Machine Learning B (2025)
Home Assignment 1

Yasin Baysal, cmv882

Contents

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

Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables with bias $p = \mathbb{P}(X = 1)$. Given a sample of size n, the empirical average is given by $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

1) When we make numerical comparisons of the relative power of various bounds on the bias p, we specifically consider the following four bounds on p from (Seldin, 2025):

A. Hoeffding's bound: By Hoeffding's inequality and Equation (2.8), it holds that

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

with probability at least $1 - \delta$. Hoeffding's bound then refers to the right hand side of the inequality above.

B. The kl inequality bound: By Theorem 2.27 and inequality (2.13), it holds that

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

with probability at least $1 - \delta$. Since the binary kl–divergence is strictly increasing in p on $[\hat{p}_n, 1]$, the inequality can be inverted to an explicit upper bound by

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

where we define the "upper inverse" of the binary kl-divergence for $\varepsilon = \frac{\ln\frac{1}{\delta}}{n}$ as

$$\mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon) := \max\{p \colon p \in [\hat{p}_n, 1] \text{ and } \mathrm{kl}(\hat{p}_n||p) \leq \varepsilon\}.$$

Later, when we actually implement this bound in code, we must compute $\mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon)$, even though there is no closed-form expression for it. We observe that for fixed $\hat{p}_n$,

$$p \mapsto \mathrm{kl}(\hat{p}_n||p)$$

is convex on $[0, 1]$, attains its minimum $\mathrm{kl}(\hat{p}_n||\hat{p}_n) = 0$ at $p = \hat{p}_n$, and then increases monotonically as p grows from $\hat{p}_n$ up to 1. Hence, the solution or point p of

$$\mathrm{kl}(\hat{p}_n||p) = \varepsilon$$

is unique and lies in the interval $[\hat{p}_n, 1]$. To compute p numerically, we implement and use a binary search algorithm on $[\hat{p}_n, 1]$, which will be further explained in question 2). In question 3), we also introduce the lower bound for the kl inequality.

C. Pinsker's relaxation of the kl **inequality**: By Corollary 2.29 with Pinsker's inequality for the binary kl-divergence, we have that

$$\mathrm{kl}(\hat{p}_n||p) \geq 2(p-\hat{p}_n)^2.$$

Then, by applying the above Corollary 2.29 to inequality (2.13), it implies that

$$\begin{aligned}2(p-\hat{p}_n)^2 \leq \mathrm{kl}(\hat{p}_n\|p) &\Leftrightarrow (p-\hat{p}_n)^2 \leq \frac{\mathrm{kl}(\hat{p}_n\|p)}{2} \\ &\Leftrightarrow p \leq \hat{p}_n + \sqrt{\frac{\mathrm{kl}(\hat{p}_n\|p)}{2}} \leq \hat{p}_n + \sqrt{\frac{1}{2}\frac{\ln\frac{1}{\delta}}{n}} = \hat{p}_n + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\end{aligned}$$

with probability at least $1-\delta$. Thus, Pinsker's relaxation shows that the kl-inequality is at least as tight as Hoeffding's bound, since we exactly get $p \leq \hat{p}_n + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}$ as in A.

D. Refined Pinsker's relaxation of the kl **inequality**: By Corollary 2.31 with the refined Pinsker's inequality, we for $p > \hat{p}_n$ have that

$$\mathrm{kl}(\hat{p}_n||p) \geq \frac{(p-\hat{p}_n)^2}{2p}.$$

Since we are interested in the upper bound on p, Corollary 2.32 applied to inequality (2.13) then gives us that

$$\begin{aligned}\mathrm{kl}(\hat{p}_n\|p) \leq \frac{\ln\frac{1}{\delta}}{n} &\Rightarrow \frac{(p-\hat{p}_n)^2}{2p} \leq \frac{\ln\frac{1}{\delta}}{n} \\ &\Leftrightarrow (p-\hat{p}_n)^2 \leq 2p\frac{\ln\frac{1}{\delta}}{n} \\ &\Leftrightarrow p^2 - 2\hat{p}_n p + \hat{p}_n^2 \leq 2p\frac{\ln\frac{1}{\delta}}{n} \\ &\Leftrightarrow p^2 - 2\left(\hat{p}_n + \frac{\ln\frac{1}{\delta}}{n}\right)p + \hat{p}_n^2 \leq 0 \\ &\Leftrightarrow p \leq \hat{p}_n + \frac{\ln\frac{1}{\delta}}{n} + \sqrt{\left(\hat{p}_n + \frac{\ln\frac{1}{\delta}}{n}\right)^2 - \hat{p}_n^2} \\ &\Leftrightarrow p \leq \hat{p}_n + \sqrt{2\hat{p}_n\frac{\ln\frac{1}{\delta}}{n}} + 2\frac{\ln\frac{1}{\delta}}{n}\end{aligned}$$

with probability at least $1-\delta$. This holds if $\mathrm{kl}(\hat{p}_n||p) \leq \frac{\ln\frac{1}{\delta}}{n}$.

2) Let $n = 1000$, $\delta = 0.01$ and $\varepsilon = \frac{\ln\frac{1}{\delta}}{n}$. For each empirical average $\hat{p}_n \in [0,1]$, we define the four bounds from Question 1) on p as a function of $\hat{p}_n$, which can be seen in Table 1:

	Name	**Bound**
A	Hoeffding's bound	$B_H(\hat{p}_n) = \hat{p}_n + \sqrt{\frac{\varepsilon}{2}}$
B	The kl–inequality bound	$B_{\mathrm{kl}}(\hat{p}_n) = \mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon)$
C	Pinsker's relaxation bound	$B_P(\hat{p}_n) = \hat{p}_n + \sqrt{\frac{\varepsilon}{2}}$
D	Refined Pinsker's relaxation bound	$B_{RP}(\hat{p}_n) = \hat{p}_n + \sqrt{2\hat{p}_n\varepsilon} + 2\varepsilon$

Table 1: Summary of the four bounds on p as a function of $\hat{p}_n \in [0, 1]$.

As quickly explained in question 1), there is no closed-form expression for computing $\mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon)$, so it has to computed numerically. We thus implement the following simple binary-search algorithm to find $p = \mathrm{kl}(\hat{p}_n, \varepsilon) = \max\{p \colon p \in [\hat{p}_n, 1] \text{ and } \mathrm{kl}(\hat{p}_n \| p) \leq \varepsilon\}$:

Algorithm 1 Compute the upper inverse $\mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon)$ by binary search

```
Require: Empirical average $\hat{p}_n \in [0,1]$, tolerance $\varepsilon > 0$, search precision $\tau > 0$
Ensure: Approximate solution $p \approx \max\{p \colon p \in [\hat{p}_n, 1] \text{ and } \mathrm{kl}(\hat{p}_n \| p) \leq \varepsilon\}$
a ← $\hat{p}_n$                      ▷ initialize lower (feasible) bound
b ← 1                                ▷ initialize upper (infeasible) bound
while $b - a > \tau$ do              ▷ iterate until desired precision
    m ← $\frac{a+b}{2}$
    D ← $\mathrm{kl}(\hat{p}_n \| m)$    ▷ compute kl-divergence
    if $D > \varepsilon$ then        ▷ midpoint violates bound → too large
        b ← m
    else                             ▷ midpoint is valid → still feasible
        a ← m
    end if
end while
return a                             ▷ final approximation within tolerance
```

In our Python implementation (see `main.py` for generations of figures and tables with implementations from `question_1.py`), we first define the `kl(p, q)` function that computes the binary kl-divergence. We then implement `kl_inverse_upper(p_hat, eps, tau)` to carry out the binary search of Algorithm 1 until the interval length is below $\tau = 10^{-9}$. Next, we apply `kl_inverse_upper` over a grid from $[0, 1]$ of empirical averages $\hat{p}_n$ with $n = 1000$ and compute the four bounds from Table 1 on p for comparison. Finally, we visualize all four bound curves clipped at 1, and the results can be seen Figure 1 below:

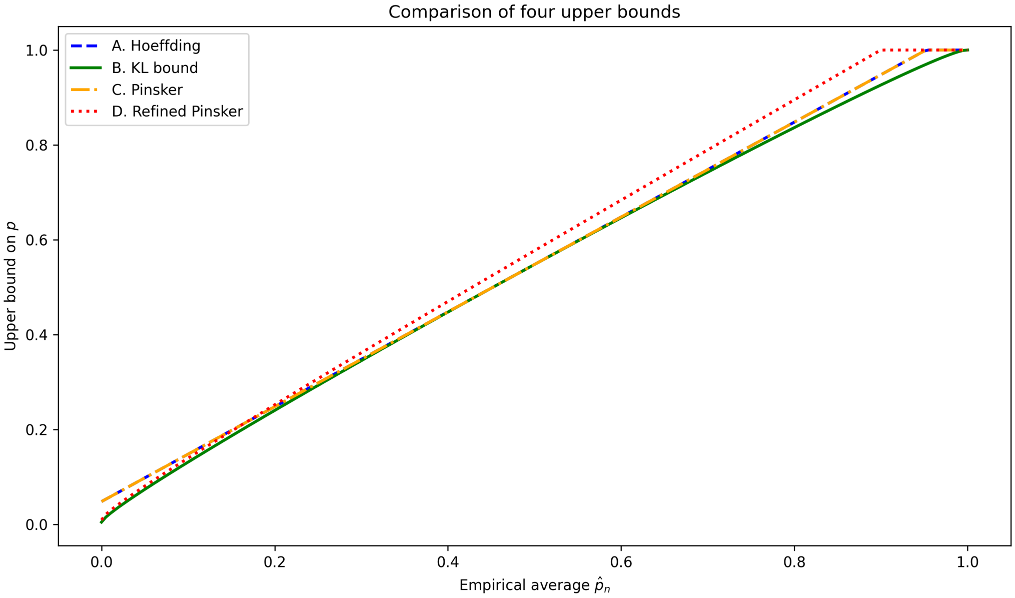

Figure 1: Comparison of four bounds on p as a function of $\hat{p}_n \in [0, 1]$ ($n = 1000$, $\delta = 0.01$).

3) Next, we generate a "zoom-in" plot of all four bounds over the interval $\hat{p}_n \in [0, 0.1]$ for $n = 1000$ and $\delta = 0.01$. We use the same code as in Question 2). However, now we no longer sample $\hat{p}_n$ across the full $[0, 1]$ range but instead restrict it to $[0, 0.1]$. The resulting plots of the four different bounds in this zoomed area can be seen in Figure 2 below:

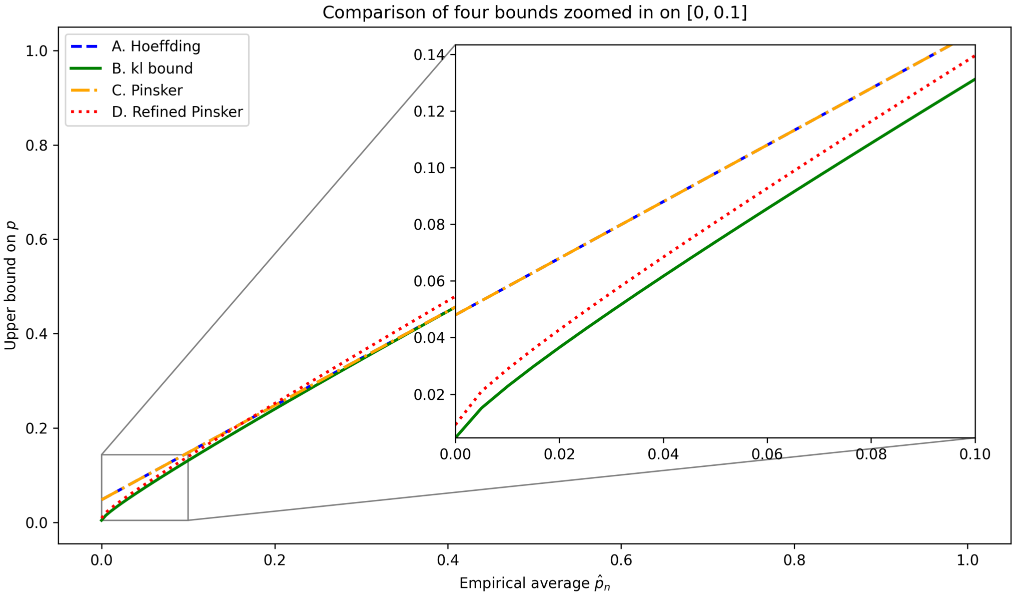

Figure 2: Comparison of the same four bounds on p from Figure 1 as a function of $\hat{p}_n \in [0, 1]$, now zoomed in for $\hat{p}_n \in [0, 0.1]$ (still with $n = 1000$ and $\delta = 0.01$).

Table 2 lists the numerical values of each of the four bounds at a few points $\hat{p}_n \in \{0.00, 0.02, \dots, 0.10\}$ in the zoom region. As will be discussed in Question 5), for small values of $\hat{p}_n$ the kl-inequality bound is significantly tighter than Hoeffding's bound (and hence also tighter than the Pinsker relaxation, which by p. 22 in Seldin (2025) recovers the same $1/\sqrt{n}$ rate as Hoeffding). This improvement reflects the so-called "fast convergence rate" of the kl bound, which for $\hat{p}_n$ near zero approaches the true p at $\mathcal{O}(1/n)$ (see p. 22 in Seldin (2025)) rather than the $\mathcal{O}(1/\sqrt{n})$ rate guaranteed by Hoeffding's inequality.

$\hat{p}_n$	A. Hoeffding	B. kl bound	C. Pinsker	D. Refined Pinsker
0.00	0.047985	0.004595	0.047985	0.009210
0.02	0.067985	0.036492	0.067985	0.042783
0.04	0.087985	0.061680	0.087985	0.068404
0.06	0.107985	0.085510	0.107985	0.092718
0.08	0.127985	0.108609	0.127985	0.116355
0.10	0.147985	0.131225	0.147985	0.139559

Table 2: Numeric values of the bounds at $\hat{p}_n \in \{0.00, 0.02, \dots, 0.10\}$ $(n = 1000, \delta = 0.01)$.

4) From Question 1, we recall that Hoeffding's upper bound on p is given by

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

with probability at least $1 - \delta$. Thus, applying the same inequality to the "flipped" deviations, we also get a lower bound on p by

$$p \geq \hat{p}_n - \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

with probability at least $1 - \delta$. Next, we (again from Question 1) recall that we have

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

with probability at least $1 - \delta$. Since the function $p \mapsto \mathrm{kl}(\hat{p}_n || p)$ is strictly decreasing on $[0, \hat{p}_n]$, we may invert it on that interval to obtain the "lower inverse" for $\varepsilon = \frac{\ln \frac{1}{\delta}}{n}$ by

$$\mathrm{kl}^{-1^-}(\hat{p}_n, \varepsilon) := \min\{p \colon p \in [0, \hat{p}_n] \text{ and } \mathrm{kl}(\hat{p}_n || p) \leq \varepsilon\}.$$

Hence, the kl-inequality yields the lower bound

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

with probability at least $1-\delta$. The two lower bounds are summarised in Table 3 below:

	Name	Bound
E	Hoeffding's lower bound	$B_{\text{l-}H}(\hat{p}_n) = \hat{p}_n - \sqrt{\frac{\varepsilon}{2}}$
F	The kl inequality lower bound	$B_{\text{l-kl}}(\hat{p}_n) = \text{kl}^{-1^-}(\hat{p}_n, \varepsilon)$

Table 3: Summary of the two lower bounds on p as a function of $\hat{p}_n \in [0,1]$.

As for the upper bound, there is no closed-form expression for $\text{kl}^{-1^-}(\hat{p}_n, \varepsilon)$. However, we observe that for fixed $\hat{p}_n$, $p \mapsto \text{kl}(\hat{p}_n||p)$ is convex on $[0,1]$, attains its minimum $\text{kl}(\hat{p}_n||\hat{p}_n) = 0$ at $p = \hat{p}_n$, and then decreases monotonically as p grows from 0 up to $\hat{p}_n$. Hence,

$$\text{kl}(\hat{p}_n||p) = \varepsilon$$

is unique and lies in the interval $[0, \hat{p}_n]$. Again, to compute p numerically, we use a binary search algorithm on $[0, \hat{p}_n]$, shrinking the interval until $\text{kl}(\hat{p}_n||p)$ is within our tolerance of ε. Specifically, we use the fact that $\text{kl}(\hat{p}_n||p) = \text{kl}(1-\hat{p}_n||1-p)$, which follows immediate from its definition in Definition 2.14. Hence, solving $\text{kl}(\hat{p}_n||p) = \varepsilon$ for $p \in [0, \hat{p}_n]$ is equivalent to solving $\text{kl}(1-\hat{p}_n||q) = \varepsilon$ for $q \in [1-\hat{p}_n, 1]$, and then setting $p = 1-q$. In other words,

$$\text{kl}^{-1^-}(\hat{p}_n, \varepsilon) = 1 - \text{kl}^{-1^+}(1-\hat{p}_n, \varepsilon).$$

Thus, rather than writing a second binary-search from scratch, we simply "flip" the inputs to our existing binary search algorithm from Algorithm 1 for the upper inverse. It is implemented in the `kl_inverse_lower(p_hat,tau)` function. Using the same values of $\hat{p}_n$, n, δ and τ as in the previous questions, we then plot the two lower bounds in Figure 3:

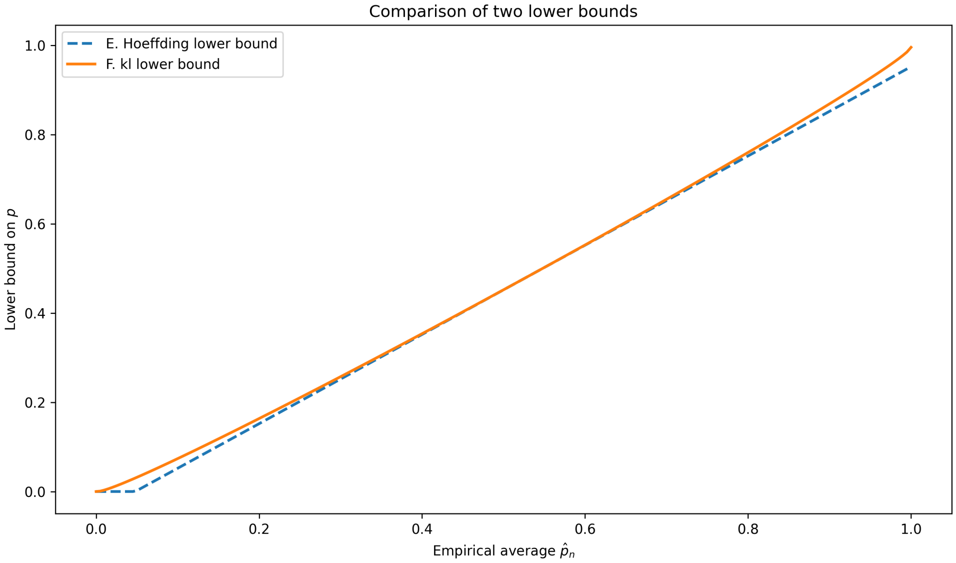

Figure 3: Lower bound comparison on p as function of $\hat{p}_n \in [0,1]$ ($n = 1000$, $\delta = 0.01$).

5) From the the experiment, we observe that the kl-inequality bound remains the most tight bound, never being overtaken by its relaxations — in particular, it sits strictly below the others at the far ends of $\hat{p}_n$ (i.e., at $\hat{p}_n = 0$ and $\hat{p}_n = 1$) in Figure 1. In the central region of approximately $\hat{p}_n \in [0.2, 0.8]$, however, Figure 1 also shows that Hoeffding's bound and Pinsker's relaxation bound collapse onto virtually the same curve as the exact kl–inequality bound, which explains why there is almost no visible gap among them here. When we "zoom in" on very small averages in Figure 2 (and inspect the precise values in Table 2), the superiority of the kl–inequality becomes clear: it decreases at the rate of $1/n$, whereas as Hoeffding's bound decreases at the slower rate of $1/\sqrt{n}$.

The refined Pinsker's relaxation bound, by contrast, only improves on Hoeffding's bound when $\hat{p}_n$ is small and close to 0, as seen in Figure 2, where its curve lies between the simple Pinsker and the kl bounds. By page 22 in Seldin (2025), this means that the refined Pinsker's relaxation bound exhibits "fast convergence rate", where it approaches p at the rate of $1/n$ rather than $1/\sqrt{n}$ for small $\hat{p}_n$ (as in Hoeffding's inequality). A similar effect is seen for the lower bounds in Figure 3. Here, the kl lower bound tightens the Hoeffding lower bound most at the extremes, while elsewhere both lower curves approximately coincide, again due to the fact that the kl bound delivers the fast $1/n$ convergence rate at the boundaries compared to the Hoeffding's bound's slow $1/\sqrt{n}$ rate uniformly.

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

1) We have to prove the following theorem for Occam's kl-razor inequality:

Theorem 3.38 (Occam's kl-razor inequality). *Let S be an i.i.d. sample of n points, let ℓ be a loss function bounded in the $[0,1]$ interval, 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,*

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

Proof. By section 1.1 in [Seldin, 2025], the empirical loss of the hypothesis $h \in \mathcal{H}$ on a sample (training set) $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of n points is given by

$$\hat{L}(h,S) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(X_i), Y_i),$$

while the expected loss of h for a loss function ℓ is given as

$$L(h) = \mathbb{E}[\ell(h(X), Y)],$$

where the expectation is taken w.r.t. a distribution $p(X, Y)$. Assume that $\ell(\cdot)$ is bounded in the $[0,1]$ interval. Let $\pi(h) \geq 0$ and $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$ for all h, which is the distribution of the confidence budget $\delta \in (0,1)$, where each h is assigned $\pi(h)$ fraction of the budget.

Fix any hypothesis $h \in \mathcal{H}$. We then invoke the kl inequality from Theorem 2.27 to relate its empirical risk $\hat{L}(h,S)$ to its true risk $L(h)$, where its assumptions are satisfied as $X_i = \ell(h(X_i), Y_i) \in [0,1]$ by assumption, and since (X_i, Y_i) pairs in S are sampled i.i.d. according to an unknown, but fixed distribution $p(X,Y)$, then the random variables $X_i = \ell(h(X_i), Y_i)$ are also i.i.d. Also, $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\sum_{i=1}^n \ell(h(X_i), Y_i) = \hat{L}(h,S)$, while $p = \mathbb{E}[\ell(h(X),Y)] = L(h)$. With assumptions satisfied, the theorem gives us that

$$\mathbb{P}\left(\mathrm{kl}(\hat{L}(h,S)||L(h)) \geq \frac{\ln\frac{1}{\eta}}{n}\right) \leq \eta, \quad \forall\delta \in (0,1),$$

where we have set $\eta(h) := \pi(h)\delta$ and assume that $0 < \pi(h) \leq 1$ and $\delta \in (0,1)$, such that it is ensured that $\eta(h) \in (0,1)$. When $\pi(h) = 0$, then also $\eta(h) = 0$, but by convention $\ln(1/\eta(h)) = \infty$ in those cases. Certainly, $\mathrm{kl}(\hat{L}(h,S)||L(h)) \geq \infty$ has probability 0, so the expression above is still valid. Thus, for this choice of η, we have that

$$\mathbb{P}\left(\mathrm{kl}(\hat{L}(h,S)||L(h)) \geq \frac{\ln\frac{1}{\pi(h)\delta}}{n}\right) \leq \pi(h)\delta, \quad \forall\delta \in (0,1).$$

Here, we used that $\pi(h)$ was chosen independently of the sample S, so that the kl-inequality may be applied with the confidence parameter $\eta = \pi(h)\delta$ not depending on S. Since the kl-inequality requires that its confidence budget η to be fixed before seeing the data. If $\pi(h)$ were allowed to depend on the sample S, then setting $\eta(h) = \pi(h)\delta$ would break this requirement and the bound would no longer hold.

Since the hypothesis set $\mathcal{H}$ is countable, we use the union bound over all $h \in \mathcal{H}$, such that

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

since we have assumed that $\sum_{h\in\mathcal{H}}\pi(h) \leq 1$. Thus, we finally obtain the desired result by

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

□

2) We have to prove the following corollary for Occam's kl-razor inequality:

Corollary 3.39 *Under the assumptions of Theorem 3.38,*

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

Proof. As we utilize the same assumptions from Theorem 3.38, we first have that

$$\exists h \in \mathcal{H}\colon \ \mathrm{kl}(\hat{L}(h,S)||L(h)) \leq \frac{\ln\frac{1}{\pi(h)\delta}}{n}$$

with probability at least $1-\delta$. Then, we note that if the binary kl-divergence $\mathrm{kl}(p||q) \leq \varepsilon$ for $p, q \in [0,1]$, it holds with probability at least $1-\delta$ that

$$q \leq p + \sqrt{2p\varepsilon} + 2\varepsilon,$$

which is the upper bound for the refined Pinsker's inequality in Corollary 2.32. In our setting, p, q and ε corresponds to

$$p = \hat{L}(h,S), \quad q = L(h), \quad \varepsilon = \frac{\ln\frac{1}{\pi(h)\delta}}{n},$$

where we from the proof of Theorem 3.38 recall that $\hat{L}(h,S), L(h) \in [0,1]$, such that Corollary 2.32 still holds here. By inserting the values into its inequality, we obtain that

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{2\hat{L}(h,S)\ln\frac{1}{\pi(h)\delta}}{n}} + \frac{2\ln\frac{1}{\pi(h)\delta}}{n}, \quad \forall h \in \mathcal{H}$$

with probability at least $1-\delta$. In other words, the probability that there exists any $h \in \mathcal{H}$ violating the bound is at most δ, that is

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

which gives us the desired result. □

3) Compared to the original Occam's razor bound from Theorem 3.3 given by

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

which always converges at rate $\mathcal{O}(1/\sqrt{n})$, Corollary 3.39 for Occam's kl-razor inequality offers key advantages: It has fast rates when $\hat{L}(h, S)$ is small. As soon as a hypothesis h achieves very low training loss, the leading term $\sqrt{(2\hat{L}(h, S)\ln(1/(\pi(h)\delta)))/n}$ decreases like $\mathcal{O}(1/n)$, i.e., with a rate of $1/n$, rather than $\mathcal{O}(1/\sqrt{n})$ with a rate of $1/\sqrt{n}$, which is the case for the original Occam's razor bound (where we also note that $\ln(1/(\pi(h)\delta))$ is a constant once $\pi(h)$ and δ are fixed values). Therefore, the leading term in Corollary 3.39 is $\mathcal{O}(1/\sqrt{n})$ when $\hat{L}(h, S)$ is bounded away from zero, but as $\hat{L}(h, S) \to 0$, it becomes $\mathcal{O}(1/n)$, such that it exhibits ”fast convergence rate” with a tighter bound than Occam's razor bound that in contrast treats all h equally, regardless of how well they fit the sample.

To visualize how the two bounds behave as n grows, we have implemented a short experiment in `main.py` with code from `question_2.py` that computes, for each n from 10 to 10^4, the term $\sqrt{(\ln(1/(\pi(h)\delta))/n)}$ from Theorem 3.3 and the term $\sqrt{(2\hat{L}(h, S)\ln(1/(\pi(h)\delta)))/n} + (2\ln(1/(\pi(h)\delta)))/n$ from Corollary 3.39, and then plots both on a log–log scale in Figure 4 below with $\delta = 0.05$, $\pi(h) = 1$ and a small fixed $\hat{L}(h, S) = 0.001$.

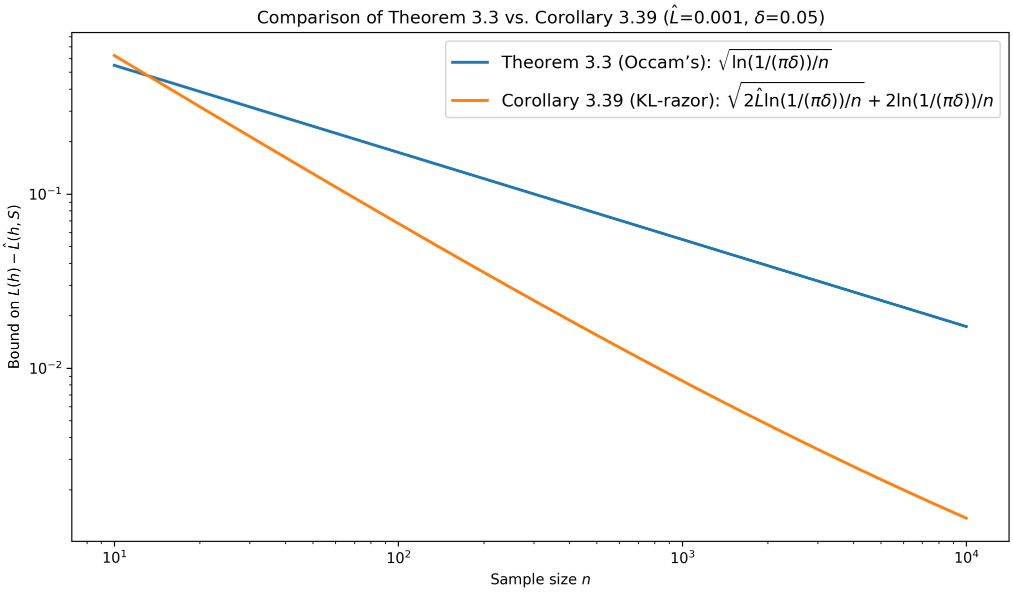

Figure 4: Generalization-gap bounds (Theorem 3.3 vs. Corollary 3.39) over sample size.

As expected, the classic Occam bound decays steadily but slowly, whereas the one from Corollary 3.39, after initially tracking it, drops much more sharply as n increases – clearly showing the fast convergence rate advantage when training loss is low.

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

We present a numerical comparison of the tightness of the kl and split-kl inequalities. To do so, we first take n i.i.d. samples $X_1, \dots, X_n$ of ternary random variables X_i for $i = 1, \dots, n$ that are taking the three values in the set $\left\{0, \frac{1}{2}, 1\right\}$. We let the probabilities be given by

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

where $p_0 + p_{\frac{1}{2}} + p_1 = 1$. Then, we set

$$p_0 = p_1 = \frac{1 - p_{\frac{1}{2}}}{2},$$

which means that the probabilities of $X = 0$ and $X = 1$ are equal, and there is just one parameter $p_{\frac{1}{2}}$ that controls the probability mass of the central value. Let $p = \mathbb{E}[X]$, such that (in this constructed setup) for any value of $p_{\frac{1}{2}}$, we have that

$$p = \mathbb{E}[X] = 0 \cdot p_0 + \frac{1}{2} \cdot p_{\frac{1}{2}} + 1 \cdot p_1 = \frac{1}{2},$$

since $p_0 = p_1$. We compare the two bounds as a function of $p_{\frac{1}{2}} \in [0, 1]$. For each value of $p_{\frac{1}{2}}$ in a grid covering the $[0, 1]$ interval, we draw a random sample $X_1, \dots, X_n$ from our constructed distribution and let $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$. We now present the two bounds:

1. **The kl bound**: By Theorem 2.27 with the kl inequality, for any $\delta \in (0, 1)$ and $\varepsilon = \frac{\ln \frac{1}{\delta}}{n}$, we have with probability at least $1 - \delta$ that

   $$\mathrm{kl}(\hat{p}_n || p) \leq \varepsilon.$$

   Since here $p = \frac{1}{2}$ is known, we (as also done in question 1 of Exercise 1) invert the monotone function $p \mapsto \mathrm{kl}(\hat{p}_n || p)$ on $[\hat{p}_n, 1]$. Define the upper inverse of kl by

   $$\mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon) := \max\{p \in [\hat{p}_n, 1] \colon \mathrm{kl}(\hat{p}_n || p) \leq \varepsilon\}.$$

   Then, we get the kl bound on $p - \hat{p}_n$ by

   $$p \leq \mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon) \Rightarrow p - \hat{p}_n \leq \mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon) - \hat{p}_n.$$

   Here, we note that in contrast to Exercise 1, we subtract the value of $\hat{p}_n$ after inversion of kl to get a bound on the difference $p - \hat{p}_n$ rather than on p.

2. **The split-**kl **bound**: By Theorem 2.34 with the split-kl inequality for discrete random variables, any $\left\{0, \frac{1}{2}, 1\right\}$-valued random variable X admits the decomposition

$$\begin{aligned} X &= b_0 + \alpha_1 X_{|1} + \alpha_2 X_{|2} \\ &= b_0 + \alpha_1 \mathbb{1}(X \geq b_1) + \alpha_2 \mathbb{1}(X \geq b_2) \\ &= b_0 + (b_1 - b_0)\mathbb{1}(X \geq b_1) + (b_2 - b_1)\mathbb{1}(X \geq b_2) \\ &= 0 + \left(\frac{1}{2} - 0\right)\mathbb{1}(X \geq \tfrac{1}{2}) + \left(1 - \frac{1}{2}\right)\mathbb{1}(X \geq 1) \\ &= \frac{1}{2}\mathbb{1}(X \geq \tfrac{1}{2}) + \frac{1}{2}\mathbb{1}(X \geq 1). \end{aligned}$$

We then define the two averages $\hat{p}_{|1}$ and $\hat{p}_{|2}$ by

$$\hat{p}_{|1} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \geq \tfrac{1}{2}), \quad \hat{p}_{|2} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \geq 1).$$

Thus, with probability at least $1 - \delta$ for $\delta \in (0, 1)$, the split-kl inequality is given by

$$p \leq 0 + \frac{1}{2}\,\mathrm{kl}^{-1^+}\left(\hat{p}_{|1}, \frac{1}{n}\ln\frac{2}{\delta}\right) + \frac{1}{2}\,\mathrm{kl}^{-1^+}\left(\hat{p}_{|2}, \frac{1}{n}\ln\frac{2}{\delta}\right).$$

Since we are interested in the bound on $p - \hat{p}_n$, subtracting $\hat{p}_n = \frac{1}{2}\hat{p}_{|1} + \frac{1}{2}\hat{p}_{|2}$ gives us

$$p - \hat{p}_n \leq \frac{1}{2}\left[\mathrm{kl}^{-1^+}\left(\hat{p}_{|1}, \frac{1}{n}\ln\frac{2}{\delta}\right) - \hat{p}_{|1}\right] + \frac{1}{2}\left[\mathrm{kl}^{-1^+}\left(\hat{p}_{|2}, \frac{1}{n}\ln\frac{2}{\delta}\right) - \hat{p}_{|2}\right]$$

as our split-kl inequality bound on $p - \hat{p}_n$ with probability at least $1 - \delta$.

The two bounds on $p - \hat{p}_n$ with $\varepsilon = \frac{\ln\frac{1}{\delta}}{n}$ and $\varepsilon' = \frac{1}{n}\ln\frac{2}{\delta}$ are summarised in Table 4 below:

	Name	**Bound**
A	kl bound	$B_{\mathrm{kl}}(\hat{p}_{\frac{1}{2}}) = \mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon) - \hat{p}_n$
B	Split-kl bound	$B_{\text{split-kl}}(\hat{p}_{\frac{1}{2}}) = \dfrac{1}{2}\left[\mathrm{kl}^{-1^+}(\hat{p}_{\|1}, \varepsilon') - \hat{p}_{\|1}\right] + \dfrac{1}{2}\left[\mathrm{kl}^{-1^+}(\hat{p}_{\|2}, \varepsilon') - \hat{p}_{\|2}\right]$

Table 4: Summary of the two bounds on $p - \hat{p}_n$ as a function of $\hat{p}_{\frac{1}{2}} \in [0, 1]$.

As also explained in question 2 in Exercise 1, there is no closed-form expression for computing the upper kl inverse $\mathrm{kl}^{-1^+}(\hat{p}_n, \varepsilon)$. We therefore use same the binary-search algorithm from Algorithm 1 to find $p = \mathrm{kl}(\hat{p}_n, \varepsilon) = \max\{p \colon p \in [\hat{p}_n, 1] \text{ and } \mathrm{kl}(\hat{p}_n || p) \leq \varepsilon\}$ numerically.

In our Python implementation (see `main.py` for generation of figure with implementations from `question_3.py`), we first set $n = 100$ and $\delta = 0.05$. Then, we compute $\varepsilon = \frac{\ln\frac{1}{\delta}}{n}$ and

$\varepsilon' = \frac{1}{n}\ln\frac{2}{\delta}$, and build a grid of $p_{\frac{1}{2}}$ values in $[0,1]$. We then define the `kl_div(p, q)` function to compute the binary kl divergence, and `kl_inv_upper(phat, eps, tau)` to numerically invert it by binary search down to tolerance $\tau = 10^{-9}$. Next, the function `simulate_bounds` loops over each $p_{\frac{1}{2}}$ in the grid, generates an i.i.d. sample $X_1, \dots, X_n \sim \left\{0, \frac{1}{2}, 1\right\}$ with the specified mixture, and computes $\hat{p}_n$, $\hat{p}_{|1}$ and $\hat{p}_{|2}$. It applies `kl_inv_upper` with ε to get the kl bound and with ε' to each average $\hat{p}_{|1}$, $\hat{p}_{|2}$, and then averages them into the split-kl bound. Lastly, we generate a figure, where we plot the kl and the split-kl bounds on $p - \hat{p}_n$ as a function of $p_{\frac{1}{2}}$ for $p_{\frac{1}{2}} \in [0,1]$ in Figure 5 below:

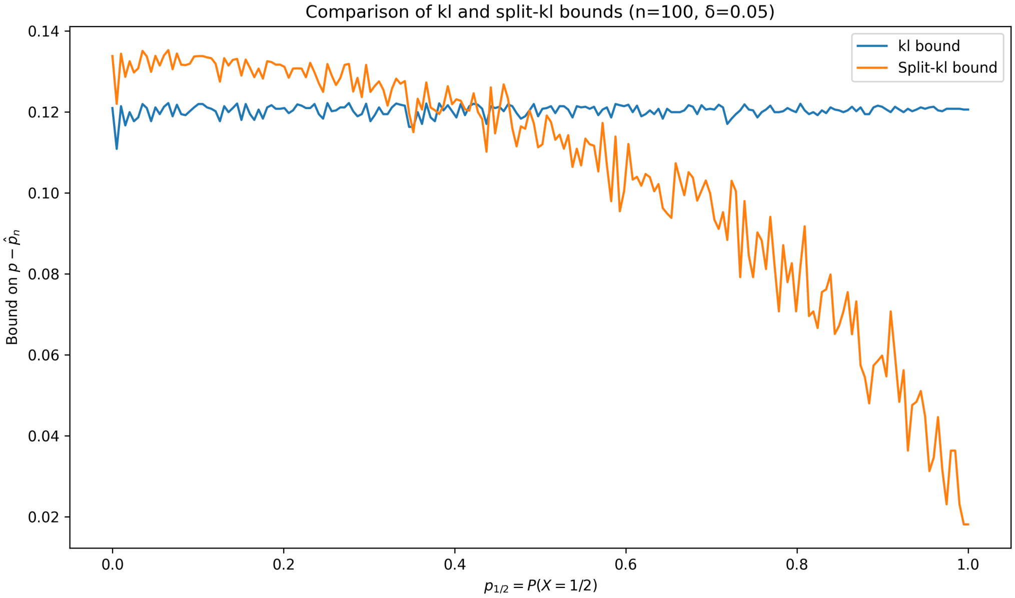

Figure 5: Comparison of bounds on $p - \hat{p}_n$ as a function of $\hat{p}_{\frac{1}{2}} \in [0,1]$ ($n = 100$, $\delta = 0.05$).

When $p_{\frac{1}{2}} = 0$, the ternary random variable X reduces to a simple Bernoulli variable and the classical kl bound is the tightest. As $p_{\frac{1}{2}}$ increases, the variance of X declines, and the standard kl bound no longer contracts accordingly, so once $p_{\frac{1}{2}}$ is above roughly 0.5, the split-kl bound becomes strictly smaller. At the extreme $p_{\frac{1}{2}} = 1$, where all mass is on $X = \frac{1}{2}$, split-kl nearly matches the exact deviation rate, whereas the ordinary kl bound remains much looser. These results show that for ternary random variables the split-kl inequality is uniformly tighter except in the pure Bernoulli limit, which makes it a powerful alternative for the standard kl inequality bound in this given setting.

References

Seldin, Y. (2025). Machine learning: The science of selection under uncertainty.