Notes on mean embeddings and covariance operators

Arthur Gretton

January 11, 2020

1 Introduction

This note contains more detailed proofs of certain results in the lecture notes on mean embeddings and covariance operators. The notes are not as complete as for lectures 1 and 2, but cover only the trickier concepts. Please let me know if there are any further parts you'd like clarified, and I'll add them to the note.

2 Mean embeddings

2.1 Proof that the mean embedding exists via Riesz

For finite dimensional feature spaces, we can define expectations in terms of inner products.

$$\phi(x) = k(\cdot, x) = \begin{bmatrix} x \\ x^2 \end{bmatrix} \qquad f(\cdot) = \begin{bmatrix} a \\ b \end{bmatrix}$$

Then

$$f(x) = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} x \\ x^2 \end{bmatrix} = \langle f, \phi(x) \rangle_{\mathcal{F}} .$$

Consider random variable $\mathsf{x} \sim \mathbf{P}$

$$\mathbf{E}_{\mathbf{P}} f(\mathsf{x}) = \mathbf{E}_{\mathbf{P}} \left(\begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} \mathsf{x} \\ \mathsf{x}^2 \end{bmatrix} \right) = \begin{bmatrix} a \\ b \end{bmatrix}^\top \begin{bmatrix} \mathbf{E}_{\mathbf{P}} \mathsf{x} \\ \mathbf{E}_{\mathbf{P}} (\mathsf{x}^2) \end{bmatrix} =: \begin{bmatrix} a \\ b \end{bmatrix}^\top \mu_{\mathbf{P}}.$$

Does this reasoning translate to infinite dimensions?

Definition 1 (Bounded operator)**.** A linear operator $A : \mathcal{F} \to \mathbb{R}$ is bounded when

$$Af \leq \lambda_A \|f\|_{\mathcal{F}} \quad \forall f \in \mathcal{F}.$$

We prove via Riesz that the mean embedding exists, and that it takes the form of the expectation of the canonical map.

Theorem 2. *[Riesz representation]In a Hilbert space $\mathcal{F}$, all bounded linear operators A can be written $\langle \cdot, g_A \rangle_{\mathcal{F}}$, for some $g_A \in \mathcal{F}$,*

$$Af = \langle f, g_A \rangle_{\mathcal{F}}$$

Now we establish the existence of the mean embedding.

Lemma 3 (Existence of mean embedding)**.** *If $\mathbf{E}_{\mathsf{P}} \sqrt{k(\mathsf{x}, \mathsf{x})} < \infty$ then $\mu_{\mathsf{P}} \in \mathcal{F}$.*

Proof. The linear operator $T_{\mathsf{P}} f := \mathbf{E}_{\mathsf{P}} f(\mathsf{x})$ for all $f \in \mathcal{F}$ is bounded under the assumption, since

$$|T_{\mathsf{P}} f| = |\mathbf{E}_{\mathsf{P}} f(\mathsf{x})| \underset{(a)}{\leq} \mathbf{E}_{\mathsf{P}} |f(\mathsf{x})| = \mathbf{E}_{\mathsf{P}} |\langle f, \phi(\mathsf{x}) \rangle_{\mathcal{F}}| \leq \mathbf{E}_{\mathsf{P}} \left(\sqrt{k(\mathsf{x}, \mathsf{x})} \|f\|_{\mathcal{F}} \right),$$

where in (a) we use Jensen's inequality. Hence by the Riesz representer theorem [6, Theorem II.4], there exists a $\mu_{\mathsf{P}} \in \mathcal{F}$ such that $T_{\mathsf{P}} f = \langle f, \mu_{\mathsf{P}} \rangle_{\mathcal{F}}$. □

If we set $f = \phi(x) = k(x, \cdot)$, we obtain $\mu_{\mathsf{P}}(x) = \langle \mu_p, k(x, \cdot) \rangle = \mathbf{E}_{\mathsf{P}} k(x, \mathsf{x})$: in other words, the mean embedding of the distribution P is the expecation under P of the canonical feature map.

2.2 Proof that MMD injective for universal kernel

First, it is clear that $\mathsf{P} = \mathsf{Q}$ implies MMD $\{\mathsf{P}, \mathsf{Q}; F\}$ is zero. We now prove the converse. By the universality of $\mathcal{F}$, for any given $\epsilon > 0$ and $f \in C(\mathcal{X})$ there exists a $g \in \mathcal{F}$ such that

$$\|f - g\|_{\infty} \leq \epsilon.$$

We will need [2, Lemma 9.3.2]:

Lemma 4. *Let $(\mathcal{X}, d)$ be a metric space, and let P, Q be two Borel probability measures defined on $\mathcal{X}$, where we define the random variables $\mathsf{x} \sim \mathsf{P}$ and $\mathsf{y} \sim \mathsf{Q}$. Then $\mathsf{P} = \mathsf{Q}$ if and only if $\mathbf{E}_{\mathsf{P}}(f(\mathsf{x})) = \mathbf{E}_{\mathsf{Q}}(f(\mathsf{y}))$ for all $f \in C(\mathcal{X})$, where $C(\mathcal{X})$ is the space of bounded continuous functions on $\mathcal{X}$.*

We now use these two results to formulate a proof. We begin with the expansion

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})| \leq |\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{P}} g(\mathsf{x})| + |\mathbf{E}_{\mathsf{P}} g(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} g(\mathsf{y})| + |\mathbf{E}_{\mathsf{Q}} g(\mathsf{y}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})|.$$

The first and third terms satisfy

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{P}} g(\mathsf{x})| \leq \mathbf{E}_{\mathsf{P}} |f(\mathsf{x}) - g(\mathsf{x})| \leq \epsilon.$$

Next, write

$$\mathbf{E}_{\mathsf{P}} g(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} g(\mathsf{y}) = \langle g, \mu_{\mathsf{P}} - \mu_{\mathsf{Q}} \rangle = 0,$$

since MMD $\{\mathsf{P}, \mathsf{Q}; F\} = 0$ implies $\mu_{\mathsf{P}} = \mu_{\mathsf{Q}}$. Hence

$$|\mathbf{E}_{\mathsf{P}} f(\mathsf{x}) - \mathbf{E}_{\mathsf{Q}} f(\mathsf{y})| \leq 2\epsilon$$

for all $f \in C(\mathcal{X})$ and $\epsilon > 0$, which implies $\mathsf{P} = \mathsf{Q}$ by Lemma 4.

3 Covariance operators

One of the most important and widely used tools in RKHS theory is the covariance operator: this is an infinite dimensional analog to the covariance matrix. This forms the backbone of kernel PCA, the kernel Fisher discriminant, kernel partial least squares, the kernel canonical correlation, and so on.

In this note, we describe the Hilbert space of Hilbert-Schmidt operators. We then introduce the covariance operator, demonstrate it is Hilbert-Schmidt, and express it in terms of kernel functions.

3.1 Hilbert-Schmidt operators

This discussion is based on [9, Section 2.1] and [8, Section A.5.2].

Let $\mathcal{F}$ and $\mathcal{G}$ be separable Hilbert spaces. Define $(e_i)_{i\in I}$ to be an orthonormal basis for $\mathcal{F}$, and $(f_j)_{j\in J}$ to be an orthonormal basis for $\mathcal{G}$. The index sets I, J are assumed to be either finite our countably infinite.[1] Define two compact linear operators $L : \mathcal{G} \to \mathcal{F}$ and $M : \mathcal{G} \to \mathcal{F}$. Define the Hilbert-Schmidt norm of the operators L, M to be

$$\begin{aligned} \|L\|_{\mathrm{HS}}^2 &= \sum_{j\in J} \|Lf_j\|_{\mathcal{F}}^2 \\ &= \sum_{i\in I}\sum_{j\in J} \left|\langle Lf_j, e_i\rangle_{\mathcal{F}}\right|^2, \end{aligned} \tag{3.1}$$

where we use Parseval's indentity on each of the norms in the first sum. The operator L is Hilbert-Schmidt when this norm is finite.

The Hilbert-Schmidt operators mapping from $\mathcal{G}$ to $\mathcal{F}$ form a Hilbert space, written $\mathrm{HS}(\mathcal{G}, \mathcal{F})$, with inner product

$$\langle L, M\rangle_{\mathrm{HS}} = \sum_{j\in J} \langle Lf_j, Mf_j\rangle_{\mathcal{F}}, \tag{3.2}$$

which is independent of the orthonormal basis chosen. It is clear the norm (3.1) is recovered from this inner product. Another form for this inner product is

$$\langle L, M\rangle_{\mathrm{HS}} = \sum_{i\in I}\sum_{j\in J} \langle Lf_j, e_i\rangle_{\mathcal{F}} \langle Mf_j, e_i\rangle_{\mathcal{F}}. \tag{3.3}$$

Proof. Since any element of $\mathcal{F}$ can be expanded in terms of its orthonormal basis, we have that this holds in the specific case of the mapping of f_j by L or M,

$$Lf_j = \sum_{i\in I} \alpha_i^{(j)} e_i \qquad Mf_j = \sum_{i'\in I} \beta_{i'}^{(j)} e_{i'}. \tag{3.4}$$

[1] Recall that a Hilbert space has a countable orthonormal basis if and only if it is separable: that is, it has a countable dense subset [6, p. 47].

Substituting these into (3.2), we obtain

$$\begin{aligned}\langle L, M\rangle_{\mathrm{HS}} &= \sum_{j\in J}\left\langle \sum_{i\in I}\alpha_i^{(j)} e_i, \sum_{i'\in I}\beta_{i'}^{(j)} e_{i'}\right\rangle_{\mathcal{F}} \\ &= \sum_{i\in I}\sum_{j\in J}\alpha_i^{(j)}\beta_i^{(j)}.\end{aligned}$$

We obtain the identical result when we substitute (3.4) into (3.3). □

3.2 Rank-one operators, tensor product space

Given $b \in \mathcal{G}$ and $a \in \mathcal{F}$, we define the tensor product $a \otimes b$ as a rank-one operator from $\mathcal{G}$ to $\mathcal{F}$,

$$(b \otimes a) f \;\mapsto\; \langle f, a\rangle_{\mathcal{F}}\, b. \tag{3.5}$$

This is a generalization of the standard outer product in linear algebra, $(ba^\top)f = (a^\top f)b$, if all three of a, b, f were vectors. First, is this operator Hilbert-Schmidt? We compute its norm according to (3.1),

$$\begin{aligned}\|a \otimes b\|_{\mathrm{HS}}^2 &= \sum_{j\in J}\|(a\otimes b) f_j\|_{\mathcal{F}}^2 \\ &= \sum_{j\in J}\left\|a\,\langle b, f_j\rangle_{\mathcal{G}}\right\|_{\mathcal{F}}^2 \\ &= \|a\|_{\mathcal{F}}^2 \sum_{j\in J}\left|\langle b, f_j\rangle_{\mathcal{G}}\right|^2 \\ &= \|a\|_{\mathcal{F}}^2 \|b\|_{\mathcal{G}}^2,\end{aligned} \tag{3.6}$$

where we use Parseval's identity. Thus, the operator is Hilbert-Schmidt.

Given a second Hilbert-Schmidt operator $L \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$, we have the result:

$$\langle L, a\otimes b\rangle_{\mathrm{HS}} = \langle a, Lb\rangle_{\mathcal{F}} \tag{3.7}$$

A particular instance of this result is

$$\langle u\otimes v, a\otimes b\rangle_{\mathrm{HS}} = \langle u, a\rangle_{\mathcal{F}}\,\langle b, v\rangle_{\mathcal{G}}\,. \tag{3.8}$$

Proof. The key result we use is the expansion of b in terms of the orthonormal basis, $b = \sum_{j\in J}\langle b, f_j\rangle_{\mathcal{G}}\, f_j$. Then

$$\begin{aligned}\langle a, Lb\rangle &= \left\langle a, L\left(\sum_j \langle b, f_j\rangle_{\mathcal{G}} f_j\right)\right\rangle_{\mathcal{F}} \\ &= \sum_j \langle b, f_j\rangle_{\mathcal{G}}\,\langle a, Lf_j\rangle_{\mathcal{F}}\end{aligned}$$

and

$$\begin{aligned}\langle a \otimes b, L\rangle_{\mathrm{HS}} &= \sum_j \langle Lf_j, (a \otimes b) f_j\rangle_{\mathcal{F}} \\ &= \sum_j \langle b, f_j\rangle_{\mathcal{G}} \langle Lf_j, a\rangle_{\mathcal{F}}.\end{aligned}$$

To show (3.8), we simply substitute $u \otimes v$ for L above, and then apply the definition (3.5),

$$\begin{aligned}\langle u \otimes v, a \otimes b\rangle_{\mathrm{HS}} \langle a, (u \otimes v)\, b\rangle_{\mathcal{F}} \\ = \langle u, a\rangle_{\mathcal{F}} \langle b, v\rangle_{\mathcal{G}}\end{aligned}$$

□

3.3 Cross-covariance operator

In this section, we define the cross-covariance operator, in the case where $\mathcal{F}$ and $\mathcal{G}$ are reproducing kernel Hilbert spaces with respective kernels k and l, and feature maps ϕ and ψ. This is a generalization of the cross-covariance matrix to infinite dimensional feature spaces. The results we want are feature space analogues to:

$$\widetilde{C}_{XY} = \mathbf{E}\left(\mathsf{x}\mathsf{y}^\top\right) \qquad f^\top \widetilde{C}_{XY} g = \mathbf{E}_{\mathsf{x}\mathsf{y}}\left[\left(f^\top \mathsf{x}\right)\left(g^\top \mathsf{y}\right)\right],$$

where we use the notation $\widetilde{C}_{XY}$ to denote a covariance operator without centering. The corresponding centered covariance is

$$C_{XY} := \widetilde{C}_{XY} - \mu_X \mu_Y^\top,$$

where $\mu_X := \mathbf{E}(\mathsf{x})$ and $\mu_Y := \mathbf{E}(\mathsf{y})$We now describe how we can get these results in feature space.

The cross product $\phi(\mathsf{x}) \otimes \psi(\mathsf{y})$ is a random variable in $\mathrm{HS}(\mathcal{G}, \mathcal{F})$: use the result in [9, p. 265] that for all $A \in \mathrm{HS}(\mathcal{G}, \mathcal{F})$, the linear form $\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A\rangle_{\mathrm{HS}}$ is measurable. For the expectation of this random variable to exist (and to be an element of $\mathrm{HS}(\mathcal{G}, \mathcal{F})$), we require the expected norm of $\phi(\mathsf{x}) \otimes \psi(\mathsf{y})$ to be bounded: in other words, $\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}}\right) < \infty$. Given the expectation exists, and writing it $\widetilde{C}_{XY}$, then this expectation is the unique element satisfying

$$\left\langle \widetilde{C}_{XY}, A\right\rangle_{\mathrm{HS}} = \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A\rangle_{\mathrm{HS}} \tag{3.9}$$

Proof. The operator

$$\begin{aligned} T_{\mathsf{x}\mathsf{y}} : \mathrm{HS}(\mathcal{G}, \mathcal{F}) &\to \mathbb{R} \\ A &\mapsto \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A\rangle_{\mathrm{HS}} \end{aligned}$$

is bounded when $\mathbf{E}_{\mathsf{x},\mathsf{y}}\left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}}\right) < \infty$, since by applying first Jensen's inequality, then Cauchy-Schwarz,

$$
\begin{aligned}
|\mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}| &\leq \mathbf{E}_{\mathsf{x},\mathsf{y}} |\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), A \rangle_{\mathrm{HS}}| \\
&\leq \|A\|_{\mathrm{HS}} \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right).
\end{aligned}
$$

Thus by the Riesz representer theorem (Theorem (2)), the covariance operator (3.9) exists. We can make a further simplification to the condition: substituting (3.6), we get the requirement

$$
\begin{aligned}
\mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x}) \otimes \psi(\mathsf{y})\|_{\mathrm{HS}} \right) &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\|\phi(\mathsf{x})\|_{\mathcal{F}} \|\phi(\mathsf{y})\|_{\mathcal{G}} \right) \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\sqrt{k(\mathsf{x},\mathsf{x}) l(\mathsf{y},\mathsf{y})} \right) < \infty.
\end{aligned}
$$

We could also use the weaker condition $\mathbf{E}_{\mathsf{x},\mathsf{y}} \left(k(\mathsf{x},\mathsf{x}) l(\mathsf{y},\mathsf{y}) \right)$, which is implied from the above by Jensen's inequality. □

We now use the particular element $f \otimes g$. Combining (3.7) and (3.9), we have the result

$$
\begin{aligned}
\left\langle f, \widetilde{C}_{XY} g \right\rangle_{\mathcal{F}} &= \left\langle \widetilde{C}_{XY}, f \otimes g \right\rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), f \otimes g \rangle_{\mathrm{HS}} \\
&= \mathbf{E}_{\mathsf{x}\mathsf{y}} \left[\langle f, \phi(\mathsf{x}) \rangle_{\mathcal{F}} \langle g, \psi(\mathsf{y}) \rangle_{\mathcal{G}} \right] \\
&= \mathbf{E}_{\mathsf{x}\mathsf{y}} \left[f(\mathsf{x}) g(\mathsf{y}) \right] = \mathrm{cov}(f, g).
\end{aligned}
$$

What does this operator look like? To see this, we apply it to $k(x, \cdot) l(y, \cdot)$ (just as we plotted the mean embedding by evaluating it on $k(x, \cdot)$).

We are given an i.i.d. sample from $\mathbf{P} = \mathbf{P}_{\mathsf{x}} \mathbf{P}_{\mathsf{y}}$, written $\mathbf{z} := ((x_1, y_1) \dots (x_n, y_n))$. Write the empirical

$$
\widehat{C}_{XY} := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i) - \hat{\mu}_x \otimes \hat{\mu}_y,
$$

where we have now included the *centering terms* $\hat{\mu}_x := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$. With some algebra, this can be written

$$
\widehat{C}_{XY} = \frac{1}{n} X H Y^\top,
$$

where $H = I_n - n^{-1} \mathbf{1}_n$, and $\mathbf{1}_n$ is an $n \times n$ matrix of ones, and

$$
X = \begin{bmatrix} \phi(x_1) & \dots & \phi(x_n) \end{bmatrix} \qquad Y = \begin{bmatrix} \psi(y_1) & \dots & \psi(y_n) \end{bmatrix}.
$$

Define the kernel matrices

$$
K_{ij} = \left(X^\top X \right)_{ij} = k(x_i, x_j) \qquad L_{ij} = l(y_i, y_j),
$$

and the kernel matrices between centred variables,

$$
\widetilde{K} = HKH \qquad \widetilde{L} = HLH
$$

(exercise: prove that the above are kernel matrices for the variables centred in feature space).

4 Using the covariance operator to detect dependence

There are two measures of dependence we consider: the constrained covariance (COCO), which is the largest singular value of the covariance operator, and the Hilbert-Schmidt Independence Criterion, which is its Hilbert-Schmidt norm.

4.1 Empirical COCO and proof

We now derive the functions satisying

$$\begin{aligned} \text{maximize} \quad & \left\langle g, \widehat{C}_{XY} f \right\rangle_{\mathcal{G}} \\ \text{subject to} \quad & \|f\|_{\mathcal{F}} = 1 \qquad (4.1) \\ & \|g\|_{\mathcal{G}} = 1 \qquad (4.2) \end{aligned}$$

We assume that

$$f = \sum_{i=1}^{n} \alpha_i \left[\phi(x_i) - \hat{\mu}_x\right] = XH\alpha \qquad g = \sum_{j=1}^{n} \beta_i \left[\psi(y_i) - \hat{\mu}_y\right] = YH\beta,$$

where

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i) \qquad \hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i).$$

The associated Lagrangian is

$$\mathcal{L}(f, g, \lambda, \gamma) = -f^\top \widehat{C}_{XY} g + \frac{\lambda}{2}\left(\|f\|_{\mathcal{F}}^2 - 1\right) + \frac{\gamma}{2}\left(\|g\|_{\mathcal{G}}^2 - 1\right),$$

where we have negated the covariance to make it a minimization problem (for consistency with the optimisation lecture later in the course), and we divide the Lagrange multipliers by 2 to simplify the discussion later . We now write this in terms of α and β:

$$\begin{aligned} f^\top \widehat{C}_{XY} g &= \frac{1}{n}\alpha^\top H X^\top \left(XHY^\top\right) YH\beta \\ &= \frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta, \end{aligned}$$

where we note that $H = HH$. Similarly

$$\|f\|_{\mathcal{F}}^2 = \alpha^\top HXX^\top H\alpha = \alpha^\top \widetilde{K}\alpha.$$

Substituting these into the Lagrangian, we get a new optimization in terms of α and β,

$$\mathcal{L}(\alpha, \beta, \lambda, \gamma) = -\frac{1}{n}\alpha^\top \widetilde{K}\widetilde{L}\beta + \frac{\lambda}{2}\left(\alpha^\top \widetilde{K}\alpha - 1\right) + \frac{\gamma}{2}\left(\beta^\top \widetilde{L}\beta - 1\right). \qquad (4.3)$$

We must minimise this wrt the primal variables α, β. Differentiating wrt α and β and setting the resulting expressions to zero,[2] we obtain

$$-\frac{1}{n}\widetilde{K}\widetilde{L}\beta + \lambda\widetilde{K}\alpha = 0 \tag{4.4}$$

$$-\frac{1}{n}\widetilde{L}\widetilde{K}\alpha + \gamma\widetilde{L}\beta = 0 \tag{4.5}$$

Multiply the first equation by $\alpha^\top$, and the second by $\beta^\top$,

$$\frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta = \lambda\alpha^\top\widetilde{K}\alpha$$
$$\frac{1}{n}\beta^\top\widetilde{L}\widetilde{K}\alpha = \gamma\beta^\top\widetilde{L}\beta$$

Subtracting the first expression from the second, we get

$$\lambda\alpha^\top\widetilde{K}\alpha = \gamma\beta^\top\widetilde{L}\beta.$$

Thus for $\lambda \neq 0$ and $\gamma \neq 0$, we conclude that $\lambda = \gamma$. Making this replacement in (4.4) and (4.5), we must find the largest[3] γ that solves the following expression wrt α, β:

$$\begin{bmatrix} 0 & \frac{1}{n}\widetilde{K}\widetilde{L} \\ \frac{1}{n}\widetilde{L}\widetilde{K} & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \gamma \begin{bmatrix} \widetilde{K} & 0 \\ 0 & \widetilde{L} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{4.6}$$

This is a generalized eigenvalue problem, and can be solved straightforwardly in Matlab. The maximum eigenvalue is indeed COCO: at the solution, $\alpha^\top\widetilde{K}\alpha = 1$ and $\beta^\top\widetilde{L}\beta = 1$, hence the two norm terms in the Lagrangian (4.3) vanish.[4]

[2] We use [5, eqs. (61) and (73)]

$$\frac{\partial a^\top U a}{\partial a} = (U + U^\top)a, \qquad \frac{\partial v^\top a}{\partial a} = \frac{\partial a^\top v}{\partial a} = v.$$

[3] Given $\lambda = \gamma$, the system of equations (4.4) and (4.5) becomes:

$$\frac{1}{n}\widetilde{K}\widetilde{L}\beta = \gamma\widetilde{K}\alpha$$
$$\frac{1}{n}\widetilde{L}\widetilde{K}\alpha = \gamma\widetilde{L}\beta$$

However we also get a valid solution by switching $\check{\beta} := -\beta$,

$$\frac{1}{n}\widetilde{K}\widetilde{L}\check{\beta} = -\gamma\widetilde{K}\alpha$$
$$\frac{1}{n}\widetilde{L}\widetilde{K}\alpha = -\gamma\widetilde{L}\check{\beta}$$

In other words, the solutions γ of the generalised eigenvalue problem (4.6) come in pairs $\pm\gamma$, depending on the relative sign of α and β.

[4] For a more roundabout way of reaching the same conclusion: pre-multiply (4.6) by $[\alpha^\top \beta^\top]$ to get the system of equations

$$\begin{bmatrix} \frac{1}{n}\alpha^\top\widetilde{K}\widetilde{L}\beta \\ \frac{1}{n}\beta^\top\widetilde{L}\widetilde{K}\alpha \end{bmatrix} = \gamma \begin{bmatrix} \frac{1}{n}\alpha^\top\widetilde{K}\alpha \\ \frac{1}{n}\beta^\top\widetilde{L}\beta \end{bmatrix} = \gamma \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

where in the final line we substitute the constraints from (4.1).

4.2 The Hilbert-Schmidt Independence Criterion

4.2.1 Population expression

What is the Hilbert-Schmidt norm of the covariance operator?[5] Consider the centered, squared norm of the RKHS covariance operator,

$$\begin{aligned} HSIC^2(\mathcal{F}, \mathcal{G}, P_{xy}) &= \|\widetilde{C}_{XY} - \mu_X \otimes \mu_Y\|^2_{\text{HS}} \\ &= \left\langle \widetilde{C}_{XY}, \widetilde{C}_{XY} \right\rangle_{\text{HS}} + \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\text{HS}} - 2 \left\langle \widetilde{C}_{XY}, \mu_X \otimes \mu_Y \right\rangle_{\text{HS}}, \end{aligned}$$

where $\widetilde{C}_{XY}$ is the uncentered covariance operator defined in (3.9). There are three terms in the expansion.

To obtain the first term, we apply (3.9) twice, denoting by $(\mathsf{x}', \mathsf{y}')$ an independent copy of the pair of variables (x, y),

$$\begin{aligned} \|\widetilde{C}_{XY}\|^2_{\text{HS}} &= \left\langle \widetilde{C}_{XY}, \widetilde{C}_{XY} \right\rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left\langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \widetilde{C}_{XY} \right\rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \phi(\mathsf{x}') \otimes \psi(\mathsf{y}') \rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \langle \phi(\mathsf{x}), [\phi(\mathsf{x}') \otimes \psi(\mathsf{y}')] \psi(\mathsf{y}) \rangle_{\mathcal{F}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} \left[\langle \phi(\mathsf{x}), \phi(\mathsf{x}') \rangle_{\mathcal{F}} \langle \psi(\mathsf{y}'), \psi(\mathsf{y}) \rangle_{\mathcal{G}} \right] \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} k(\mathsf{x}, \mathsf{x}') l(\mathsf{y}, \mathsf{y}') \\ &=: A \end{aligned}$$

Similar reasoning can be used to show

$$\begin{aligned} \langle \mu_X \otimes \mu_Y, \mu_X \otimes \mu_Y \rangle_{\text{HS}} &= \langle \mu_X, \mu_X \rangle_{\mathcal{F}} \langle \mu_Y, \mu_Y \rangle_{\mathcal{G}} \\ &= \mathbf{E}_{\mathsf{x}\mathsf{x}'} k(\mathsf{x}, \mathsf{x}') \mathbf{E}_{\mathsf{y}\mathsf{y}'} l(\mathsf{y}, \mathsf{y}') \\ &=: D, \end{aligned}$$

and for the cross-terms,

$$\begin{aligned} \left\langle \widetilde{C}_{XY}, \mu_X \otimes \mu_Y \right\rangle_{\text{HS}} &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \langle \phi(\mathsf{x}) \otimes \psi(\mathsf{y}), \mu_X \otimes \mu_Y \rangle_{\text{HS}} \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\langle \phi(\mathsf{x}), \mu_X \rangle_{\mathcal{F}} \langle \phi(\mathsf{y}), \mu_Y \rangle_{\mathcal{G}} \right) \\ &= \mathbf{E}_{\mathsf{x},\mathsf{y}} \left(\mathbf{E}_{\mathsf{x}'} k(\mathsf{x}, \mathsf{x}') \mathbf{E}_{\mathsf{y}'} l(\mathsf{y}, \mathsf{y}') \right) \\ &=: B. \end{aligned}$$

[5] Other norms of the operator may also be used in determining dependence, e.g. the spectral norm from the previous section. Another statistic on the kernel spectrum is the Kernel Mutual Information, which is an upper bound on the true mutual information near independence, but is otherwise difficult to interpret [4]. One can also define independence statistics on the correlation operator [1], which may be better behaved for small sample sizes, although the asymptotic behavior is harder to analyze.

4.2.2 Biased estimate

A biased estimate of HSIC was given in [3]. We observe a sample $Z := \{(x_1, y_1), \dots (x_n, y_n)\}$ drawn independently and identically from P_{xy}, we wish to obtain empirical expressions for HSIC,

$$HSIC^2(\mathcal{F}, \mathcal{G}, Z) := \widehat{A} - 2\widehat{B} + \widehat{D}.$$

A direct approach would be to replace the population uncentred covariance operator $\widetilde{C}_{XY}$ with an empirical counterpart,

$$\check{C}_{XY} = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i),$$

and the population mean embeddings with their respective empirical estimates,

$$\hat{\mu}_x = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i), \qquad \hat{\mu}_y = \frac{1}{n} \sum_{i=1}^{n} \psi(y_i),$$

however the resulting estimates are biased (we will show the amount of bias in the next section). The first term is

$$\begin{aligned} \widehat{A}_b = \left\| \check{C}_{XY} \right\|^2 &= \left\langle \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i), \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i) \right\rangle_{\text{HS}} \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k_{ij} l_{ij} = \frac{1}{n^2} \text{tr}(KL), \end{aligned}$$

we use the shorthand $k_{ij} = k(x_i, x_j)$, and the subscript b to denote a biased estimate. The expression is not computationally efficient, and is written this way for later use - in practice, we would never take the matrix product if the intent was then to compute the trace. Next,

$$\begin{aligned} \widehat{B}_b = \left\langle \check{C}_{XY}, \hat{\mu}_X \otimes \hat{\mu}_Y \right\rangle &= \left\langle \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \psi(y_i), \left(\frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \right) \otimes \left(\frac{1}{n} \sum_{i=1}^{n} \psi(y_i) \right) \right\rangle_{\text{HS}} \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\langle \phi(x_i), \left(\frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \right) \right\rangle_{F} \left\langle \psi(y_i), \left(\frac{1}{n} \sum_{i=1}^{n} \psi(y_i) \right) \right\rangle_{\mathcal{G}} \\ &= \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{q=1}^{n} k_{ij} l_{iq} \\ &= \frac{1}{n^3} 1_n^\top KL 1_n = \frac{1}{n^3} 1_n^\top LK 1_n \end{aligned}$$

(we will use both forms to get our final biased estimate of HSIC), and

$$\begin{aligned}\widehat{D}_b = \langle \hat{\mu}_X \otimes \hat{\mu}_Y, \hat{\mu}_X \otimes \hat{\mu}_Y \rangle &= \left\langle \left(\frac{1}{n}\sum_{i=1}^n \phi(x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \psi(y_i)\right), \left(\frac{1}{n}\sum_{i=1}^n \phi(x_i)\right) \otimes \left(\frac{1}{n}\sum_{i=1}^n \psi(y_i)\right) \right\rangle_{\text{HS}} \\ &= \frac{1}{n^4}\left(\sum_{i=1}^n \sum_{j=1}^n k(x_i, x_j)\right)\left(\sum_{i=1}^n \sum_{j=1}^n l(y_i, y_j)\right) \\ &= \frac{1}{n^4}\left(1_n^\top K 1_n\right)\left(1_n^\top L 1_n\right)\end{aligned}$$

We now combine these terms, to obtain the biased estimate

$$\begin{aligned}HSIC_b^2(\mathcal{F}, \mathcal{G}, Z) &= \frac{1}{n^2}\left(\text{tr}(KL) - \frac{2}{n} 1_n^\top KL 1_n + \frac{1}{n^2}\left(1_n^\top K 1_n\right)\left(1_n^\top L 1_n\right)\right) \\ &= \frac{1}{n^2}\left[\text{tr}(KL) - \frac{1}{n}\text{tr}\left(1_n 1_n^\top KL\right) - \frac{1}{n}\text{tr}\left(K 1_n 1_n^\top L\right) + \frac{1}{n^2}\text{tr}\left(1_n 1_n^\top K 1_n 1_n^\top L\right)\right] \\ &= \frac{1}{n^2}\text{tr}\left[\left(I - \frac{1}{n} 1_n 1_n^\top\right) K \left(I - \frac{1}{n} 1_n 1_n^\top\right) L\right] \\ &= \frac{1}{n^2}\text{tr}(KHLH)\end{aligned}$$

where we define

$$H := I - \frac{1}{n} 1_n 1_n^\top$$

as a centering matrix (when pre-multiplied by a matrix it centers the rows; when post-multiplied, it centers the columns).

4.2.3 Unbiased estimate

An unbiased estimate of $A := \|\widetilde{C}_{XY}\|_{\text{HS}}^2$ is

$$\widehat{A} := \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n k_{ij} l_{ij} = \frac{1}{(n)_2} \sum_{(i,j) \in \mathbf{i}_2^n} k_{ij} l_{ij},$$

where $\mathbf{i}_p^n$ is the set of all p-tuples drawn from $\{1, \ldots, n\}$, and

$$(n)_p = \frac{n!}{(n-p)!} = \frac{1}{n(n-1)\ldots(n-p+1)}.$$

Note that $\mathbf{E}(\widehat{A}) = \mathbf{E}_{\mathsf{x},\mathsf{y}} \mathbf{E}_{\mathsf{x}',\mathsf{y}'} k(\mathsf{x}, \mathsf{x}') l(\mathsf{y}, \mathsf{y}')$, which is not true of the biased expression (which does not properly treat the independent copies x' of x and y' of

y). The difference between the biased and unbiased estimates is

$$\begin{aligned}\widehat{A}_b - \widehat{A} &= \frac{1}{n^2}\sum_{i,j=1}^{n} k_{ij}l_{ij} - \frac{1}{n(n-1)}\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} k_{ii}l_{ii} + \left(\frac{1}{n^2} - \frac{1}{n(n-1)}\right)\left(\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij}\right) \\ &= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} k_{ii}l_{ii} - \frac{1}{n(n-1)}\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij}l_{ij}\right),\end{aligned}$$

thus the *expectation* of this difference (i.e., the bias) is $O(n^{-1})$.

The unbiased estimates of the remaining two terms are

$$\widehat{B} := \frac{1}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij}l_{iq}$$

and

$$\widehat{D} := \frac{1}{(n)_4}\sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij}l_{qr}.$$

While these expressions are unbiased, they are at first sight much more expensive to compute than the respective biased estimates, with $\widehat{B}$ costing $O(n^3)$ and $\widehat{D}$ costing $O(n^4)$. It is possible, however, to obtain these unbiased estimates in $O(n^2)$, i.e., the same cost as the biased estimates, as shown by [7, Theorem 2]. First, we note that diagonal entries of the kernel matrices K and L never appear in the sums, hence we immediately replace these matrices with $\widetilde{K}$ and $\widetilde{L}$ having the diagonal terms set to zero. The term $\widehat{A}$ can be written concisely in matrix form as

$$\widehat{A} = \frac{1}{(n)_2}\left(\widetilde{K}\odot\widetilde{L}\right)_{++} = \frac{1}{(n)_2}\mathrm{trace}\left(\widetilde{K}\widetilde{L}\right),$$

where $\odot$ is the entrywise matrix product and $(A)_{++}$ is the sum of all the entries in A. Looking next at the term $\widehat{B}$, and defining as 1_n the $n\times 1$ vector of ones, we have

$$\begin{aligned}\widehat{B} = \frac{1}{(n)_3}\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij}l_{iq} &= \frac{1}{(n)_3}\left[\sum_{i,j=1}^{n}\sum_{q\neq(i,j)}^{n} k_{iq}l_{qj} - \sum_{i=1}^{n}\sum_{q\neq i}^{n} k_{iq}l_{iq}\right] \\ &= \frac{1}{(n)_3}1_n^\top\begin{bmatrix} \sum_{j=2}^{n} k_{1q}l_{q1} & \cdots & \sum_{q\neq(i,j)}^{n} k_{iq}l_{qj} & \cdots \\ & \ddots & \vdots & \end{bmatrix}1_n \\ &\quad - \frac{1}{(n)_3}\left(\widetilde{K}\odot\widetilde{L}\right)_{++} \\ &= \frac{1}{(n)_3}1_n^\top\tilde{K}\tilde{L}1_n - \frac{1}{(n)_3}\left(\widetilde{K}\odot\widetilde{L}\right)_{++}.\end{aligned}$$

The first expression in the final line can be computed in time $O(n^2)$, as long as the matrix-vector products are taken first. Finally, looking at the fourth term,[6]

$$\begin{aligned}
\widehat{D} &= \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij} l_{qr} = \frac{1}{(n)_4} \Bigg[\sum_{(i,j)\in \mathbf{i}_2^n} \sum_{(q,r)\in \mathbf{i}_2^n} k_{ij} l_{qr} \\
&\quad - \underbrace{\sum_{(i,j,r)\in \mathbf{i}_3^n} k_{ij} l_{ir}}_{q=i} - \underbrace{\sum_{(i,j,r)\in \mathbf{i}_3^n} k_{ij} l_{jr}}_{q=j} - \underbrace{\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij}}_{(q=i,r=j)\equiv(q=j,r=i)} \\
&\quad - \underbrace{\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij} l_{iq}}_{r=i} - \underbrace{\sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij} l_{jq}}_{r=j} - \underbrace{\sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij}}_{(r=i,q=j)\equiv(r=j,q=i)} \Bigg] \\
&= \frac{1}{(n)_4} \left[\left(\sum_{i=1}^n \sum_{j\neq i}^n k_{ij} \right) \left(\sum_{i=1}^n \sum_{j\neq i}^n l_{ij} \right) - 4 1_n^\top \tilde{K} \tilde{L} 1_n + 2 \left(\widetilde{K} \odot \widetilde{L} \right)_{++} \right] \\
&= \frac{1}{(n)_4} \left[\left(1_n^\top \widetilde{K} 1_n \right) \left(1_n^\top \widetilde{L} 1_n \right) - 4 1_n^\top \tilde{K} \tilde{L} 1_n + 2 \left(\widetilde{K} \odot \widetilde{L} \right)_{++} \right],
\end{aligned}$$

which can also be computed in $O(n^2)$. We now establish the net contribution of each term:

$$\begin{aligned}
\left(\widetilde{K} \odot \widetilde{L} \right)_{++} : & \frac{1}{(n)_2} + \frac{2}{(n)_3} + \frac{2}{(n)_4} \\
&= \frac{(n-2)(n-3) + (2n-6) + 2}{(n)_4} \\
&= \frac{(n-2)(n-1)}{(n)_4}
\end{aligned}$$

and

$$\begin{aligned}
1_n^\top \tilde{K} \tilde{L} 1_n \ : & \frac{-2}{(n)_3} - \frac{4}{(n)_4} \\
&= \frac{-2(n-3) - 4}{(n)_4} = \frac{-2(n-1)}{(n)_4}.
\end{aligned}$$

Thus, we have our empirical unbiased HSIC expression,

$$HSIC^2(\mathcal{F}, \mathcal{G}, Z) := \frac{1}{n(n-3)} \left[\left(\widetilde{K} \odot \widetilde{L} \right)_{++} - \frac{2}{(n-2)} 1_n^\top \tilde{K} \tilde{L} 1_n + \frac{1}{(n-1)(n-2)} \left(1_n^\top \widetilde{K} 1_n \right) \left(1_n^\top \widetilde{L} 1_n \right) \right]$$

[6]The equivalences $\equiv$ in the first line below indicate that both index matching constraints amount to the same thing, hence these terms appear only once.

5 HSIC for feature selection

As we saw in the previous section, a biased estimate for the centred HSIC can be written

$$\text{HSIC} := \frac{1}{n^2}\text{trace}(KHLH).$$

Consider the case where we wish to find a subset of features that maximizes HSIC with respect to some set of labels. Assume we have a sample $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$, and binary class labels. We choose a particular form for the class labels: $y_i \in \left\{n_+^{-1}, -n_-^{-1}\right\}$, where n_+ is the number of positive labels and n_- is the number of negative labels.

We denote by $x_i[\ell]$ the ℓth coordinate of x_i, and write

$$x[\ell] := \left[\begin{array}{ccc} x_1[\ell] & \ldots & x_n[\ell] \end{array}\right]^\top$$

the column vector of the ℓth coordinate of *all* samples. If we use a linear kernel on the x_i, then

$$K_{i,j} = x_i^\top x_j = \sum_{\ell=1}^{d} x_i[\ell]x_j[\ell].$$

It follows we can write the kernel as the sum of kernels on individual dimensions,

$$K = \sum_{\ell=1}^{d} K_\ell,$$

where $K_\ell := x[\ell]x[\ell]^\top$. In this case, HSIC is the sum of HSIC values for each such kernel,

$$\text{HSIC} := \frac{1}{n^2}\sum_{\ell=1}^{d}\text{trace}(K_\ell HLH).$$

What happens when we choose a linear kernel on the labels? Assuming the classes are grouped together,

$$L = yy^\top = \left[\begin{array}{cc} n_+^{-2}\mathbf{I} & -n_+n_-\mathbf{I} \\ -n_+n_-\mathbf{I} & n_-^{-2}\mathbf{I} \end{array}\right],$$

where y is the vector of all class labels. Note further than

$$\sum_{i=1}^{n} y_i = 0,$$

and hence $HLH = L$. Finally, using $\mathrm{trace}(AB) = \mathrm{trace}(BA)$,

$$\begin{aligned}
\mathrm{HSIC} &= \frac{1}{n^2}\sum_{\ell=1}^{d}\mathrm{trace}(K_\ell L) \\
&= \frac{1}{n^2}\sum_{\ell=1}^{d}\mathrm{trace}(x[\ell]x[\ell]^\top yy^\top) \\
&= \frac{1}{n^2}\sum_{\ell=1}^{d}\left(\frac{1}{n_+}\sum_{i=1}^{n_+}x_i[l] - \frac{1}{n_-}\sum_{i=n_++1}^{n}x_i[l]\right)^2
\end{aligned}$$

6 Acknowledgments

Thanks to Aaditya Ramdas, Wittawat Jitkrittum, and Dino Sejdinovic for corrections and improvements to these notes.

References

[1] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

[2] R. M. Dudley. *Real analysis and probability.* Cambridge University Press, Cambridge, UK, 2002.

[3] A. Gretton, O. Bousquet, A. J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference*, pages 63–78, 2005.

[4] A. Gretton, R. Herbrich, A. J. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[5] K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2008. Version 20081110.

[6] M. Reed and B. Simon. *Methods of modern mathematical physics. Vol. 1: Functional Analysis.* Academic Press, San Diego, 1980.

[7] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *JMLR*, 13:1393–1434, 2012.

[8] Ingo Steinwart and Andreas Christmann. *Support Vector Machines.* Information Science and Statistics. Springer, 2008.

[9] L. Zwald, O. Bousquet, and G. Blanchard. Statistical properties of kernel principal component analysis. In *Proc. Annual Conf. Computational Learning Theory*, 2004.