}{\ell}\left(\mathbb{E}_{S\sigma}\left[\sum_{i,j=1}^\ell \sigma_i\sigma_j\kappa(\mathbf{x}_i,\mathbf{x}_j)\right]\right)^{1/2} \\ &= \frac{2}{\ell}\left(\mathbb{E}_S\left[\sum_{i=1}^\ell \kappa(\mathbf{x}_i,\mathbf{x}_i)\right]\right)^{1/2} \leq \frac{2R}{\sqrt{\ell}}. \qquad (14) \end{aligned}$$

We now give an explanation of the stages in this derivation. The second equality introduces a second random sample $\tilde{S}$ of the same size drawn according to the same distribution. Hence the expectation of its centre of mass is indeed the true expectation of the random vector. The expectation over $\tilde{S}$ can now be moved outwards in two stages, the second of which follows from an application of the triangle inequality. The next equality makes use of the independence of the generation of the individual examples to introduce random exchanges of the corresponding points in the two samples. The random variables

$\sigma = \{\sigma_1, \ldots, \sigma_\ell\}$ assume values -1 and $+1$ independently with equal probability 0.5, hence either leaving the effect of the examples $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ as it was or effectively interchanging them. Since the points are generated independently such a swap gives an equally likely configuration and averaging over all possible swaps leaves the overall expectation unchanged. The next steps split the sum and again make use of the triangle inequality together with the fact that the generation of S and $\tilde{S}$ is identical. The movement of the square root function through the expectation follows from Jensen's inquality and the concavity of the square root, while the disappearance of the mixed terms $\sigma_i\sigma_j\kappa(\mathbf{x}_i, \mathbf{x}_j)$ for $i \neq j$ follows from the fact that the four possible combinations of -1 and $+1$ have equal probability with two of the four having the opposite sign and hence cancelling out.

Hence, setting the right hand side of equation (11) equal to δ, solving for ϵ, and combining with equation (14) gives the following result.

Theorem 3 *Let S be an ℓ sample generated independently at random according to a distribution P. Then with probability at least $1-\delta$ over the choice of S, we have*

$$\left\|\bar{\phi}_S - \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{x})]\right\| \leq \frac{R}{\sqrt{\ell}}\left(2 + \sqrt{2\ln\frac{1}{\delta}}\right). \qquad (15)$$

This shows that with high probability our sample does indeed give a good estimate of $\mathbb{E}[\phi(\mathbf{x})]$ in a way that does not depend on the dimension of the feature space. Note that the introduction of the random $\{-1, +1\}$ variables σ_i play a key role. Such random numbers are known as Rademacher variables. They allow us to move from an expression involving two samples in equation (12) to twice an expression involving one sample modified by the Rademacher numbers in equation (13).

3 First Applications

We have shown that the centre of mass of the training sample is indeed a good estimator for the true mean. We first give an example of using this result to motivate a simple novelty detection algorithm that checks if a new datapoint is further from the true mean than the furthest training point. The chances of this happening for data generated from the same distribution can be shown to be small, hence when such points are found there is a high probability that they are outliers.

Let $\mathbf{d} = (d_1, \ldots, d_\ell)$ be a vector of ℓ real numbers. We introduce the notation

$$\text{percent}(\mathbf{d}, \alpha) = \text{argmin}\left\{d_i : |\{j : d_j \leq d_i\}| \geq \alpha\ell\right\}.$$

This is the α percentile of the sequence of numbers, so that for example

$$\text{percent}(\mathbf{d}, 1) = \max \mathbf{d},$$

and $\text{percent}(\mathbf{d}, 0.5)$ is the median of the sequence.

If we consider the training set as a sample of points that provide an estimate of the distances $d_1, \ldots, d_\ell$ from the point $\mathbb{E}[\phi(\mathbf{x})]$, where

$$d_i = \|\phi(\mathbf{x}_i) - \mathbb{E}[\phi(\mathbf{x})]\|, \qquad (16)$$

we can bound the probability that a new random point $\mathbf{x}_{\ell+1}$ satisfies

$$d_{\ell+1} = \|\phi(\mathbf{x}_{\ell+1}) - \mathbb{E}[\phi(\mathbf{x})]\| > \text{percent}(\mathbf{d}, \alpha), \qquad (17)$$

with

$$\begin{aligned} P\left\{\|\phi(\mathbf{x}_{\ell+1}) - \mathbb{E}[\phi(\mathbf{x})]\| \right. &> \left. \text{percent}(\mathbf{d}, \alpha)\right\} \\ &= P\left\{d_{\ell+1} > \text{percent}(\mathbf{d}, \alpha)\right\} \\ &\leq \frac{\ell + 1 - \alpha\ell}{\ell+1} = 1 - \alpha + \frac{\alpha}{\ell+1}, \end{aligned}$$

by the symmetry of the i.i.d. assumption. Though we cannot compute the distance to the point $\mathbb{E}[\phi(\mathbf{x})]$, we can compute

$$\left\|\phi(\mathbf{x}) - \bar{\phi}_S\right\| = \sqrt{\kappa(\mathbf{x}, \mathbf{x}) + \frac{1}{\ell^2}\sum_{i,j=1}^{\ell}\kappa(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{\ell}\sum_{i=1}^{\ell}\kappa(\mathbf{x}, \mathbf{x}_i)}. \qquad (18)$$

Then we can with probability $1-\delta$ estimate $\|\phi(\mathbf{x}_{\ell+1}) - \mathbb{E}[\phi(\mathbf{x})]\|$ using the triangle inequality and equation (15)

$$\begin{aligned} d_{\ell+1} &= \|\phi(\mathbf{x}_{\ell+1}) - \mathbb{E}[\phi(\mathbf{x})]\| \\ &\geq \left\|\phi(\mathbf{x}_{\ell+1}) - \bar{\phi}_S\right\| - \left\|\bar{\phi}_S - \mathbb{E}[\phi(\mathbf{x})]\right\| \\ &= \left\|\phi(\mathbf{x}_{\ell+1}) - \bar{\phi}_S\right\| - \sqrt{\frac{2R^2}{\ell}}\left(1 + \sqrt{\ln\frac{1}{\delta}}\right). \end{aligned}$$

Similarly, we have that for $i = 1, \ldots, \ell$,

$$d_i = \|\phi(\mathbf{x}_i) - \mathbb{E}[\phi(\mathbf{x})]\| \leq \left\|\phi(\mathbf{x}_i) - \bar{\phi}_S\right\| + \left\|\bar{\phi}_S - \mathbb{E}[\phi(\mathbf{x})]\right\|. \qquad (19)$$

Hence, with probability $1-\delta$

$$\begin{aligned} P\Bigg\{\left\|\phi(\mathbf{x}_{\ell+1}) - \bar{\phi}_S\right\| &> \text{percent}\left(\left\|\phi(\mathbf{x}_i) - \bar{\phi}_S\right\|, \alpha\right) \qquad (20) \\ &\quad + 2\sqrt{\frac{2R^2}{\ell}}\left(1 + \sqrt{\ln\frac{1}{\delta}}\right)\Bigg\} \leq \\ &\leq P\left\{d_{\ell+1} > \text{percent}(\mathbf{d}, \alpha)\right\} \qquad (21) \\ &\leq 1 - \alpha + \frac{\alpha}{\ell+1}. \qquad (22) \end{aligned}$$

We have therefore proven the following theorem.

Theorem 4 *Fix $\delta > 0$. Let S be a sample drawn independently according to a distribution P with support in the sphere of radius R about the origin in feature space. Then with probability at least $1-\delta$ over the draw of the sample S, points drawn at random according to P will satisfy the following bound on their likelihood as a function of their distance from the empirical mean*

$$1 - \alpha + \frac{\alpha}{\ell+1}, \text{ where } \alpha\ell =$$

$$\left|\left\{i : \left\|\phi(\mathbf{x}) - \bar{\phi}_S\right\| > \left\|\phi(\mathbf{x}_i) - \bar{\phi}_S\right\| + 2R\sqrt{\frac{2}{\ell}}\left(1 + \sqrt{\ln\frac{\ell}{\delta}}\right)\right\}\right|$$

Proof. We apply the argument given above the theorem for ℓ values of α, $\alpha = j/\ell$, $j = 1, \ldots, \ell$ using δ/ℓ in place of δ in each case. Hence, with probability at least $1-\delta$ all of the bounds hold for the different values of α. The theorem states the smallest bound arising from the different applications. ■

Our aim is not only to provide applications of the base theorem but to show how it gives rise to bounds for higher order moments though a connection to the polynomial kernel. To this end we consider the second moment and introduce some additional notation.

The next corollary of Theorem 3 shows that the bound for the mean can also be applied to the second moment. Recall that the second moment correlation matrix is defined as

$$\mathbf{C} = \mathbb{E}\left[\phi(\mathbf{x})\phi(\mathbf{x})'\right].$$

Let the empirical estimate of this quantity be

$$\hat{\mathbf{C}} = \hat{\mathbb{E}}\left[\phi(\mathbf{x})\phi(\mathbf{x})'\right] = \frac{1}{\ell}\sum_{i=1}^{\ell}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)'.$$

For two matrices $\mathbf{A}$ and $\mathbf{B}$ with the same dimension $n\times m$, we use the notation $\mathbf{A}\circ\mathbf{B}$ to denote the Frobenius inner product

$$\mathbf{A}\circ\mathbf{B} = \sum_{i,j=1}^{n,m}\mathbf{A}_{ij}\mathbf{B}_{ij}.$$

Note that if $\mathbf{B}$ is the rank one matrix $\mathbf{uu}'$ then

$$\mathbf{A}\circ\mathbf{B} = \sum_{i,j=1}^{n,m}\mathbf{A}_{ij}u_iu_j = \mathbf{u}'\mathbf{A}\mathbf{u}.$$

Hence, the Frobenius norm $\|\cdot\|_F$ of a matrix A is given by

$$\|A\|_F = \sqrt{A\circ A}.$$

Corollary 5 *Let S be an ℓ sample generated independently at random according to a distribution P. Then with probability at least $1-\delta$ over the choice of S, we have*

$$\left\|\hat{\mathbf{C}}-\mathbf{C}\right\|_F \le \frac{R^2}{\sqrt{\ell}}\left(2+\sqrt{2\ln\frac{1}{\delta}}\right), \tag{23}$$

where R is the radius of the ball in the feature space containing the support of the distribution.

Proof. We apply Theorem 3 to the mapping

$$\hat{\phi}\colon \mathbf{x}\longmapsto \phi(\mathbf{x})\phi(\mathbf{x})'.$$

Clearly we have

$$\hat{\mathbf{C}} = \hat{\mathbb{E}}\left[\hat{\phi}(\mathbf{x})\right] \text{ and } \mathbf{C} = \mathbb{E}\left[\hat{\phi}(\mathbf{x})\right].$$

Applying the theorem the result follows since

$$\begin{aligned}\sup_{\mathbf{x}\in\mathrm{supp}(P)}\|\hat{\phi}(\mathbf{x})\| &= \sup_{\mathbf{x}\in\mathrm{supp}(P)}\|\phi(\mathbf{x})\phi(\mathbf{x})'\|_F\\ &= \sup_{\mathbf{x}\in\mathrm{supp}(P)}\sqrt{\phi(\mathbf{x})\phi(\mathbf{x})'\circ\phi(\mathbf{x})\phi(\mathbf{x})'}\\ &= \sup_{\mathbf{x}\in\mathrm{supp}(P)}\sqrt{\left(\phi(\mathbf{x})'\phi(\mathbf{x})\right)^2} \le R^2.\end{aligned}$$

■

Next consider the covariance matrix defined as

$$\mathbf{\Sigma} = \mathbb{E}\left[(\phi(\mathbf{x})-\bar{\phi})(\phi(\mathbf{x})-\bar{\phi})'\right] = \mathbf{C}-\bar{\phi}\bar{\phi}'.$$

Let the empirical estimate of this quantity be

$$\hat{\mathbf{\Sigma}} = \hat{\mathbb{E}}\left[(\phi(\mathbf{x})-\bar{\phi}_S)(\phi(\mathbf{x})-\bar{\phi}_S)'\right] = \hat{\mathbf{C}}-\bar{\phi}_S\bar{\phi}_S'.$$

A similar corollary applies to the covariances.

Corollary 6 *Let S be an ℓ sample generated independently at random according to a distribution P. Then with probability at least $1-\delta$ over the choice of S, we have*

$$\left\|\hat{\mathbf{\Sigma}}-\mathbf{\Sigma}\right\|_F \le \frac{2R^2}{\sqrt{\ell}}\left(2+\sqrt{2\ln\frac{2}{\delta}}\right), \tag{24}$$

where R is the radius of the ball in the feature space containing the support of the distribution and provided

$$\ell \ge \left(2+\sqrt{2\ln\frac{2}{\delta}}\right)^2.$$

Proof. Consider the effect of shifting the origin of the feature space by a fixed translation vector ψ prior to computing the mean and covariance. Hence the new mean will be

$$\tilde{\phi}_S = \bar{\phi}_S - \psi,$$

while the new empirical covariance matrix will be

$$\begin{aligned}\tilde{\mathbf{\Sigma}} &= \hat{\mathbb{E}}\left[(\phi(\mathbf{x})-\psi-\tilde{\phi}_S)(\phi(\mathbf{x})-\psi-\tilde{\phi}_S)'\right]\\ &= \hat{\mathbb{E}}\left[(\phi(\mathbf{x})-\psi-\bar{\phi}_S+\psi)(\phi(\mathbf{x})-\psi-\bar{\phi}_S+\psi)'\right]\\ &= \hat{\mathbf{\Sigma}}.\end{aligned}$$

Hence, we may assume that the origin has been moved to the centre of mass of the distribution. Applying Corollary 5 and Theorem 3 each with δ replaced by $\delta/2$ we have

$$\begin{aligned}\left\|\mathbf{\Sigma}-\hat{\mathbf{\Sigma}}\right\|_F &\le \left\|\mathbf{C}-\hat{\mathbf{C}}\right\|_F + \left(\bar{\phi}\bar{\phi}'-\bar{\phi}_S\bar{\phi}_S'\right)\circ\left(\bar{\phi}\bar{\phi}'-\bar{\phi}_S\bar{\phi}_S'\right)\\ &\le \left\|\mathbf{C}-\hat{\mathbf{C}}\right\|_F + \left(\bar{\phi}_S\bar{\phi}_S'\right)\circ\left(\bar{\phi}_S\bar{\phi}_S'\right)\\ &\le \frac{R^2}{\sqrt{\ell}}\left(2+\sqrt{2\ln\frac{2}{\delta}}\right) + \left\|\bar{\phi}_S\right\|^2\\ &\le \frac{R^2}{\sqrt{\ell}}\left(2+\sqrt{2\ln\frac{2}{\delta}}\right) + \frac{R^2}{\ell}\left(2+\sqrt{2\ln\frac{2}{\delta}}\right)^2\\ &\le \frac{2R^2}{\sqrt{\ell}}\left(2+\sqrt{2\ln\frac{2}{\delta}}\right),\end{aligned}$$

where the last inequality follows from the lower bound on ℓ. ■

Note that the condition on ℓ is fairly benign. Even if we take $\delta = 0.01$, the bound is equivalent to $\ell > 27$. When we apply this theorem in later results we will assume that this condition holds for clarity of presentation.

Our last result of this section again uses the polynomial kernel trick but this time to obtain a generalisation of Theorem 4.

Corollary 7 *Fix $\delta > 0$. Let S be a sample drawn independently according to a distribution P with support in the sphere of radius R about the origin in a feature space defined by a kernel $\kappa(\mathbf{x},\mathbf{z})$. Then with probability at least $1-\delta$ over the draw of the sample S, points drawn at random according to P will satisfy the following bound on their likelihood*

$$1-\alpha+\frac{\alpha}{\ell+1}, \text{ where}$$

$$\alpha\ell = \left|\left\{i : \frac{2}{\ell}\phi(\mathbf{x})'\hat{\mathbf{C}}\phi(\mathbf{x}) - \|\phi(\mathbf{x})\|^2 > \frac{2}{\ell}\phi(\mathbf{x}_i)'\hat{\mathbf{C}}\phi(\mathbf{x}_i)\right.\right.$$

$$- \|\phi(\mathbf{x}_i)\|^2 + 2R\sqrt{\frac{2}{\ell}}\left(1 + \sqrt{\ln\frac{\ell}{\delta}}\right)\Bigg\}\Bigg|$$

Proof. The result follows from applying Theorem 4 in the feature space defined by the polynomial kernel of degree 2 over the base kernel. As observed above the equivalent of $\bar{\phi}_S$ becomes

$$\frac{1}{\ell}\hat{\mathbf{C}}.$$

Furthermore,

$$\hat{\mathbf{C}} \circ \phi(\mathbf{x})\phi(\mathbf{x})' = \phi(\mathbf{x})'\hat{\mathbf{C}}\phi(\mathbf{x}).$$

The result follows. ■

The interesting thing about the bound is that it computes the degree to which a new point is unusual relative to the covariance matrix, hence taking into account the different degree to which distribution fills space in different directions in contrast to standard novelty detection algorithms which consider a circular region in feature space.

The expressions can of course be evaluated implicitly using the kernel since

$$\begin{aligned}
\|\phi(\mathbf{x})\|^2 &= \kappa(\mathbf{x},\mathbf{x})^2 \\
\text{while } \phi(\mathbf{x})'\hat{\mathbf{C}}\phi(\mathbf{x}) &= \sum_{i=1}^{\ell} \phi(\mathbf{x})'\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)'\phi(\mathbf{x}) \\
&= \sum_{i=1}^{\ell} \left(\phi(\mathbf{x})'\phi(\mathbf{x}_i)\right)^2 \\
&= \sum_{i=1}^{\ell} \kappa(\mathbf{x},\mathbf{x}_i)^2
\end{aligned}$$

4 Robust Minimax Classification

Our next application is connected with a classification algorithm developed by Lanckriet *et al.* [1]. The basis for the approach is the following Lemma.

Lemma 8 *Let $\bar{\phi}$ be the mean of a distribution and $\mathbf{\Sigma}$ its covariance matrix, $\mathbf{a} \neq 0$, b given, such that $\mathbf{a}'\bar{\phi} \leq b$ and $\alpha \in [0,1)$, then if*

$$b - \mathbf{a}'\bar{\phi} \geq \kappa(\alpha)\sqrt{\mathbf{a}'\mathbf{\Sigma}\mathbf{a}},$$

where $\kappa(\alpha) = \sqrt{\frac{\alpha}{1-\alpha}}$, then

$$P\left(\mathbf{a}'\phi(\mathbf{x}) \leq b\right) \geq \alpha$$

They also show that the condition becomes an equivalence if we take the infimum of the probability over all distributions having the same mean and covariance. Lanckriet *et al.*then derive an algorithm that chooses the vector $\mathbf{a}$ and threshold b to minimise the misclassification probability for a classification problem where the positive examples have mean and covariance

$$\left(\bar{\phi}_{\mathbf{x}}, \mathbf{\Sigma}_{\mathbf{x}}\right)$$

and the negative examples mean and covariance

$$\left(\bar{\phi}_{\mathbf{y}}, \mathbf{\Sigma}_{\mathbf{y}}\right).$$

The algorithm uses the sample based estimates of these quantities and outputs an error bound that holds under the assumption that these are the true means and covariances.

We now apply the machinery developed in this paper to derive a bound on the true error in terms of the error estimate output by the algorithm. The optimisation problem they solve to optimise the bound is

$$\begin{aligned}
\kappa_\star^{-1} &= \min_{\mathbf{a}} \sqrt{\mathbf{a}'\hat{\mathbf{\Sigma}}_{\mathbf{x}}\mathbf{a}} + \sqrt{\mathbf{a}'\hat{\mathbf{\Sigma}}_{\mathbf{y}}\mathbf{a}} \qquad (25) \\
\text{subject to} \quad & \mathbf{a}'(\bar{\phi}_{S_{\mathbf{x}}} - \bar{\phi}_{S_{\mathbf{y}}}) = 1,
\end{aligned}$$

where we have used $S_{\mathbf{x}}$ ($S_{\mathbf{y}}$) to denote the set of positive (negative) training examples. The value of the threshold is then determined as

$$b_\star = \mathbf{a}_\star'\bar{\phi}_{\mathbf{x}} - \kappa_\star\sqrt{\mathbf{a}_\star'\hat{\mathbf{\Sigma}}_{\mathbf{x}}\mathbf{a}_\star}$$

and the resulting bound on the misclassification probability (assuming the empirical mean and covariance are the true mean and covariance) is

$$1 - \alpha_\star = \frac{1}{1+\kappa_\star^2} = \frac{\left(\sqrt{\mathbf{a}_\star'\hat{\mathbf{\Sigma}}_{\mathbf{x}}\mathbf{a}_\star} + \sqrt{\mathbf{a}_\star'\hat{\mathbf{\Sigma}}_{\mathbf{y}}\mathbf{a}_\star}\right)^2}{1 + \left(\sqrt{\mathbf{a}_\star'\hat{\mathbf{\Sigma}}_{\mathbf{x}}\mathbf{a}_\star} + \sqrt{\mathbf{a}_\star'\hat{\mathbf{\Sigma}}_{\mathbf{y}}\mathbf{a}_\star}\right)^2}.$$

In order to provide a true error bound we must bound the difference between this estimate and the value that would have been obtained had the true mean and covariance been used.

We now prove a version of Lemma 8 involving the empirical mean and covariance.

Lemma 9 *Let $\bar{\phi}_S$ be the mean of a sample of ℓ points drawn independently according to a probability distribution P and $\hat{\mathbf{\Sigma}}$ its empirical covariance matrix, $\mathbf{a} \neq 0$ with norm 1, and b given, such that $\mathbf{a}'\bar{\phi} \leq b$ and $\alpha \in [0,1)$. Then with probability $1-\delta$ over the draw of the random sample, if*

$$b - \mathbf{a}'\bar{\phi}_S \geq \kappa\sqrt{\mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a}},$$

then

$$P\left(\mathbf{a}'\phi(\mathbf{x}) \leq b\right) \geq \alpha.$$

where α solves the equation

$$\frac{\alpha}{1-\alpha}\left(\frac{2R^2}{\sqrt{\ell}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right) + \mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a}\right) = \kappa^2\mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a} - \frac{4R^2}{\sqrt{\ell}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right).$$

Proof. Define

$$T = \frac{4R^2}{\sqrt{\ell}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right) + \frac{\alpha 2R^2}{(1-\alpha)\sqrt{\ell}}\left(2 + \sqrt{2\ln\frac{2}{\delta}}\right).$$

We will show the auxiliary result that with probability $1-\delta$ if

$$\left(b - \mathbf{a}'\bar{\phi}_S\right)^2 - \kappa(\alpha)^2\mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a} \geq T \qquad (26)$$

then

$$\left(b - \mathbf{a}'\bar{\phi}\right)^2 - \kappa(\alpha)^2\mathbf{a}'\mathbf{\Sigma}\mathbf{a} \geq 0. \qquad (27)$$

We first show that this will imply the lemma. Applying Lemma 8 we have

$$P\left(\mathbf{a}'\phi(\mathbf{x}) \leq b\right) \geq \alpha.$$

where $\kappa(\alpha)^2 = \frac{\alpha}{1-\alpha}$. But we can express $\kappa(\alpha)$ in terms of κ as

$$\kappa(\alpha)^2 \mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a} + T = \kappa^2 \mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a}.$$

Substituting for $\kappa(\alpha)$ and T gives the result. It therefore only remains to prove the auxiliary result. It is sufficient to bound with high probability the difference between the left hand sides of the two inequalities (26) and (27) by T.

$$\begin{aligned}
&\left|\left(b - \mathbf{a}'\bar{\phi}_S\right)^2 - \kappa(\alpha)^2 \mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a} - \left(b - \mathbf{a}'\bar{\phi}\right)^2 + \kappa(\alpha)^2 \mathbf{a}'\mathbf{\Sigma}\mathbf{a}\right| \\
\leq\ & \|\bar{\phi} - \bar{\phi}_S\| \left(2b + \|\bar{\phi} + \bar{\phi}_S\|\right) + \kappa(\alpha)^2 \left|\mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a} - \mathbf{a}'\mathbf{\Sigma}\mathbf{a}\right| \\
\leq\ & \|\bar{\phi} - \bar{\phi}_S\| 4R + \kappa(\alpha)^2 \left|\left(\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right) \circ \mathbf{a}\mathbf{a}'\right| \\
\leq\ & \|\bar{\phi} - \bar{\phi}_S\| 4R + \kappa(\alpha)^2 \left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F \|\mathbf{a}\mathbf{a}'\|_F \\
\leq\ & \|\bar{\phi} - \bar{\phi}_S\| 4R + \kappa(\alpha)^2 \left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|_F .
\end{aligned}$$

Now we apply Theorem 3 and Corollary 6 each with δ replaced by $\delta/2$. Substituting the resulting bounds on the difference between empirical and true means and covariances gives

$$\begin{aligned}
&\left|\left(b - \mathbf{a}'\bar{\phi}_S\right)^2 - \kappa(\alpha)^2 \mathbf{a}'\hat{\mathbf{\Sigma}}\mathbf{a} - \left(b - \mathbf{a}'\bar{\phi}\right)^2 + \kappa(\alpha)^2 \mathbf{a}'\mathbf{\Sigma}\mathbf{a}\right| \\
&\quad \leq \frac{4R^2}{\sqrt{\ell}} \left(2 + \sqrt{2 \ln \frac{2}{\delta}}\right) + \frac{\alpha 2R^2}{(1-\alpha)\sqrt{\ell}} \left(2 + \sqrt{2 \ln \frac{2}{\delta}}\right) \\
&\quad = T,
\end{aligned}$$

as required. ■

We are now in a position to derive a formula that will give a bound on the generalisation error of the minimax classification function in terms of the parameters of the resulting hyperplane and the empirical covariance matrices.

Proposition 10 *Let* $\mathbf{a}$*,* b*, be the (normalised) weight vector and associated threshold returned by the minimax algorithm when presented with a training set* S*. Furthermore, let* $\hat{\mathbf{\Sigma}}_{\mathbf{x}}$ *(*$\hat{\mathbf{\Sigma}}_{\mathbf{y}}$*) be the empirical covariance matrices associated with the positive and negative examples. Then with probability at least* $1-\delta$ *over the draw of the random training set* S *of* $\ell^{\mathbf{x}}$ *positive and* $\ell^{\mathbf{y}}$ *negative training examples, the generalisation error* ϵ *is bounded by*

$$\epsilon \leq \max(1 - \alpha^{\mathbf{x}}, 1 - \alpha^{\mathbf{y}})$$

where $\alpha^{\mathbf{i}}$*,* $\mathbf{i} = \mathbf{x}, \mathbf{y}$ *solves the equations*

$$\begin{aligned}
&\frac{\alpha^{\mathbf{i}}}{1 - \alpha^{\mathbf{i}}} \left(\frac{2R^2}{\sqrt{\ell^{\mathbf{i}}}} \left(2 + \sqrt{2 \ln \frac{4}{\delta}}\right) + \mathbf{a}'\hat{\mathbf{\Sigma}}_{\mathbf{i}}\mathbf{a}\right) \\
&\quad = \kappa_\star^2 \mathbf{a}'\hat{\mathbf{\Sigma}}_{\mathbf{i}}\mathbf{a} - \frac{4R^2}{\sqrt{\ell^{\mathbf{i}}}} \left(2 + \sqrt{2 \ln \frac{4}{\delta}}\right),
\end{aligned}$$

and $\kappa_\star$ *is the value of the optimum obtained in the minimax solution of (25).*

Proof. The bound comes from two applications of Lemma 9 again with δ replaced by $\delta/2$, one for the positive training examples and one for the negatives. The overall error is bounded by the maximum of the probability of misclassifying a positive or negative example. ■

It would be interesting to see if this bound could motivate an improved version of the minimax algorithm by effectively taking into account the different accuracy of approximation between positive and negative examples.

5 Conclusions

We have proven a general result bounding the norm of the difference between the true mean of a distribution and its estimation from a finite sample. The result does not involve the dimension of the feature space, but rather in line with methods for analysing kernel methods involves the radius of the ball in the feature space containing the support of the distribution.

We combine this result with a link between higher order moments and polynomial kernels to obtain corresponding bounds on the errors of estimating higher order moments. The important feature of these results is that they are not killed by the high dimensionality typical of higher order moments.

As an application we derive a new novelty detection algorithm and consider its implications when applied in the quadratic kernel feature space.

A further application involves the derivation of rigorous bounds on the generalisation error of a recently proposed classification algorithm known as the Robust Minimax algorithm.

We believe that the paper opens up an exciting new avenue of investigation of kernel method algorithms involving the use of different moments of the input distribution all estimated empirically. Many of these bounds may give rise to novel algorithms through optimising the derived expression for the generalisation. For example our bound on the error of the Robust Minimax algorithm involves data-dependent features that vary between the positive and negative examples. It is likely that the algorithm can be adapted to balance the errors made on positive and negative examples to take account of these estimation errors, hence giving a tighter upper bound on the error of the resulting hypothesis. The derivation of these results and algorithms is beyond the scope of this paper.

References

[1] Gert R.G. Lanckriet, Laurent El Ghaoui, Chiranjib Bhattacharyya, and Michael I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2002.

[2] C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, 1989.