Problem 1 is important to look at, to build some familiarity and conceptual understanding of the result. Problems 2 to 4 do not really require you to have read the lecture notes in detail or watched the lecture (just look at the corresponding definitions). They are not difficult, but somewhat on the technical side, and can be skipped, though Problem 2 is good practice and the idea of Problem 4 is important (though the proof is a little tedious).

Problem 5 is good practice, but not crucial and quite long, and can be skipped during the tutorial if you are short on time. Go back to it later on for practice, on your own time.

Problem 6 is just a "sanity check" to see while an obvious approach does not work. Think of it as optional, no need to focus on it in a first pass.

Problem 6 is on the more difficult side, but worth doing to understand why the MG algorithms is a sketching algorithm.

Problem 7 is important, as it applies to a practical (simple) example the algorithms seen in the lecture.

Problem 8 (Advanced) is interesting "to go further", and will help you build some understanding of the techniques used in this area. Worth doing after going through the others.

Warm-up

Problem 1. Check your understanding: how does the Pearson–Neyman lemma (Lemma 49.1) imply that Alice-Bob game interpretation?

Solution 1. The probability that Bob loses is

$$\begin{aligned}
\Pr[\text{Bob loses}] &= \Pr[\text{Bob says }\mathsf{Heads} \mid \mathsf{Tails}]\cdot\Pr[\mathsf{Tails}] + \Pr[\text{Bob says }\mathsf{Tails} \mid \mathsf{Heads}]\cdot\Pr[\mathsf{Heads}] \\
&= \Pr_{x\sim\mathbf{q}}[\text{Bob says }\mathsf{Heads}]\cdot\frac{1}{2} + \Pr_{x\sim\mathbf{p}}[\text{Bob says }\mathsf{Tails}]\cdot\frac{1}{2} \\
&= (\text{Type I error})\cdot\frac{1}{2} + (\text{Type II error})\cdot\frac{1}{2} \qquad (\star)\\
&= \frac{\text{Type I error} + \text{Type II error}}{2} \\
&\geq \frac{1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q})}{2}
\end{aligned}$$

the last line by the Pearson–Neyman lemma, and $(\star)$ by seeing Bob as a distinguisher between $\mathbf{p}$ and $\mathbf{q}$ (i.e., a function Bob: $\mathcal{X} \to \{\mathsf{Heads}, \mathsf{Tails}\}$ where Heads corresponds to $\mathbf{p}$ and Tails to $\mathbf{q}$).

Problem 2. Prove the upper bound of Corollary 50.1 directly, via Hoeffding.

Solution 2. Letting $\widehat{p} := \frac{1}{n}\sum_{i=1}^{n} x_i$ be the empirical estimator of p, we have that $\mathbb{E}[\widehat{p}] = p$ and $x_1, \dots, x_n \in \{0,1\}$ are i.i.d. random variables with mean p. By

Hoeffding's inequality (Corollary 12.1), for $\varepsilon \in (0,1]$, we habe

$$\Pr[\,|\widehat{p} - p| > \varepsilon\,] \leq 2e^{-2\varepsilon^2 n}$$

To have the RHS be at most $\delta \in (0,1]$, it suffices to have $n \geq \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta}$, so for instance $n = \lceil \frac{1}{2\varepsilon^2}\ln\frac{2}{\delta} \rceil = O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ suffices.

Problem 3. Show that ℓ_2 and ℓ_∞ distances between distributions:

$$\ell_2(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_2 = \sqrt{\sum_{x\in\mathcal{X}}(\mathbf{p}(x)-\mathbf{q}(x))^2}, \quad \ell_\infty(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_\infty = \max_{x\in\mathcal{X}}|\mathbf{p}(x)-\mathbf{q}(x)|$$

do not satisfy the Data Processing Inequality.

Solution 3. Suppose $k \geq 4$ is even, and let $\mathbf{p}$ be the uniform distribution over $\mathcal{X} = \{1,2,\dots,k\}$, while $\mathbf{q}$ is uniform over $\{1,2,\dots,k/2\}$ (and puts probability zero on $\{k/2+1,\dots,k\}$). One can check that

$$\ell_2(\mathbf{p},\mathbf{q}) = \left(\frac{k}{2}\cdot\left(\frac{2}{k}-\frac{1}{k}\right)^2 + \frac{k}{2}\cdot\left(0-\frac{1}{k}\right)^2\right)^{1/2} = \frac{1}{\sqrt{k}}$$

Now, let $f\colon \mathcal{X} \to \mathcal{X}$ be defined as

$$f(x) = \begin{cases} 1 & \text{if } x \leq k/2 \\ 2 & \text{if } x > k/2 \end{cases}$$

and $\mathbf{p}'$ (resp. $\mathbf{q}'$) be the distribution of $f(x)$ when $x \sim \mathbf{p}$ (resp. $x \sim \mathbf{q}$). Then $\mathbf{p}'$ is uniform on $\{1,2\}$ (and 0 elsewhere), while $\mathbf{q}'$ puts probability mass one (all of it) on element 1: $\mathbf{q}'(1) = 1$. It follows that

$$\ell_2(\mathbf{p}',\mathbf{q}') = \left((\mathbf{p}'(1)-\mathbf{q}'(1))^2 + (\mathbf{p}'(2)-\mathbf{q}'(2))^2\right)^{1/2} = \left(\left(\tfrac{1}{2}-1\right)^2 + \left(\tfrac{1}{2}-0\right)^2\right)^{1/2} = \tfrac{1}{\sqrt{2}}$$

showing that $\ell_2(\mathbf{p}',\mathbf{q}') > \ell_2(\mathbf{p},\mathbf{q})$. The same counter-example works for ℓ_∞, as $\ell_\infty(\mathbf{p},\mathbf{q}) = \frac{1}{k}$, while $\ell_\infty(\mathbf{p}',\mathbf{q}') = \frac{1}{2}$.

Problem 4. Prove Scheffé's lemma. *(Hint: consider the set $S = \{x \in \mathcal{X} : \mathbf{p}(x) > \mathbf{q}(x)\}$.)*

Solution 4.

- Consider the suggested set $S^* := \{x \in \mathcal{X} : \mathbf{p}(x) > \mathbf{q}(x)\}$. For this set, we have

$$\begin{aligned}
\mathbf{p}(S^*) - \mathbf{q}(S^*) &= \sum_{x\in S^*}\mathbf{p}(x) - \sum_{x\in S^*}\mathbf{q}(x) \\
&= \sum_{x\in S^*}(\mathbf{p}(x)-\mathbf{q}(x)) \\
&= \sum_{x\in S^*}|\mathbf{p}(x)-\mathbf{q}(x)| && (\text{as } \mathbf{p}(x)-\mathbf{q}(x) > 0 \text{ for } x \in S^*) \\
&= \sum_{x\in S^*}|\mathbf{p}(x)-\mathbf{q}(x)| && (\dagger)
\end{aligned}$$

We also have that

$$\begin{aligned}
\sum_{x\notin S^*} |\mathbf{p}(x)-\mathbf{q}(x)| &= \sum_{x\notin S^*} (\mathbf{q}(x)-\mathbf{p}(x)) \qquad \text{(as } \mathbf{p}(x)-\mathbf{q}(x)\le 0 \text{ for } x\notin S^*\text{)}\\
&= \sum_{x\notin S^*} \mathbf{q}(x) - \sum_{x\notin S^*} \mathbf{p}(x))\\
&= \left(1-\sum_{x\in S^*}\mathbf{q}(x)\right) - \left(1-\sum_{x\in S^*}\mathbf{p}(x)\right)\\
&\qquad \text{(as } \textstyle\sum_{x\in\mathcal{X}}\mathbf{p}(x)=\sum_{x\in\mathcal{X}}\mathbf{q}(x)=1\text{)}\\
&= \sum_{x\in S^*}\mathbf{p}(x) - \sum_{x\in S^*}\mathbf{q}(x)\\
&= \sum_{x\in S^*}|\mathbf{p}(x)-\mathbf{q}(x)|
\end{aligned}$$

and so

$$\begin{aligned}
\sum_{x\in S^*}|\mathbf{p}(x)-\mathbf{q}(x)| &= \frac{1}{2}\left(\sum_{x\in S^*}|\mathbf{p}(x)-\mathbf{q}(x)| + \sum_{x\notin S^*}|\mathbf{p}(x)-\mathbf{q}(x)|\right)\\
&= \frac{1}{2}\left(\sum_{x\in\mathcal{X}}|\mathbf{p}(x)-\mathbf{q}(x)|\right) = \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_1 .
\end{aligned}$$

Along with (†), this shows $\sup_{S\subseteq\mathcal{X}}(\mathbf{p}(S)-\mathbf{q}(S)) \ge \mathbf{p}(S^*)-\mathbf{q}(S^*) = \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_1$. Overall, this establishes that

$$\mathbf{p}(S^*)-\mathbf{q}(S^*) = \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_1 . \tag{‡}$$

- We can use this to prove $\sup_{S\subseteq\mathcal{X}}(\mathbf{p}(S)-\mathbf{q}(S)) = \frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_1$:

  Take any set $T\subseteq\mathcal{X}$. We can write both T and S^* as the union of 2 disjoint sets, $T=(T\setminus S^*)\cup(T\cap S^*)$ and $S^*=(S^*\setminus T)\cup(T\cap S^*)$. Note that by definition of S^*, and the fact that $T\setminus S^*\subseteq\mathcal{X}\setminus S^*$,

  $$\mathbf{p}(T\setminus S^*)\le\mathbf{q}(T\setminus S^*), \qquad \mathbf{p}(S^*\setminus T)\ge\mathbf{q}(S^*\setminus T)$$

  (since the inequalities hold for each element x of these sets). This implies that

  $$\begin{aligned}
  \mathbf{p}(T)-\mathbf{q}(T) &= (\mathbf{p}(T\setminus S^*)+\mathbf{p}(T\cap S^*)) - (\mathbf{q}(T\setminus S^*)+\mathbf{q}(T\cap S^*))\\
  &= \mathbf{p}(T\cap S^*)-\mathbf{q}(T\cap S^*) + \underbrace{(\mathbf{p}(T\setminus S^*)-\mathbf{q}(T\setminus S^*))}_{\le 0}\\
  &\le \mathbf{p}(T\cap S^*)-\mathbf{q}(T\cap S^*)\\
  &\le \mathbf{p}(T\cap S^*)-\mathbf{q}(T\cap S^*) + \underbrace{(\mathbf{p}(S^*\setminus T)-\mathbf{q}(S^*\setminus T))}_{\ge 0}\\
  &= (\mathbf{p}(S^*\setminus T)+\mathbf{p}(T\cap S^*)) - (\mathbf{q}(S^*\setminus T)+\mathbf{q}(T\cap S^*))\\
  &= \mathbf{p}(S^*)-\mathbf{q}(S^*) .
  \end{aligned}$$

  Since $\mathbf{p}(T)-\mathbf{q}(T)\le\mathbf{p}(S^*)-\mathbf{q}(S^*)$ for every $T\subseteq\mathcal{X}$, the inequality holds for the supremum, showing

  $$\sup_{S\subseteq\mathcal{X}}(\mathbf{p}(S)-\mathbf{q}(S)) \le \mathbf{p}(S^*)-\mathbf{q}(S^*) .$$

and so (since clearly $\sup_{S\subseteq\mathcal{X}}(\mathbf{p}(S)-\mathbf{q}(S)) \geq \mathbf{p}(S^*)-\mathbf{q}(S^*)$, as the supremum is an upper bound over all sets)

$$\sup_{S\subseteq\mathcal{X}}(\mathbf{p}(S)-\mathbf{q}(S)) = \mathbf{p}(S^*)-\mathbf{q}(S^*),$$

which combined with (‡) proves Scheffé's lemma.

Problem solving

Problem 5. Prove the two "suboptimal" sample complexities for learning distributions. For the second, explain how to get rid of the assumption on $\min_i p_i$ (possibly losing some constant factors in the sample complexity).

Solution 5. For both, we are analysing the usual *empirical estimator*, defined by

$$\widehat{\mathbf{p}}(i) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{x_j=i}, \qquad i\in\mathcal{X},$$

where $\mathcal{X}$ is a known discrete domain of size k. Note that, for any fixed i and any $1\leq j\leq n$, $\mathbb{E}[\mathbf{1}_{x_j=i}] = \Pr[\, x_j = i\,] = \mathbf{p}(i)$.

- The first one requires to choose n such that, for every $i\in\mathcal{X}$,

$$\Pr\left[\, |\widehat{\mathbf{p}}(i)-\mathbf{p}(i)| > \frac{2\varepsilon}{k}\,\right] \leq \frac{\delta}{k} \tag{$*$}$$

since then, by a union bound, we get

$$\Pr\left[\, \forall i\in\mathcal{X},\ |\widehat{\mathbf{p}}(i)-\mathbf{p}(i)| \leq \frac{2\varepsilon}{k}\,\right] \geq 1-k\cdot\frac{\delta}{k} = 1-\delta\,,$$

and so, with probability at least $1-\delta$,

$$\mathrm{d_{TV}}(\mathbf{p},\widehat{\mathbf{p}}) = \frac{1}{2}\sum_{i\in\mathcal{X}}|\mathbf{p}(i)-\widehat{\mathbf{p}}(i)| \leq \frac{1}{2}\sum_{i\in\mathcal{X}}\frac{2\varepsilon}{k} = \varepsilon\,.$$

So finding n such that $(*)$ holds is *sufficient* to learn to TV distance ε with probability $1-\delta$. How big n must be for $(*)$? From the same analysis as learning the bias of a coin (Corollary 50.1), i.e., by a Hoeffding bound (or directly using that result, since for fixed i we *are* estimating the bias $\mathbf{p}(i)$ of a "coin" from n samples), we need

$$n = O\left(\frac{1}{(\varepsilon/k)^2}\log\frac{1}{(\delta/k)}\right) = \boxed{O\left(\frac{k^2}{\varepsilon^2}\log\frac{k}{\delta}\right)}.$$

- The second one requires to choose n such that, for every $i \in \mathcal{X}$,

$$\Pr[\, |\widehat{\mathbf{p}}(i) - \mathbf{p}(i)| > 2\varepsilon \cdot \mathbf{p}(i)\,] \leq \frac{\delta}{k} \tag{$**$}$$

since then, by a union bound, we get

$$\Pr[\, \forall i \in \mathcal{X},\ |\widehat{\mathbf{p}}(i) - \mathbf{p}(i)| \leq 2\varepsilon \cdot \mathbf{p}(i)\,] \geq 1 - k \cdot \frac{\delta}{k} = 1 - \delta\,,$$

and so, with probability at least $1 - \delta$,

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \widehat{\mathbf{p}}) = \frac{1}{2}\sum_{i \in \mathcal{X}} |\mathbf{p}(i) - \widehat{\mathbf{p}}(i)| \leq \frac{1}{2}\sum_{i \in \mathcal{X}} 2\varepsilon \cdot \mathbf{p}(i) = \varepsilon\,,$$

using $\sum_{i \in \mathcal{X}} \mathbf{p}(i) = 1$. So finding n such that $(**)$ holds is *sufficient* to learn to TV distance ε with probability $1 - \delta$. How big n must be for $(**)$? Without any further assumption, we cannot get *any* bound on this. If $\mathbf{p}(1) = 1/2^{2^{2^k}}$, then we cannot get a multiplicative estimate of it (which $(**)$ asks for) unless we take $n = \Omega(2^{2^{2^k}})$ samples: before that, with overwheliming probability we wouldn't see "1" even once in our samples, and so $\widehat{\mathbf{p}}(1) = 0$.

This is why we make the assumption that $\min_{i \in \mathcal{X}} \mathbf{p}(i) \geq \tau = \frac{\varepsilon}{k}$ (why this particular value for τ? Essentially, as we will see it's because that's a value we can guarantee via a simple "trick", and we cannot guarantee anything better).

For convenience, and also "without loss of generality" we will also assume $\varepsilon \leq 1/2$: if it is bigger, say 0.99 learn to distance $1/2$ instead, this gives a better guarantee and loses only a constant factor in the sample complexity. Then, by a Chernoff bound (Theorem 13), for any fixed i we have

$$\Pr[\, |\widehat{\mathbf{p}}(i) - \mathbf{p}(i)| > 2\varepsilon \cdot \mathbf{p}(i)\,] \leq 2e^{-4\varepsilon^2 n \mathbf{p}(i)} \leq 2e^{-4\varepsilon^2 n \tau}$$

and so, for this to be at most $\frac{\delta}{k}$, we need

$$n \geq \frac{1}{4\varepsilon^2 \tau} \ln \frac{2k}{\delta} = \frac{k}{4\varepsilon^3} \ln \frac{2k}{\delta}$$

and so $\boxed{n = O\left(\frac{k}{\varepsilon^3} \ln \frac{k}{\delta}\right)}$ suffices.

- *Removing that assumption (up to a constant factor somewhere).* The issue with this assumption is that it is not true that all probability distributions have some probability at least $\tau > 0$ on each domain element. What we can do, however, is "mix" the unknown distribution $\mathbf{p}$ with the uniform distribution $\mathbf{u}_k$: define $\mathbf{p}'$ as

$$\mathbf{p}' = (1 - \alpha) \cdot \mathbf{p} + \alpha \cdot \mathbf{u}_k$$

the distribution obtained by the following process:

1: Flip a coin with bias α.
2: If it landed Heads, draw $x \sim \mathbf{u}_k$
3: Else, draw $x \sim \mathbf{p}$
4: **return** x

We can easily get n i.i.d. samples from $\mathbf{p}'$ given n i.i.d. samples from $\mathbf{p}$ (we most likely not even use them all), as long as we also have our own randomness (to flip the coin and, sometimes, sample from $\mathbf{u}_k$). We also have

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}') \leq \alpha$$

since

$$\frac{1}{2}\sum_{x\in\mathcal{X}} |\mathbf{p}'(x) - \mathbf{p}(x)| = \frac{1}{2}\sum_{x\in\mathcal{X}} \alpha|\mathbf{u}_k(x) - \mathbf{p}(x)| = \alpha \cdot \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_k) \leq \alpha\,.$$

And by the triangle inequality, if we learn $\mathbf{p}'$ to some distance parameter ε' (and get $\widehat{\mathbf{p}}$) then

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \widehat{\mathbf{p}}) \leq \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}') + \mathrm{d}_{\mathrm{TV}}(\mathbf{p}', \widehat{\mathbf{p}}) \leq \alpha + \varepsilon'$$

If we want this to be at most ε, then we can choose for instance

$$\alpha = \varepsilon' = \frac{\varepsilon}{2}$$

and then *learning* $\mathbf{p}'$ *with n samples to distance* $\frac{\varepsilon}{2}$ *implies learning* $\mathbf{p}$ *to distance ε with n samples, and the same probability of success.*

But why did we do all this? Well, now, for every i,

$$\mathbf{p}'(i) = (1-\alpha)\cdot\mathbf{p}(i) + \alpha\cdot\mathbf{u}_k(i) \geq \alpha\cdot\mathbf{u}_k(i) = \frac{\varepsilon}{2}\cdot\frac{1}{k}$$

and so we satisfy the assumption with $\tau = \frac{\varepsilon}{2k}$.

Problem 6. Instead of looking at all $\binom{n}{2}$ possible pairs of samples in Algorithm 21 for uniformity testing, describe and analyse the tester which partitions the n samples into $\frac{n}{2}$ (independent) pairs of samples, and use them to estimate $\Pr[\,X = Y\,]$. What is the resulting sample complexity?

Solution 6. The algorithm is as follows: given n i.i.d. samples from $\mathbf{p}$ (n is assumed even without loss of generality), get $n/2$ disjoint pairs of the form (x_{2i-1}, x_{2i}) for $1 \leq i \leq n/2$, and for set

$$y_i := \mathbf{1}_{x_{2i-1}=x_{2i}}$$

We have that $y_1, \dots, y_{n/2}$ are i.i.d. Bernoulli (coin tosses) with

$$\mathbb{E}[y_i] = \Pr[\,x_{2i-1} = x_{2i}\,] = \|p\|_2^2\,.$$

We want to distinguish between $\|p\|_2^2 = \frac{1}{k}$ and $\|p\|_2^2 > \frac{1+4\varepsilon^2}{k}$ (see discussion just after Remark 55.1), say with constant failure probability $\delta = 1/3$, which by Theorem 52 will lead to

$$\frac{n}{2} = O\left(\frac{1}{(1/k)\cdot(\varepsilon^2)^2}\log\frac{1}{\delta}\right) = O\left(\frac{k}{\varepsilon^4}\right)$$

which is even worse than what we would need to learn the distribution! What went wrong? Instead of looking at all possible things which could give us a collision (all $\binom{n}{2} = \Theta(n^2)$ pairs of samples), we ended up only looking at $n/2$ pairs. It is much simpler to analyse, but we lost a quadratic factor in n by doing so, which is intuitively why we end up with k/ε^4 instead of $\sqrt{k}/\varepsilon^2$.

Problem 7. *This is a programming exercise, to be done in, e.g., a Jupyter notebook.*

a) Write a function which, given two probability distributions represented as two arrays of the same size, computes their total variation distance.

b) Implement the empirical estimator seen in class: given the domain size k and a multiset of n numbers in $\{1, 2, \dots, k\}$, return the empirical probability distribution over $\{1, 2, \dots, k\}$.

c) Implement the uniformity testing algorithm (Algorithm 21).

d) Import the Canada's 6/49 lotto dataset (from `https://www.kaggle.com/datasets/dataScienceai/lottery-dataset`, available on Ed).

e) Learn the distribution of the first number, from the $n = 3,665$ samples. Plot the result.

f) Test whether the distribution of the "bonus number" is uniform, from the $n = 3,665$ samples, for $\varepsilon \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Report the results.

g) Learn the distribution of the "bonus number", from the $n = 3,665$ samples, and compute the total variation distance between the resulting $\widehat{\mathbf{p}}$ and the uniform distribution on $\{1, 2, \dots, 49\}$.

Advanced

Problem 8. Consider the following alternative approach to learn a probability distribution over a domain $\mathcal{X}$ of size k:

1. Take n i.i.d. samples from $\mathbf{p}$

2. Compute, for every domain element $i \in \mathcal{X}$, the number n_i of times it appears among the n samples.

3. For every $i \in \mathcal{X}$, let

$$\widehat{\mathbf{p}}(i) = \frac{n_i + 1}{n + k}$$

4. return $\widehat{\mathbf{p}}$

(This is called the *Laplace estimator*. Note that, in contrast to the empirical estimator, it assigns non-zero probability to every element of the domain, even those that do not appear in the samples.)

a) Show that $\widehat{\mathbf{p}}$ is a probability distribution.

b) Define the *chi-squared divergence* between probability distributions as

$$\chi^2(\mathbf{p}\|\mathbf{q}) = \sum_{x\in\mathcal{X}} \frac{(\mathbf{p}(x)-\mathbf{q}(x))^2}{\mathbf{q}(x)}$$

(Note that this is not symmetric, and not bounded!) Show that $d_{\mathrm{TV}}(\mathbf{p},\mathbf{q})^2 \leq \frac{1}{4}\chi^2(\mathbf{p}\|\mathbf{q})$ for every $\mathbf{p},\mathbf{q}$.

c) Show that $\mathbb{E}[\chi^2(\mathbf{p}\|\widehat{\mathbf{p}})] \leq \frac{k-1}{n+1}$.

d) Conclude on the value of n sufficient to learn $\mathbf{p}$ to total variation distance ε using the Laplace estimator.

Solution 8.

a) Since $\widehat{\mathbf{p}}$ is non-negative, it suffices that it sums to 1:

$$\sum_{i\in\mathcal{X}} \widehat{\mathbf{p}}(i) = \sum_{i\in\mathcal{X}} \frac{n_i+1}{n+k} = \frac{\sum_{i\in\mathcal{X}}(n_i+1)}{n+k} = \frac{n+k}{n+k} = 1$$

since $|\mathcal{X}| = k$ and $\sum_{i\in\mathcal{X}} n_i = n$.

b) By Cauchy–Schwarz,

$$\begin{aligned} d_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) &= \frac{1}{2}\sum_{x\in\mathcal{X}} |\mathbf{p}(x)-\mathbf{q}(x)| \\ &\leq \frac{1}{2}\sqrt{\sum_{x\in\mathcal{X}} \frac{(\mathbf{p}(x)-\mathbf{q}(x))^2}{\mathbf{q}(x)}}\sqrt{\sum_{x\in\mathcal{X}} \mathbf{q}(x)} \\ &= \frac{1}{2}\sqrt{\chi^2(\mathbf{p}\|\mathbf{q})}. \end{aligned}$$

c) First, we can expand the χ^2 divergence to get

$$\chi^2(\mathbf{p}\|\mathbf{q}) = \sum_{x\in\mathcal{X}} \frac{\mathbf{p}(x)^2 - 2\mathbf{p}(x)\mathbf{q}(x) + \mathbf{q}(x)^2}{\mathbf{q}(x)} = \sum_{x\in\mathcal{X}} \frac{\mathbf{p}(x)^2}{\mathbf{q}(x)} - 1$$

after simplifying and summing, using that $\sum_{x\in\mathcal{X}} \mathbf{p}(x) = \sum_{x\in\mathcal{X}} \mathbf{q}(x) = 1$. While $n_1, \dots, n_k$ are not independent, we can still use linearity of expectation:

$$\mathbb{E}[\chi^2(\mathbf{p}\|\widehat{\mathbf{p}})] = -1 + \sum_{x\in\mathcal{X}} \mathbb{E}\left[\frac{\mathbf{p}(x)^2}{\widehat{\mathbf{p}}}\right] = -1 + \sum_{x\in\mathcal{X}} \mathbf{p}(x)^2(n+k)\mathbb{E}\left[\frac{1}{n_x+1}\right]$$

Since $n_x \sim \mathrm{Bin}(n, \mathbf{p}(x))$, a "simple calculation" involving manipulating Binomial coefficients shows that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n_x+1}\right] &= \sum_{\ell=0}^{n} \binom{n}{\ell} \frac{\mathbf{p}(x)^\ell(1-\mathbf{p}(x))^{n-\ell}}{\ell+1} \\
&= \frac{1}{\mathbf{p}(x)(n+1)} \sum_{\ell=0}^{n} \binom{n+1}{\ell+1} \mathbf{p}(x)^{\ell+1}(1-\mathbf{p}(x))^{n+1-(\ell+1)} \\
&= \frac{1}{\mathbf{p}(x)(n+1)} \sum_{m=1}^{n+1} \binom{n+1}{m} \mathbf{p}(x)^{m}(1-\mathbf{p}(x))^{n+1-m} \\
&= \frac{1-(1-\mathbf{p}(x))^{n+1}}{\mathbf{p}(x)(n+1)} \\
&\le \frac{1}{\mathbf{p}(x)(n+1)}
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{E}[\chi^2(\mathbf{p}\|\widehat{\mathbf{p}})] &\le -1 + \sum_{x\in\mathcal{X}} \mathbf{p}(x)^2 \frac{n+k}{\mathbf{p}(x)(n+1)} \\
&= -1 + \frac{n+k}{n+1} \\
&= \frac{k-1}{n+1}
\end{aligned}$$

d) By b), to learn to TV ε it is enough to learn to χ^2 divergence $4\varepsilon^2$. By c), to learn to *expected* χ^2 divergence $O(\varepsilon^2)$ it is enough to have $n = O(k/\varepsilon^2)$. Combining this with Markov's inequality, to learn to TV ε with probability at least $9/10$ it is enough to have *expected* χ^2 divergence $4\varepsilon^2/10$, and so it is enough to have $\boxed{n \geq \frac{10k}{4\varepsilon^2}}$.

In more detail, here are two possible approaches (the second giving a slightly worse bound). The first:

$$\begin{aligned}
\Pr[d_{\mathrm{TV}}(\mathbf{p}, \hat{\mathbf{p}}) > \varepsilon] &= \Pr[d_{\mathrm{TV}}(\mathbf{p}, \hat{\mathbf{p}})^2 > \varepsilon^2] \\
&\leq \frac{\mathbb{E}[d_{\mathrm{TV}}(\mathbf{p}, \hat{\mathbf{p}})^2]}{\varepsilon^2} && \text{(Markov)} \\
&\leq \frac{\mathbb{E}[\frac{1}{4}\chi^2(\mathbf{p} \| \hat{\mathbf{p}})]}{\varepsilon^2} tagbyb) \\
&= \frac{\frac{k-1}{n+1}}{4\varepsilon^2} && \text{(by c))} \\
&\leq \frac{k}{4n\varepsilon^2}
\end{aligned}$$

and for this to be at most $1/10$, we set $n \geq \frac{10}{4\varepsilon^2}$.

Another (slightly worse!) option uses Jensen's inequality in the middle to handle the square root, instead of getting rid of it before Markov's inequality):

$$\begin{aligned}
\Pr[d_{\mathrm{TV}}(\mathbf{p}, \hat{\mathbf{p}}) > \varepsilon] &\leq \frac{\mathbb{E}[d_{\mathrm{TV}}(\mathbf{p}, \hat{\mathbf{p}})]}{\varepsilon} && \text{(Markov)} \\
&\leq \frac{\mathbb{E}[\frac{1}{2}\sqrt{\chi^2(\mathbf{p} \| \hat{\mathbf{p}})}]}{\varepsilon} \\
&\leq \frac{\frac{1}{2}\sqrt{\mathbb{E}[\chi^2(\mathbf{p} \| \hat{\mathbf{p}})]}}{\varepsilon} && \text{(Jensen's inequality)} \\
&= \sqrt{\frac{\frac{k-1}{n+1}}{4\varepsilon^2}} \\
&\leq \sqrt{\frac{k}{4n\varepsilon^2}}
\end{aligned}$$

and for this to be at most $1/10$, we set $n \geq \frac{100}{4\varepsilon^2} = \frac{25}{\varepsilon^2}$.