

# Machine Learning B (2025)

## Home Assignment 1

Carsten Jørgensen, student ID: skj730

### Contents

1	Numerical comparison of kl inequality with its relaxations and with Hoeffding's inequality (40 points) [Yevgeny]	2
2	Occam's razor with kl inequality (30 points) [Yevgeny]	5
3	Numerical comparison of the kl and split-kl inequalities (30 points) [Yevgeny]	8

# 1 Numerical comparison of kl inequality with its relaxations and with Hoeffding's inequality (40 points) [Yevgeny]

## Bounds

We shall evaluate the following four bounds on  $p$ :

- Hoeffding:  $\hat{p}_n + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$
- The kl inequality:  $\text{kl}^{-1+}(\hat{p}_n, \epsilon) = \max\{p : \text{kl}(\hat{p}_n \| p) \leq \epsilon\}$
- Pinsker's relaxation: identical to Hoeffding according to eq. (2.12) in the lecture notes
- Refined Pinsker's:  $\hat{p}_n + \sqrt{\frac{2\hat{p}_n \ln \frac{1}{\delta}}{n}} + \frac{2 \ln \frac{1}{\delta}}{n}$

## Plot of upper bounds

In figure 1 we plot the upper bounds.

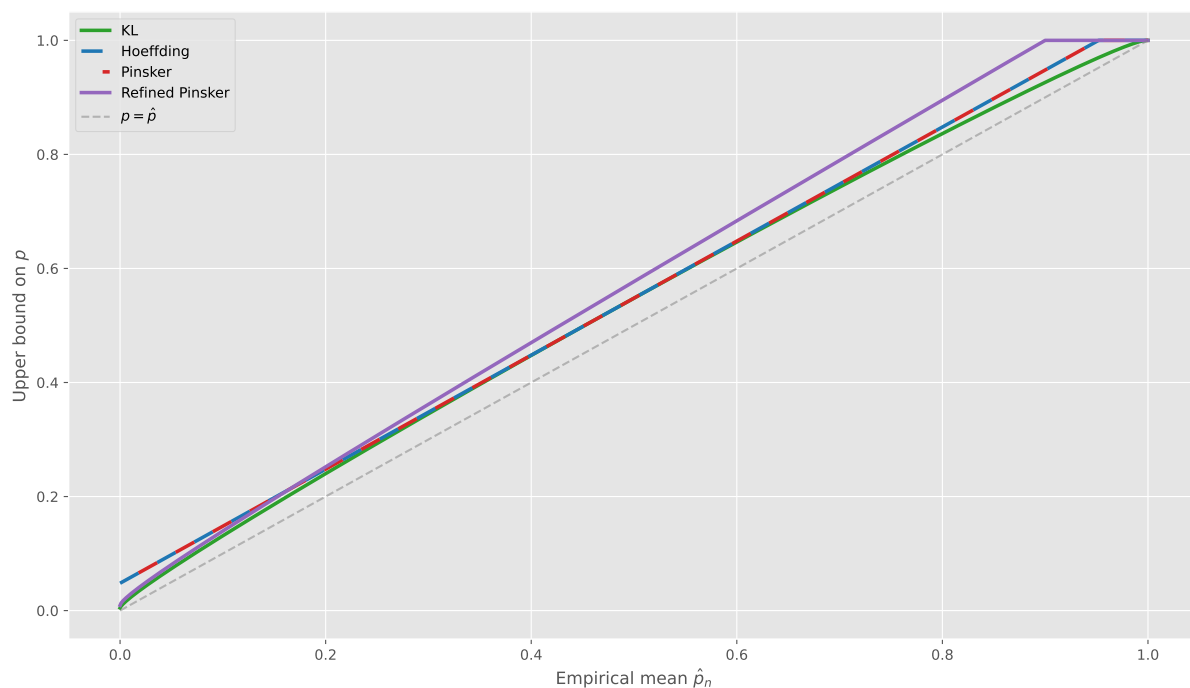


Figure 1: Upper bounds for  $\hat{p}_n \in [0, 1]$

and in figure 2 we plot the same upper bounds "zoomed in" on  $\hat{p}_n \in [0, 0.1]$ .

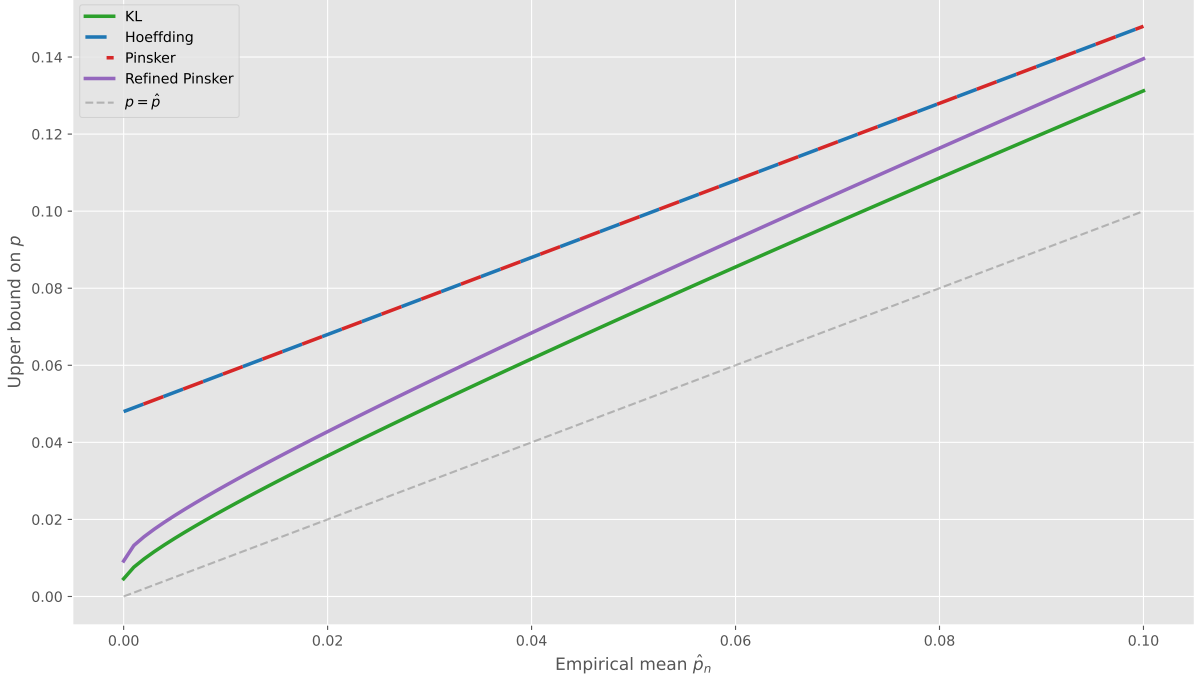


Figure 2: Zoomed upper bounds for  $\hat{p}_n \in [0, 0.1]$

## Plot of lower bounds

The lower bounds are shown in figure 3.

## Computation of upper and lower inverse of kl

Please see 2 for the implementation of the upper inverse of kl. To compute the lower inverse of kl, we note that  $\text{kl}(a||b) = \text{kl}(1-a||1-b)$  for all  $a, b \in [0, 1]$ . So  $\text{kl}(\hat{p}||p) \leq \epsilon$  is equivalent to  $\text{kl}(1-\hat{p}||1-p) \leq \epsilon$ . For  $q = 1-p$  the sets  $p : \text{kl}(\hat{p}||p) \leq \epsilon$  and  $q : \text{kl}(1-\hat{p}||q) \leq \epsilon$  are identical.

By the definition of the upper inverse we have

$$q^+ = \text{kl}^{-1+}(1-\hat{p}, \epsilon) = \max\{q \geq 1-\hat{p} : \text{kl}(1-\hat{p}||q) \leq \epsilon\}$$

We now replace  $q$  with  $1-p$  to obtain

$$p^- = 1 - q^+ = 1 - \text{kl}^{-1+}(1-\hat{p}, \epsilon) = \min\{p \leq \hat{p} : \text{kl}(\hat{p}||p) \leq \epsilon\} = \text{kl}^{-1-}(\hat{p}, \epsilon)$$

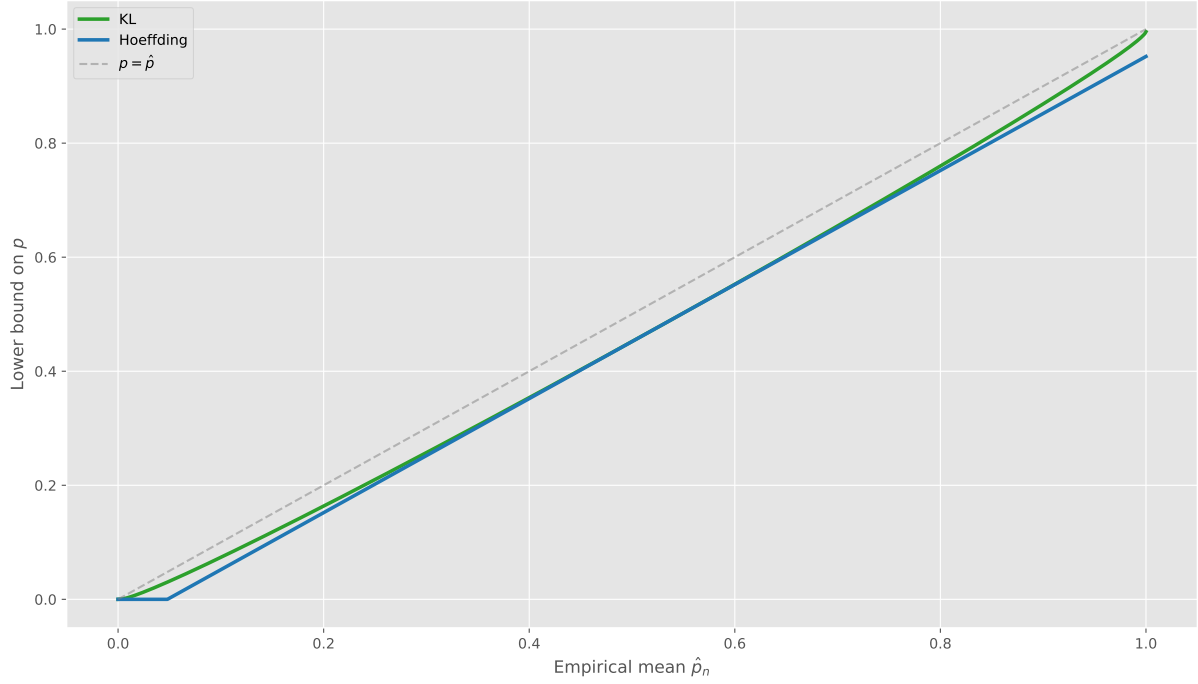


Figure 3: Lower bounds for  $\hat{p}_n \in [0, 1]$

So the lower inverse can be obtained from the upper inverse through the code:

```
def kl_lower_bound_symm(p_hat, n, delta, **kw):
    return 1.0 - kl_upper_bound(1.0 - p_hat, n, delta, **kw)
```

Listing 1: Lower inverse

I also implemented the lower inverse using the same numerical approach as for the upper inverse and used property testing to verify that the two approaches produce same results.

## Conclusion

kl is the tightest bound in the whole interval  $[0, 1]$ . As long as we are close to 0, refined Pinsker is only slightly worse than kl. Once we pass approximately  $\hat{p}_n = 0.2$  Hoeffding is actually better than Refined Pinsker.

## 2 Occam's razor with kl inequality (30 points) [Yevgeny]

I have not be able to provide a direct proof of Occam's razor with kl inequality. As an alternative I used a "backward" approach going from the desired result and move backwards. For this approach I end up with the following in-equality

$$\mathbb{P}(\text{kl}(\hat{L}(h, S) \| L(h)) \geq \varepsilon) \leq e^{-n\varepsilon} \quad (1)$$

that should hold for any  $\varepsilon > 0$ .

It looks like a KL-version of Chernoff's bound but I have not be able to proof that it is correct. Assuming eq. 1 is correct the proof goes like this.

**Theorem 1** (Occam's kl-razor inequality). *Let  $S$  be an i.i.d. sample of  $n$  points, let  $\ell$  be a loss function bounded in the interval  $[0, 1]$ , let  $\mathcal{H}$  be countable and let  $\pi(h)$  be such that it is independent of the sample  $S$  and satisfies  $\pi(h) \geq 0$  for all  $h$  and  $\sum_{h \in \mathcal{H}} \pi(h) \leq 1$ . Let  $\delta \in (0, 1)$ . Then*

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

*Proof.* Define for each hypothesis  $h$ :

$$\varepsilon_h = \frac{\ln \frac{1}{\pi(h)\delta}}{n}$$

Using eq. 1 this gives us:

$$\mathbb{P} \left( \text{kl}(\hat{L}(h, S) \| L(h)) \geq \frac{\ln \frac{1}{\pi(h)\delta}}{n} \right) \leq e^{-n \cdot \frac{\ln \frac{1}{\pi(h)\delta}}{n}} = e^{-\ln \frac{1}{\pi(h)\delta}} = \pi(h)\delta$$

Now we apply the union bound over all  $h \in \mathcal{H}$ :

$$\begin{aligned} \mathbb{P} \left( \exists h \in \mathcal{H} : \text{kl}(\hat{L}(h, S) \| L(h)) \geq \frac{\ln \frac{1}{\pi(h)\delta}}{n} \right) &\leq \sum_{h \in \mathcal{H}} \mathbb{P} \left( \text{kl}(\hat{L}(h, S) \| L(h)) \geq \frac{\ln \frac{1}{\pi(h)\delta}}{n} \right) \\ &\leq \sum_{h \in \mathcal{H}} \pi(h)\delta = \delta \sum_{h \in \mathcal{H}} \pi(h) \leq \delta \cdot 1 = \delta \end{aligned}$$

where the second last inequality follows from the condition that  $\sum_{h \in \mathcal{H}} \pi(h) \leq 1$ .  $\square$

### Importance of $\pi(h)$ being independent of $S$

The critical step where we use the independence of  $\pi(h)$  from the sample  $S$  is when applying the union bound. If  $\pi(h)$  were to depend on  $S$ , we could not treat it as a fixed quantity when calculating the probability. Without independence,  $\pi(h)$  becomes a random variable that depends on the same sample  $S$  we are using to compute  $\hat{L}(h, S)$ .

**Corollary 2.** *Under the assumptions of Theorem 3.38 (Occam's kl-razor inequality), the following holds:*

$$\mathbb{P} \left( \exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{2\hat{L}(h, S) \ln \frac{1}{\pi(h)\delta}}{n}} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \right) \leq \delta. \quad (2)$$

*Proof.* From Theorem 3.38, with probability at least  $1 - \delta$ , for all  $h \in \mathcal{H}$ :

$$\text{kl}(\hat{L}(h, S) \| L(h)) < \frac{\ln \frac{1}{\pi(h)\delta}}{n}$$

We need to convert this KL-divergence bound into an explicit bound on  $L(h)$ . For  $p, q \in [0, 1]$  with  $p \leq q$ , we can use the following lower bound on KL-divergence:

$$\text{kl}(p \| q) \geq \frac{(q - p)^2}{2q}.$$

This is from corollary 2.31 (Refined Pinsker's inequality) in the lecture notes.

Here we are interested in the case where  $\hat{L}(h, S) \leq L(h)$ , we can apply this with  $p = \hat{L}(h, S)$  and  $q = L(h)$ .

To simplify notation, we denote  $\hat{L} = \hat{L}(h, S)$  and  $L = L(h)$ .

$$\text{kl}(\hat{L} \| L) \geq \frac{(L - \hat{L})^2}{2L}$$

Now theorem 3.38 give us, with probability at least  $1 - \delta$ :

$$\frac{(L - \hat{L})^2}{2L} < \frac{\ln \frac{1}{\pi(h)\delta}}{n}$$

Solving for  $L$ :

$$(L - \hat{L})^2 < \frac{2L \ln \frac{1}{\pi(h)\delta}}{n}$$

This is a quadratic inequality in  $L$ , that we can re-write to:

$$L^2 - L \left( 2\hat{L} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \right) + \hat{L}^2 < 0$$

Using the quadratic formula, the solutions to  $aL^2 + bL + c = 0$  with

$$a = 1, b = - \left( 2\hat{L} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \right), \text{ and } c = \hat{L}^2$$

are:

$$L = \hat{L} + \frac{\ln \frac{1}{\pi(h)\delta}}{n} \pm \sqrt{\frac{2\hat{L} \ln \frac{1}{\pi(h)\delta}}{n} + \frac{(\ln \frac{1}{\pi(h)\delta})^2}{n^2}}$$

For a quadratic inequality of the form  $aL^2 + bL + c < 0$  with  $a > 0$ , the solution is between the two roots. We are looking for an upper bound for  $L$ , which will be the larger root:

$$L < \hat{L} + \frac{\ln \frac{1}{\pi(h)\delta}}{n} + \sqrt{\frac{2\hat{L} \ln \frac{1}{\pi(h)\delta}}{n} + \frac{(\ln \frac{1}{\pi(h)\delta})^2}{n^2}}$$

Using that for non-negative  $a$  and  $b$  we have  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ :

$$L < \hat{L} + \frac{\ln \frac{1}{\pi(h)\delta}}{n} + \sqrt{\frac{2\hat{L} \ln \frac{1}{\pi(h)\delta}}{n}} + \frac{\ln \frac{1}{\pi(h)\delta}}{n} = \hat{L} + \sqrt{\frac{2\hat{L} \ln \frac{1}{\pi(h)\delta}}{n}} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n}$$

Hence with probability at least  $1 - \delta$ , for all  $h \in \mathcal{H}$ :

$$L(h) < \hat{L}(h, S) + \sqrt{\frac{2\hat{L}(h, S) \ln \frac{1}{\pi(h)\delta}}{n}} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n}$$

Using the complement event:

$$\mathbb{P} \left( \exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{2\hat{L}(h, S) \ln \frac{1}{\pi(h)\delta}}{n}} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \right) \leq \delta$$

which is exactly the statement of Corollary 3.39.  $\square$

## Discussion of the Corollary

This corollary provides important advantages over the original KL-divergence formulation:

1. It provides an explicit upper bound on the true loss  $L(h)$  in terms of the empirical loss  $\hat{L}(h, S)$
2. It clearly shows the convergence rate through the terms  $\sqrt{\frac{2\hat{L}(h, S) \ln \frac{1}{\pi(h)\delta}}{n}}$  and  $\frac{2 \ln \frac{1}{\pi(h)\delta}}{n}$
3. The first term scales with  $\sqrt{\frac{\hat{L}(h, S)}{n}}$ , showing faster convergence for hypotheses with lower empirical error

### 3 Numerical comparison of the kl and split-kl inequalities (30 points) [Yevgeny]

We consider a ternary random variable  $X$  taking values  $X \in \{0, \frac{1}{2}, 1\}$ . Let

$$p_0 = \mathbb{P}(X = 0), \quad p_{\frac{1}{2}} = \mathbb{P}(X = \frac{1}{2}), \quad p_1 = \mathbb{P}(X = 1).$$

and set  $p_0 = p_1 = (1 - p_{\frac{1}{2}})/2$ , so that the probabilities of  $X = 0$  and  $X = 1$  are equal, and there is only one parameter  $p_{\frac{1}{2}}$ , which controls the probability mass of the central value.

We now want to compare the two bounds kl and split-kl as a function of  $p_{\frac{1}{2}} \in [0, 1]$ . The upper kl bound for  $p - \hat{p}_n$  is given by

$$\text{kl}^{-1+}\left(\hat{p}_n, \frac{\ln \frac{n+1}{\delta}}{n}\right) - \hat{p}_n$$

and the split-kl bound is

$$b_0 + \sum_{j=1}^K \left( \alpha_j k_1^{-1,+}\left(\hat{p}_{|j}, \frac{1}{n} \ln \frac{K}{\delta}\right) - \hat{p}_{|j} \right), \quad (3)$$

where  $\hat{p}_{|j} = \frac{1}{n} \sum_{i=1}^n X_{i|j}$  and  $X_{i|j} = \mathbb{1}(X_i \geq b_j)$  denotes the elements of the binary decomposition of  $X_i$ .

For this experiment, we find that the domain is  $b_0 = 0, b_1 = 0.5, b_2 = 1$  and the  $K = 2$  segments are  $\alpha_1 = \alpha_2 = 1/2$ . So eq. 3 simplifies to:

$$\frac{1}{2} \left( \text{kl}^{-1+}(\hat{p}_{|1}, \epsilon) - \hat{p}_{|1} \right) + \frac{1}{2} \left( \text{kl}^{-1+}(\hat{p}_{|2}, \epsilon) - \hat{p}_{|2} \right),$$

where  $\epsilon = \frac{1}{n} \ln \frac{K}{\delta} = \frac{1}{n} \ln \frac{2}{\delta}$ . In figure 3 we see a plot of kl and split-kl.

#### Brief discussion

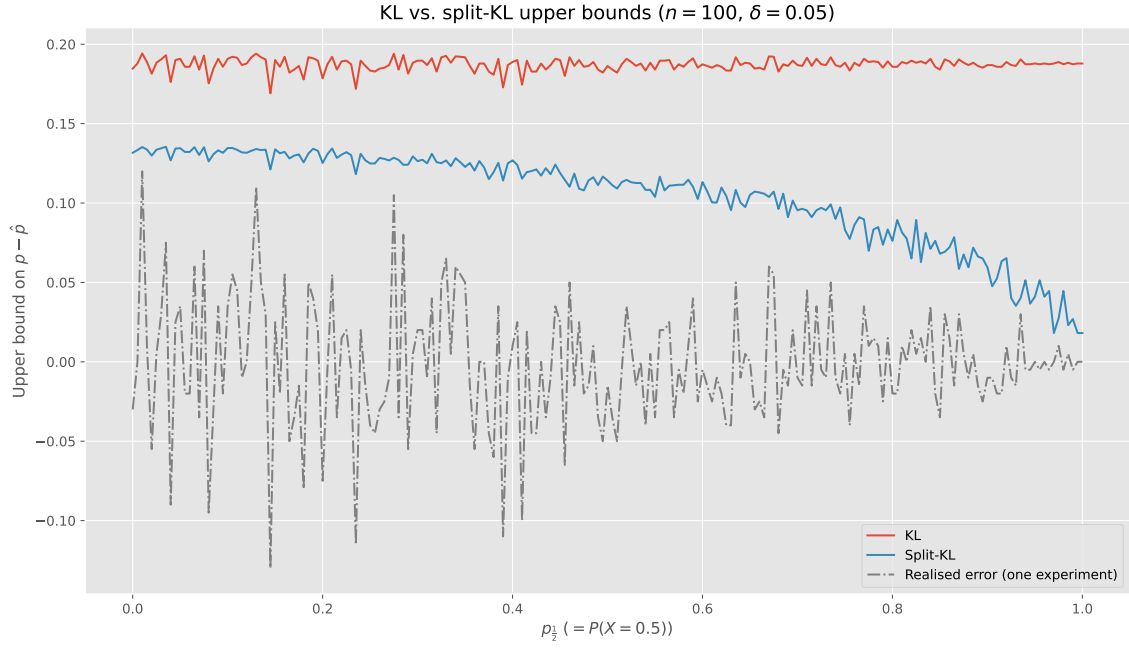
Because the true mean is fixed at 0.5 for every choice of  $p_{1/2}$ , the empirical mean fluctuates around 0.5, with variance depending on the distribution's shape.

The dashed curve is the realized error in the single sample that was drawn for each  $p_{1/2}$ . Let us consider what happens when  $p_{\frac{1}{2}} \rightarrow 1$ . In that case most of the probability mass is pushed to the center value 0.5, and the two outcomes (0 and 1) become very rare since their probabilities  $p_0 = p_1 = (1 - p_{\frac{1}{2}})/2 \rightarrow 0$ . Consequently the variance of  $X$  given by  $\text{Var}(X) = 0.25(1 - p_{\frac{1}{2}}) \rightarrow 0$ . Hence the empirical mean is in practice locked very tightly around the true mean 0.5 once  $p_{\frac{1}{2}}$  get closer to 1.

Split-kl decomposes  $X$  into the two binary indicators:

$$X_{|1} = \mathbb{1}(X \geq 0.5), \quad X_{|2} = \mathbb{1}(X \geq 1).$$





Their expectations are

$$p_{|1} = \mathbb{P}(X \geq 0.5) = 1 - p_0 \rightarrow 1,$$

$$p_{|2} = \mathbb{P}(X \geq 1.0) = p_1 \rightarrow 0$$

as  $p_{\frac{1}{2}} \rightarrow 1$ .

When a Bernoulli parameter is very close to 1 or 0, the kl-inverse difference  $\text{kl}^{-1+}(\hat{p}, \epsilon) - \hat{p}$  is proportional to  $\hat{p}(1 - \hat{p})$  and therefore goes to zero.

For the split-kl bound these two tiny differences are multiplied by  $\alpha_1 = \alpha_2 = 0.5$  and then added, so the whole bound collapses toward 0 as soon as  $X$  almost never takes the extreme values. That is why in figure 3 observe split-kl close to 0 for  $p_{\frac{1}{2}}$  close to 1.

## Python code

```
def kl_upper_bound(p_hat, n, delta, tol=1e-12, max_iter=100):
    """
    Calculate the upper confidence bound using the Kullback-Leibler
    divergence.

    This function computes an upper bound p such that
     $KL(p\_hat, p) \leq \log(1/\delta)/n$ , where p_hat is the empirical
    probability estimate, n is the sample size, and delta is the confidence level.

    Parameters
    -----
    p_hat : float
        The empirical probability estimate in [0, 1].
    n : int
        Number of samples.
    delta : float
        Confidence level parameter in (0, 1).
        The bound holds with probability 1-delta.
    tol : float, optional
        Tolerance for the binary search convergence. Default is 1e-12.
    max_iter : int, optional
        Maximum number of iterations for the binary search. Default is
        100.

    Returns
    -----
    float
        The upper confidence bound p such that  $KL(p\_hat, p) \leq \log(1/\delta)/n$ .

    Notes
    -----
    Uses binary search to find the upper bound. Special cases are handled
    for
    p_hat = 0 and p_hat = 1.
    """
    eps = np.log(1.0 / delta) / n
    if p_hat <= 0.0:
        return 1.0 - np.exp(-eps)
    if p_hat >= 1.0:
        return 1.0
    lo, hi = p_hat, 1.0
    for _ in range(max_iter):
        mid = 0.5 * (lo + hi)
        if kl(p_hat, mid) > eps:
            hi = mid
        else:
            lo = mid
```

```

    if hi - lo < tol:
        break
    return hi

```

Listing 2: Upper inverse

## Detailed derivation

Solving for  $L$ :

$$(L - \hat{L})^2 < \frac{2L \ln \frac{1}{\pi(h)\delta}}{n}$$

This is a quadratic inequality in  $L$ . Rewriting:

$$L^2 - 2\hat{L}L + \hat{L}^2 - \frac{2L \ln \frac{1}{\pi(h)\delta}}{n} < 0$$

$$L^2 - L \left( 2\hat{L} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \right) + \hat{L}^2 < 0$$

Using the quadratic formula, the solutions to  $aL^2 + bL + c = 0$  with  $a = 1$ ,  $b = - \left( 2\hat{L} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \right)$ , and  $c = \hat{L}^2$  are:

$$L = \frac{\left( 2\hat{L} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \right) \pm \sqrt{\left( 2\hat{L} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \right)^2 - 4\hat{L}^2}}{2}$$

Simplifying the discriminant:

$$\begin{aligned} \left( 2\hat{L} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \right)^2 - 4\hat{L}^2 &= 4\hat{L}^2 + \frac{8\hat{L} \ln \frac{1}{\pi(h)\delta}}{n} + \frac{4(\ln \frac{1}{\pi(h)\delta})^2}{n^2} - 4\hat{L}^2 \\ &= \frac{8\hat{L} \ln \frac{1}{\pi(h)\delta}}{n} + \frac{4(\ln \frac{1}{\pi(h)\delta})^2}{n^2} \end{aligned}$$

So the roots are:

$$L = \frac{2\hat{L} + \frac{2 \ln \frac{1}{\pi(h)\delta}}{n} \pm \sqrt{\frac{8\hat{L} \ln \frac{1}{\pi(h)\delta}}{n} + \frac{4(\ln \frac{1}{\pi(h)\delta})^2}{n^2}}}{2}$$

Simplifying:

$$L = \hat{L} + \frac{\ln \frac{1}{\pi(h)\delta}}{n} \pm \sqrt{\frac{2\hat{L} \ln \frac{1}{\pi(h)\delta}}{n} + \frac{(\ln \frac{1}{\pi(h)\delta})^2}{n^2}}$$