Approximate Hypothesis Testing

By Nicolas Le Gouic. Advisers: Robert Graczyk and Stefan Moser.

Abstract

We observe samples $\boldsymbol{X} = X_1, \ldots, X_n$ that are IID according to some unknown distribution $P \in \mathcal{H}$. How many samples are needed to guess a $\hat{P}$ "close" to P?

We seek the approximate sample complexity

$$n_{\varepsilon,\delta}(\mathcal{H}) := \min\left\{ n : \inf_{\hat{P}} \sup_{P \in \mathcal{H}} \mathbb{P}_{\boldsymbol{X} \sim \text{IID } P}\left[D\left(P, \hat{P}(\boldsymbol{X})\right) > \varepsilon \right] \leq \delta \right\}$$

Main Result (Informal)

The sample complexity is at most

$$n_{\varepsilon,\delta}(\mathcal{H}) = O\left(\frac{\ln\left(\frac{1}{\delta}\right) + \ln(\mathcal{C})}{\inf_{C, C' \in \mathcal{C}} d(C, C')} \right).$$

The sample complexity is at least

$$n_{\varepsilon,\delta}(\mathcal{H}) = \Omega\left(\frac{\ln\left(\frac{1}{\delta}\right)}{\inf_{C, C' \in \mathcal{C}} d(C, C')} \right).$$

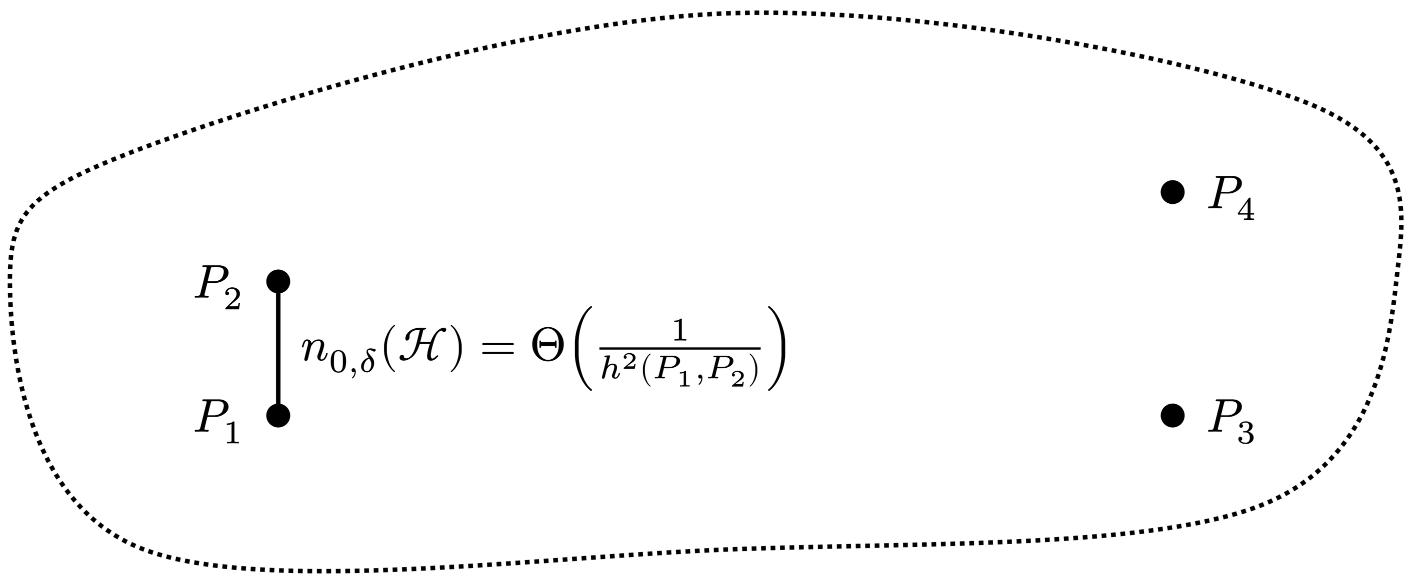
Here, $\mathcal{C}$ is a collection of (D, ε)-**dependent clusters that cover** $\mathcal{H}$, and $d(C, C')$ is the distance between clusters C and C' (defined below).

Classical Hypothesis Testing

Corresponds to $D\left(\hat{P}, P\right) = \mathbb{I}\left(\hat{P} = P\right)$ and $\varepsilon = 0$. Here, the sample complexity is characterized by the least squared Hellinger distance on $\mathcal{H}$,

$$n_{0,\delta}(\mathcal{H}) = \Theta\left(\frac{1}{\inf_{P, P' \in \mathcal{H}} h^2(P, P')} \right).$$

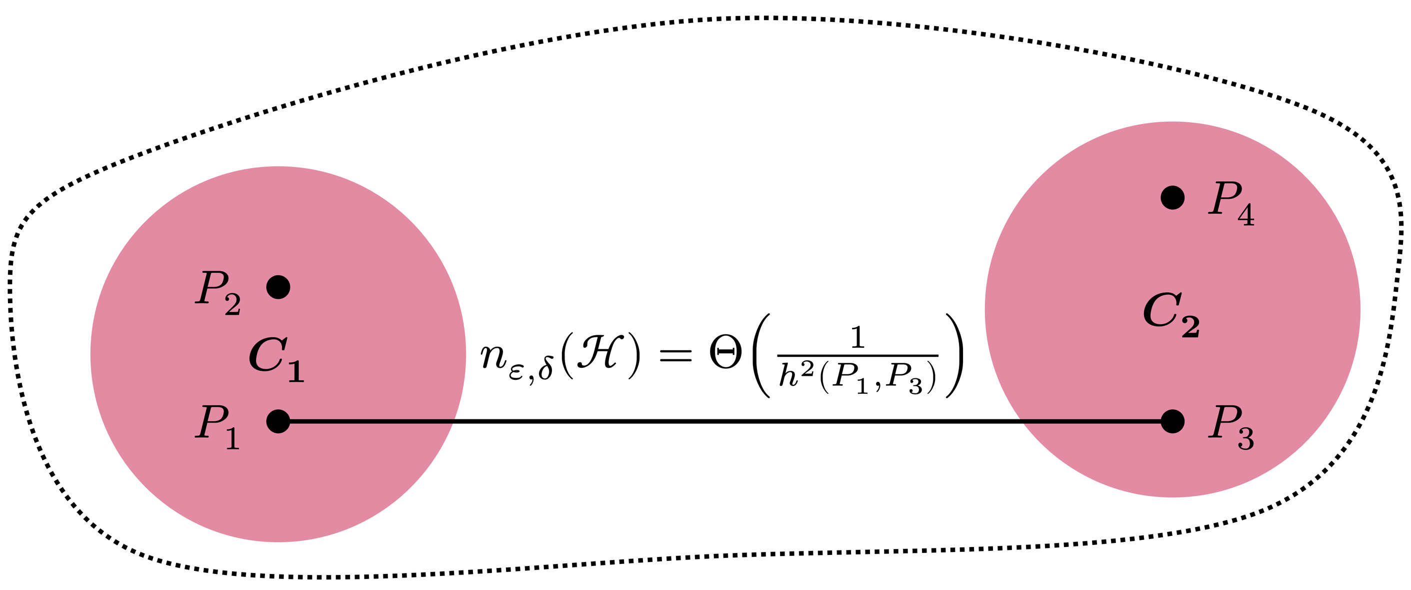
For example:



$\mathcal{H} = \{P_1, P_2, P_3, P_4\}$

Approximate Hypothesis Testing

We need no longer distinguish between P and P' that are ε-close!



$\mathcal{H} = C_1 \cup C_2$ where $C_1 = \{P_1, P_2\}, C_2 = \{P_3, P_4\}$

Key Insight

Cover $\mathcal{H}$ with clusters $\mathcal{C} = \{C_1, C_2, \ldots\}$ and define the distance

$$d(C, C') := \inf_{P \in C, P' \in C'} h^2(P, P').$$

Treat approximate hypothesis testing as **classical hypothesis testing on clusters** with a sample complexity that scales as $\Theta\left(1 / \inf_{C, C' \in \mathcal{C}} d(C, C')\right)$.

Proof Sketch: Upper Bound

First consider only distinguishing between two clusters C and C'. To that end, use a LLR test for suitable mixtures $\int_{P \in C} P \mu(P)$ vs. $\int_{P' \in C'} P' \mu'(P')$:

$$\begin{aligned}
&\sup_{C} \sup_{P \in C} \mathbb{P}_{\boldsymbol{X} \sim \text{IID } P}\left[\hat{C}(\boldsymbol{X}) \neq C\right] \\
&\quad \leq \sup_{P \in C} \mathbb{P}_{\boldsymbol{X} \sim \text{IID } P}\left[\hat{C}(\boldsymbol{X}) \neq C\right] + \sup_{P' \in C'} \mathbb{P}_{\boldsymbol{X} \sim \text{IID } P'}\left[\hat{C}(\boldsymbol{X}) \neq C'\right] \\
&\quad \leq 1 - \inf_{P \in C, P' \in C'} \text{TV}\left(P^{\otimes n}, P'^{\otimes n}\right) \\
&\quad \leq 1 - \inf_{P \in C, P' \in C'} h^2\left(P^{\otimes n}, P'^{\otimes n}\right) \\
&\quad \leq \left(1 - \inf_{P \in C, P' \in C'} h^2(P, P')\right)^n \\
&\quad = (1 - d(C, C'))^n \\
&\quad \leq e^{-n d(C, C')}
\end{aligned}$$

To distinguish between all clusters, apply the above test for every pair (C_i, C_j), and take a majority vote: the cluster containing the data-generating distribution wins if it is voted for when compared to any other cluster. By the union bound, the majority vote is not won with probability at most:

$$\sup_{C} \sup_{P \in C} \mathbb{P}_{\boldsymbol{X} \sim \text{IID } P}\left[\hat{C}(\boldsymbol{X})\right] \leq |\mathcal{C}| e^{-n \inf_{C, C' \in \mathcal{C}} d(C, C')}.$$

Proof Sketch: Lower Bound

Observe that

- distinguishing between fixed clusters C and C' is easier than distinguishing between all clusters.
- distinguishing between fixed distributions $P \in C$ and $P' \in C'$ is easier than distinguishing between all distributions in C and C'.

Thus, for any $C, C' \in \mathcal{C}$, and $P \in C, P' \in C'$:

$$\begin{aligned}
&\sup_{C} \sup_{P \in C} \mathbb{P}_{\boldsymbol{X} \sim \text{IID } P}\left[\hat{C}(\boldsymbol{X})\right] \\
&\quad \geq \max\left(\mathbb{P}_{\boldsymbol{X} \sim \text{IID } P}\left[\hat{P}(\boldsymbol{X}) \neq P\right], \mathbb{P}_{\boldsymbol{X} \sim \text{IID } P'}\left[\hat{P}(\boldsymbol{X}) \neq P'\right] \right) \\
&\quad \geq \frac{1}{2} - \frac{\text{TV}\left(P^{\otimes n}, P'^{\otimes n}\right)}{2} \\
&\quad > \frac{1}{4}\left(1 - h^2(P, P')\right)^{4n} \\
&\quad \geq \frac{1}{4} e^{-8 n h^2(P, P')}
\end{aligned}$$

Since this bound holds for any $P \in C, P' \in C'$,

$$\sup_{C} \sup_{P \in C} \mathbb{P}_{\boldsymbol{X} \sim \text{IID } P}\left[\hat{C}(\boldsymbol{X})\right] > \frac{1}{4} e^{-8 n d(C, C')}.$$

And since it also holds for any $C, C' \in \mathcal{C}$,

$$\sup_{C} \sup_{P \in C} \mathbb{P}_{\boldsymbol{X} \sim \text{IID } P}\left[\hat{C}(\boldsymbol{X})\right] > \frac{1}{4} e^{-8 n \inf_{C, C' \in \mathcal{C}} d(C, C')}.$$

Outlook

Key Paradigm

Approximate hypothesis testing is a flexible tool to understand data!

We plan to:

- study the minimal cluster covering $\inf_{\mathcal{C}} |\mathcal{C}|$ as a complexity measure of the hypothesis class $\mathcal{H}$.
- relate the distance $D(P, P')$ to real-world applications through

$$D(P, P') = \left|\mathbb{E}_{X \sim P}[c(X)] - \mathbb{E}_{X \sim P'}[c(X)]\right|^{\rho},$$

  where $c(X)$ is the cost of outcome X (e.g., loss due to rising stock price).
- run simulations on synthetic and real data.

nicolas.legouic@ip-paris.fr, robert.graczyk@telecom-paris.fr, moser@isi.ee.ethz.ch