Assignment 2

1 Kernels (50 points)

1.1 (Distance in feature space)

Using the fact that the dot-product is a symmetric bilinear form, we have that

$$\begin{aligned}||\phi(x)-\phi(z)||^2 &= \langle\phi(x)-\phi(z),\phi(x)-\phi(z)\rangle\\ &= \langle\phi(x),\phi(x)-\phi(z)\rangle-\langle\phi(z),\phi(x)-\phi(z)\rangle\\ &= \langle\phi(x),\phi(x)\rangle-\langle\phi(x),\phi(z)\rangle-\langle\phi(z),\phi(x)\rangle+\langle\phi(z),\phi(z)\rangle\\ &= \langle\phi(x),\phi(x)\rangle-2\langle\phi(x),\phi(z)\rangle+\langle\phi(z),\phi(z)\rangle\end{aligned}$$

Now using the fact that k is a kernel for ϕ, such that $k(x,x')=\langle\phi(x),\phi(x')\rangle$ and taking square roots in the above equation, yields that,

$$||\phi(x)-\phi(z)|| = \sqrt{k(x,x)-2k(x,z)+k(z,z)}$$

as desired.

1.2 (Sum of kernels)

Let K_1 and K_2 be the Gram matrices corresponding to k_1 and k_2. Then the Gram matrix corresponding to $k(x,z)=ak_1(x,z)+bk_2(x,z)$, where $a,b\in\mathbb{R}^+$. The corresponding Gram matrix K is thus given by

$$K = aK_1+bK_2$$

By the distributive and associative properties of matrix multiplication,

$$x^TKx = x^T(aK_1+bK_2)x = x^T(aK_1x+bK_2x) = x^TaK_1x+x^TbK_2x = a(x^TK_1x)+b(x^TK_2x)$$

Now using the fact that a and b are positive, and that K_1 and K_2 are positive definite by assumption, then

$$x^TKx = a(x^TK_1x)+b(x^TK_2x)\geq 0$$

Which shows that K is positive definite, and thus that k is a positive definite kernel.

1.3 (Rank of Gram matrix)

Let $X\in\mathbb{R}^{d\times m}$ be defined as the matrix with x_i as the i'th column. Then $X^T\in\mathbb{R}^{m\times d}$, has x_i as the i'th row. Then the Gram matrix K, induced by $k(x,z)=x^Tz$, on $x_1,\cdots,x_m$ can be written as

$$X^TX = \begin{pmatrix}x_1^Tx_1 & \cdots & x_1^Tx_m\\ \vdots & \ddots & \vdots\\ x_m^Tx_1 & \cdots & x_m^Tx_m\end{pmatrix} = \begin{pmatrix}k(x_1,x_1) & \cdots & k(x_1,x_m)\\ \vdots & \ddots & \vdots\\ k(x_m,x_1) & \cdots & k(x_m,x_m)\end{pmatrix} = K$$

We show that $\mathrm{Rank}(X^T X) = \mathrm{Rank}(X) = \mathrm{Rank}(K)$. Let $N(X) = \{y \in \mathbb{R}^m | Xy = 0\}$, be the null space of the matrix X, and let $\mathrm{Null}(X)$ denote the nullity of X - that is the dimension of the null space. We show that $N(X) = N(X^T X)$. First, let $y \in N(X)$. Then

$$Xy = 0 \Rightarrow X^T Xy = 0$$

Which shows that $y \in N(X^T X)$ and thus $N(X) \subset N(X^T X)$. Now, let $y \in N(X^T X)$. Then,

$$X^T Xy = 0$$

Multiplying both sides from the left with y^T, and using the fact that $AB = (B^T A^T)^T$

$$0 = y^T X^T Xy = (Xy)^T (Xy) = \langle Xy, Xy \rangle$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^d$. We know the above equality holds if and only $Xy = 0$, which shows that $y \in N(X)$, and thus $N(X^T X) \subset N(X)$, and finally $N(X) = N(X^T X)$. It follows that the nullity of $X^T X$ and X are the same, we can now compute the rank of $X^T X$ by applying the rank-nullity theorem. We have that,

$$\mathrm{Rank}(X^T X) = m - \mathrm{Null}(X^T X) \quad \text{and} \quad \mathrm{Rank}(X) = m - \mathrm{Null}(X)$$

Using the fact that the nullities are equal it follows that $\mathrm{Rank}(K) = \mathrm{Rank}(X^T X) = \mathrm{Rank}(X)$. The rank of X is upper-bounded by its' number of rows d and its number of columns m. We thus see that, $\mathrm{Rank}(K) = \mathrm{Rank}(X) \leq \min\{d, m\}$, which is our upper bound on the Gram matrix.

2. Numerical comparison of kl inequality with its relaxations and with Hoeffding's inequality (25 points)

1.

We are evaluating the four bounds,

A.

Hoeffdings inequality,

$$p \leq \hat{p} + \sqrt{\frac{\log 1/\delta}{2n}}$$

B.

The kl inequality,

$$p \leq \mathrm{kl}^{-1^+}(\hat{p}_n, \frac{\log \frac{n+1}{\delta}}{n})$$

The upper inverse is found by binary search.

C.

Pinsker's relaxation of the kl inequality,

$$p \leq \hat{p} + \sqrt{\frac{\log \frac{n+1}{\delta}}{2n}}$$

This is not exactly the way it is stated in the lecture notes, but we have the trivial bound $p - \hat{p} \leq |p - \hat{p}|$ which gives us the above bound.

D.

Refined Pinsker's relaxation,

$$p \leq \hat{p} + \sqrt{\frac{2\hat{p}\log\frac{n+1}{\delta}}{n}} + \frac{2\log\frac{n+1}{\delta}}{n}$$

2.

We get the following plot with the four bounds

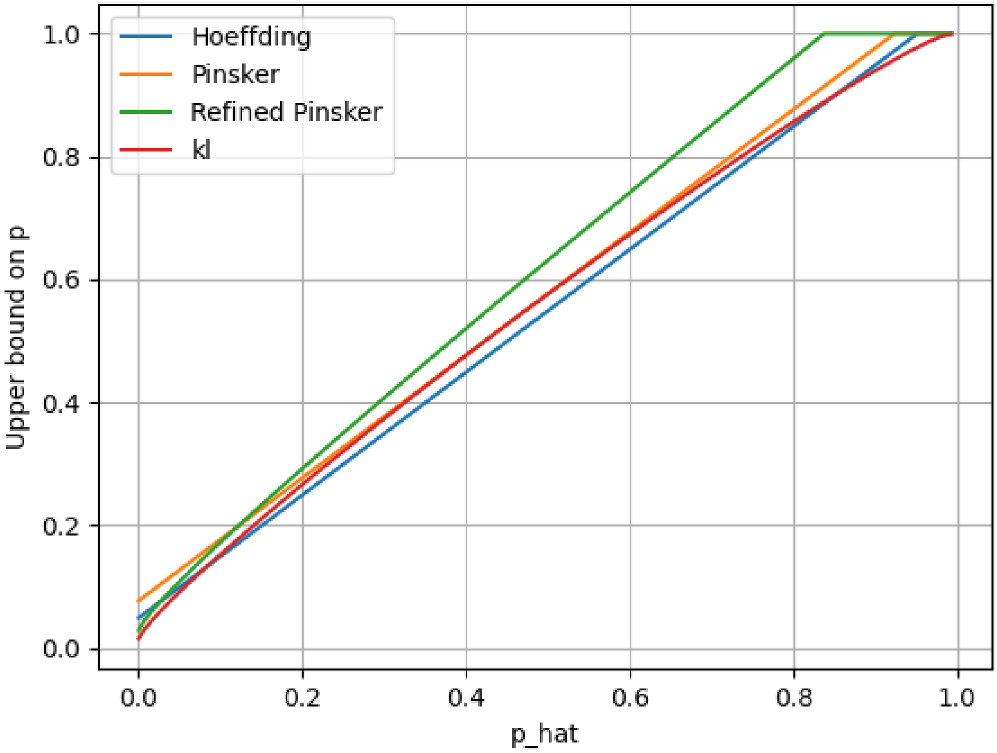

Figure 1: Upper bounds for p as a function of $\hat{p}$

Note that if the bound is greater than 1, we simply replace the bound with 1.

3.

We plot a zoomed in version of the above plot, plotting in the range $\hat{p}_n \in [0, 0.1]$,

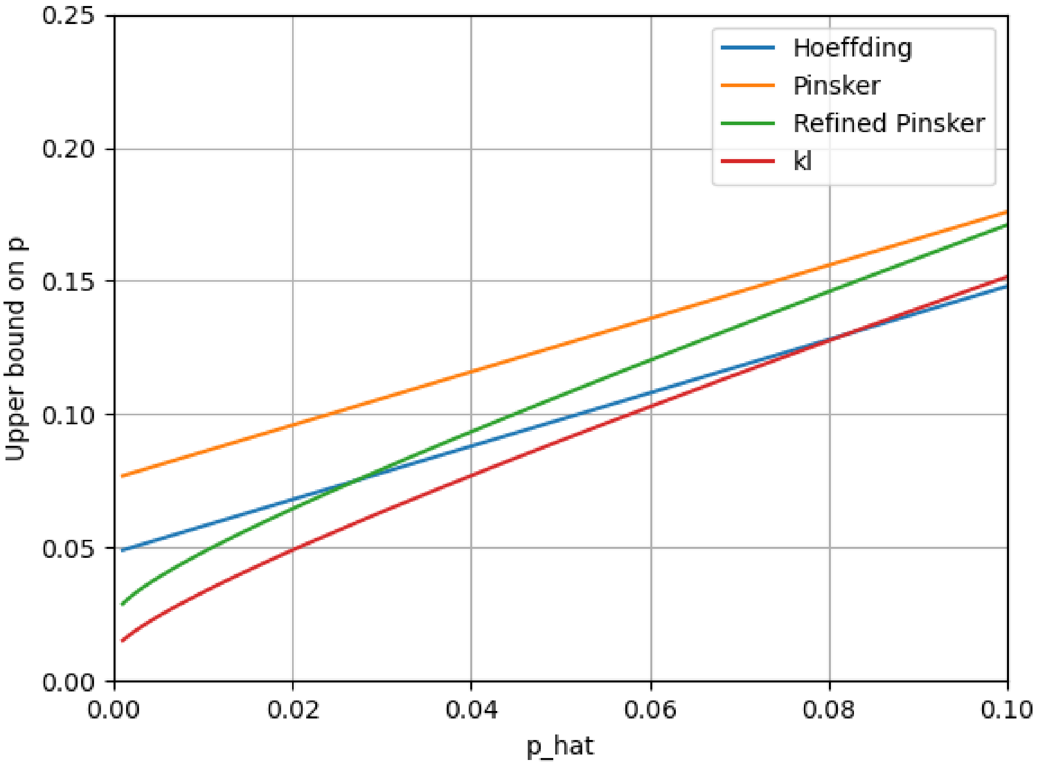

Figure 2: Upper bounds for p as a function of $\hat{p}$, when $\hat{p} \in [0, 0.1]$

4.

We can compute the lower inverse of kl, by using the fact (presented at lecture) that for a fixed $\hat{p_n}$, $\mathrm{kl}(p||\hat{p}_n)$ is symmetric around $\hat{p}_n$. Therefore we have that,

$$\mathrm{kl}^{-1^-}(\hat{p}_n, \frac{\log \frac{n+1}{\delta}}{n}) = 2\hat{p} - \mathrm{kl}^{-1^+}(\hat{p}_n, \frac{\log \frac{n+1}{\delta}}{n})$$

We plot the Hoeffding and kl lower bounds,

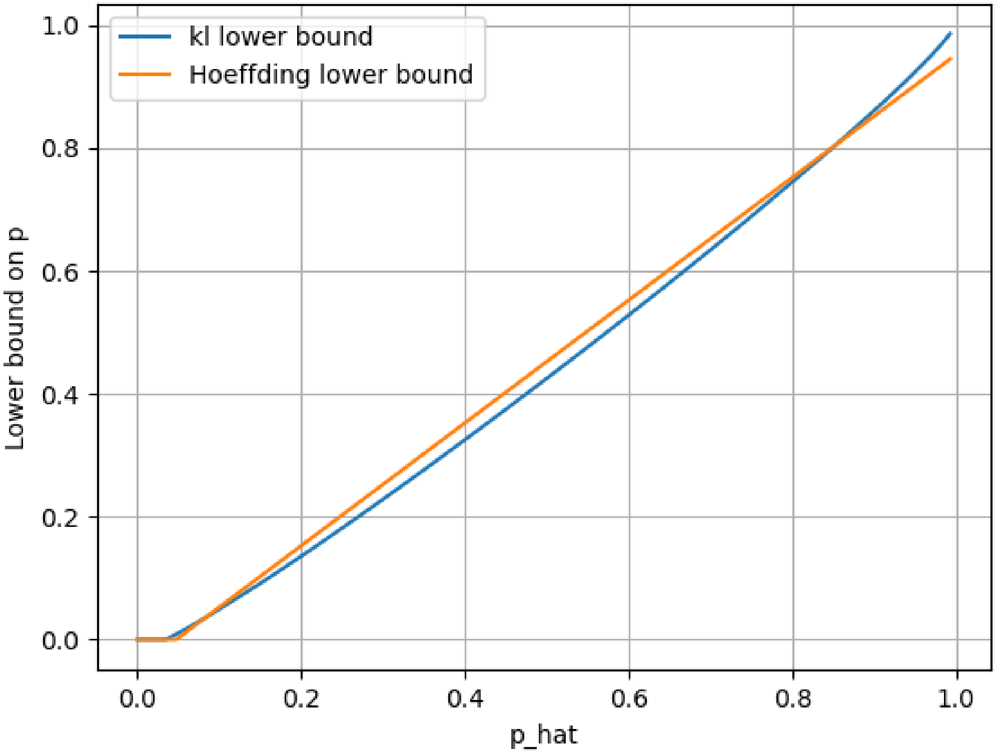

Figure 3: Lower bounds for p as a function of $\hat{p}$

5.

The tightest upper bound is the kl-bound, when $\hat{p}$ is small. For small values of $\hat{p}$ the refined Pinsker inequality is slightly tighter than Hoeffding's inequality and slightly looser than kl (which makes sense, as Pinsker is a relaxation of kl). For larger values of $\hat{p}_n$, Pinsker, Hoeffding, and the kl bound follow each other closely - while the refined Pinsker bound is pretty loose in comparison. We also note that there is not much difference between the kl lower bound and the Hoeffding lower bound. All in all, for small values of $\hat{p}_n$ the kl-bounds and its' relaxation in form of the Pinsker-bound can be significantly tighter than Hoeffding.

Occam's razor with kl inequality (15 points)

By taking a union bound, we have that

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \mathrm{kl}(\hat{L}(h,S)||L(h)) \geq \frac{\log \frac{n+1}{\pi(h)\delta}}{n}\right) \leq \sum_{h \in \mathcal{H}} \mathbb{P}\left(\mathrm{kl}(\hat{L}(h,S)||L(h)) \geq \frac{\log \frac{n+1}{\pi(h)\delta}}{n}\right)$$

Note that under our assumptions, the empirical loss is the sample average of i.i.d. Bernoulli variables. We can therefore apply theorem 2.15, to the above probability. However, one has to be careful, if π was not independet of S, we would not be able to apply the theorem, as the right hand side of the inequality inside the probability, would depend on the sample, which is not allowed when applying theorem 2.15, as the ϵ would not be a fixed number. Anyway, since we have assumed π independent of S, applying theorem 2.15 yields,

$$\sum_{h\in\mathcal{H}} \mathbb{P}\left(\mathrm{kl}(\hat{L}(h,S)||L(h)) \geq \frac{\log \frac{n+1}{\pi(h)\delta}}{n}\right) \leq \sum_{h\in\mathcal{H}} (n+1)\exp(-n\frac{\log \frac{n+1}{\pi(h)\delta}}{n}) = \sum_{h\in\mathcal{H}} \pi(h)\delta$$

And using the assumption that $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$, we finally have that,

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \mathrm{kl}(\hat{L}(h,S)||L(h)) \geq \frac{\log \frac{n+1}{\pi(h)\delta}}{n}\right) \leq \sum_{h\in\mathcal{H}} \pi(h)\delta \leq \delta$$

Which is what we wanted to show.

Refined Pinsker's Lower Bound (10 points)

Assume that $\mathrm{kl}(p||q) \leq \epsilon$. By applying lemma 2.18, we see that,

$$\epsilon \geq \mathrm{kl}(p||q) \geq \frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{1-p,1-q\}}$$

Assume first that $p \geq q$, as both terms in the above sum is positive,

$$\frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{1-p,1-q\}} = \frac{(p-q)^2}{2p} + \frac{(p-q)^2}{2(1-q)} \geq \frac{(p-q)^2}{2p}$$

And we thus have,

$$\epsilon \geq \frac{(p-q)^2}{2p} \Leftrightarrow q \geq p - \sqrt{2p\epsilon}$$

Assume now that $p \leq q$, then it trivially holds that,

$$q \geq p - \sqrt{2p\epsilon}$$

In both cases we have that,

$$q \geq p - \sqrt{2p\epsilon}$$

Which is what we wanted to show.