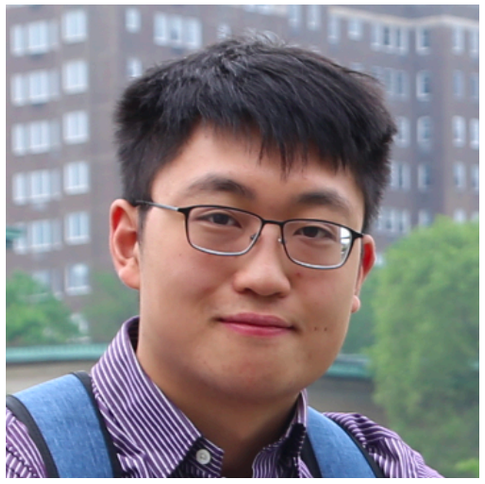
Online multiple testing with e-values

https://arxiv.org/pdf/2311.06412

E-readers
June 14th, 2024



Ziyu Xu (Neil)
(CMU)

Joint work w/
Aaditya Ramdas (CMU)

Carnegie Mellon University

Outline

- Online multiple testing
- Doubly sequential inference
- Prior work
- Our method: e-LOND (and proof of FDR control)
- Empirical results
- Stochastic rounding (randomization) for more power
- Weighted conformal selection

Online multiple testing problem

Stream of hypotheses we wish to test:

H_1 H_2 H_3 H_4 …

$\mathcal{N} \subseteq \mathbb{N}$ is the set of hypotheses are truly null (e.g., treatment has no effect)

Data for each hypothesis (e.g., from an experiment).

X_1 X_2 X_3 X_4 …

P_1 P_2 P_3 P_4 …

P-values: statistics that are superuniform under the null.
$\mathbb{P}(P_t \leq s) \leq s$ for all $s \in [0,1]$ if $H_t \in \mathcal{N}$

α_1 α_2 α_3 α_4 …

Make a discovery if $P_t \leq \alpha_t$, i.e., on hypotheses we believe are non-null

R_1 R_2 R_3 R_4 …

Discovery sets
$R_t := \{i \in [t] : P_i \leq \alpha_i\}$

False discovery proportion (FDP)

$$\text{FDP}(R_t) := \frac{|\mathcal{N} \cap R_t|}{|R_t| \vee 1}$$

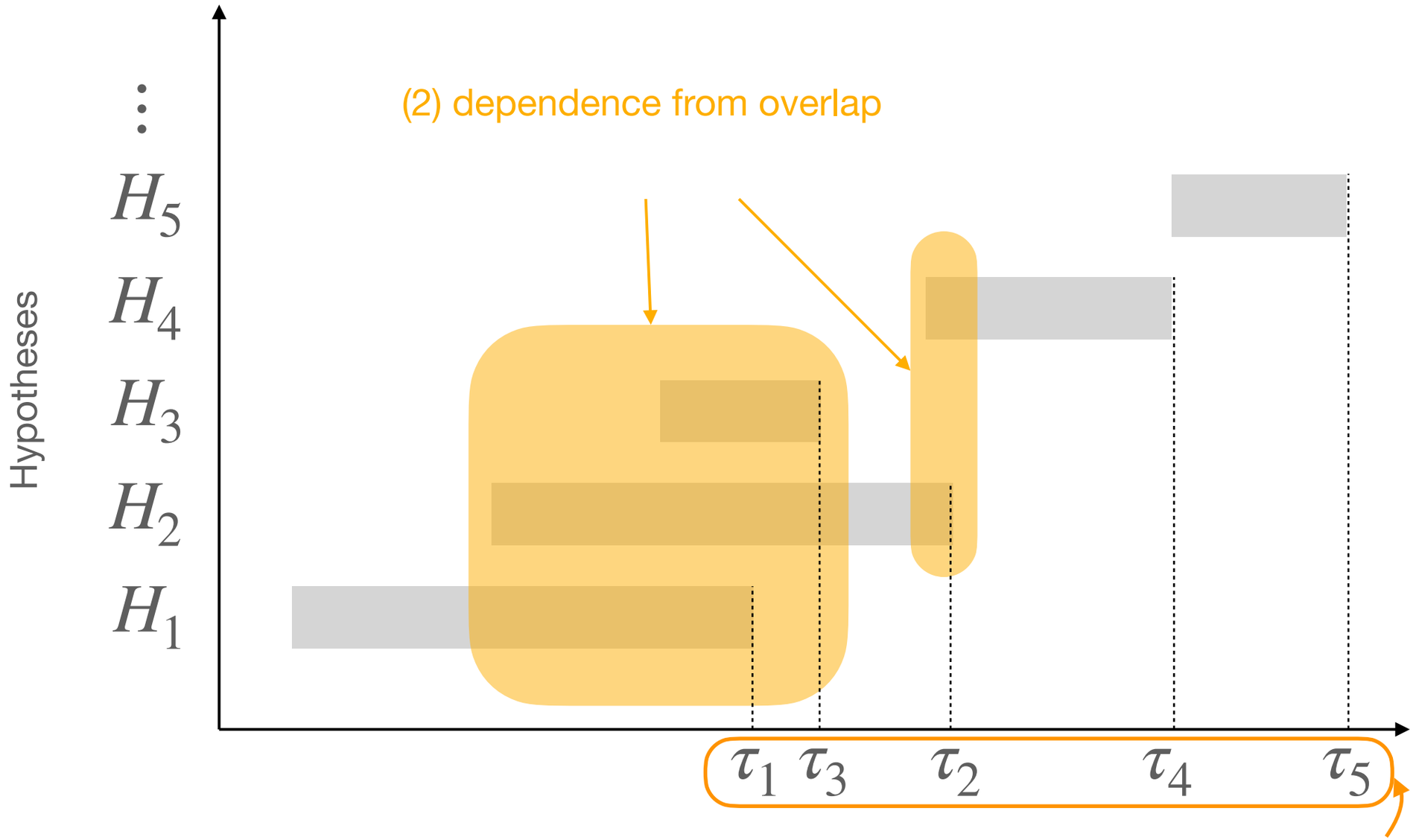
False discovery rate (FDR)

$\text{FDR}(R_t) := \mathbb{E}[\text{FDP}(R_t)]$

Goal: Ensure $\text{FDR}(R_t) \leq \alpha$ for all $t \in \mathbb{N}$ for a fixed $\alpha \in [0,1]$

Doubly sequential inference

Imagine we have teams of data scientists doing A/B testing…



Desiderata: a method that is robust to
(1) data-dependent stopping times
(2) dependence among hypotheses

E-values are the predominant method for solving (1), and our new method addresses (2)

Sequential testing requires e-processes

Consider a *stream of data* $X_1, X_2, \ldots,$
and we construct a p-value P_i at each step i from $X_1, \ldots, X_i$

We wish to have an *anytime-valid p-value:*
$\mathbb{P}(\exists i \in \mathbb{N} : P_i \leq s) \leq s$ for all $s \in [0,1]$

a fixed time (standard) p-value:
$\mathbb{P}(P_i \leq s) \leq s$ for all $s \in [0,1]$ and $i \in \mathbb{N}$

This means that rejecting the null at step $\tau = \min\{i : P_i \leq \alpha\}$, i.e., *early stopping*, is a valid test.

Anytime-valid p-values are frequently used in A/B testing [1] and traditionally has used likelihood ratios (e.g. mSPRT [2]).

An e-process is a nonnegative process (E_i) where E_τ is an e-value for any stopping time τ.
All known anytime-valid p-values can be written as $P_i = 1/\max_{j \leq i} E_i$
(or is dominated by such an anytime-valid p-value).

Practically, we always have access to e-processes in the sequential regime.

[1] Johari, Pekelis, Walsh. Always Valid Inference: Continuous Monitoring of A/B Tests. *Operations Research (2022)*
[2] Robbins. Statistical Methods Related to the Law of the Iterated Logarithm. *Ann. Math. Stat (1970)*
[3] Ramdas, Ruf, Larsson, Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv (2023)*

Overview of online multiple testing

LORD algorithm

Javanmard and Montanari (2018)
Ramdas, Yang, Wainwright, Jordan (2017)

$$\text{Ensures: } \widehat{\text{FDP}}_t := \frac{\sum_{i=1}^{t} \alpha_t}{|R_{t-1}| + 1} \leq \alpha$$

SAFFRON algorithm

Ramdas, Zrnic,
Wainwright, Jordan (2019)

$$\text{Ensures: } \widehat{\text{FDP}}_t := \frac{\sum_{i=1}^{t} \alpha_t \cdot \frac{\mathbf{1}\{P_t > \lambda_t\}}{1-\lambda_t}}{|R_{t-1}| + 1} \leq \alpha$$

Foster and Stine (2008)
Saharoni and Rosset (2014)

$$\text{mFDR}(R_T) := \frac{\mathbb{E}[|\mathcal{N} \cap R_T|]}{\mathbb{E}[|R_T|]}$$

mFDR control at stopping times (not just fixed times).

Tian and Ramdas (2019) — ADDIS = SAFFRON + discarding of conservative nulls

Weinstein and Ramdas (2020) — Online control FCR for selective confidence intervals

All of the above require independent or conditional (on the past) p-values (or confidence intervals).

Zrnic et. al (2020, 2021) — Extends LORD and SAFFRON for batches, under local dependence and arbitrary dependence

Bao et al. (2024) — Online FCR control for selective conformal prediction intervals (subsequent)

See Robertson et al. (2023) for a comprehensive survey

Prior work

r-LOND algorithm [3]

$$\alpha_t^{\text{r-LOND}} := \alpha\gamma_t \cdot (|R_{t-1}| + 1)/\ell_t$$

where (γ_t) are nonnegative and sum to 1

$$\ell_t := \Sigma_{i=1}^{t} 1/i \approx \log t$$

Theorem [3]: FDR$(R_t) \leq \alpha$ for all $t \in \mathbb{N}$ for *arbitrarily dependent p-values* (P_t) is guaranteed by *r-LOND*.

LOND algorithm [4]

$$\alpha_t^{\text{LOND}} := \alpha\gamma_t(|R_{t-1}| + 1)$$

Theorem : FDR$(R_t) \leq \alpha$ for all $t \in \mathbb{N}$ for *independent* [4] *or positively dependent PRDS p-values* [3] (P_t) is guaranteed by *LOND*.

Our contribution: With e-values, we can use the more the powerful LOND, even under arbitrary dependence.

[1] Blanchard and Roquain. Two simple sufficient conditions for FDR control. *Elec. J. Stat. (2008)*
[2] Benjamini and Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. of Stat.* (2001)
[3] Zrnic, Ramdas, Jordan. Asynchronous online testing of multiple hypotheses. *JMLR (2021)*
[4] Javanmard and Montanari. Online rules for control of false discovery rate and false discovery exceedance. *Ann. of Stat. (2018)*

e-LOND: FDR control under arbitrary dependence

e-LOND algorithm

$$\alpha_t^{\text{e-LOND}} := \alpha\gamma_t \cdot (|R_{t-1}| + 1)$$

Theorem (ours): $\text{FDR}(R_t) \leq \alpha$ for all $t \in \mathbb{N}$ for *arbitrarily dependent e-values* (E_t) is guaranteed by *e-LOND*, and $\alpha_t^{\text{e-LOND}} > \alpha_t^{\text{r-LOND}}$ for all $t \in \mathbb{N}$, i.e., strictly dominates r-LOND.

Proof

$$\text{FDR}(R_t) = \mathbb{E}\left[\frac{|\mathcal{N} \cap R_t|}{|R_t| \vee 1}\right] = \sum_{i \in \mathcal{N}} \mathbb{E}\left[\frac{\mathbf{1}\{E_i \geq 1/(\alpha_i^{\text{e-LOND}})\}}{|R_t| \vee 1}\right]$$

$$\leq \sum_{i \in \mathcal{N}} \mathbb{E}\left[\frac{\mathbf{1}\{E_i \geq 1/(\alpha_i^{\text{e-LOND}})\}}{|R_{i-1}| + 1}\right] \leq \sum_{i \in \mathcal{N}} \mathbb{E}\left[\frac{\alpha_i^{\text{e-LOND}} \cdot E_i}{|R_{i-1}| + 1}\right]$$

$$\leq \sum_{i \in \mathcal{N}} \mathbb{E}\left[\frac{\alpha\gamma_i \cdot (|R_{i-1}| + 1) \cdot E_i}{|R_{i-1}| + 1}\right] \leq \sum_{i \in \mathcal{N}} \alpha\gamma_i \mathbb{E}[E_i] \leq \alpha$$

More power through randomization

Let (U_t) be a sequence of random variables that are uniformly distributed on [0,1] and independent of (E_t) or (P_t)

Ue-LOND algorithm

$$\alpha_t^{\text{Ue-LOND}} := \alpha\gamma_t \cdot (|R_{t-1}| + 1)/U_t$$

Idea: Let $\hat{\alpha} \in [0,1]$ and $E > 0$ be dependent RVs.
Define the stochastically rounded e-value [1]

$$S_{\hat{\alpha}}(E) := \hat{\alpha}^{-1} \cdot \mathbf{1}\{U \leq \hat{\alpha}E\}$$

Has bounded expectation

$$\mathbb{E}[E] \geq \mathbb{E}[S_{\hat{\alpha}}(E)]$$

Discovery threshold property

$$\mathbf{1}\{S_{\hat{\alpha}}(E) \geq \hat{\alpha}^{-1}\} = \mathbf{1}\{E \geq U \cdot \hat{\alpha}^{-1}\}$$

Ur-LOND algorithm

$$\alpha_t^{\text{Ur-LOND}} := \alpha\gamma_t \cdot (\lfloor (|R_{t-1}| + 1)/(\ell_t \cdot U_t) \rfloor \wedge t)$$

Similar ideas can be applied to improve r-LOND
(by its equivalence to being calibrated p-value + e-LOND)

Theorem (ours):

1. FDR$(R_t) \leq \alpha$ for all $t \in \mathbb{N}$ for *arbitrarily dependent e-values* (E_t) is guaranteed by *Ue-LOND,* and $\alpha_t^{\text{Ue-LOND}} > \alpha_t^{\text{e-LOND}}$ for all $t \in \mathbb{N}$ almost surely.

2. FDR$(R_t) \leq \alpha$ for all $t \in \mathbb{N}$ for *arbitrarily dependent p-values* (P_t) is guaranteed by *Ur-LOND,* and has strictly greater power than *r-LOND.*

[1] **Xu** and Ramdas. More powerful multiple testing under dependence via randomization. *arXiv (2023)*

Empirical results

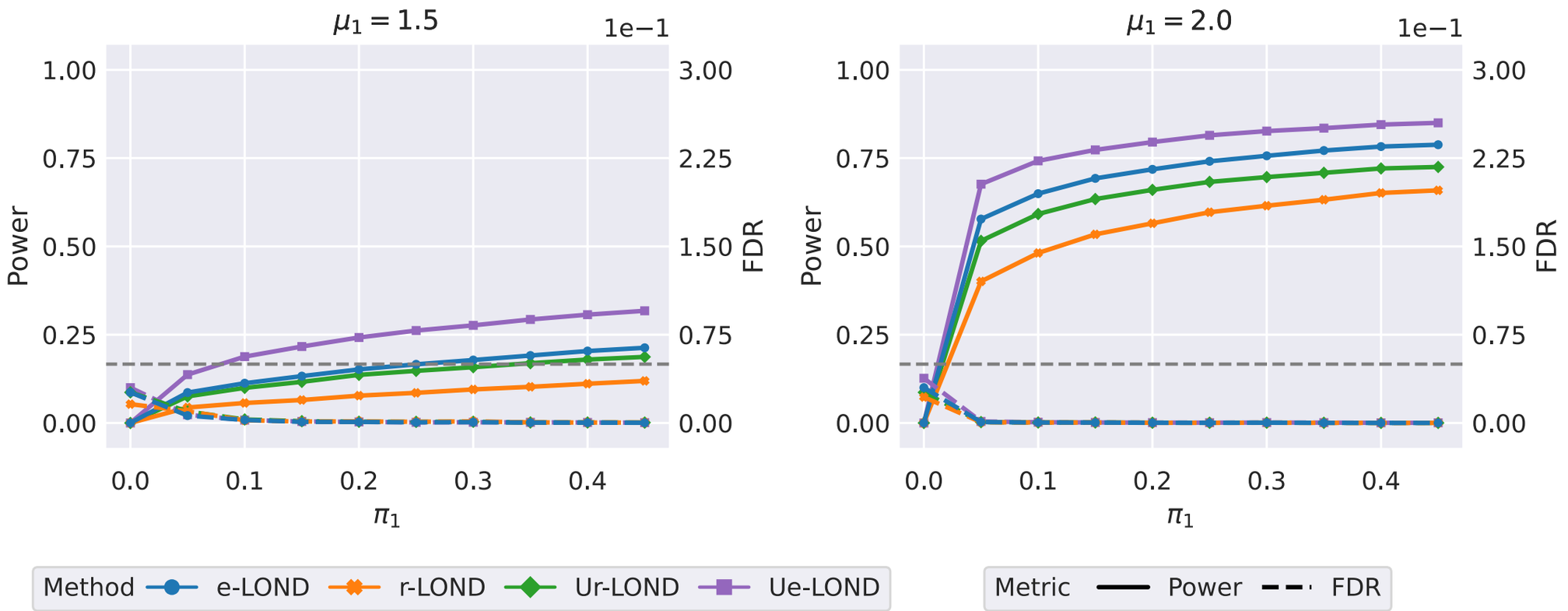
Simulation: testing $\mu = \mathbb{E}[X]$ where $X \in [\pm 4]$ is bounded.

$H_0 : \mu \leq 0$ vs. $H_1 : \mu > 0$

Data for each hypothesis was a sample without replacement from the same population.

$\Downarrow$

Negative dependence among (E_t)



(More simulation results in paper)

Selection by conformal prediction (Jin and Candès 2023)

Assumption: exchangeability

We have a calibration dataset of size n: $(X_1, Y_1), \ldots, (X_n, Y_n)$

We also receive a batch of m test points: $X_{n+1}, \ldots, X_{n+m}$

Assume that $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ are exchangeable (calibration dataset + stream are all exchangeable).

Assumption: monotonic score function

Let V be a scoring function that is monotonic in its second argument, i.e., $y \le c \Rightarrow V(x, y) \le V(x, c)$ for all x

Example: $V(x, y) := y - \hat{\mu}(x)$

$V_i := V(X_i, Y_i)$ for $i \in [n]$, $\hat{V}_{n+t} := V(X_{n+t}, c_{n+t})$ for $t \in \{1,2,\ldots\}$

For each $t \in \{1,2,\ldots\}$, our goal is to test the stochastic null hypothesis, i.e.,

$$H_{0,t} : Y_{n+t} \le c_{n+t}$$

where c_{n+t} is a known fixed or data dependent threshold.

Goal: Large rejection sets R with FDR control.

Conformalized selection (batch, no weighting)

1. Construct conformal p-values $P_t := \dfrac{\sum_{i=1}^{n} \mathbf{1}\{V_i \le \hat{V}_{n+t}\}}{n+1}$.
2. Apply the BH (Benjamini-Hochberg) procedure, i.e.,

$$\hat{k} := \max\left\{ k \in [m] : \sum_{i=1}^{m} \mathbf{1}\{P_t \le \alpha k/m\} \ge k \right\}, \text{ reject}$$

$\hat{k}$ smallest p-values

Proof idea:
Consider "oracle" p-values:

$$\bar{P}_t := \frac{\sum_{i=1}^{n} \mathbf{1}\{V_i \le V_{n+t}\}}{n+1}$$

$\{H_{0,t}$ is true $\} \Rightarrow \{\bar{P}_t \le P_t\}$ (by monotonicity)

$P_1, \ldots, \bar{P}_t, \ldots, P_m$ are positively dependent, i.e., PRDS (Bates et. al. 2021).

BH provides FDR control for PRDS p-values.

Weighted conformalized selection (Jin and Candès 2023)

Assumption: known covariate shift (weighted exchangeability)

The calibration dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. draws from $\mathscr{P}$

We also receive a batch of m test points: $X_{n+1}, \ldots, X_{n+m}$ where $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$ are i.i.d. draws from $\mathscr{Q}$

Covariate shift is known, i.e., $w(x) := \dfrac{d\mathscr{Q}}{d\mathscr{P}}(x, y)$

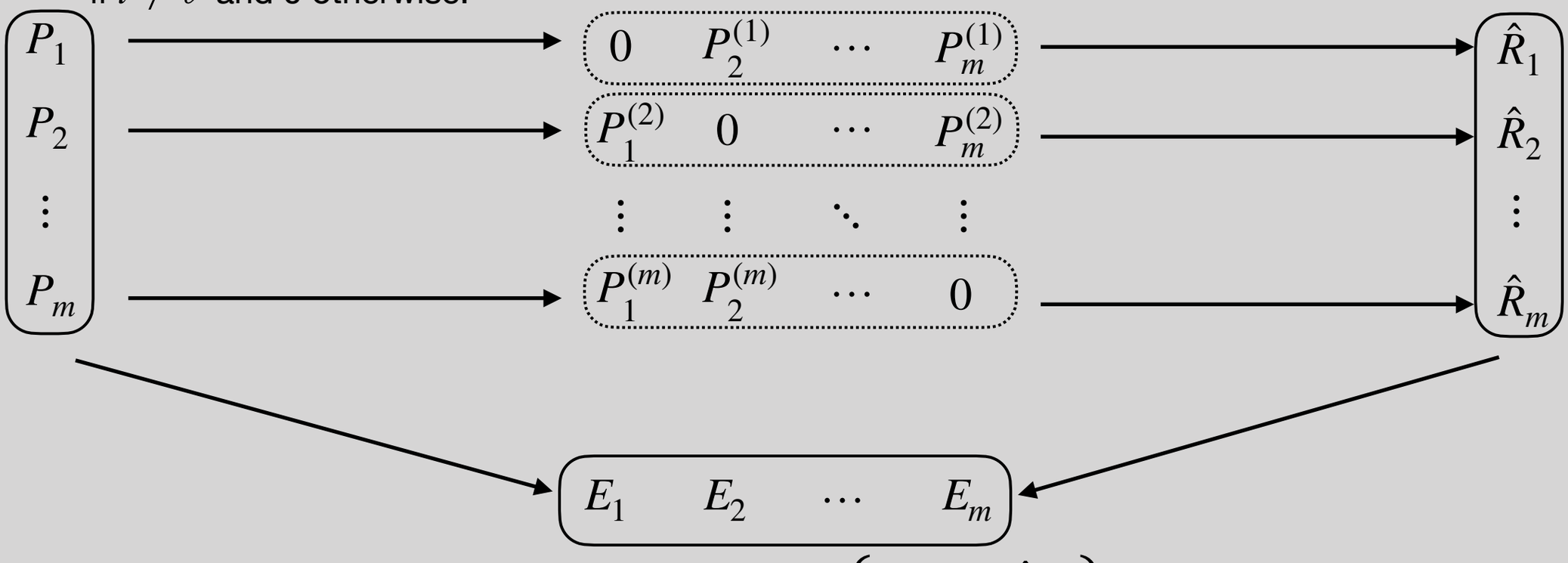
Weighted conformalized selection (WCS batch)

Construct <u>multiple</u> weighted conformal p-values for each $t \in [m]$:

$$P_\ell^{(t)} := \frac{w(X_{n+t})\mathbf{1}\{\hat{V}_{n+t} \le \hat{V}_{n+\ell}\} + \sum_{i=1}^{n} w(X_i)\mathbf{1}\{V_i \le \hat{V}_{n+\ell}\}}{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)}$$

if $t \neq \ell$ and 0 otherwise.

Apply BH procedure to $(P_\ell^{(t)})_{\ell \in [m]}$ to get rejection set $\hat{R}_t$.



Apply e-BH to $E_t := \dfrac{m}{|\hat{R}_t|\alpha} \cdot \mathbf{1}\left\{P_t \le \dfrac{\alpha|\hat{R}_t|}{m}\right\}$ to get R.

FDR control for WCS batch

Naive proof idea:
Natural weighted p-values:

$$P_t := \frac{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)\mathbf{1}\{V_i \le \hat{V}_{n+t}\}}{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)}$$

are not PRDS —- $w(X_{n+t})$ and V_{n+t} could be arbitrarily dependent.

Define the following weighted oracle p-values:

$$\bar{P}_t := \frac{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)\mathbf{1}\{V_i \le \hat{V}_{n+t}\}}{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)},$$

$$\bar{P}_\ell^{(t)} := \frac{w(X_{n+t})\mathbf{1}\{V_{n+t} \le \hat{V}_{n+\ell}\} + \sum_{i=1}^{n} w(X_i)\mathbf{1}\{V_i \le \hat{V}_{n+\ell}\}}{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)}$$ if $\ell \neq t$ and otherwise 0

(validity sketch)

$\hat{F}_{[n]\cup\{n+t\}} := \frac{\sum_{i=1}^{n} w(X_i)\cdot\delta_{V_i} + w(X_{n+t})\cdot\delta_{V_{n+t}}}{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)}$ is the weighed cdf of the n calibration points combined with tth test point.

Note that $\bar{P}_\ell^{(t)}$ is a function of $\hat{V}_{n+\ell}$ and $\hat{F}_{[n]\cup\{n+t\}}$. Since $\hat{V}_{n+\ell} \perp V_{n+t}$, we have that $\bar{P}_t \perp \bar{P}_\ell^{(t)} \mid \hat{F}_{[n]\cup\{n+t\}}$.
In addition, $\bar{P}_t \mid \hat{F}_{[n]\cup\{n+t\}}$ is superuniform.

Let $\bar{R}_t$ the rejection set of BH applied to $(\bar{P}_\ell^{(t)})_{\ell\in[m]}$. Then, $\bar{E}_t := \frac{K}{\alpha|\bar{R}_t|}\mathbf{1}\left\{\bar{P}_t \le \frac{\alpha|\bar{R}_t|}{K}\right\}$ is an e-value by $\bar{P}_t \perp \bar{R}_t \mid \hat{F}_{[n]\cup\{n+t\}}$.

Note that $\{H_{0,t} \text{ is true}\} \Rightarrow \{\bar{P}_t \le P_t\}$.
$\{H_{0,t} \text{ is true}, P_t \le \alpha|\hat{R}_t|/K\} \Rightarrow \{\hat{R}_t = \bar{R}_t\}$
(key cases are $\{\hat{V}_{n+t} \le \hat{V}_{n+\ell}\} \Rightarrow \{P_\ell^{(t)} = \bar{P}_\ell^{(t)}\}$ and $\{\hat{V}_{n+t} > \hat{V}_{n+\ell}\} \Rightarrow \{P_\ell^{(t)}, \bar{P}_\ell^{(t)} \le P_t\}$)

Hence $E_t := \frac{K}{\alpha|\hat{R}_t|}\mathbf{1}\left\{P_t \le \frac{\alpha|\hat{R}_t|}{K}\right\}$ satisfies $\mathbb{E}[E_t \cdot \mathbf{1}\{H_{0,t} \text{ is true}\}] \le \mathbb{E}[\bar{E}_t] \le 1$

(power)
$P_\ell^{(t)} \approx P_t \approx P_j^{(t)}$, so $|\hat{R}_t| \approx |R|$
(can deploy stochastic rounding to have power when $|\hat{R}_t| < |R|$)

Default approach to online weighted conformalized selection

E-value desiderata: Create (dependent) e-values

$$E_t := \frac{\mathbf{1}\{P_t \le \alpha\gamma_t(|\hat{R}_{t-1}|+1)\}}{\alpha\gamma_t(|\hat{R}_{t-1}|+1)}$$

where we want the following to be true

(validity) $P_t \perp \hat{R}_{t-1} \mid \hat{F}_{[n]\cup\{n+t\}}$

(power) $|\hat{R}_{t-1}| \approx |R_{t-1}|$

Online weighted conformalized selection (attempt)

Do the following for each $t \in \{1,2,\dots\}$

1. Construct <u>multiple</u> weighted conformal p-values (i.e., for each $\ell \in [t]$):

$$P_\ell^{(t)} := \frac{w(X_{n+t})\mathbf{1}\{\hat{V}_{n+t} \le \hat{V}_{n+\ell}\} + \sum_{i=1}^{n} w(X_i)\mathbf{1}\{V_i \le \hat{V}_{n+\ell}\}}{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)}.$$

2. Apply LOND to $(P_\ell^{(t)})_{\ell\in[t-1]}$ to derive $\hat{R}_{t-1}$.

3. Let $E_t := \frac{\mathbf{1}\{P_t \le \alpha\gamma_t(|\hat{R}_{t-1}|+1)\}}{\alpha\gamma_t(|\hat{R}_{t-1}|+1)}$.

Apply e-LOND or (Ue-LOND) to (E_t) to get (R_t)

Where does this go wrong?

Let $\bar{R}_{t-1}$ be the rejection set from running LOND on $(\bar{P}_\ell^{(t)})_{\ell\in[t-1]}$.

$\bar{E}_t := \frac{\mathbf{1}\{P_t \le \alpha\gamma_t(|\bar{R}_{t-1}|+1)\}}{\alpha\gamma_t(|\bar{R}_{t-1}|+1)}$ is a valid e-value,

but we can no longer relate $\hat{R}_{t-1}$ to $\bar{R}_{t-1}$!

This is because LOND is not step-up like BH —- it does not reject all p-values below a threshold.

Our approach to online weighted conformalized selection

Online weighted conformalized selection (ours)

Do the following for each $t \in \{1,2,\dots\}$

1. Construct multiple pairs of weighted conformal p-values (i.e., for each $\ell \in [t-1]$):

$$P_\ell^{(t),-} := \frac{\sum_{i=1}^{n} w(X_i)\mathbf{1}\{V_i \le \hat{V}_{n+\ell}\}}{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)}, P_\ell^{(t),+} := \frac{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)\mathbf{1}\{V_i \le \hat{V}_{n+\ell}\}}{w(X_{n+t}) + \sum_{i=1}^{n} w(X_i)},.$$

2. Apply LOND to $(P_\ell^{(t),-})_{\ell\in[t-1]}$ and $(P_\ell^{(t),+})_{\ell\in[t-1]}$ to derive $\hat{R}_{t-1}^-$ and $\hat{R}_{t-1}^+$

3. Let $E_t := \dfrac{\mathbf{1}\left\{P_t \le \alpha\gamma_t(|\hat{R}_{t-1}^+|+1)\right\}}{\alpha\gamma_t(|\hat{R}_{t-1}^-|+1)}$ (note that $|\hat{R}_{t-1}^+| \le |\bar{R}_{t-1}| \le |\hat{R}_{t-1}^-|$).

Apply **e-LOND** (or **Ue-LOND**) to (E_t) to get (R_t)

Theorem (ours): FDR$(R_t) \le \alpha$ for all $t \in \mathbb{N}$ when e-LOND or Ue-LOND is applied to (E_t) as defined above.

And we can also show a result for arb. dep. p-values as well:

Theorem (ours): FDR$(R_t) \le \alpha$ for all $t \in \mathbb{N}$ when r-LOND or Ur-LOND is applied to (P_t), i.e., the standard weighted p-values.

Results on drug discovery dataset

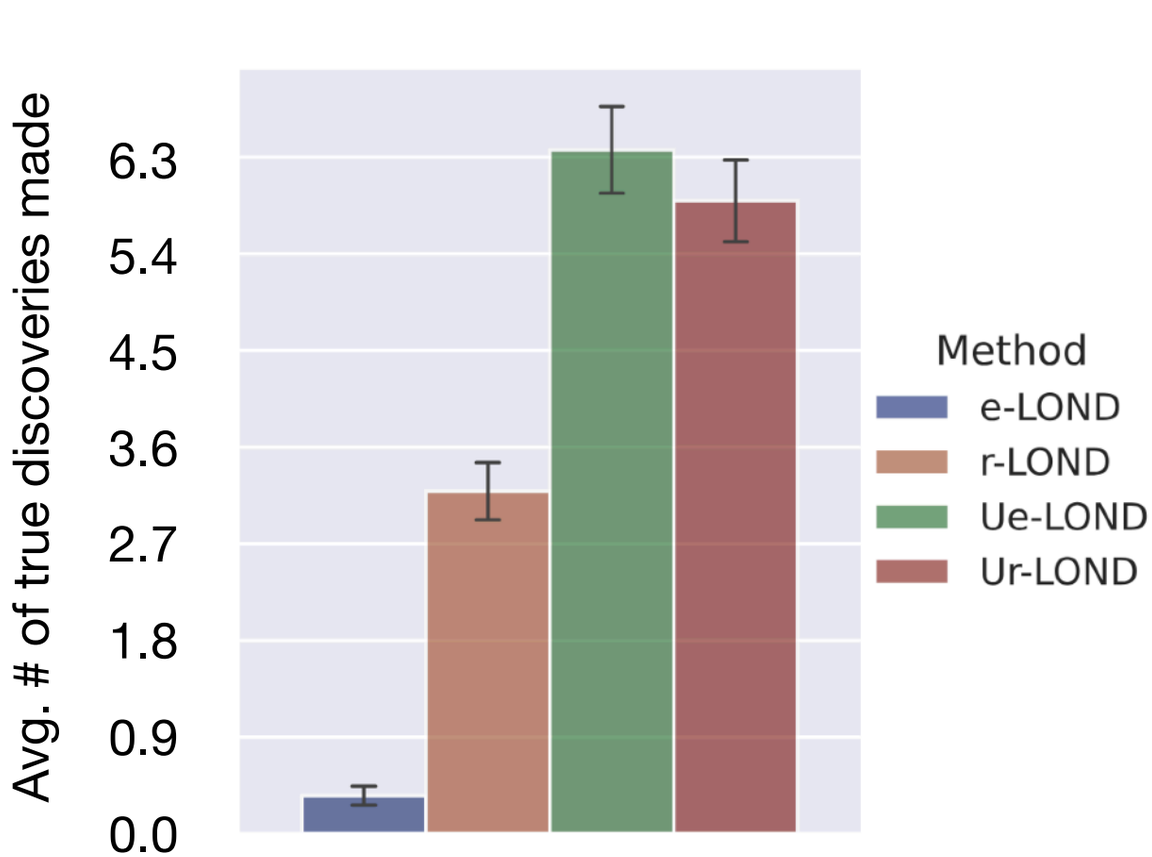
Drug discovery dataset of X = chemical structure, Y = binds to target protein

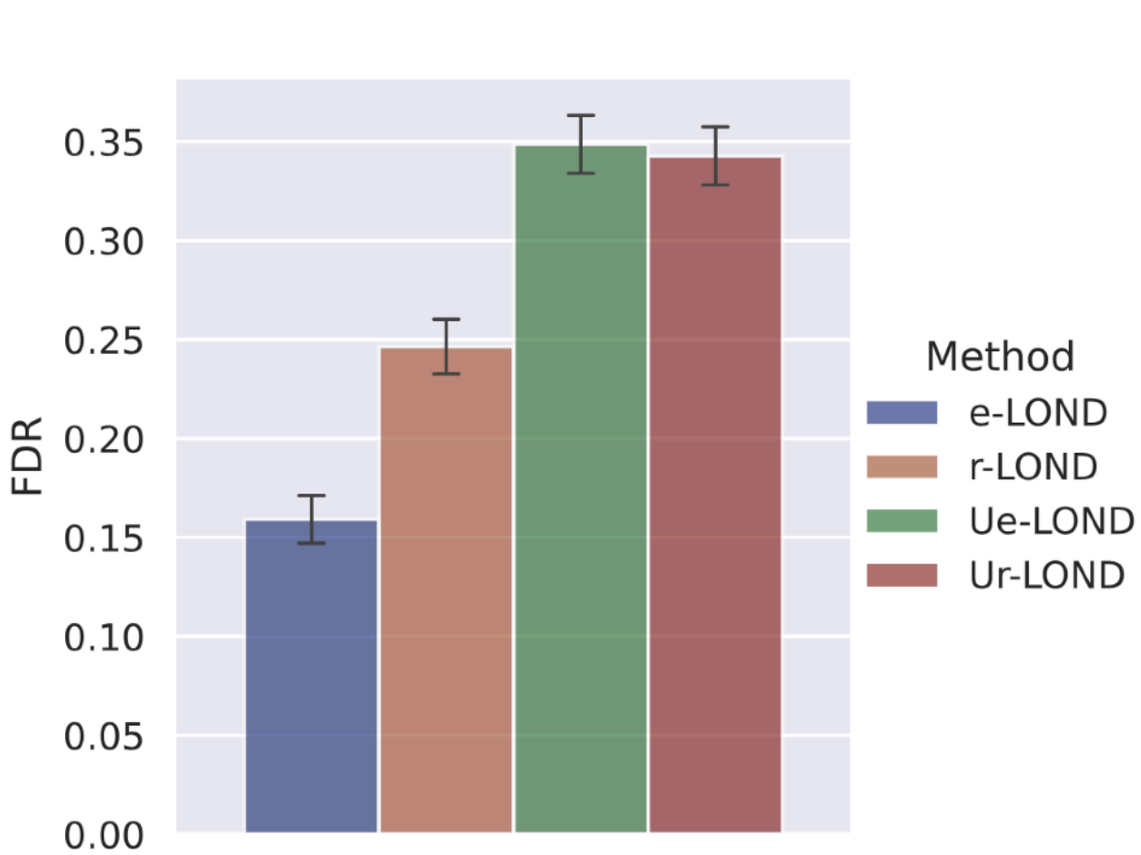
We have access to neural network predictor $\widehat{\mu}$ and $\bar{\mu} = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} \widehat{\mu}(X_i)$

Sample tth drug (i.e., run an experiment to learn Y_{n+t}) to determine if it binds with probability

$$p(X_{n+t}) := \text{sigmoid}(\widehat{\mu}(X_{n+t}) - \bar{\mu}) \wedge 0.8$$

Resulting covariate shift of $w(x) \propto 1/p(x)$.





Extensions and conclusion

Takeaways:

- Unknown dependence is commonplace in the doubly sequential framework modern data science falls into.
- With e-values, we can avoid correction when controlling the FDR under unknown or arbitrary dependence.
- We can utilize randomization to maximize power for e-LOND and r-LOND
- We can develop an online weighted conformal selection procedure using e-values to handle dependence between weighted conformal p-values.
- We can also control the false coverage rate (FCR) of e-value based CIs for any selection rule using a CI analog of e-LOND (see paper)