Contextual Online False Discovery Rate Control

Shiva Kasiviswanathan

Amazon Research

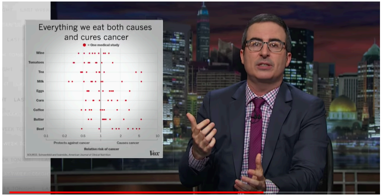
Joint work with: Shiyun Chen (UC San Diego)

May 9, 2019

Problem of False Discoveries

The Problem of False Discovery
MANY SCIENTIFIC RESULTS CAN'T BE REPLICATED, LEADING TO SERIOUS QUESTIONS ABOUT WHAT'S TRUE AND FALSE IN THE WORLD OF RESEARCH.

"Trouble at the Lab" – The Economist

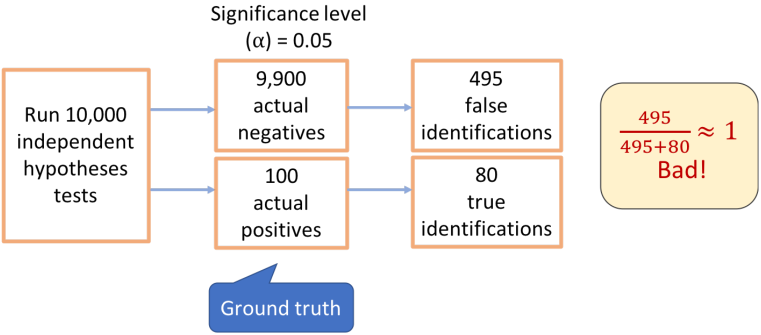


Problem of False Discoveries

Modern Scientific Analysis $=$ Lots of Hypothesis Tests

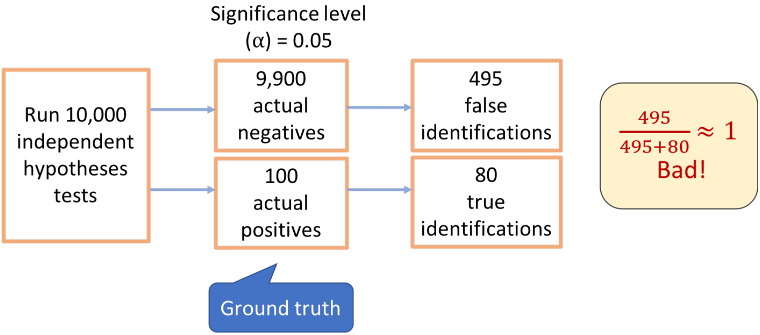
Example: A typical microarray experiment might result in performing $10,000$ separate hypothesis tests.

Problem of False Discoveries



This problem will occur when you run multiple tests, even if hypotheses, tests, and data are all independent!

Problem of False Discoveries



This problem will occur when you run multiple tests, even if hypotheses, tests, and data are all independent!

Question: How to control the number of spurious discoveries?

About 25 years ago: False Discovery Rate Control (BH95)

A Tale of Prefixes

About 25 years ago: False Discovery Rate Control (BH95)

About 10 years ago: Online False Discovery Rate Control (FS08)

A Tale of Prefixes

About 25 years ago: False Discovery Rate Control (BH95)

About 10 years ago: Online False Discovery Rate Control (FS08)

This work: Contextual Online False Discovery Rate Control

Offline Multiple Testing

Setting: n hypotheses $H_1, \ldots, H_n$ with p-values $\mathbf{P} = (P_1, \ldots, P_n)$

Offline Multiple Testing

Setting: n hypotheses $H_1, \ldots, H_n$ with p-values $\mathbf{P} = (P_1, \ldots, P_n)$

A **multiple testing procedure** $\mathcal{R}$ is of form

$$\mathcal{R} : \mathbf{P} \mapsto \mathcal{R}(\mathbf{P}) \subset [n]$$

taking the p-values $\mathbf{P}$ and returning a subset of $[n] := 1, \ldots, n$ representing the null hypotheses to be rejects.

Possible Outcomes

	Accept null	Reject null	Total
Null true	U	V	n_0
Alternative true	T	S	n_1
	W	R	n

Table: Outcomes from n hypothesis tests

False Discovery Rate (FDR)

	Accept null	Reject null	Total
Null true	U	V	n_0
Alternative true	T	S	n_1
	W	R	n

Given a multiple hypothesis procedure $\mathcal{R}$, the false discovery rate is defined as the expected fraction of mistaken rejections (BH95)

$$\text{FDR}(\mathcal{R}) = \mathbb{E}[\text{FDP}(\mathcal{R})], \text{ and } \quad \text{FDP}(\mathcal{R}) := \frac{V}{R \vee 1}.$$

FDR is expected proportion of *Type I error* of a test procedure

False Discovery Rate (FDR)

	Accept null	Reject null	Total
Null true	U	V	n_0
Alternative true	T	S	n_1
	W	R	n

Given a multiple hypothesis procedure $\mathcal{R}$, the false discovery rate is defined as the expected fraction of mistaken rejections (BH95)

$$\textsc{fdr}(\mathcal{R}) = \mathbb{E}[\textsc{fdp}(\mathcal{R})], \text{ and } \quad \textsc{fdp}(\mathcal{R}) := \frac{V}{R \vee 1}.$$

FDR is expected proportion of *Type I error* of a test procedure

In the offline setting, *Benjamini-Hochberg* (BH) procedure is a popular way to control FDR

Other Side: Statistical Power

	Accept null	Reject null	Total
Null true	U	V	n_0
Alternative true	T	S	n_1
	W	R	n

True discovery proportion and rate (power) are defined as

$$\text{TDR}(\mathcal{R}) = \mathbb{E}[\text{TDP}(\mathcal{R})], \text{ and } \quad \text{TDP}(\mathcal{R}) := \frac{S}{n_1}.$$

Real World is Online

Online Multiple Testing

In real world, hypotheses are not all available, but come over time

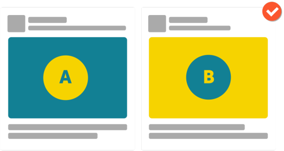
Online Multiple Testing

In real world, hypotheses are not all available, but come over time

Take an example of a company performing A/B testing:

Stream of Hypotheses

H_1:

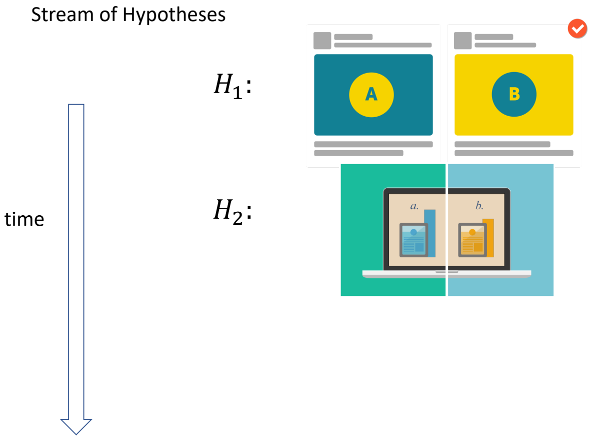


time

Online Multiple Testing

In real world, hypotheses are not all available, but come over time

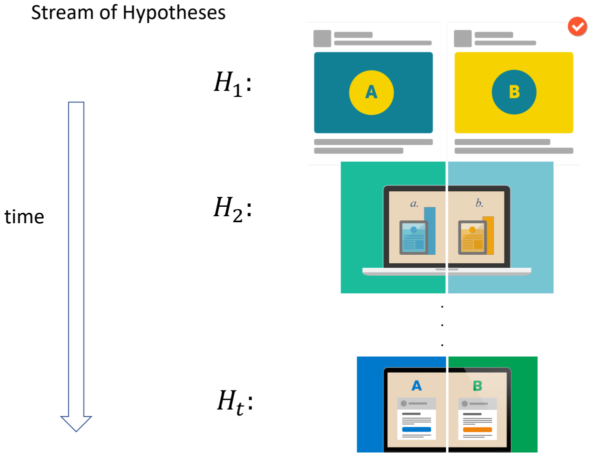
Take an example of a company performing A/B testing:



Online Multiple Testing

In real world, hypotheses are not all available, but come over time

Take an example of a company performing A/B testing:



Online Multiple Testing

Offline $\Rightarrow$ Online (FS08)

Offline ⇒ Online (FS08)

Setting: A sequence of ordered, possibly infinite hypotheses $H_1, H_2, \ldots,$ arriving in a stream with corresponding p-values $P_1, P_2, \ldots$

At each step, an investigator must decide whether to reject the current null hypothesis, without having access to the number of hypotheses or the future p-values

Offline ⇒ Online (FS08)

Setting: A sequence of ordered, possibly infinite hypotheses $H_1, H_2, \ldots,$ arriving in a stream with corresponding p-values $P_1, P_2, \ldots$

At each step, an investigator must decide whether to reject the current null hypothesis, without having access to the number of hypotheses or the future p-values

Goal: Control False Discovery Rate

Online Multiple Testing

An online testing procedure provides a sequence of significance levels α_t, with decision rule:

$$R_t = \begin{cases} 1 & P_t \leq \alpha_t, \quad \text{reject } H_t, \\ 0 & \text{otherwise}, \quad \text{accept } H_t. \end{cases}$$

Significance levels are the functions of prior outcomes:

$$\alpha_t = \alpha_t(R_1, \ldots, R_{t-1})$$

Online Multiple Testing

An online testing procedure provides a sequence of significance levels α_t, with decision rule:

$$R_t = \begin{cases} 1 & P_t \leq \alpha_t, \quad \text{reject } H_t, \\ 0 & \text{otherwise}, \quad \text{accept } H_t. \end{cases}$$

Significance levels are the functions of prior outcomes:

$$\alpha_t = \alpha_t(R_1, \dots, R_{t-1})$$

Let $R(t)$ be number of rejections made by the algorithm till time t
Let $V(t)$ be the number of false rejections till time t

$$\mathrm{FDR}(t) = \mathbb{E}[\mathrm{FDP}(t)], \quad \mathrm{FDP}(t) := \frac{V(t)}{R(t) \vee 1}$$

Goal: $\sup_{T \in \mathbb{N}} \mathrm{FDR}(T) \leq \alpha$

Online Multiple Testing

An online testing procedure provides a sequence of significance levels α_t, with decision rule:

$$R_t = \begin{cases} 1 & P_t \leq \alpha_t, \quad \text{reject } H_t, \\ 0 & \text{otherwise}, \quad \text{accept } H_t. \end{cases}$$

Significance levels are the functions of prior outcomes:

$$\alpha_t = \alpha_t(R_1, \ldots, R_{t-1})$$

Let $R(t)$ be number of rejections made by the algorithm till time t
Let $V(t)$ be the number of false rejections till time t

$$\mathrm{FDR}(t) = \mathbb{E}[\mathrm{FDP}(t)], \quad \mathrm{FDP}(t) := \frac{V(t)}{R(t) \vee 1}$$

Goal: $\sup_{T \in \mathbb{N}} \mathrm{FDR}(T) \leq \alpha$

Similarly, we can define $\mathrm{TDR}(T)$ in an online setting

Prior Work: Online Multiple Testing

Generalized Alpha Investing (GAI) Rules (AR14):

Example: Levels based On Recent Discovery (LORD) (JM18)
Example: Improved Levels based On Recent Discovery (LORD++)(RYWJ17)

Prior Work: Online Multiple Testing

Generalized Alpha Investing (GAI) Rules (AR14):

Example: Levels based On Recent Discovery (LORD) (JM18)
Example: Improved Levels based On Recent Discovery (LORD++)(RYWJ17)

Slightly Different: **SAFFRON** (RZWJ18):

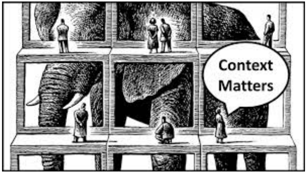
Adaptively estimates the proportion of true nulls like in *Storey's* procedure (Sto02)

Using Contextual Information

Typically, in addition to the p-value, each hypothesis can also have a set of features which encode contextual (side) information related to the tested hypothesis, which is also referred as **contextual information**.

Using Contextual Information

Typically, in addition to the p-value, each hypothesis can also have a set of features which encode contextual (side) information related to the tested hypothesis, which is also referred as **contextual information**.



Think of contextual information as containing some indirect information about the likelihood of a hypothesis being false, but the relationship is not known ahead of time

Some Examples

Problem	Example "Context" Info.
A/B testing of webpage	Size of the banner ad, content of text on each page
Gene association with a trait	Location of each gene, counts of each gene
Disease prediction	Biographical information of each patient

Long line of work in the offline setting in utilizing contextual information with testing (IKZH16; GRW06; LB16; RBWJ17; XZZT17; LF18)...

Contextual Online Multiple Testing

- Setting: A sequence of ordered hypotheses $H_1, H_2, \ldots$ arrives in a stream. Each hypothesis H_i is associated with a p-value $P_i \in (0, 1)$ and a vector of contextual features $X_i \in \mathcal{X}$, thus can be represented by a tuple (H_i, P_i, X_i)

Contextual Online Multiple Testing

- Setting: A sequence of ordered hypotheses $H_1, H_2, \ldots$ arrives in a stream. Each hypothesis H_i is associated with a p-value $P_i \in (0, 1)$ and a vector of contextual features $X_i \in \mathcal{X}$, thus can be represented by a tuple (H_i, P_i, X_i)

- At each step i, decide whether to reject H_i having only access to previous decisions and contextual information so far

Contextual Online Multiple Testing

- Setting: A sequence of ordered hypotheses $H_1, H_2, \ldots$ arrives in a stream. Each hypothesis H_i is associated with a p-value $P_i \in (0,1)$ and a vector of contextual features $X_i \in \mathcal{X}$, thus can be represented by a tuple (H_i, P_i, X_i)

- At each step i, decide whether to reject H_i having only access to previous decisions and contextual information so far

- Overall Goal: Control online FDR under a given level α

Contextual Online Multiple Testing

- Setting: A sequence of ordered hypotheses $H_1, H_2, \ldots$ arrives in a stream. Each hypothesis H_i is associated with a p-value $P_i \in (0,1)$ and a vector of contextual features $X_i \in \mathcal{X}$, thus can be represented by a tuple (H_i, P_i, X_i)
- At each step i, decide whether to reject H_i having only access to previous decisions and contextual information so far
- Overall Goal: Control online FDR under a given level α and improve the number of useful discoveries by using contextual information

Contextual Online Multiple Testing

In online testing with contextual information, the significance levels can be functions of prior results and the contextual features seen so far:

$$\alpha_t = \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t).$$

Contextual Online Multiple Testing

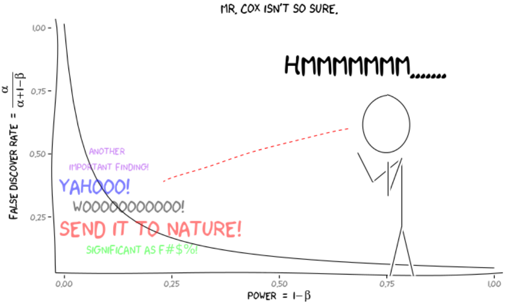
In online testing with contextual information, the significance levels can be functions of prior results and the contextual features seen so far:

$$\alpha_t = \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t).$$

$$R_t = \begin{cases} 1 & P_t \leq \alpha_t = \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t) \qquad \text{reject } H_t, \\ 0 & \text{otherwise} \qquad \text{accept } H_t. \end{cases}$$

Reminder of this Talk: Our Results

1. Online FDR Control with Contextual Information
2. Power Analysis with Contextual Features
   - Increase in Statistical Power
3. Experimental Results



1 Online FDR Control with Contextual Information

2 Power Analysis with Contextual Features
- Increase in Statistical Power

3 Experimental Results

Contextual Generalized Alpha-investing Rules

Starting Point: Generalized Alpha Investing (GAI) Rules (AR14)

Contextual Generalized Alpha-investing Rules

Starting Point: Generalized Alpha Investing (GAI) Rules (AR14)

We propose a new class of online testing rules called **Contextual Generalized Alpha-investing Rules** by modifying **Generalized Alpha-investing Rules** (AR14; RYWJ17)

Starting Point: Generalized Alpha Investing (GAI) Rules (AR14)

We propose a new class of online testing rules called **Contextual Generalized Alpha-investing Rules** by modifying **Generalized Alpha-investing Rules** (AR14; RYWJ17)

So what are "Generalized Alpha-investing Rules"?

Generalized Alpha-investing Rules in Pictures

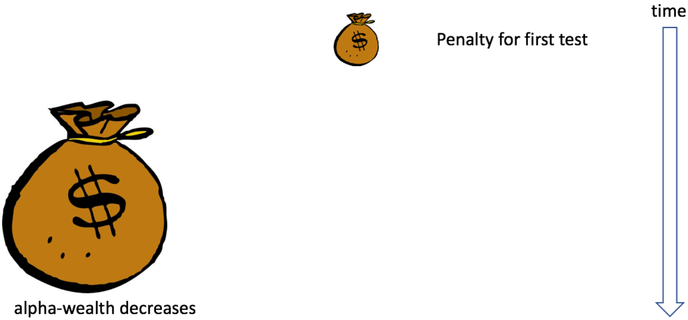
Error budget or alpha-wealth

Generalized Alpha-investing Rules in Pictures

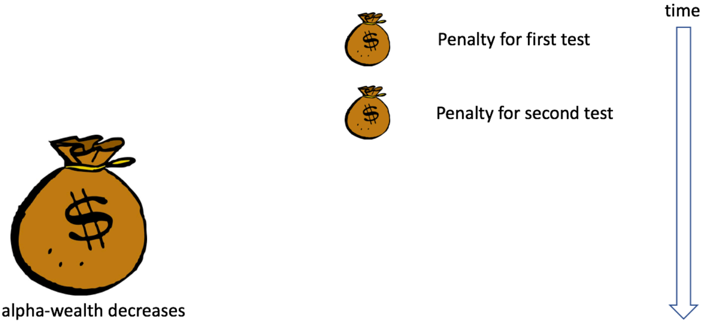
Error budget or alpha-wealth

Tests use wealth

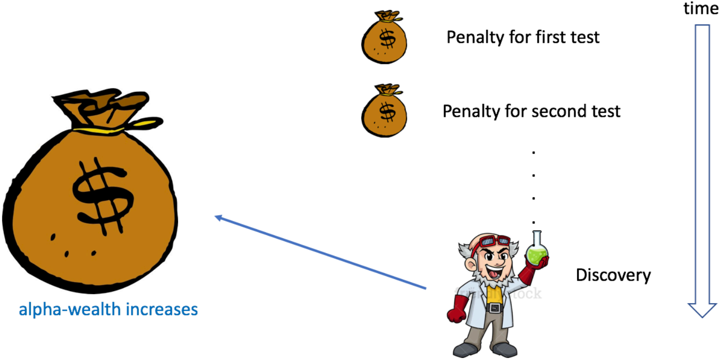
Generalized Alpha-investing Rules in Pictures



Generalized Alpha-investing Rules in Pictures



Generalized Alpha-investing Rules in Pictures



Discoveries increases wealth

Generalized Alpha-investing Rules Mathematically

1. Penalty function: ϕ_t
2. Reward function: ψ_t
3. Significance level: α_t

Generalized Alpha-investing Rules:

$$\textbf{Initial Wealth: } W(0) = w_0, \text{with } 0 < w_0 < \alpha,$$

Generalized Alpha-investing Rules Mathematically

1. Penalty function: ϕ_t
2. Reward function: ψ_t
3. Significance level: α_t

Generalized Alpha-investing Rules:

Initial Wealth: $W(0) = w_0, \text{with } 0 < w_0 < \alpha,$

Wealth Update: $W(t) = W(t-1) - \phi_t + R_t \cdot \psi_t,$

Generalized Alpha-investing Rules Mathematically

1. Penalty function: ϕ_t
2. Reward function: ψ_t
3. Significance level: α_t

Generalized Alpha-investing Rules:

$$\textbf{Initial Wealth: } W(0) = w_0, \text{with } 0 < w_0 < \alpha,$$
$$\textbf{Wealth Update: } W(t) = W(t-1) - \phi_t + R_t \cdot \psi_t,$$
$$\textbf{Non-negativity: } \phi_t \leq W(t-1),$$

Generalized Alpha-investing Rules Mathematically

1. Penalty function: ϕ_t
2. Reward function: ψ_t
3. Significance level: α_t

Generalized Alpha-investing Rules:

$$\textbf{Initial Wealth: } W(0) = w_0, \text{with } 0 < w_0 < \alpha,$$
$$\textbf{Wealth Update: } W(t) = W(t-1) - \phi_t + R_t \cdot \psi_t,$$
$$\textbf{Non-negativity: } \phi_t \leq W(t-1),$$
$$\textbf{Upper Bound on Reward: } \psi_t \leq \min\{\phi_t + b_t, \tfrac{\phi_t}{\alpha_t} + b_t - 1\},$$
$$\text{where } b_t = \alpha - w_0 \mathbb{1}\{\rho_1 > t-1\} (\rho_1 \text{ is time of first discovery})$$

where $\alpha_t, \phi_t, \psi_t \in \sigma(R_1, \ldots, R_{t-1})$.

How to Incorporate Contextual Information?

1. Penalty function: ϕ_t
2. Reward function: ψ_t
3. Significance level: α_t

Contextual Generalized Alpha-investing Rules:

$$\textbf{Initial Wealth: } W(0) = w_0, \text{with } 0 < w_0 < \alpha,$$
$$\textbf{Wealth Update: } W(t) = W(t-1) - \phi_t + R_t \cdot \psi_t,$$
$$\textbf{Non-negativity: } \phi_t \leq W(t-1),$$
$$\textbf{Upper Bound on Reward: } \psi_t \leq \min\{\phi_t + b_t, \tfrac{\phi_t}{\alpha_t} + b_t - 1\},$$
$$\text{where } b_t = \alpha - w_0 \mathbb{1}\{\rho_1 > t-1\} (\rho_1 \text{ is time of first discovery})$$

where $\alpha_t, \phi_t, \psi_t \in \sigma(R_1, \ldots, R_{t-1})$. → $\sigma(\sigma(R_1,\ldots,R_{t-1}) \cup \sigma(X_1,\ldots,X_t))$

Monotone Contextual Generalized Alpha-investing Rules

A Contextual Generalized Alpha-investing rule is **monotone** if we have $\tilde{R}_i \leq R_i$ for all $i \leq t-1$, then we have

$$\alpha_t(\tilde{R}_1, \ldots, \tilde{R}_{t-1}, X_1, \ldots, X_t) \leq \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t),$$

for any fixed $\mathbf{X}^t = (X_1, \ldots, X_t)$

Monotone Contextual Generalized Alpha-investing Rules

A Contextual Generalized Alpha-investing rule is **monotone** if we have $\tilde{R}_i \leq R_i$ for all $i \leq t-1$, then we have

$$\alpha_t(\tilde{R}_1, \ldots, \tilde{R}_{t-1}, X_1, \ldots, X_t) \leq \alpha_t(R_1, \ldots, R_{t-1}, X_1, \ldots, X_t),$$

for any fixed $\mathbf{X}^t = (X_1, \ldots, X_t)$

"Significance level is higher with more rejections"

Our FDR Result

Theorem

If for all timesteps t, the p-values P_t's are independent, and P_t's and X_t's are independent under the null, then for any Monotone Contextual Generalized Alpha-investing rule, we have

$$\sup_{T \in \mathbb{N}} \mathrm{FDR}(T) \leq \alpha.$$

Note that P_t's could be related to X_t's (via some unknown function) under alternate

Our FDR Result

Theorem

If for all timesteps t, the p-values P_t's are independent, and P_t's and X_t's are independent under the null, then for any Monotone Contextual Generalized Alpha-investing rule, we have

$$\sup_{T \in \mathbb{N}} \mathrm{FDR}(T) \leq \alpha.$$

Note that P_t's could be related to X_t's (via some unknown function) under alternate

Additional Results:

- Results on *modified FDR* (FS08) control under weaker assumption on p-values
- Results for dependent p-values

Proof Idea

- Let $\mathcal{H}^0$ denote the indices of true nulls
- Number of false discoveries: $V(T) = \sum_{t=1}^{T} R_t \mathbb{1}\{t \in \mathcal{H}^0\}$
- Wealth: $W(T) = w_0 + \sum_{t=1}^{T}(-\phi_t + R_t\psi_t)$

Proof Idea

- Let $\mathcal{H}^0$ denote the indices of true nulls
- Number of false discoveries: $V(T) = \sum_{t=1}^{T} R_t \mathbb{1}\{t \in \mathcal{H}^0\}$
- Wealth: $W(T) = w_0 + \sum_{t=1}^{T}(-\phi_t + R_t\psi_t)$

$$\begin{aligned}
\mathrm{FDR}(T) := \mathbb{E}\left[\frac{V(T)}{R(T) \vee 1}\right] &\leq \mathbb{E}\left[\frac{V(T) + W(T)}{R(T) \vee 1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\frac{R_t \mathbb{1}\{t \in \mathcal{H}^0\} + \frac{w_0}{T} - \phi_t + R_t\psi_t}{R(T) \vee 1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t}{R(T) \vee 1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t}{R(T) \vee 1}\middle|\sigma(\sigma(R_1, \ldots, R_{t-1}) \cup \sigma(X_1, \ldots, X_t))\right]\right]
\end{aligned}$$

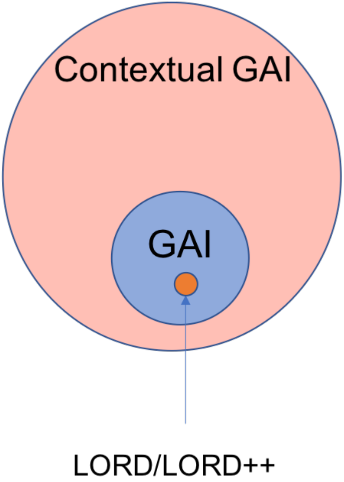
Proof Idea

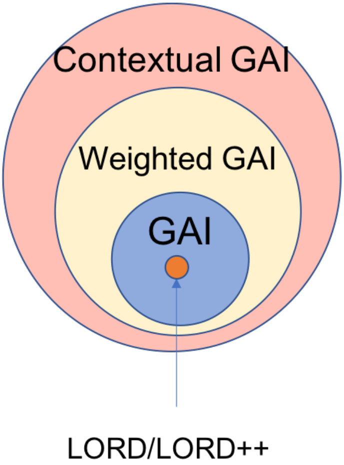
- Let $\mathcal{H}^0$ denote the indices of true nulls
- Number of false discoveries: $V(T) = \sum_{t=1}^{T} R_t \mathbb{1}\{t \in \mathcal{H}^0\}$
- Wealth: $W(T) = w_0 + \sum_{t=1}^{T}(-\phi_t + R_t\psi_t)$

$$
\begin{aligned}
\mathrm{FDR}(T) &:= \mathbb{E}\left[\frac{V(T)}{R(T) \vee 1}\right] \leq \mathbb{E}\left[\frac{V(T) + W(T)}{R(T) \vee 1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\frac{R_t \mathbb{1}\{t \in \mathcal{H}^0\} + \frac{w_0}{T} - \phi_t + R_t\psi_t}{R(T) \vee 1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t}{R(T) \vee 1}\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\frac{\frac{w_0}{T} + R_t(\psi_t + \mathbb{1}\{t \in \mathcal{H}^0\}) - \phi_t}{R(T) \vee 1}\middle| \sigma(\sigma(R_1, \ldots, R_{t-1}) \cup \sigma(X_1, \ldots, X_t))\right]\right]
\end{aligned}
$$

Two cases (use the reward bounds):

1. $t \in \mathcal{H}^0$: We use $\psi_t \leq \frac{\phi_t}{\alpha_t} + b_t - 1$
2. $t \notin \mathcal{H}^0$: We use $\psi_t \leq \phi_t + b_t$


Contextual GAI
GAI
LORD/LORD++


Contextual GAI
Weighted GAI
GAI
LORD/LORD++

1 Online FDR Control with Contextual Information

2 Power Analysis with Contextual Features
- Increase in Statistical Power

3 Experimental Results

Question: Can contextual information help with increasing the statistical power?

Question: Can contextual information help with increasing the statistical power?

Answer: Yes*

Increase in Statistical Power in Online Setting

First Piece:
Online FDR Rules

+

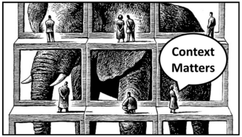
Second Piece:
Model

+

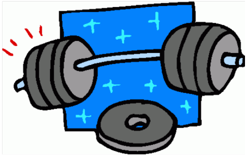
Third Piece:
Conditions

Increase in Statistical Power in Online Setting

Our Idea: Use current context to weigh the significance level



X_t



$\omega(X_t)$

Set: $\alpha_t = \alpha_t(R_1, \ldots, R_{t-1})\omega(X_t)$

$$R_t = \begin{cases} 1 & P_t \leq \alpha_t = \alpha_t(R_1, \ldots, R_{t-1})\omega(X_t) \quad \text{reject } H_t, \\ 0 & \text{otherwise} \quad \text{accept } H_t. \end{cases}$$

Increase in Statistical Power in Online Setting

First Piece: Online FDR Rules + Second Piece: Model + Third Piece: Conditions

First Piece: Online FDR Rules

LORD (JM18): A popular subclass of Generalized Alpha-investing rules

Any sequence of nonnegative numbers $\gamma = (\gamma_t)_{t=1}^{\infty}$, which is monotonically non-increasing with $\sum_{t=1}^{\infty} \gamma_t = 1$.

$$W(0) = \frac{\alpha}{2},$$

$$\text{Penalty:} \quad \phi_t = \alpha_t = \gamma_{t-\tau_t} \frac{\alpha}{2},$$

$$\text{Reward:} \quad \psi_t = \frac{\alpha}{2},$$

where τ_t is the last time a discovery was made before t.

Second Piece: Mixture Model

Let $H_t = 0$ (denote null) and $H_t = 1$ (denote alternate)

For any $t \in \mathbb{N}$, let

$$H_1, \ldots, H_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$

Second Piece: Mixture Model

Let $H_t = 0$ (denote null) and $H_t = 1$ (denote alternate)

For any $t \in \mathbb{N}$, let

$$H_1, \dots, H_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$
$$X_t \mid H_t = 0 \sim \mathcal{L}_0(\mathcal{X}), \quad X_t \mid H_t = 1 \sim \mathcal{L}_1(\mathcal{X}),$$

where $0 < \pi_1 < 1$ and where $\mathcal{L}_0(\mathcal{X})$, $\mathcal{L}_1(\mathcal{X})$ are two probability distribution on the contextual feature space $\mathcal{X}$

Second Piece: Mixture Model

Let $H_t = 0$ (denote null) and $H_t = 1$ (denote alternate)

For any $t \in \mathbb{N}$, let

$$H_1, \dots, H_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$
$$X_t \mid H_t = 0 \sim \mathcal{L}_0(\mathcal{X}), \quad X_t \mid H_t = 1 \sim \mathcal{L}_1(\mathcal{X}),$$
$$\text{Under Null: } P_t \mid H_t = 0, X_t \sim \text{Uniform}(0,1),$$

where $0 < \pi_1 < 1$ and where $\mathcal{L}_0(\mathcal{X})$, $\mathcal{L}_1(\mathcal{X})$ are two probability distribution on the contextual feature space $\mathcal{X}$

Second Piece: Mixture Model

Let $H_t = 0$ (denote null) and $H_t = 1$ (denote alternate)

For any $t \in \mathbb{N}$, let

$$H_1, \ldots, H_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$
$$X_t \mid H_t = 0 \sim \mathcal{L}_0(\mathcal{X}), \quad X_t \mid H_t = 1 \sim \mathcal{L}_1(\mathcal{X}),$$
$$\text{Under Null: } P_t \mid H_t = 0, X_t \sim \text{Uniform}(0, 1),$$
$$\text{Under Alternate: } P_t \mid H_t = 1, X_t \sim F_1(p \mid X_t).$$

where $0 < \pi_1 < 1$ and where $\mathcal{L}_0(\mathcal{X})$, $\mathcal{L}_1(\mathcal{X})$ are two probability distribution on the contextual feature space $\mathcal{X}$

Mixture Model: An Example

Let $H_t = 0$ (denote null) and $H_t = 1$ (denote alternate)

For any $t \in \mathbb{N}$, let

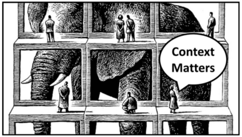
$$H_1, \ldots, H_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$
$$X_t \mid H_t = 0 \sim \mathcal{L}_0(\mathcal{X}), \quad X_t \mid H_t = 1 \sim \mathcal{L}_1(\mathcal{X}),$$
$$\text{Under Null: } P_t \mid H_t = 0, X_t \sim \text{Uniform}(0,1),$$
$$\text{Under Alternate: } P_t \mid H_t = 1, X_t \sim F_1(p \mid X_t).$$

Normal Means Model

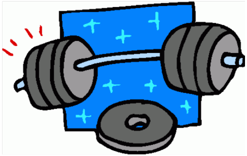
For any $t \in \mathbb{N}$, let

$$H_1, \ldots, H_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$
$$X_t \mid H_t = 0 \sim \mathcal{L}_0(\mathcal{X}), \quad X_t \mid H_t = 1 \sim \mathcal{L}_1(\mathcal{X}),$$
$$\text{Null: } \mu_t = 0, \quad \text{Alternate: } \mu_t = \mu(X_t),$$
$$\text{Test Statistic: } Z_t = \mathcal{N}(\mu_t, 1),$$
$$P_t = 2\Phi(-|Z_t|).$$

Third Piece: Conditions



X_t



$\omega(X_t)$

Assume for any $t \in \mathbb{N}$,

1. $\omega_t = \omega(X_t)$ is a random variable with different distributions under null and alternate
2. Weighting is **informative**[1], in that the weights under alternate is more likely to be larger than that under the null

[1] Similar notion used by (GRW06) for studying weighted Benjamini-Hochberg procedure in the offline setting.

Statistical Power Increase with Weighting

Unweighted Case

Given a sequence of p-values $(P_1, P_2, \dots)$ from the mixture model, apply LORD procedure on this sequence.

Theorem (JM18): Tight bound on average power

Statistical Power Increase with Weighting

Unweighted Case

Given a sequence of p-values $(P_1, P_2, \dots)$ from the mixture model, apply LORD procedure on this sequence.

Theorem (JM18): Tight bound on average power

Weighted Case

Given a sequence of p-values $(P_1, P_2, \dots)$ from the mixture model, and a sequence of informative weights $(\omega_1, \omega_2, \dots)$ (based on contextual features), apply LORD procedure on the sequence $(P_1/\omega_1, P_2/\omega_2, \dots)$.

Theorem: Lower bound on average power

Statistical Power Increase with Weighting

Unweighted Case

Given a sequence of p-values $(P_1, P_2, \dots)$ from the mixture model, apply LORD procedure on this sequence.

Theorem (JM18): Tight bound on average power

Weighted Case

Given a sequence of p-values $(P_1, P_2, \dots)$ from the mixture model, and a sequence of informative weights $(\omega_1, \omega_2, \dots)$ (based on contextual features), apply LORD procedure on the sequence $(P_1/\omega_1, P_2/\omega_2, \dots)$.

Theorem: Lower bound on average power

Comparing the above power bounds gives a necessary condition under which a separation in power holds

Under some reasonable assumptions, contextual features could help with increasing the power of the online testing rules (without affecting the FDR control)

1 Online FDR Control with Contextual Information

2 Power Analysis with Contextual Features
- Increase in Statistical Power

3 Experimental Results

Modeling the Weight Function

Input: Sequence of p-values, contextual features pairs: $(P_1, X_1), (P_2, X_2), .$

Decision Rule:

$$R_t = \begin{cases} 1, & P_t \leq \alpha_t = \alpha_t(R_1, \dots, R_{t-1})\omega(X_t) \qquad \text{reject } H_t, \\ 0, & \text{otherwise} \qquad \text{accept } H_t. \end{cases}$$

Question: How do we define the weight function $\omega(\cdot)$?

Modeling the Weight Function

Input: Sequence of p-values, contextual features pairs: $(P_1, X_1), (P_2, X_2), .$

Decision Rule:

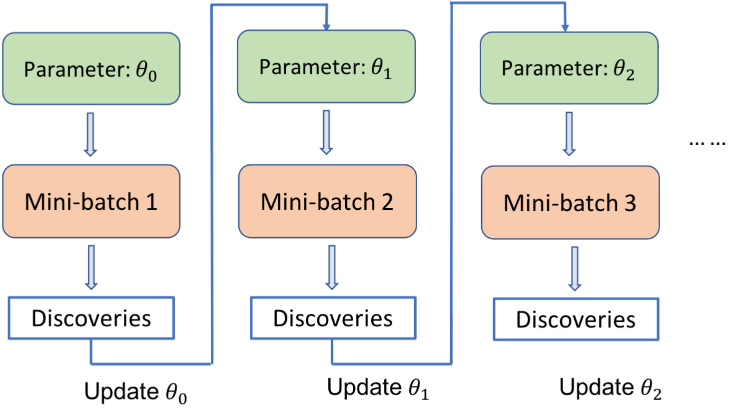
$$R_t = \begin{cases} 1, & P_t \leq \alpha_t = \alpha_t(R_1, \ldots, R_{t-1})\omega(X_t) \qquad \text{reject } H_t, \\ 0, & \text{otherwise} \qquad \text{accept } H_t. \end{cases}$$

Question: How do we define the weight function $\omega(\cdot)$?
Answer: We use a neural network to model $\omega(\cdot)$.

- $\omega(X_t) = \omega(X_t; \theta)$ where θ are parameters of a neural network
- Training of the network to maximize the number of **empirical discoveries**, subject to **FDR control**

Training the Network

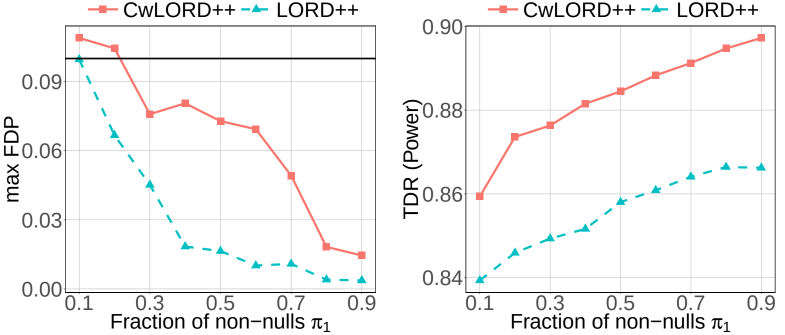
Training Procedure: Learn parameters in an online fashion to maximize empirical discoveries subject to FDR control



Experiments on Synthetic Data

Normal Means Model:

$$H_1, \ldots, H_t \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_1),$$
$$\text{Null: } \mu_t = 0, \quad \text{Alternate: } \mu_t = \langle \beta, X_t \rangle,$$
$$\text{Test Statistic: } Z_t = \mathcal{N}(\mu_t, 1),$$
$$P_t = 2\Phi(-|Z_t|).$$



(a) FDR Plot (b) Power Plot

Our Algorithm: CwLORD++. Baseline: LORD++ (RYWJ17)

Overlays

Diabetes Detection Dataset: Kaggle Dataset. Biographical information used as contextual information

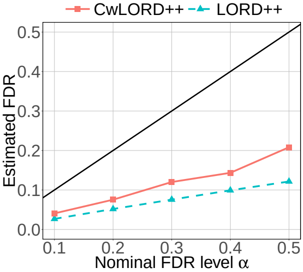
Online Testing Class	FDR ($\alpha = 0.2$)	Power
LORD++	0.147	0.384
Ours (CwLORD++)	0.176	0.580

Overlays

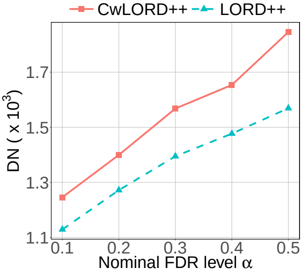
Diabetes Detection Dataset: Kaggle Dataset. Biographical information used as contextual information

Online Testing Class	FDR ($\alpha = 0.2$)	Power
LORD++	0.147	0.384
Ours (CwLORD++)	0.176	0.580

Airway RNA-Seq Dataset: log count for each gene used as contextual information



(a) FDR Plot



(b) Power Plot

Concluding Remarks

- Introduced the problem of contextual online FDR control
- Proposed a new class of online FDR control rules
- Theoretical analysis: FDR control, Power Improvement (under *informative weighting*)
- Better empirical performance

Concluding Remarks

- Introduced the problem of contextual online FDR control
- Proposed a new class of online FDR control rules
- Theoretical analysis: FDR control, Power Improvement (under *informative weighting*)
- Better empirical performance

Open Questions

- Can we check for informative weighting in practice?
- Theoretical properties of the neural network based online testing procedure?

Reference

[AR14] Ehud Aharoni and Saharon Rosset. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794, 2014.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[FS08] Dean P Foster and Robert A Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444, 2008.

[GRW06] Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[IKZH16] Nikolaos Ignatiadis, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577, 2016.

[JM18] Adel Javanmard and Andrea Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics*, 46(2):526–554, 2018.

[LB16] Ang Li and Rina Foygel Barber. Multiple testing with the structure adaptive benjamini-hochberg algorithm. *arXiv preprint arXiv:1606.07926*, 2016.

[LF18] Lihua Lei and William Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679, 2018.

[RBWJ17] Aaditya Ramdas, Rina Foygel Barber, Martin J Wainwright, and Michael I Jordan. A unified treatment of multiple testing with prior knowledge using the p-filter. *arXiv preprint arXiv:1703.06222*, 2017.

Third Piece: Conditions

Let $\omega : \mathcal{X} \to \mathbb{R}$ be a weight function.
Define weight distributions Q_0 and Q_1 as:

$$Q_0 = \omega(X) \text{ with } X \sim \mathcal{L}_0$$
$$Q_1 = \omega(X) \text{ with } X \sim \mathcal{L}_1$$

For any $t \in \mathbb{N}$, we assume $\omega(X_t)$ is drawn from either of these distributions

$$\omega(X_t) = \omega_t \sim Q_0 \mid H_t = 0$$
$$\omega(X_t) = \omega_t \sim Q_1 \mid H_t = 1$$

Informative: $u_0 = \mathbb{E}[Q_0], u_1 = \mathbb{E}[Q_1]$, and $u_0 < 1$ and $u_1 > 1$ (weight under alternative is more likely to be larger than that under the null)

Theorem

Define $D(t) = \Pr[P/\omega \leq t]$*. Then, the average power of contextual weighted LORD rule is almost surely bounded as follows:*

$$\liminf_{T\to\infty} \mathrm{TDR}(T) \geq \Big(\sum_{m=1}^{\infty} \prod_{j=1}^{m} (1 - D(b_0\gamma_j))\Big)^{-1}$$