Credal Two-sample Tests of Epistemic Uncertainty

Siu Lun Chau[1], Antonin Schrab[2], Arthur Gretton[2], Dino Sejdinovic[4], Krikamol Muandet[1]

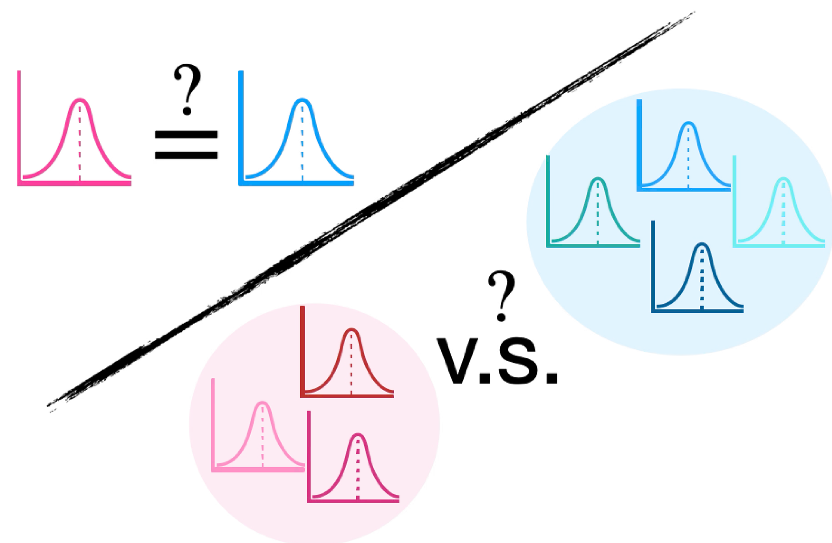
[1]CISPA Helmholtz Center for Information Security, Germany [2]University College London, UK [3]University of Adelaide, Australia

01 Motivation: How to Test under Dataset Uncertainty?

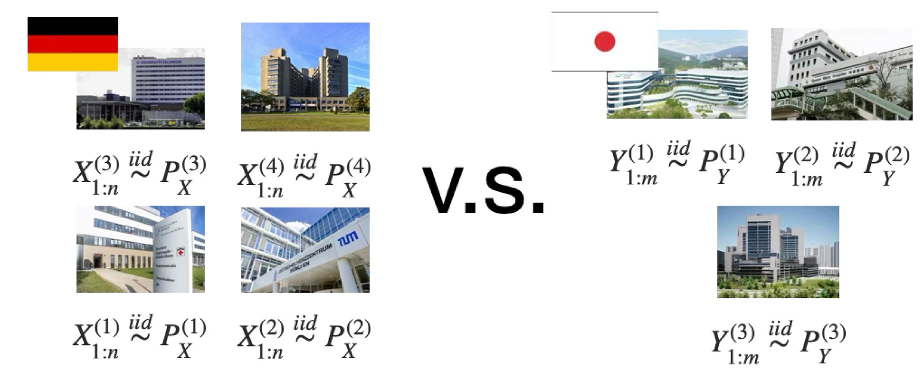
Precise Testing: Compare Aleatoric Uncertainty

V.S.

$X_1, \ldots, X_n \overset{iid}{\sim} P_X$ $Y_1, \ldots, Y_m \overset{iid}{\sim} P_Y$



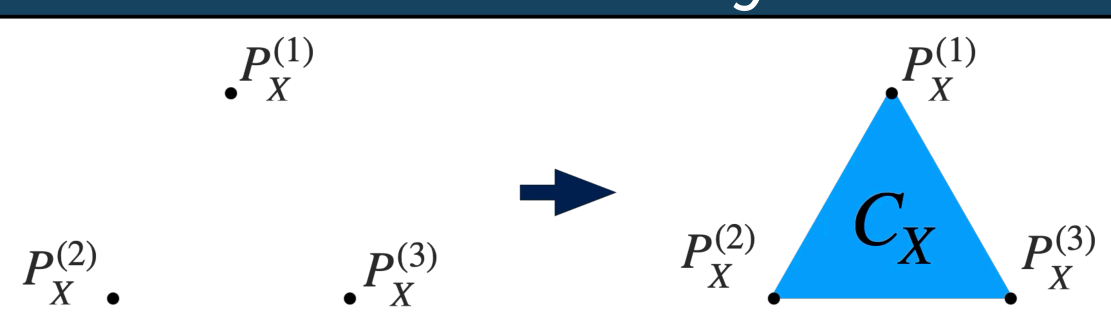
Credal Testing: Compare **Epistemic Uncertainty**



"Proxies" for P_X "Proxies" for P_Y

Main Research Question: Can we still reason statistically about P_X and P_Y in light of Dataset Uncertainty?

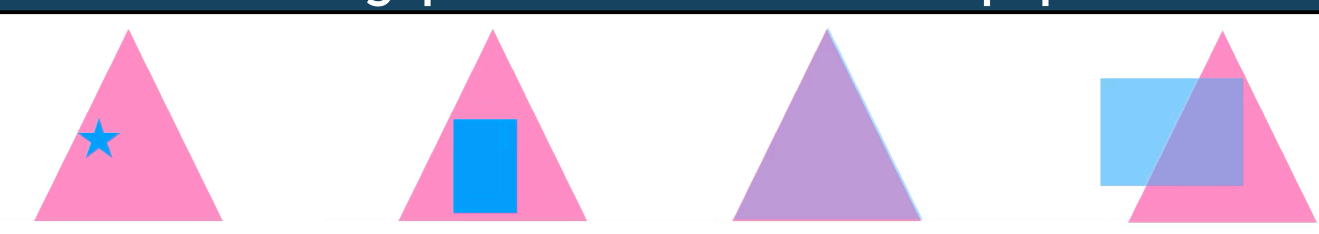
02 Dataset Uncertainty -> Credal Set



Turning Physical Probabilities into Belief Probabilities

- **Credal Set** $C_X = \mathrm{CH}\{P_X^{(1)}, \ldots, P_X^{(m)}\}$ represents Epistemic Dataset Uncertainty
- Championed by Imprecise Probabilists, Quasi-Bayesian Decision Theorists, Robust Bayesians, Formal Epistemologists

03 Credal Hypotheses & Applications



Specification	Inclusion	Equality	Plausibility
$H_{0,\in}: P_X \in \mathcal{C}_Y$	$H_{0,\subseteq}: \mathcal{C}_X \subseteq \mathcal{C}_Y$	$H_{0,=}: \mathcal{C}_X = \mathcal{C}_Y$	$H_{0,\cap}: \mathcal{C}_X \cap \mathcal{C}_Y \neq \emptyset$
$H_{A,\in}: P_X \notin \mathcal{C}_Y$	$H_{A,\subseteq}: \mathcal{C}_X \not\subseteq \mathcal{C}_Y$	$H_{A,=}: \mathcal{C}_X \neq \mathcal{C}_Y$	$H_{A,\cap}: \mathcal{C}_X \cap \mathcal{C}_Y = \emptyset$

- **Specification:** *Testing finite mixtures, Credal predictor calibration, Credal set verification...*
- **Inclusion:** *Uncertainty comparison...*
- **Equality:** *Testing treatment effect under ambiguity...*
- **Plausibility:** *Distributionally robust two-sample test...*

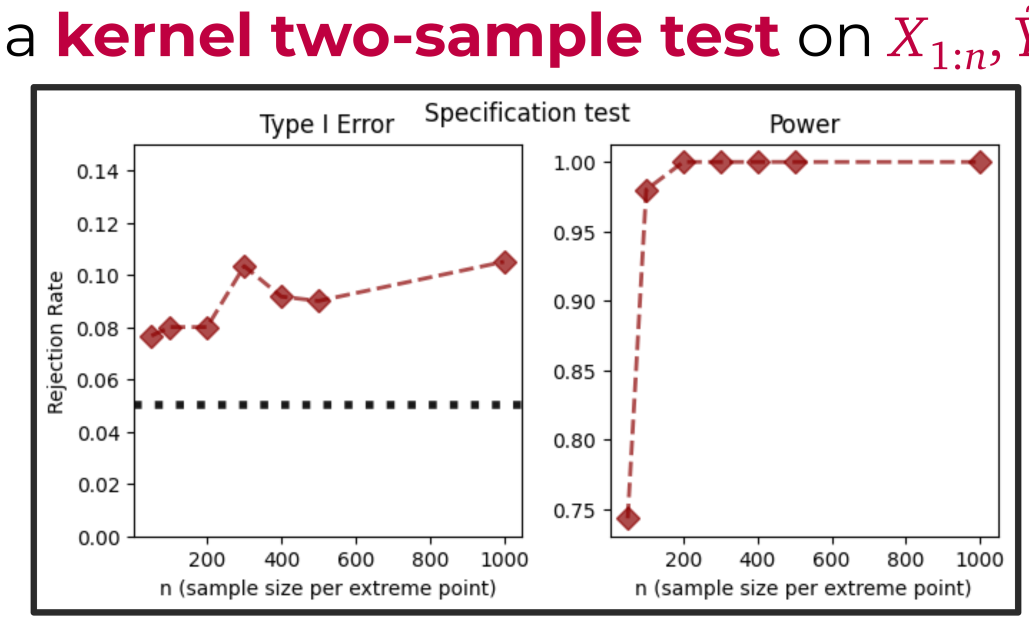
04 How do We Do it? (Specification)

Notice $P_X \in C_Y \iff \exists \eta_0 \in \Delta_{r-1},\ P_X = \eta_0^\top \mathbf{P}_Y)$

1. We start with $X_{1:n} \overset{iid}{\sim} P_X$, $Y_{1:m}^{(j)} \overset{iid}{\sim} P_Y^{(j)}$ for $j = 1, \ldots, r$,
2. Split data for estimation and testing based on ratio ρ.
3. Perform **Epistemic Alignment**:

$$\hat{\eta} = \arg\min_{\eta \in \Delta_{r-1}} \widehat{\mathrm{MMD}}^2(P_X, \eta^\top \mathbf{P}_Y)$$

4. Sample pseudo observations: $\tilde{Y}_{1:n} \sim \hat{\eta}^\top \hat{\mathbf{P}}_Y$
5. Perform a **kernel two-sample test** on $X_{1:n}, \tilde{Y}_{1:n}$.



Fixed sample splitting —> **Inflated** Type I control (**Invalid** testing procedure)

05 What's wrong & How to fix it?

Asymptotic Validity under Adaptive Splitting

Under $H_{0,\in}$ and regularity assumptions, when n is large,

$$\left| n_t \widehat{\mathrm{MMD}}^2(P_X, \hat{\eta}^\top \mathbf{P}_Y) - n_t \widehat{\mathrm{MMD}}^2(P_X, \eta_0^\top \mathbf{P}_Y) \right| = O\left(\sqrt{\frac{n_t}{n_e}}\right)$$

Therefore, if ρ is chosen such that $\frac{n_t}{n_2} \to 0$, then

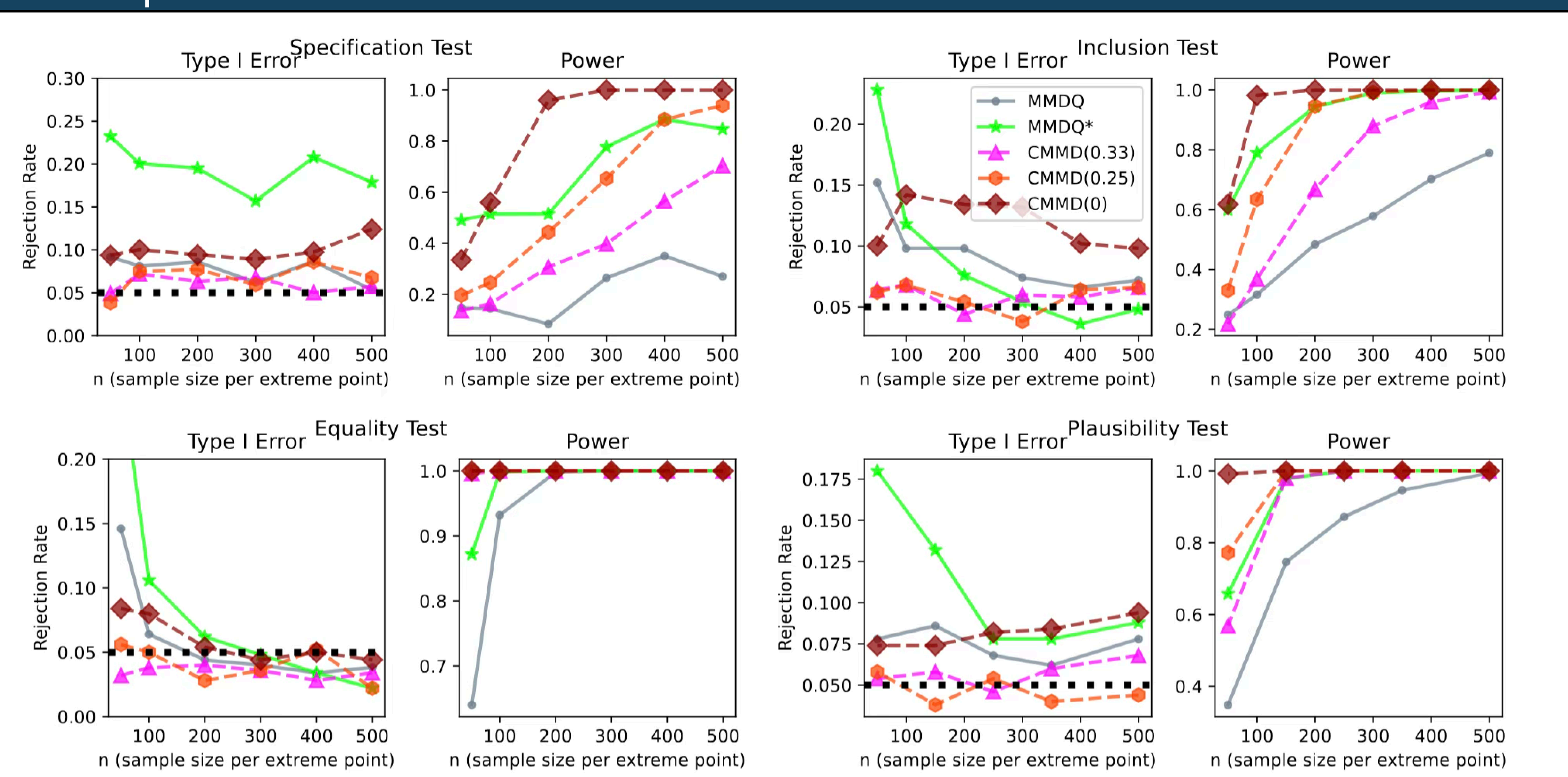
$$n_t \widehat{\mathrm{MMD}}^2(P_X, \hat{\eta}^\top \mathbf{P}_Y) \overset{D}{\to} \sum_{i=1}^{\infty} \zeta_i Z_i^2$$

that is, the same limiting distribution as if no estimation has happened. Furthermore, under $H_{A,\in}$,

$$n_t \widehat{\mathrm{MMD}}^2(P_X, \hat{\eta}^\top \mathbf{P}_Y) \to \infty$$

- **Intuition:** As n grows, estimation error decreases, but tests get more powerful.
- **Solution:** Set the "right" balance between estimation accuracy and test power with adaptive splitting, e.g. $n_t = \sqrt{n_e}$, $n_t = n_e^{1/3}$

06 Experiments with MNIST data



Adaptive splitting strategy yields valid test procedures

Check out more!

Ask me

- What is imprecise probabilistic machine learning?
- How to conduct the rest of the tests?
- How is this different to Bayesian test?
- Where can I read about your work:

RATIONAL INTELLIGENCE

CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

UCL

THE UNIVERSITY of ADELAIDE