Inference in Non-parametric Settings with Generalised Likelihood Ratios

Wouter M. Koolen

CWI and University of Twente

ML-ST-N&O-afternoon, CWI, December 13, 2024

Goal

In this talk we look at statistically rejecting hypotheses.

Goal

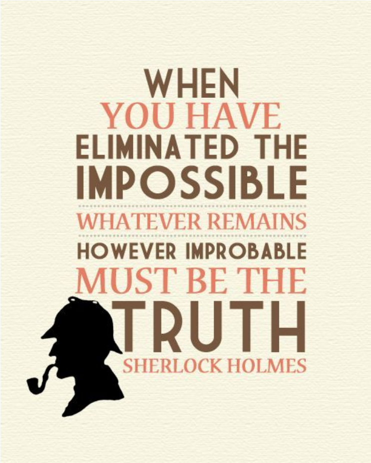
In this talk we look at statistically rejecting hypotheses.

Why is that **interesting**? Don't we want to learn the truth?

Goal

In this talk we look at statistically rejecting hypotheses.

Why is that **interesting**? Don't we want to learn the truth?



Setup

We look at a sequence of scalar outcomes $X_1, X_2, \ldots$ revealed to us sequentially.

Setup

We look at a sequence of scalar outcomes $X_1, X_2, \ldots$ revealed to us sequentially.

We have some hypothesis that X_i are i.i.d. from P.

Setup

We look at a sequence of scalar outcomes $X_1, X_2, \ldots$ revealed to us sequentially.

We have some hypothesis that X_i are i.i.d. from P.

We do not trust this hypothesis.

Setup

We look at a sequence of scalar outcomes $X_1, X_2, \ldots$ revealed to us sequentially.

We have some hypothesis that X_i are i.i.d. from P.

We do not trust this hypothesis.

So we want to reject P. Ideally fast.

Simple vs Simple

Go-to-setting

Say we do not believe P is the case. Instead, we think Q is a better explanation.

If we are right and data come from Q, how long until we can reject P?

Definition

Fix a confidence level $\delta \in (0,1)$. A stopping time τ against P is **δ-correct** if

$$P\{\tau < \infty\} \leq \delta.$$

Among all δ-correct τ stopping times, we like to minimise expected stopping time $\mathbb{E}_Q[\tau]$.

Simple vs Simple result

The optimal expected stopping time is

$$\min_{\substack{\tau \text{ a stopping time} \\ \text{that is } \delta\text{-correct against } P}} \mathbb{E}_Q[\tau]$$

Simple vs Simple result

The optimal expected stopping time is

$$\min_{\substack{\tau \text{ a stopping time} \\ \text{that is } \delta\text{-correct against } P}} \mathbb{E}_Q[\tau]$$

In the simple vs simple case, this is

$$\min_{\substack{\tau \text{ a stopping time} \\ \text{that is } \delta\text{-correct against } P}} \mathbb{E}_Q[\tau] \;=\; \frac{\ln \frac{1}{\delta}}{\mathrm{KL}(Q \| P)}$$

Lower bound by KL Compression

Theorem

Any δ-correct stopping time τ against P has expected stopping time at least

$$\mathbb{E}_Q[\tau] \;\geq\; \frac{\ln \frac{1}{\delta}}{\mathsf{KL}(Q\|P)}$$

Lower bound by KL Compression

Theorem

Any δ-correct stopping time τ against P has expected stopping time at least

$$\mathbb{E}_Q[\tau] \;\geq\; \frac{\ln \frac{1}{\delta}}{\mathrm{KL}(Q\|P)}$$

Proof.

By KL contraction and δ-correctness, we have

$$\mathbb{E}_Q[\tau]\,\mathrm{KL}(Q\|P) \;=\; \mathrm{KL}(Q^\tau\|P^\tau) \;\geq\; \mathrm{kl}\left(Q\left\{\tau < \infty\right\}, P\left\{\tau < \infty\right\}\right) \;\geq\; \ln\frac{1}{\delta}.$$

□

Upper bound by likelihood ratio stopping

Let's consider the likelihood ratio for data $X_1, \ldots, X_n$

$$\frac{dQ}{dP}(X^n) = \prod_{t=1}^{n} \frac{dQ}{dP}(X_t)$$

and the associated **likelihood ratio stopping time**

$$\tau := \inf\left\{ n \middle| \frac{dQ}{dP}(X^n) \geq \frac{1}{\delta} \right\}.$$

Likelihood ratio stopping works

Theorem

The likelihood ratio stopping time τ

- *is* δ*-correct*
- *ensures* $\mathbb{E}_Q[\tau] = \frac{\ln \frac{1}{\delta}}{KL(Q \| P)}$.

Likelihood ratio stopping works

Theorem

The likelihood ratio stopping time τ

- *is* δ*-correct*
- *ensures* $\mathbb{E}_Q[\tau] = \frac{\ln \frac{1}{\delta}}{KL(Q\|P)}$.

Proof.

- By Ville's Inequality, $P\left\{\tau < \infty\right\} = P\left\{\exists n : \frac{dQ}{dP}(X^n) \geq \frac{1}{\delta}\right\} \leq \delta$.
- By Wald's Equality, assuming $Q\left\{\tau < \infty\right\} = 1$, we have,

$$\ln \frac{1}{\delta} \approx \mathbb{E}_Q\left[\sum_{t=1}^{\tau} \ln \frac{dQ}{dP}(X_t)\right] = \mathbb{E}_Q\left[\sum_{t=1}^{\tau} \mathrm{KL}(Q\|P)\right] = \mathbb{E}_Q[\tau]\,\mathrm{KL}(Q\|P)$$

□

Summary

Consider two distributions P and Q.

We have a stopping time such that

- (Safety) If we are in P, we will only reject it with small probability.
- (Power) If we are in Q, we will reject P with about $\frac{\ln \frac{1}{\delta}}{\mathrm{KL}(Q \| P)}$ samples.

Summary

Consider two distributions P and Q.

We have a stopping time such that

- (Safety) If we are in P, we will only reject it with small probability.
- (Power) If we are in Q, we will reject P with about $\frac{\ln \frac{1}{\delta}}{\mathrm{KL}(Q \| P)}$ samples.

Application: we can do this in parallel with P and Q reversed, to figure out in which of the two we are.

Summary

Consider two distributions P and Q.

We have a stopping time such that

- (Safety) If we are in P, we will only reject it with small probability.
- (Power) If we are in Q, we will reject P with about $\frac{\ln \frac{1}{\delta}}{\operatorname{KL}(Q \| P)}$ samples.

Application: we can do this in parallel with P and Q reversed, to figure out in which of the two we are.

Problem: we typically want to reject many P and we may not know a good Q.

Composite Null and Alternative

Let's go composite

Let's study probability distributions on the interval $[0, 1]$. For $m \in [0, 1]$, consider

$$\mathcal{H}_m \ := \ \{P \text{ on } [0, 1] | \mathbb{E}_P[X] = m\} \, .$$

Let us try to reject the composite null $\mathcal{H}_m$.

Let's go composite

Let's study probability distributions on the interval $[0, 1]$. For $m \in [0, 1]$, consider

$$\mathcal{H}_m \; := \; \{P \text{ on } [0, 1] | \mathbb{E}_P[X] = m\} \,.$$

Let us try to reject the composite null $\mathcal{H}_m$.

Definition

We say stopping time τ against $\mathcal{H}_m$ is **δ-correct** if

$$\forall P \in \mathcal{H}_m : \quad P\{\tau < \infty\} \leq \delta$$

Let's go composite

Let's study probability distributions on the interval $[0, 1]$. For $m \in [0, 1]$, consider

$$\mathcal{H}_m \ := \ \{P \text{ on } [0, 1] | \mathbb{E}_P[X] = m\} \, .$$

Let us try to reject the composite null $\mathcal{H}_m$.

Definition

We say stopping time τ against $\mathcal{H}_m$ is **δ-correct** if

$$\forall P \in \mathcal{H}_m : \quad P\{\tau < \infty\} \leq \delta$$

Suppose data come from $Q \notin \mathcal{H}_m$. How may samples will it take to reject $\mathcal{H}_m$?

Sample complexity

By the same KL compression lower bound, for any $P \in \mathcal{H}_m$,

$$\mathbb{E}_Q[\tau] \;\geq\; \frac{\ln \frac{1}{\delta}}{\mathrm{KL}(Q\|P)}$$

or equivalently,

$$\mathbb{E}_Q[\tau] \;\geq\; \frac{\ln \frac{1}{\delta}}{\mathrm{KLinf}(Q\|m)} \qquad \text{where} \qquad \mathrm{KLinf}(Q\|m) \;:=\; \inf_{P \in \mathcal{H}_m} \mathrm{KL}(Q\|P)$$

Sample complexity

By the same KL compression lower bound, for any $P \in \mathcal{H}_m$,

$$\mathbb{E}_Q[\tau] \;\geq\; \frac{\ln \frac{1}{\delta}}{\mathrm{KL}(Q \| P)}$$

or equivalently,

$$\mathbb{E}_Q[\tau] \;\geq\; \frac{\ln \frac{1}{\delta}}{\mathrm{KLinf}(Q \| m)} \qquad \text{where} \qquad \mathrm{KLinf}(Q \| m) \;:=\; \inf_{P \in \mathcal{H}_m} \mathrm{KL}(Q \| P)$$

Question: is that also an **upper bound**?

Duality for KLinf (Honda and Takemura, 2010)

optimisation

Can we understand that KLinf? Well,

$$
\begin{aligned}
\mathrm{KLinf}(Q\|m) &= \inf_{P \in \mathcal{H}_m} \mathrm{KL}(Q\|P) \\
&= \min_{\substack{P \text{ prob } [0,1] \\ \mathbb{E}_P[X]=m}} \mathrm{KL}(Q\|P) \\
&= \max_{\lambda,\nu} \min_{P \text{ meas } [0,1]} \mathrm{KL}(Q\|P) + \lambda\, \mathbb{E}_P[X-m] + \nu(\mathbb{E}_P[1]-1) \\
&= \max_{\substack{\lambda,\nu \\ \forall x\in[0,1]: \nu+\lambda(x-m)\geq 0}} \mathbb{E}_Q\left[\ln\left(\nu+\lambda(X-m)\right)\right] + 1 - \nu \\
&= \max_{\substack{\lambda \\ \forall x\in[0,1]: 1+\lambda(x-m)\geq 0}} \mathbb{E}_Q\left[\ln\left(1+\lambda(X-m)\right)\right]
\end{aligned}
$$

Duality for KLinf (Honda and Takemura, 2010)

optimisation

Can we understand that KLinf? Well,

$$
\begin{aligned}
\operatorname{KLinf}(Q\|m) &= \inf_{P \in \mathcal{H}_m} \operatorname{KL}(Q\|P) \\
&= \min_{\substack{P \text{ prob } [0,1] \\ \mathbb{E}_P[X]=m}} \operatorname{KL}(Q\|P) \\
&= \max_{\lambda,\nu} \min_{P \text{ meas } [0,1]} \operatorname{KL}(Q\|P) + \lambda\, \mathbb{E}_P[X-m] + \nu(\mathbb{E}_P[1]-1) \\
&= \max_{\substack{\lambda,\nu \\ \forall x\in[0,1]: \nu+\lambda(x-m)\geq 0}} \mathbb{E}_Q\left[\ln\left(\nu+\lambda(X-m)\right)\right] + 1 - \nu \\
&= \max_{\substack{\lambda \\ \forall x\in[0,1]: 1+\lambda(x-m)\geq 0}} \mathbb{E}_Q\left[\ln\left(1+\lambda(X-m)\right)\right]
\end{aligned}
$$

The optimal choice is

$$
P^* = \frac{Q}{\nu+\lambda(X-m)} \qquad \text{and} \qquad \nu^* = 1
$$

with possibly some extra mass at either endpoint 0 or 1 of the domain.

Martingale

e-values

We proved

$$\mathrm{KLinf}(Q\|m) = \max_{\lambda\in\left[\frac{-1}{1-m},\frac{1}{m}\right]} \mathbb{E}_Q\left[\ln\left(1+\lambda(X-m)\right)\right]$$

In fact, for every $\lambda \in \left[\frac{-1}{1-m}, \frac{1}{m}\right]$ the expression $1+\lambda(X-m)$ is a

- multiplicative increment of a non-negative martingale
- e-value
- **likelihood ratio**
- **Bayes factor**

against P for **every** $P \in \mathcal{H}_m$.

Martingale

e-values

We proved

$$\mathrm{KLinf}(Q\|m) = \max_{\lambda \in \left[\frac{-1}{1-m}, \frac{1}{m}\right]} \mathbb{E}_Q\left[\ln\left(1 + \lambda(X - m)\right)\right]$$

In fact, for every $\lambda \in \left[\frac{-1}{1-m}, \frac{1}{m}\right]$ the expression $1 + \lambda(X - m)$ is a

- multiplicative increment of a non-negative martingale
- e-value
- **likelihood ratio**
- **Bayes factor**

against P for **every** $P \in \mathcal{H}_m$.

Suggests the "likelihood ratio" statistic

$$\sum_{t=1}^{n} \ln(1 + \lambda_Q(X_t - m))$$

where λ_Q is the $\arg\max_\lambda$ of the $\mathrm{KLinf}(Q\|m)$.

Likelihood ratio

Let us stop when

$$\tau \;:=\; \inf\left\{ n \middle| \sum_{t=1}^{n} \ln(1 + \lambda_Q(X_t - m)) \geq \ln\frac{1}{\delta} \right\}.$$

Likelihood ratio

Let us stop when

$$\tau \;:=\; \inf\left\{ n \,\middle|\, \sum_{t=1}^{n} \ln(1+\lambda_Q(X_t - m)) \geq \ln\frac{1}{\delta} \right\}.$$

This is δ-correct under $\mathcal{H}_0$, again by Ville's Inequality. Moreover, by Wald's Equality

$$\mathbb{E}_Q[\tau]\,\mathbb{E}_Q[\ln(1+\lambda_Q(X-m))] \;=\; \mathbb{E}_Q\left[\sum_{t=1}^{\tau} \ln(1+\lambda_Q(X_t-m))\right] \;\approx\; \ln\frac{1}{\delta}$$

What if we do not know Q?

We talked about rejecting $\mathcal{H}_m$ using Q. That lead to the recipe of using a fixed λ_Q. What if we do not know an a-priori suitable Q?

What if we do not know Q?

We talked about rejecting $\mathcal{H}_m$ using Q. That lead to the recipe of using a fixed λ_Q. What if we do not know an a-priori suitable Q?

We need to somehow learn the alternative Q.

What if we do not know Q?

We talked about rejecting $\mathcal{H}_m$ using Q. That lead to the recipe of using a fixed λ_Q. What if we do not know an a-priori suitable Q?

We need to somehow learn the alternative Q.

Simple idea: fit λ to the data.

- Good: it will converge to actual Q
- Bad: it over-fits the data

What if we do not know Q?

We talked about rejecting $\mathcal{H}_m$ using Q. That lead to the recipe of using a fixed λ_Q. What if we do not know an a-priori suitable Q?

We need to somehow learn the alternative Q.

Simple idea: fit λ to the data.

- Good: it will converge to actual Q
- Bad: it over-fits the data

Technically, we will use the statistic

$$n \operatorname{KLinf}(\hat{P}_n \| m) = \max_{\lambda \in \left[\frac{-1}{1-m}, \frac{1}{m}\right]} \sum_{t=1}^{n} \ln\left(1 + \lambda(X_t - m)\right)$$

What if we do not know Q?

We talked about rejecting $\mathcal{H}_m$ using Q. That lead to the recipe of using a fixed λ_Q. What if we do not know an a-priori suitable Q?

We need to somehow learn the alternative Q.

Simple idea: fit λ to the data.

- Good: it will converge to actual Q
- Bad: it over-fits the data

Technically, we will use the statistic

$$n \,\mathrm{KLinf}(\hat{P}_n \| m) \;=\; \max_{\lambda \in \left[\frac{-1}{1-m}, \frac{1}{m}\right]} \sum_{t=1}^{n} \ln\left(1 + \lambda(X_t - m)\right)$$

In contrast to the fixed λ case, this is not (the logarithm of) a martingale. Endangers δ-correctness.

Taming the over-fitting

online learning

What is the probability under P that

$$n\,\mathrm{KLinf}(\hat{P}_n \| m) = \max_{\lambda \in \left[\frac{-1}{1-m}, \frac{1}{m}\right]} \sum_{t=1}^{n} \ln\left(1 + \lambda(X_t - m)\right)$$

exceeds some given threshold?

Taming the over-fitting

online learning

What is the probability under P that

$$n\,\mathrm{KLinf}(\hat{P}_n \| m) \;=\; \max_{\lambda \in \left[\frac{-1}{1-m}, \frac{1}{m}\right]} \sum_{t=1}^{n} \ln\left(1 + \lambda(X_t - m)\right)$$

exceeds some given threshold?

Idea: We can relate the max to an average.

Theorem

$$n\,\mathrm{KLinf}(\hat{P}_n \| m) \;\le\; \ln \int_{\frac{-1}{1-m}}^{\frac{1}{m}} e^{\sum_{t=1}^{n} \ln(1+\lambda(X_t - m))} m(1-m)\,\mathrm{d}\lambda + \ln n + O(1)$$

Taming the over-fitting

online learning

What is the probability under P that

$$n\,\mathrm{KLinf}(\hat{P}_n \| m) \;=\; \max_{\lambda \in \left[\frac{-1}{1-m}, \frac{1}{m}\right]} \sum_{t=1}^{n} \ln\left(1 + \lambda(X_t - m)\right)$$

exceeds some given threshold?

Idea: We can relate the max to an average.

Theorem

$$n\,\mathrm{KLinf}(\hat{P}_n \| m) \;\le\; \ln \int_{\frac{-1}{1-m}}^{\frac{1}{m}} e^{\sum_{t=1}^{n} \ln(1+\lambda(X_t - m))} m(1-m)\,\mathrm{d}\lambda + \ln n + O(1)$$

Proof.

Invoke worst-case regret bound for exp-concave losses. □

Upshot

Under any $P \in \mathcal{H}_m$, we have

$$P\left\{\exists n : n\,\mathsf{KLinf}(\hat{P}_n \| m) \geq \ln\frac{1}{\delta} + \ln n\right\} \leq \delta$$

which witnesses δ-correctness of the stopping time

$$\tau \;:=\; \inf\left\{n \middle| n\,\mathsf{KLinf}(\hat{P}_n \| m) \geq \ln\frac{1}{\delta} + \ln n\right\}.$$

Upshot

Under any $P \in \mathcal{H}_m$, we have

$$P\left\{\exists n : n \operatorname{KLinf}(\hat{P}_n \| m) \geq \ln \frac{1}{\delta} + \ln n\right\} \leq \delta$$

which witnesses δ-correctness of the stopping time

$$\tau \;:=\; \inf\left\{n \,\middle|\, n \operatorname{KLinf}(\hat{P}_n \| m) \geq \ln \frac{1}{\delta} + \ln n\right\}.$$

As for the power, we have

$$\mathbb{E}_Q[\tau] \;\leq\; \frac{\ln \frac{1}{\delta}}{\operatorname{KLinf}(Q \| m)} + \ln \frac{\ln \frac{1}{\delta}}{\operatorname{KLinf}(Q \| m)}$$

Asymptotic optimality in $\delta \to 0$.

Extensions

How general is this KLinf idea?

Moment-constrained classes. Let's look at e.g.

$$\mathcal{H}^{\epsilon}_{B,m} = \left\{ P \text{ on } \mathbb{R} \;\middle|\; \mathbb{E}_P[X] = m, \mathbb{E}_P\left[|X|^{1+\epsilon}\right] \le B \right\}$$

How general is this KLinf idea?

Moment-constrained classes. Let's look at e.g.

$$\mathcal{H}_{B,m}^{\epsilon} \;=\; \left\{ P \text{ on } \mathbb{R} \;\middle|\; \mathbb{E}_P[X] = m, \mathbb{E}_P\left[|X|^{1+\epsilon}\right] \leq B \right\}$$

Going through duality, we end up with two Lagrange multipliers:

$$\operatorname{KLinf}(Q\|m) \;=\; \max_{\substack{\lambda_1 \in \mathbb{R}, \lambda_2 \geq 0 \\ \forall x \in \mathbb{R}: 1+\lambda_1(X-m)+\lambda_2\left(|X|^{1+\epsilon}-B\right) \geq 0}} \mathbb{E}_Q\left[\ln\left(1+\lambda_1(X-m)+\lambda_2\left(|X|^{1+\epsilon}-B\right)\right)\right]$$

Online learning regret now $2 \ln n$. In general with d constraints, $d \ln n$.

How general is this KLinf idea?

Moment-constrained classes. Let's look at e.g.

$$\mathcal{H}_{B,m}^{\epsilon} \;=\; \left\{ P \text{ on } \mathbb{R} \;\middle|\; \mathbb{E}_P[X] = m, \mathbb{E}_P\left[|X|^{1+\epsilon}\right] \le B \right\}$$

Going through duality, we end up with two Lagrange multipliers:

$$\operatorname{KLinf}(Q\|m) \;=\; \max_{\substack{\lambda_1 \in \mathbb{R}, \lambda_2 \ge 0 \\ \forall x \in \mathbb{R}: 1+\lambda_1(X-m)+\lambda_2\left(|X|^{1+\epsilon}-B\right) \ge 0}} \mathbb{E}_Q\left[\ln\left(1+\lambda_1(X-m)+\lambda_2\left(|X|^{1+\epsilon}-B\right)\right)\right]$$

Online learning regret now $2\ln n$. In general with d constraints, $d \ln n$.

Application: anytime-valid confidence intervals for heavy-tailed distributions. (Agrawal, Juneja, and Koolen, 2021)

Questions

Questions

- What about infinitely many constraints? E.g.
  - Sub-Gaussian class
  $$\mathcal{H} = \left\{ P \text{ on } \mathbb{R} \middle| \forall \eta \in \mathbb{R} : \mathbb{E}_P[e^{\eta X}] \leq e^{\frac{1}{2}\eta^2} \right\}$$
    (project with Shubhada Agrawal)
  - Monontone densities (project with Yunda Hao)
- Is that **regret** step tight? (project with Rémy Degenne, Timothée Mathieu, Shubhada Agarwal)
- What about centred moment-constrained classes? Adversarially **corrupted** distributions? (project with Debabrota Basu)
- In bandit applications often want to learn (i.e. reject) relations between two arms
  - Multi-objective best arm, Pareto front (Crepon, Garivier, and Koolen, 2024)
  - What about **constrained** best arm under dependence (project with Tyron Lardy and Christina Katsimerou)

Conclusion

Conclusion

We discussed KLinf, one of my favourite mathematical objects.

Let's talk!

References i

Agrawal, S., S. Juneja, and W. M. Koolen (Aug. 2021). **“Regret Minimization in Heavy-Tailed Bandits”.** In: *Proceedings of the 34th Annual Conference on Learning Theory (COLT)*.

Crepon, É., A. Garivier, and W. M. Koolen (Feb. 2024). **“Sequential Learning of the Pareto Front for Multi-objective Bandits”.** In: *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Vol. 238. Proceedings of Machine Learning Research.

Honda, J. and A. Takemura (2010). **“An Asymptotically Optimal Bandit Algorithm for Bounded Support Models.”.** In: *COLT*.