Machine Learning B
Home Assignment 4

Jakob Schauser, pwn274

December 2021

1 Numerical comparison of kl inequality with its relaxations and with Hoeffding's inequality

1.1

The four bounds are:
Good ol' Hoeffding:

$$p \leq \hat{p}_n + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}} \tag{1}$$

The kl upper bound (here i have found what I believe to be the correct z in the lecture notes):

$$\sup \left\{ p : \mathrm{kl}\left(\hat{p}_n \| p\right) \leq \frac{ln(n+1)/\delta}{n} \right\} \tag{2}$$

Pinsker's inequality:

$$p \leq \hat{p} + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}} \tag{3}$$

The refined Pinsker's inequality:

$$p \leq \hat{p} + \sqrt{\frac{2\hat{p} \ln \frac{n+1}{\delta}}{n}} + \frac{2 \ln \frac{n+1}{\delta}}{n} \tag{4}$$

1.2

For plotting the four bounds, my code is kind of boring. The only thing that isn't simply a direct translation of the bound is the binary search. I have kept everything vectorized, so the search looks about as follows:

```
while True:
        p = (pmin + pmax)/2
        diff = kl(p_emp,p) - z

        close = np.isclose(diff,0)

        if np.all(close):
            return p
        else:
            pmin[diff < 0] = p[diff < 0]
            pmax[diff > 0] = p[diff > 0]
```

There are some problems in this that I have solved in the code but not included in the snipped. This assignment was in general not very focused on the programming side of the course.

The four bounds, all plottet together, looks as follows:

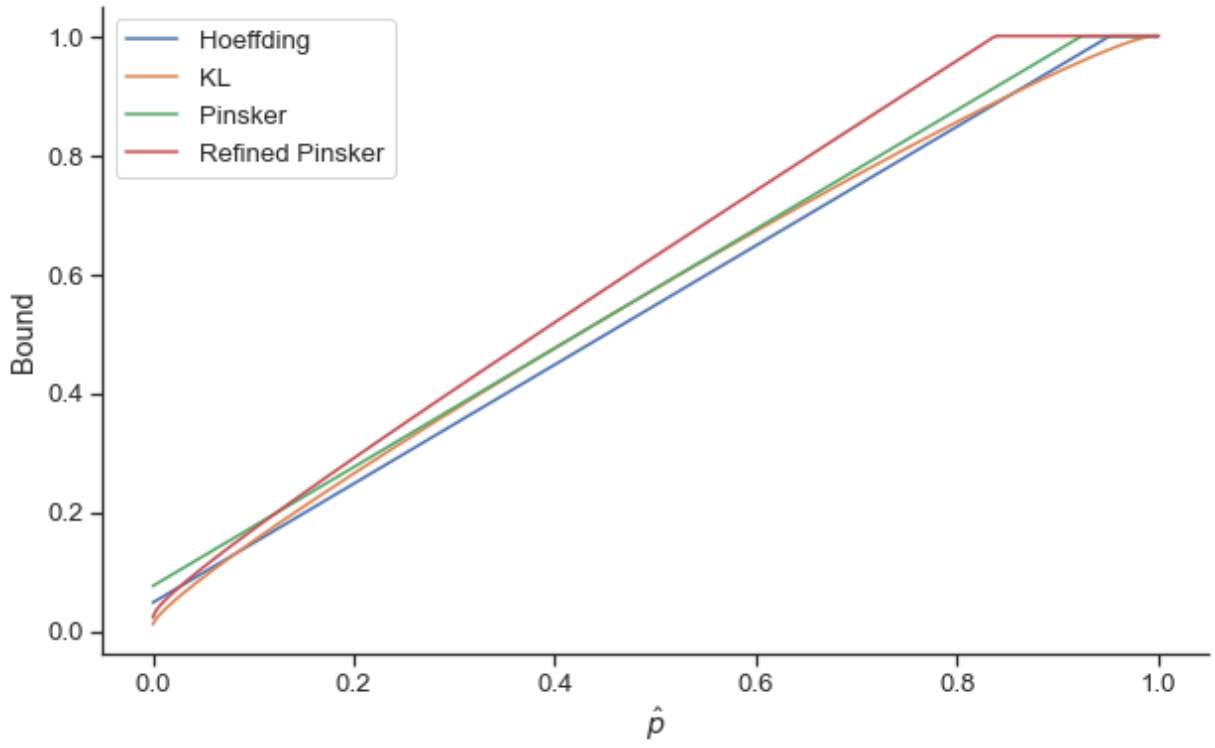

Figure 1: The four bounds

1.3

As can clearly be seen in the last plot, the better bound changes as a function of $\hat{p}$. To better see, the following plot focuses on the small-$\hat{p}$:

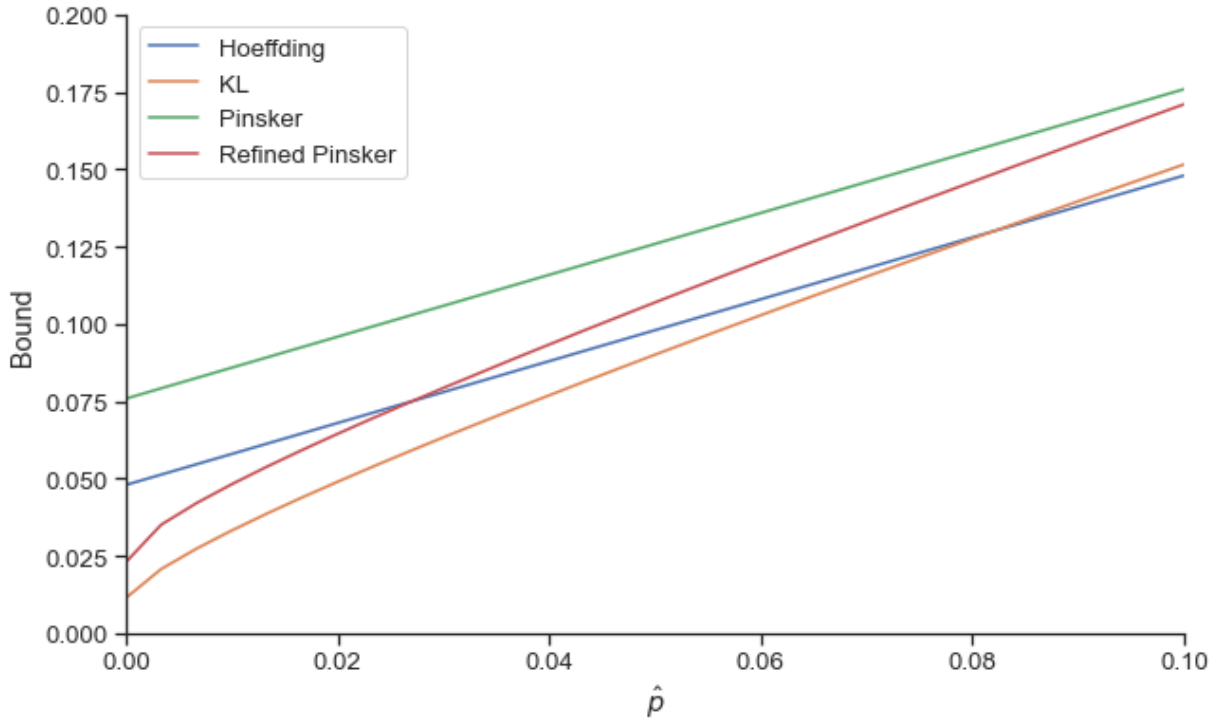

Figure 2: zoomed in version of figure 1

1.4

When looking at the lower bound, only a few corrections were needed for the binary search to work. These consisted of defining new upper/lower bounds for the search and flipping the sign of the inequality-checks. The Hoeffding lower bound and the kl lower bound can be seen here:

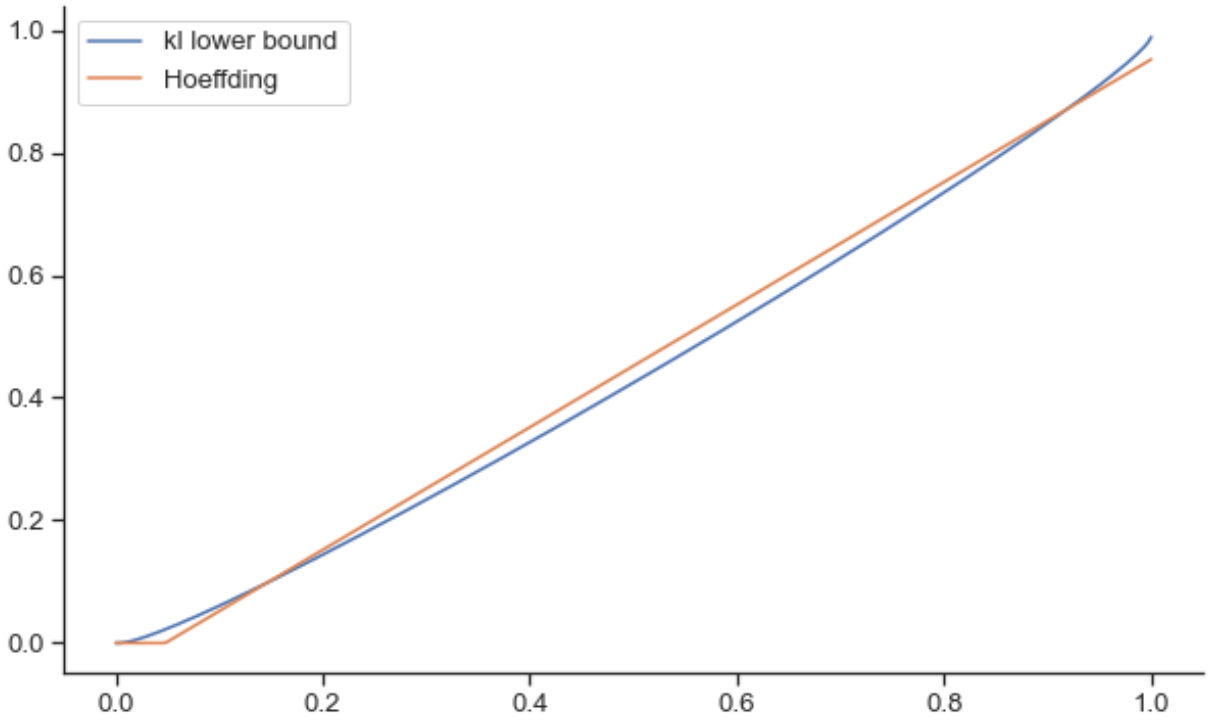

Figure 3: The lower bounds, as asked for

1.5

For small $\hat{p}$, the kl-bound is definitely the smarter choice. This changes for the Pinsker bound at around $\hat{p} = 0.08$. After 0.9 this changes back.
About the same thing happens for the lower bound, except the kl bound is beaten by Hoeffding for non-extreme values of $\hat{p}$.

Generally the kl bound is slightly better in a few specific intervals at the edges of $\hat{p}$.

2 Occam's razor with kl inequality

Need to prove that:

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \mathrm{kl}(\hat{L}(h,S) \| L(h)) \geq \frac{\ln \frac{n+1}{\pi(h)\delta}}{n}\right) \leq \delta \tag{5}$$

I am starting from Pinsker, that states, that with probability greater than $1-\delta$:

$$\mathrm{kl}(\hat{p} \| p) \leq \frac{\ln \frac{n+1}{\delta}}{n} \tag{6}$$

Using the fact that our 0-1-loss meets the requirements of the probability p and a simple negation, this expression it can be rewritten as:

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \mathrm{kl}(\hat{L}(h,S)\|L(h)) \geq \frac{\ln \frac{n+1}{\delta}}{n}\right) \leq \delta \tag{7}$$

Now I can just follow the same logic as we did in the proof for **Theorem 3.3** except for basing it around the kl-inequality instead of Hoeffding's inequality:

$$\begin{aligned}
&\mathbb{P}\left(h \in \mathcal{H} : \mathrm{kl}(\hat{L}\|L) \geq \frac{\ln \frac{n+1}{\pi(h)\delta}}{n}\right) \\
&\leq \sum_{h\in\mathcal{P}} \mathbb{P}\left(\mathrm{kl}(\hat{L}\|L) \geq \frac{\ln \frac{n+1}{\pi(h)\delta}}{n}\right) \\
&\leq \sum_{h\in\mathcal{H}} \frac{n+1}{\exp(\ln((n+1)/(\pi(h)\delta))} = \sum_{h\in\mathcal{H}} \pi(h)\delta \leq \delta
\end{aligned} \tag{8}$$

where I have used the union bound for getting to the second row and both **Lemma 2.14** and the independence of $\pi(h)$ in the third row of the equations.

3 Refined Pinsker's Lower Bound

Prove that if $\mathrm{kl}(p\|q) \leq \varepsilon$ then $q \geq p - \sqrt{2p\varepsilon}$. Taking the hint, I start from **Lemma 2.18**:

$$\mathrm{kl}(p\|q) \geq \frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{(1-p),(1-q)\}} \tag{9}$$

If $\mathrm{kl}(p\|q) \leq \varepsilon$, this also means the right hand side is less than ε:

$$\frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{(1-p),(1-q)\}} \leq \varepsilon \tag{10}$$

Now we have three cases. I start out by looking at $p > q$:

$$\frac{(p-q)^2}{p} + \frac{(p-q)^2}{(1-q)} \leq 2\varepsilon \tag{11}$$

Using the fact that we are looking at an inequality, we can rewrite as follows:

$$\frac{(p-q)^2}{p} \leq 2\varepsilon \tag{12}$$

Now it is just finding the solutions of a quadratic equation:

$$q \geq p - \sqrt{2\varepsilon p} \tag{13}$$

For $q > p$ we require a little more cleverness:

$$\frac{(p-q)^2}{q} + \frac{(p-q)^2}{(1-p)} \leq 2\varepsilon \tag{14}$$

Again, making a looser bound:

$$2\varepsilon \geq \frac{(p-q)^2}{(1-p)} \tag{15}$$

Now, solving for q gives:

$$q \geq p - \sqrt{2\varepsilon(1-p)} \geq p - \sqrt{2\varepsilon p} \tag{16}$$

where I in the last inequality used that we do not want any imaginary solutions and that $p - \sqrt{2\varepsilon} \leq p - \sqrt{2\varepsilon p}$ for $p \in (0, 1)$

The last case, where $q = p$ is trivial as $q \geq q - \sqrt{2q\varepsilon}$ for all $q \in (0, 1)$

4 Bayesian Inference

4.1

We assume a Normal conjugate prior with θ_0 and σ_0 and have from the slides that:

$$f_{\Theta|X}\left(\theta \mid x_1, \ldots, x_n\right) \propto f_{X|\Theta}\left(x_1, \ldots, x_n \mid \theta\right) f_{\Theta}(\theta) \tag{17}$$

Now, from Sadegh's slides:

$$\theta_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\theta_0 + \frac{\sigma_0^2}{n\sigma_0^2 + \sigma^2}\sum_{i=1}^{n} x_i \tag{18}$$

And

$$\sigma_n^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} \tag{19}$$

These two pieces of information should be enough to characterize the distribution, but I am unsure if this solves the problem adequately.

4.2

4.2.1

We have the binomial:

$$p_X(n, k, p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \tag{20}$$

Placing this in the calculations for the Bayesian posterior we get the following:

$$\frac{\left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k}}{\int_p \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k}} = \frac{p^k(1-p)^{n-k}}{\int_p p^k(1-p)^{n-k}} \tag{21}$$

Now, looking at the beta distribution:

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} \tag{22}$$

We can clearly see, that setting $\alpha = k+1$ and $\beta = n-k+1$ the posterior is a beta distribution.

4.2.2

Sorry - I had no time to finish this assignment