(Some) new statistical concepts that could be useful for the sciences

Aaditya Ramdas

Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

(Always looking for postdocs and funding)

Multi-part "overview" talk

1. ***Universal inference (hypothesis testing)***
2. Online control of false discoveries (multiple hypothesis testing)
3. E-values as an alternative to p-values
4. Confidence sequences (peeking at your data as you collect it)
5. Masking (interacting with data after collecting it)
6. Conformal prediction (black-box uncertainty quantification in ML)

For regular models, LRT is easy

Consider $Y_1, \ldots, Y_n \sim p_{\theta^*}$ for some $\theta^* \in \Theta$.

Test $H_0 : \theta^* \in \Theta_0$ vs. $H_1 : \theta^* \in \Theta_1 \supset \Theta_0$

Wilk's Thm (regular models): $2 \log \frac{\mathscr{L}(\hat{\theta}_1)}{\mathscr{L}(\hat{\theta}_0)} \to \chi^2_d$ under H_0.

$\mathscr{L}(\theta) := \prod_{i=1}^{n} p_\theta(Y_i)$ is the likelihood function. $\hat{\theta}_{0/1}$ is MLE under $\Theta_{0/1}$.

d is difference in dimensionality between Θ_0, Θ_1.

LRT rejects if $2 \log \frac{\mathscr{L}(\hat{\theta}_1)}{\mathscr{L}(\hat{\theta}_0)} \geq c_{\alpha,d}$.

$(1-\alpha)$ quantile of χ^2_d

Under regularity conditions, $\Pr_{H_0}(\text{rejection}) \leq \alpha + o_p(1)$.

Irregular composite testing problems are common

1. (Mixtures) $H_0 : p_{\theta^*}$ is a mixture of k Gaussians
2. (Nonparametric shape constraints) $H_0 : p$ is log-concave
3. (Dependence) $H_0 : p_{\theta^*}$ is Gaussian MTP_2 or Ising MTP_2
4. (Linear model) $H_0 : \theta^*$ is k-sparse
5. (Factor models or HMMs) $H_0 : \theta^*$ has k hidden factors/states
6. (Gaussian CI testing) $H_0 : X_1 \perp X_2, \;\; H_1 : X_1 \perp X_2 \mid X_3$

In all cases, LR limiting distribution and a level-α test are unknown.

In all these cases, we can (approximately) calculate MLE under null.

Our proposal: split LRT

(regular models) LRT rejects if $2\log\frac{\mathscr{L}(\hat{\theta}_1)}{\mathscr{L}(\hat{\theta}_0)} \geq c_{\alpha,d}$.

$(1-\alpha)$ quantile of χ^2_d

(any model) split data into two parts D_0, D_1.

split LRT rejects if $2\log\frac{\mathscr{L}_0(\hat{\theta}_1)}{\mathscr{L}_0(\hat{\theta}_0)} \geq 2\log(1/\alpha)$.

$\mathscr{L}_0(\theta) := \prod_{i\in D_0} p_\theta(Y_i)$ is the likelihood on D_0.

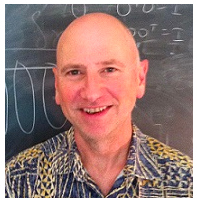
$\hat{\theta}_1$ is any estimator (MLE/Bayes/robust) under Θ_1 on D_1.

$\hat{\theta}_0$ is MLE under Θ_0 on D_0.

Under no regularity conditions $\Pr_{H_0}(\text{rejection}) \leq \alpha$.

Extensions in the paper

1. Profile split-likelihood
2. Robust 'powered' split-likelihood
3. Smoothed split-likelihood
4. Conditional split-likelihood
5. Relaxed maximum likelihood
6. Testing to model selection using sieves
7. Derandomization via averaging
8. Universal confidence sets
9. Extension to sequential settings




Universal inference

PNAS, 2020

Multi-part "overview" talk

1. Universal inference (hypothesis testing)

2. ***Online control of false discoveries (multiple hypothesis testing)***

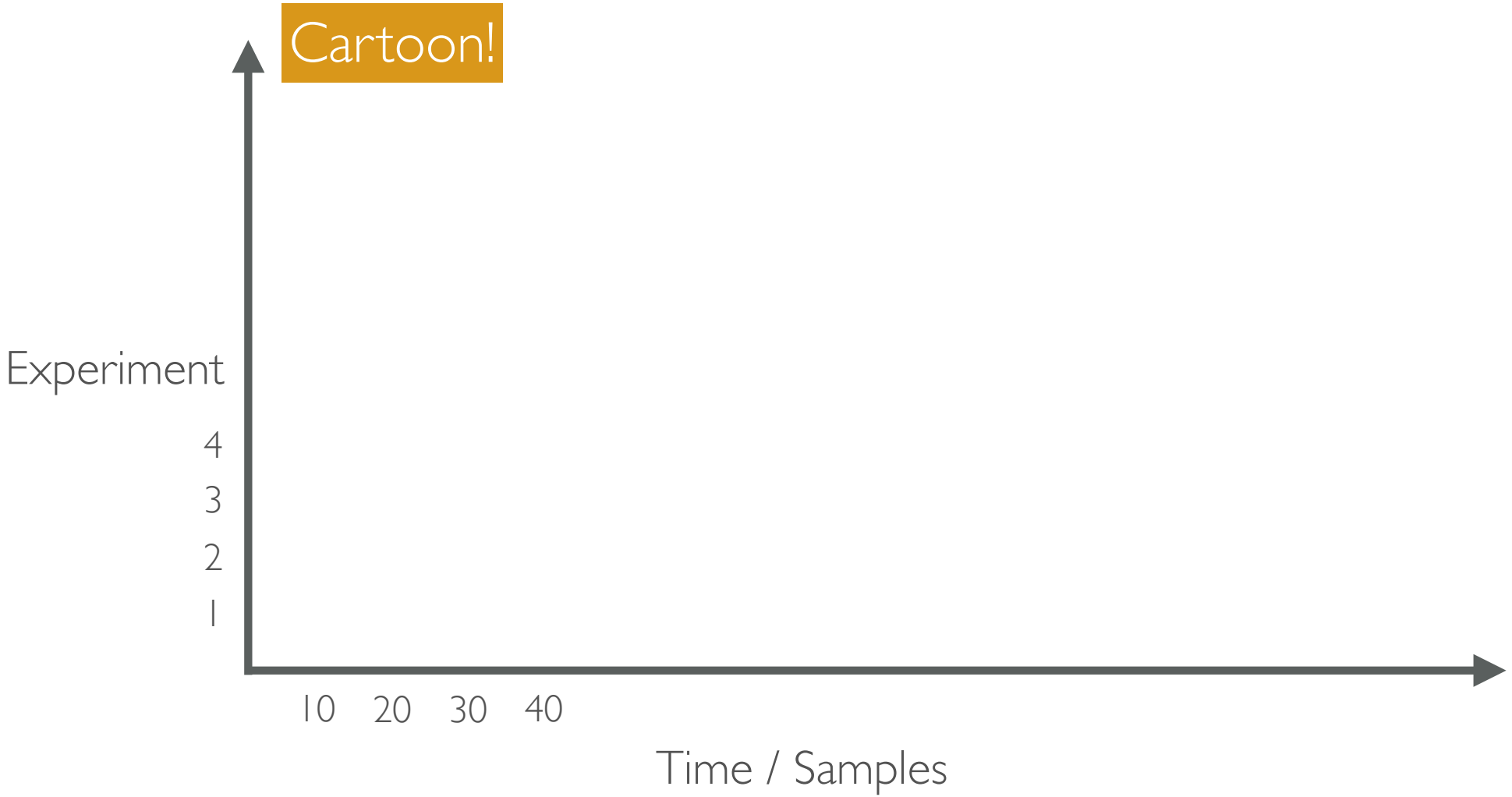
3. E-values as an alternative to p-values

4. Confidence sequences (peeking at your data as you collect it)

5. Masking (interacting with data after collecting it)

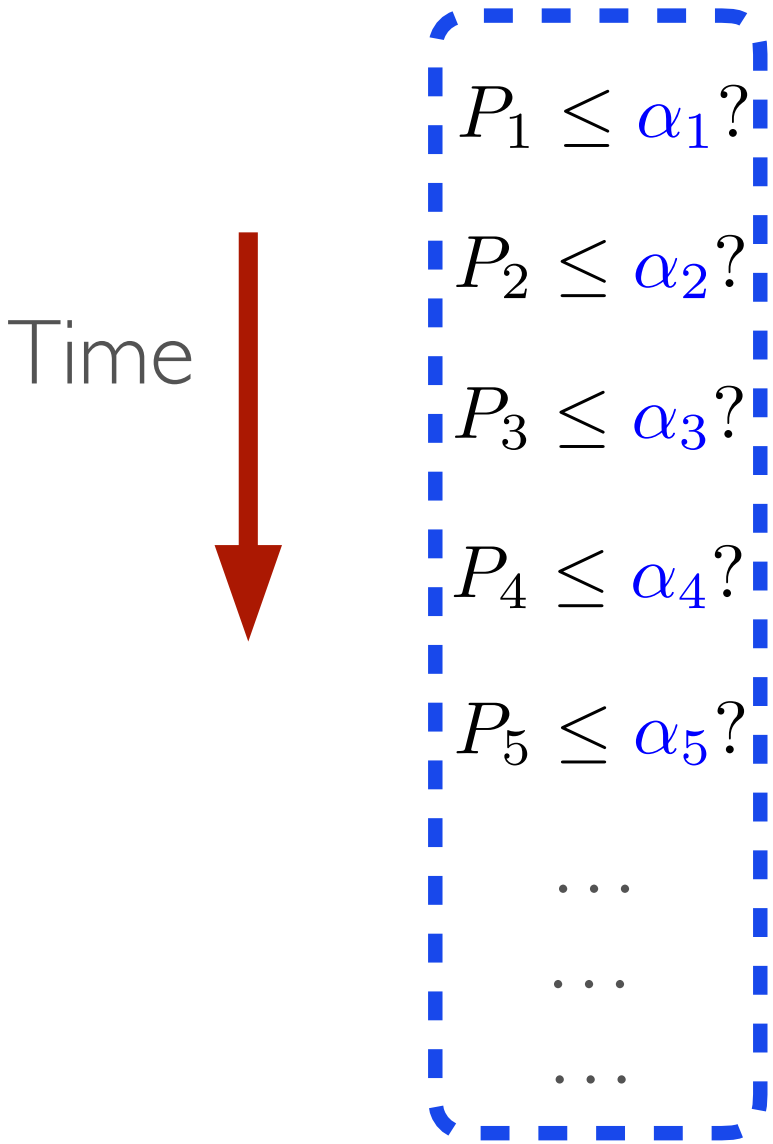
6. Conformal prediction (black-box uncertainty quantification in ML)

A sequence of sequential experiments (science?)



No temporal or longitudinal effects in this talk

Given a possibly infinite sequence of tests (p-values), can we control the FDR in a fully online fashion?



$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{false discoveries}}{\#\text{proclaimed discoveries}}\right]$$

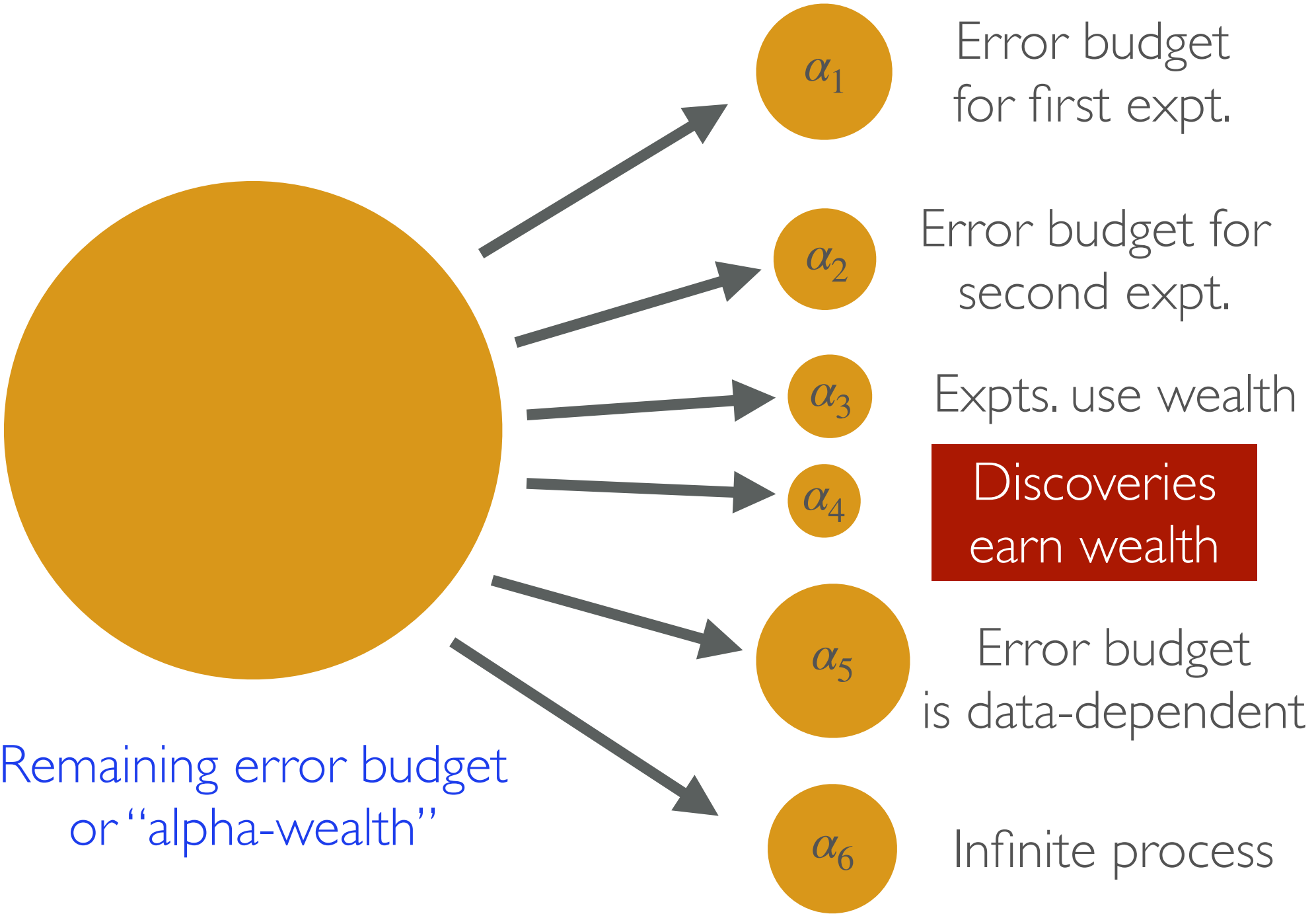
How do we set each error level to control FDR at any time?

Goal: $\forall t \in \mathbb{N}, \text{FDR}(t) \leq \alpha$.

(decisions are irrevocable)

Offline FDR methods do not yield this guarantee.

Online FDR control : high-level picture



Extensions in the paper(s)

1. Familywise error rate
2. False coverage rate
3. False sign rate
4. Post-hoc bounds
5. Powerful, adaptive algorithms

R package called "**onlineFDR**"
Collaboration led by David Robertson (Cambridge)

Some questions

(A) *In any scientific subfield, do the relevance and importance of claimed or published "discoveries" depend on the order in which hypotheses are tested by various scientists over time? (and should they?)*

(B) *For large publicly-shared scientific datasets (genetics/neuroscience/heart/...), is the statistical validity of proclaimed discoveries affected by how many people have downloaded the dataset and tested hypotheses on it? Can the dataset get "stale"? What happens when some downloads lead nowhere, while others lead to interesting findings and are published?*

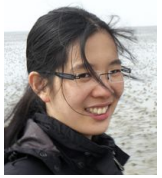
Some (more) questions

(C) *We often correct for multiple testing at a particular "scale" (eg: within a single paper, or within one simulation within a paper), but should we care about other granularities? (lab-level? field level? journal level?) If yes, then how would one do this? If not, why not?*

(D) *Instead of testing 100 hypotheses today, and thus being forced to correct for multiple testing, what if we randomly ordered the hypotheses, and tested one today, and one tomorrow, and one the day after, and so on for 100 days---since each day we only test one hypothesis, are we allowed to do so without any multiplicity correction?*

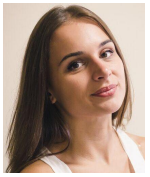
Online control of the false discovery rate with decaying memory

NeurIPS 2017



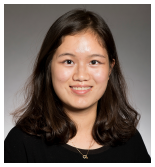
SAFFRON: an adaptive algorithm for online FDR control

ICML 2018



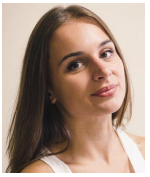
ADDIS: online FDR control with conservative nulls

NeurIPS 2019



The power of batching in multiple hypothesis testing

AISTATS 2020



Online control of the false coverage rate and false sign rate

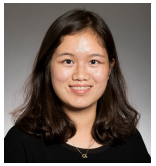
ICML 2020

Simultaneous high probability bounds on the FDP

Annals of Statistics, 2020

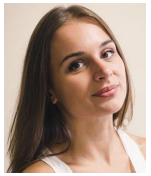
Online control of the familywise error rate

Statistical Methods in Medical Research, 2021



Asynchronous testing of multiple hypotheses

Journal of Machine Learning Research, 2021



Multi-part "overview" talk

1. Universal inference (hypothesis testing)
2. Online control of false discoveries (multiple hypothesis testing)
3. ***E-values as an alternative to p-values***
4. Confidence sequences (peeking at your data as you collect it)
5. Masking (interacting with data after collecting it)
6. Conformal prediction (black-box uncertainty quantification in ML)

P-values

A p-value is a random variable P such that, under the null, $\Pr(P \leq t) \leq t$ for any $t \in [0,1]$.

Small p-values are evidence against the null: Reject if $P \leq \alpha$.

Suppose $P_{\text{Jan-Jun}} > 0.05$ and $P_{\text{Jul-Dec}} < 0.05$, where $\Pr(P_{\text{Jul-Dec}} \leq t \,|\, D_{\text{Jan-Jun}}) \leq t$.

How do you combine them?

E-values

An e-value is a random variable E such that $E \geq 0$, and under the null, $\mathbb{E}[E] \leq 1$.

Large e-values are evidence against the null: Reject if $E \geq 1/\alpha$.

Suppose $E_{\text{Jan-Jun}} < 20$ and $E_{\text{Jul-Dec}} > 20$, where $\mathbb{E}[E_{\text{Jul-Dec}} \,|\, E_{\text{Jan-Jun}}] \leq 1$.

$E_{tot} = E_{\text{Jan-Jun}} E_{\text{Jul-Dec}}$.

P-values	E-values
Need full distributional information of test statistic to calculate p-value.	Need moment bounds of test statistic to calculate e-value.
Often cannot calculate in "irregular" testing problems	The split likelihood ratio is an e-value
Often resort to asymptotics	Typically valid in finite samples
Often need stronger dependence assumptions on data	Can be constructed for weakly dependent data
Need corrections for dependence in multiple testing	No corrections for dependence in multiple testing

Multiple testing under arbitrary dependence

The e-Benjamini-Hochberg procedure:

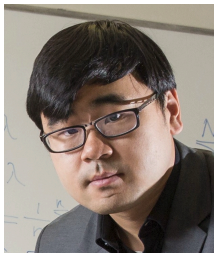
Given $E_1, \ldots, E_K$ for K hypotheses, define

$$k^* := \max \left\{ k : E_{[k]} \geq \frac{K}{k\alpha} \right\}.$$

Reject the k^* hypotheses with largest e-values.

Theorem: The e-BH procedure controls the FDR at level α under arbitrary dependence between the e-values.

False discovery rate control with e-values



Wang, Ramdas (arXiv:2009.02824)

Multi-part "overview" talk

1. Universal inference (hypothesis testing)
2. Online control of false discoveries (multiple hypothesis testing)
3. E-values as an alternative to p-values
4. ***Confidence sequences (peeking at your data as you collect it)***
5. Masking (interacting with data after collecting it)
6. Conformal prediction (black-box uncertainty quantification in ML)

A "confidence sequence" for a parameter θ is a sequence of confidence intervals (L_n, U_n) with a uniform (simultaneous) coverage guarantee.

$$\mathbb{P}(\forall n \geq 1 : \theta \in (L_n, U_n)) \geq 1 - \alpha\,.$$

Sample size

Darling, Robbins '67
Lai '84

Much stronger than the pointwise (fixed-sample) confidence intervals guarantee:

$$\forall n \geq 1, \mathbb{P}(\theta \in (\tilde{L}_n, \tilde{U}_n)) \geq 1 - \alpha\,.$$

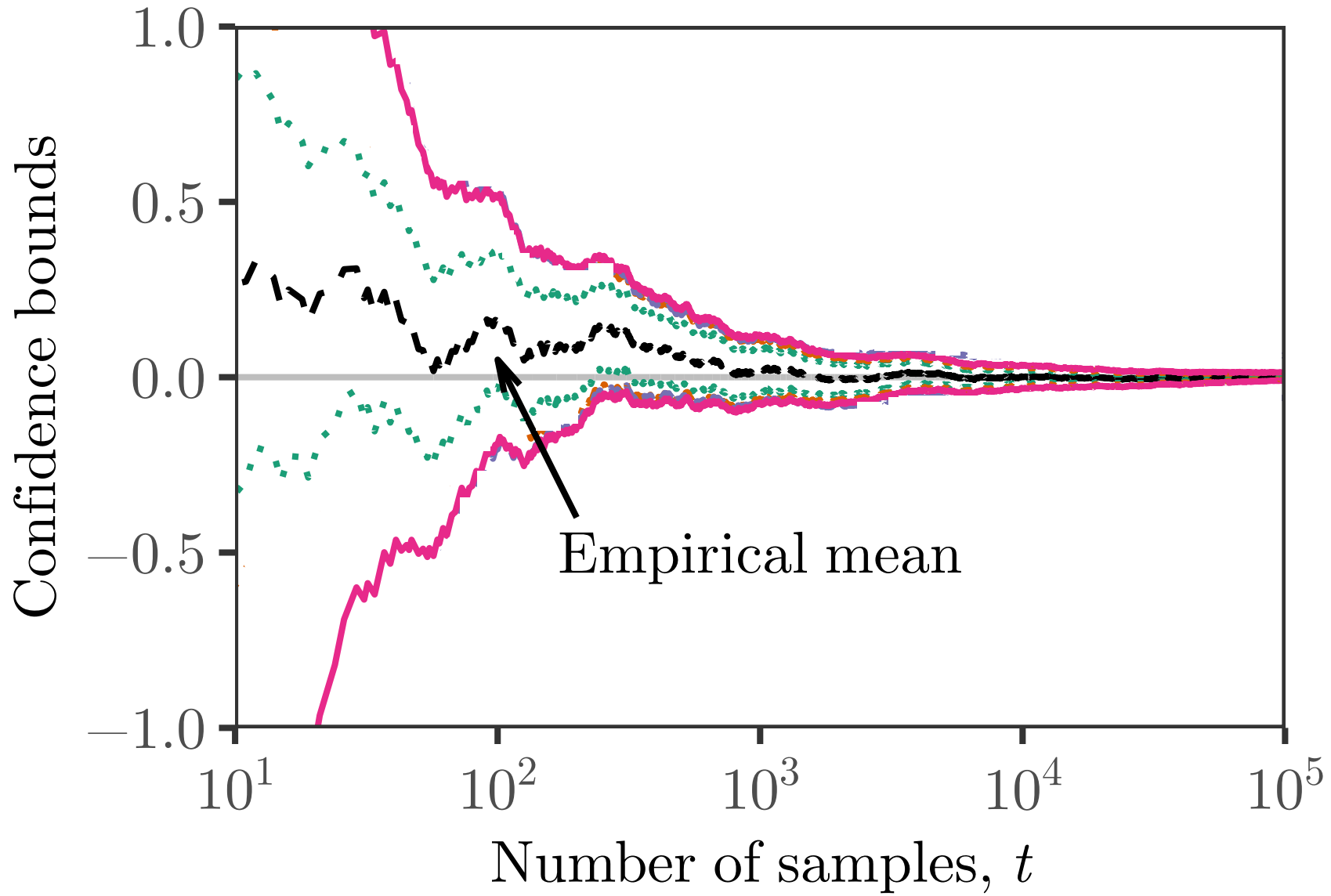
Example: tracking the mean of a Gaussian or Bernoulli from i.i.d. observations.

$$X_1, X_2, \ldots \sim N(\theta, 1) \text{ or } Ber(\theta)$$

Producing a confidence *interval* at a fixed time is elementary statistics (~100 years old).

How do we produce a confidence sequence? (which is like a confidence band over time)

(Fair coin)



Pointwise CI (CLT)

Anytime CI

Eg: If X_i is Gaussian, or bounded in $[-1,1]$, then

$$\frac{\sum_{i=1}^{n} X_i}{n} \pm 1.71\sqrt{\frac{\log\log(2n) + 0.72\log(5.19/\alpha)}{n}}$$

is a $(1-\alpha)$ confidence sequence for its mean μ.

$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} \{\theta \notin (L_n, U_n)\}\Big) \leq \alpha\,.$$

Some implications:

1. Valid inference at any time, even stopping times:
   For any stopping time τ : $\mathbb{P}(\theta \notin (L_\tau, U_\tau)) \leq \alpha\,.$

2. Valid post-hoc inference (in hindsight):
   For any random time T : $\mathbb{P}(\theta \notin (L_T, U_T)) \leq \alpha\,.$

3. No pre-specified sample size:
   can extend or stop experiments adaptively.

Learn more about supermartingales, CSs

Time-uniform Chernoff bounds via nonneg. supermg.

Probability Surveys, 2020



Annals of Statistics, 2021

Time-uniform, nonparametric, nonasymptotic CSs

Quantile CSs for A/B testing and bandits

(major revision, Bernoulli)

CSs for sampling w/o replacement

NeurIPS, 2020

Uncertainty quantification using martingales for misspecified Gaussian processes

ALT, 2021

Sequential nonparametric testing with the law of the iterated logarithm

UAI, 2016

Multi-part "overview" talk

1. Universal inference (hypothesis testing)
2. Online control of false discoveries (multiple hypothesis testing)
3. E-values as an alternative to p-values
4. Confidence sequences (peeking at your data as you collect it)
5. ***Masking (interacting with data after collecting it)***
6. Conformal prediction (black-box uncertainty quantification in ML)

If you blur the data (appropriately) and then peek all you want, you can be protected from overfitting and selection bias. We call this masking.

The user can progressively unmask the data,
one point at a time, thus gaining power,
as long as the appropriate martingale test is used.

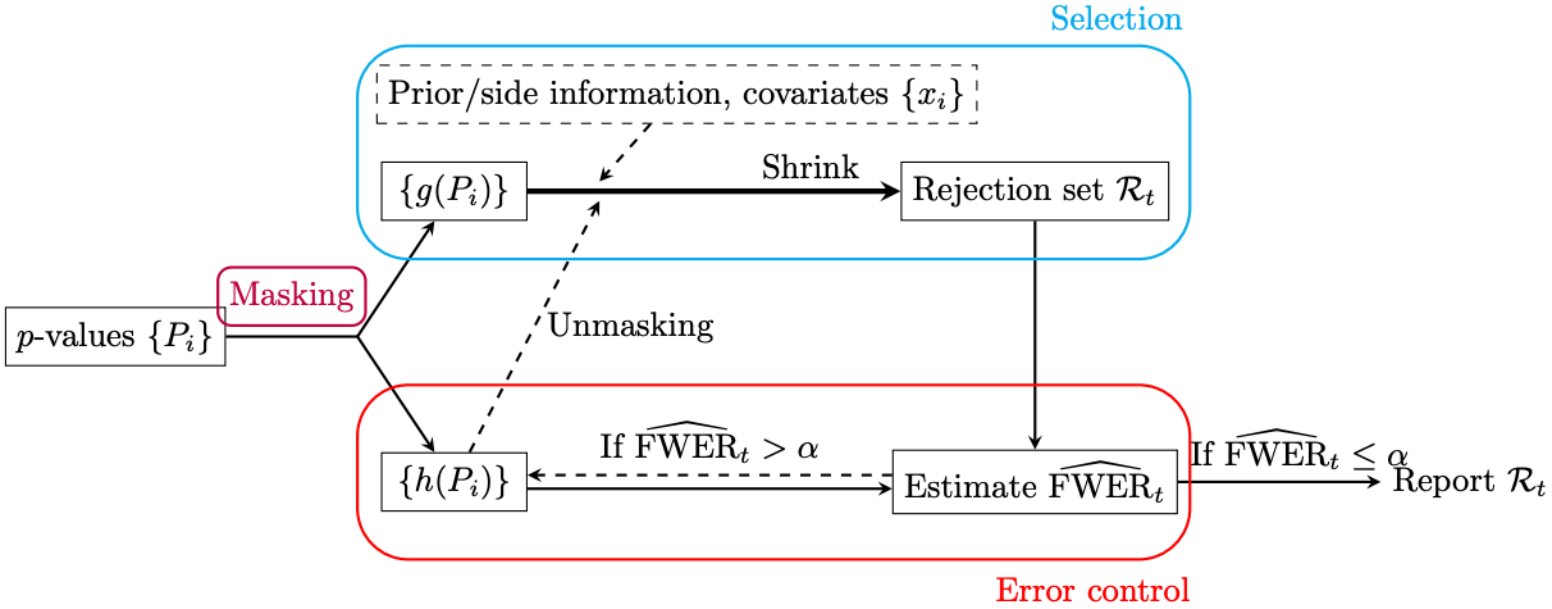
One example : "masking the p-values" enables exploration while avoiding selection bias

Data "carving"

$$P_i$$

"masked p-value" $g(P_i) = \min(P_i, 1 - P_i)$ used for "selection"

"missing bit" $h(P_i) = 2I(P_i > 1/2)$ used for "inference"

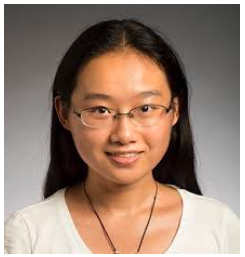
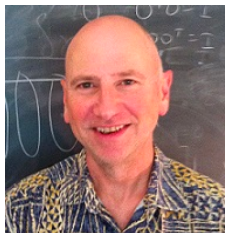


A general interactive framework for FDR control under structural constraints

Biometrika, 2020

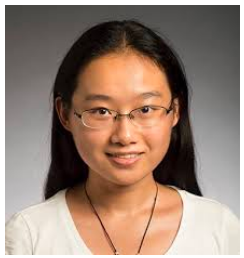
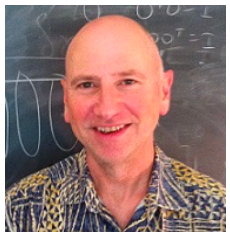

Interactive martingale tests for the global null

Electronic Journal of Statistics, 2020


Familywise error rate control by interactive unmasking

ICML, 2020

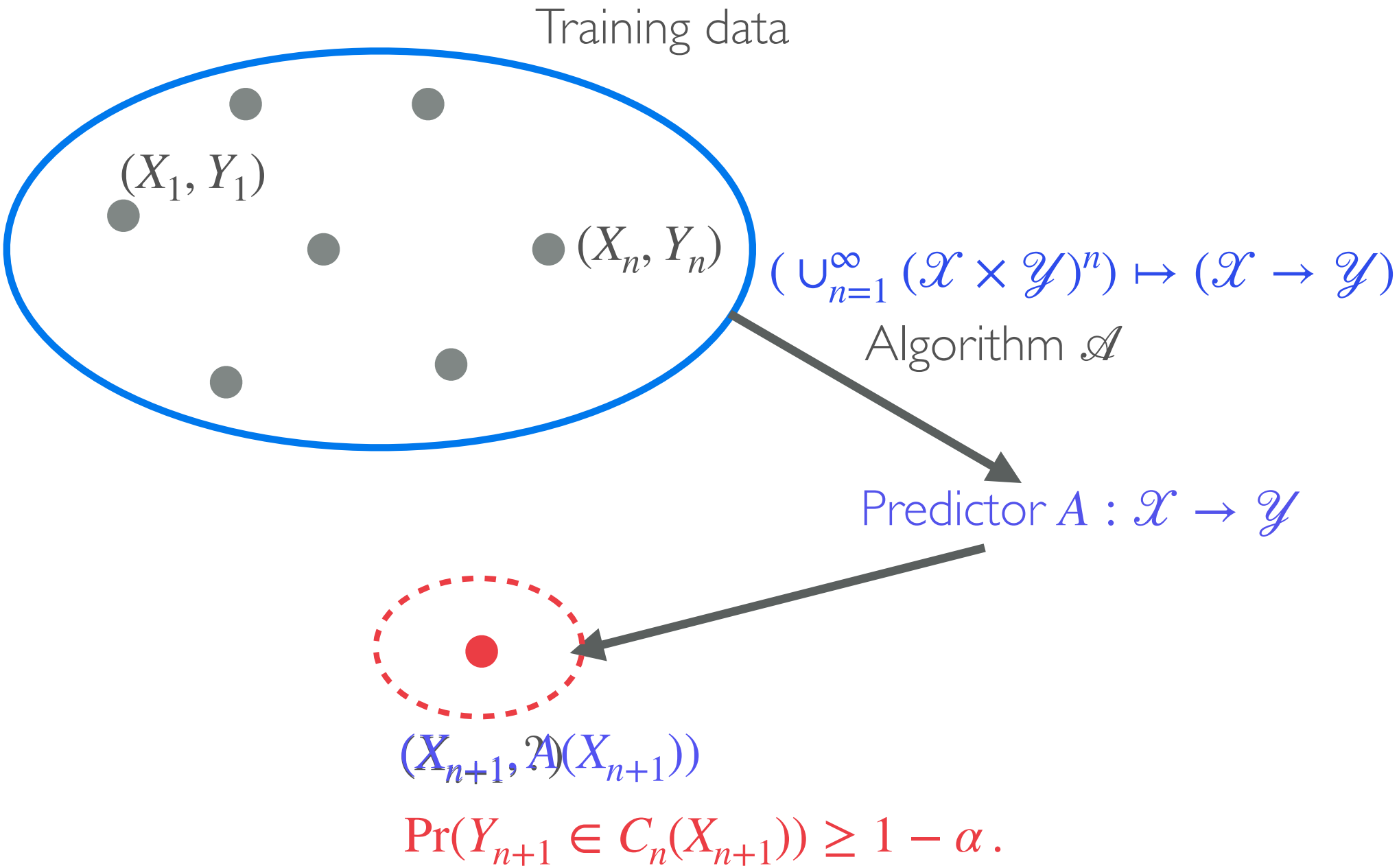
Which Wilcoxon should we use? An interactive rank test and other alternatives

(Biometrical Journal, minor revision)

Multi-part "overview" talk

1. Universal inference (hypothesis testing)
2. Online control of false discoveries (multiple hypothesis testing)
3. E-values as an alternative to p-values
4. Confidence sequences (peeking at your data as you collect it)
5. Masking (interacting with data after collecting it)
6. ***Conformal prediction (black-box uncertainty quantification in ML***

Prediction vs "Predictive Inference"



Distribution-free Predictive Inference

Given data $D_n \equiv (X_1, Y_1), \ldots, (X_n, Y_n) \sim P_X \times P_{Y|X} \equiv P_{XY}$,

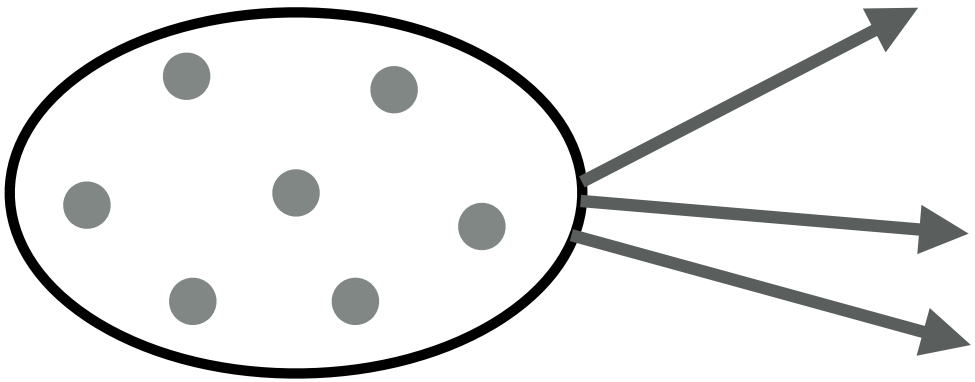
any algorithm $\mathcal{A} : (\cup_{n=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^n) \mapsto (\mathcal{X} \to \mathcal{Y})$,

and $X_{n+1} \sim P_X$, produce a set $C(X_{n+1}) \equiv C_{\mathcal{A},D_n}(X_{n+1})$ s.t.

for all P_{XY}, algorithms $\mathcal{A}$, $\Pr(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$.

Vovk, Shafer, and others, 2000-19

The jackknife+



Predictor $A_{-1} = \mathscr{A}(D \setminus (X_1, Y_1))$

$\vdots$

Predictor $A_{-i} = \mathscr{A}(D \setminus (X_i, Y_i))$

Predictor $A_{-n} = \mathscr{A}(D \setminus (X_n, Y_n))$

LOO scores $R_i \equiv |Y_i - A_{-i}(X_i)|$. Let $\bar{R} = \{R_i\}_{i \in [n]}$.

$q_{1-\alpha}(\bar{R}) :=$ $(1-\alpha)$ quantile of $\{R_1, \ldots, R_n\}$.

Jackknife+: $C_{J+}(X_{n+1}) \equiv [q_\alpha(\{A_{-i}(X_{n+1}) - R_i\}), q_{1-\alpha}(\{A_{-i}(X_{n+1}) + R_i\})]$.

Then, $\Pr(Y_{n+1} \in C_{J+}(X_{n+1})) \geq 1 - 2\alpha$.

BaCaRaTi'20

Extensions in the paper(s)

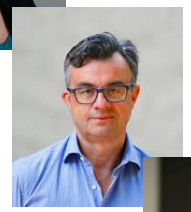
1. Handling covariate and label shift
2. Ensemble quantile methods
3. Conformal classification
4. Distribution-free classifier calibration
5. Random effect models
6. Conditional predictive inference

Predictive inference with the jackknife+

Annals of Statistics, 2021

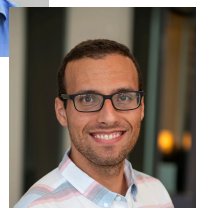
The limits of distribution-free conditional predictive inference

Information and Inference, 2020



Conformal prediction under covariate shift

NeurIPS, 2019



Distribution-free binary classification: prediction sets, confidence intervals and calibration

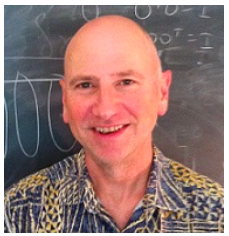
NeurIPS, 2020

Nested conformal prediction and quantile out-of-bag ensemble methods

(Pattern Recognition, minor revision)

Distribution-free prediction sets with random effects

(JASA, minor revision)



Multi-part "overview" talk

1. Universal inference (hypothesis testing)
2. Online control of false discoveries (multiple hypothesis testing)
3. E-values as an alternative to p-values
4. Confidence sequences (peeking at your data as you collect it)
5. Masking (interacting with data after collecting it)
6. Conformal prediction (black-box uncertainty quantification in ML)

(Some) new statistical concepts that could be useful for the sciences

Aaditya Ramdas

Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

(Always looking for postdocs and funding)