Real-Time Selection Under General Constraints via Predictive Inference

Lin Lu

School of Statistics and Data Sciences, Nankai University, Tianjin, China

Email: linlu102099@gmail.com

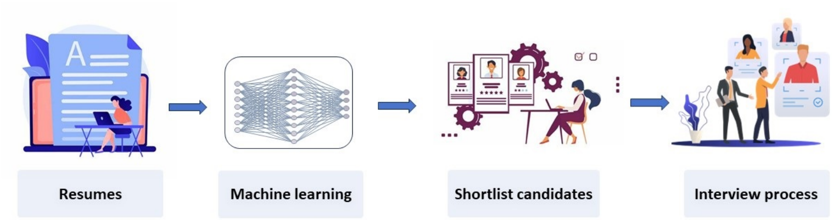
Joint work with
Yuyang Huo, Haojie Ren and Changliang Zou

NeurIPS 2024

November 10, 2024

Outline

Motivating example: online recruitment



- How to reduce the cost of the subsequent interview process for unsuitable candidates?
  - Control the False Selection Rate to avoid containing too many unsuitable candidates.
- How to contain suitable candidates with different backgrounds to enrich the diversity?
  - Control the Similarity among candidates at a desired level.

General constrained online sample selection

Problem statement:

- A sequence of i.i.d. unlabeled data $\mathbf{X}_1, \mathbf{X}_2, \cdots$ arrives in a stream with their responses $Y_1, Y_2, \cdots$ unobserved all the time.
- **Task**: to sequentially select samples whose unobserved responses Y_t' s are in the specified target region $\mathcal{A}$ until a specific stopping time T.
  - $\mathcal{A} = \{1\}$ or $\{0\}$ for classification. In recruitment, $\mathcal{A} = \{\text{the class of suitable candidate}\}$.
  - $\mathcal{A} = [a, b]$ or $[b, +\infty)$ for regression.
- At each time t, we make decision $\delta_t = 0/1$ on whether to select $\mathbf{X}_t$ until stopping.
- We call the selection $\delta_t = 1$ is correct if $\theta_t = \mathbb{I}\{Y_t \in \mathcal{A}\} = 1$.
- **Our goal**: design a real-time selection procedure that can control various constraints at the user-specified stopping time T.

Challenges

Natural idea:

- $\mu(\mathbf{x}) := Y \mid \mathbf{X} = \mathbf{x}$: regression or classification model with $(\mathbf{X}, Y)$
- Estimate $\mu(\mathbf{x})$ on some offline labelled data
- Obtain predicated value $\widehat{Y}_j = \widehat{\mu}(\mathbf{X}_j)$ for $\mathbf{X}_j$
- Use $\mathbb{I}\{\widehat{Y}_j \in \mathcal{A}\}$ to approximate $\theta_j = \mathbb{I}\{Y_j \in \mathcal{A}\}$

Challenges:

$$\{\widehat{Y}_j \in \mathcal{A}\} \neq \{Y_j \in \mathcal{A}\}$$

- How to quantify the prediction uncertainty to make the subsequent decisions reliable?
- How to make the selected subset more informative/representative?
- How to make sample selection in the online setting?

Our solutions: II-COS

Individual and Interactive Constrained Online Selection:

- (1) Quantify the uncertainty of response predictions using **predictive inference**;
- (2) Focus on two general types of constraints;
- (3) Address individual and interactive constraints at **each time**.

Focus on two general types of constraints

1. **Individual constraints** $C_1(\delta^t)$:

   each selected sample has a cost associated with θ_j and $\mathbf{X}_j$.

2. **Interactive constraints** $C_2(\delta^t)$:

   capture interactive influence among correctly selected samples.

Individual and interactive constraints

- Individual constraints

$$C_1(\boldsymbol{\delta}^t) = \mathbb{E}\left[\frac{\sum_{i\le t}\{(1-\theta_i)G_0(\mathbf{X}_i)+\theta_i G_1(\mathbf{X}_i)\}\delta_i}{(\sum_{i\le t}\delta_i)\vee 1}\right] \le \alpha. \tag{2.1}$$

  - **E.g.**: If $G_0(\mathbf{X}) = 1$ and $G_1(\mathbf{X}) = 0$, the individual constraint is false selection rate (FSR)

$$C_1(\boldsymbol{\delta}^t) = \mathrm{FSR}(\boldsymbol{\delta}^t) = \mathbb{E}\left[\frac{\sum_{i\le t}(1-\theta_i)\delta_i}{(\sum_{i\le t}\delta_i)\vee 1}\right].$$

- Interactive constraints

$$C_2(\boldsymbol{\delta}^t) = \frac{\mathbb{E}\left[\sum\sum_{1\le i<j\le t} g(\mathbf{X}_i,\mathbf{X}_j)\theta_i\theta_j\delta_i\delta_j\right]}{\mathbb{E}\left[\sum\sum_{1\le i<j\le t}\theta_i\theta_j\delta_i\delta_j\right]} \le K. \tag{2.2}$$

Involves choosing more preferable samples relying on the interaction among correctly selected samples.

Online multiple testing (online FDR control)

- Generalized α-investing framework: Foster and Stine [2008]; Aharoni and Rosset [2014]; LOND [Javanmard and Montanari, 2015], SAFFRON [Ramdas et al., 2018], ADDIS [Tian and Ramdas, 2019].
- Structure-adaptive sequential testing: SAST [Gang et al., 2023].

Conformal/predictive inference

- Conformal inference: Vovk et al. [2005], Romano et al. [2019], Chernozhukov et al. [2021], Vovk [2015], Barber et al. [2021].
- Conformal p-values: Bates et al. [2023]; Jin and Candès [2023].
- Prediction-assisted subsampling: Wu et al. [2023].

Goal and formulation

Our goal for online selection

To select samples of interest by a decision rule δ^T controlling both the individual and interactive constraints at any time t until reaching the stopping time T, i.e.,

$$C_1(\delta^t) \leq \alpha \quad \text{and} \quad C_2(\delta^t) \leq K.$$

- $W_t = \widehat{\mu}(\mathbf{X}_t)$ is a predicted value of Y_t and assume that $\widehat{\mu}(\cdot)$ is bijection almost surely.
- $\theta_t = \mathbb{I}(Y_t \in \mathcal{A})$ is Bernoulli(q) distributed with $q = \Pr(Y_t \in \mathcal{A})$.
- W_t can be viewed as generated from two-group model

$$W_t \mid \theta_t \sim (1-\theta_t)f_0 + \theta_t f_1,$$

where f_0 and f_1 denote pdf of $W_t \mid \theta_t = 0$ and $W_t \mid \theta_t = 1$.

- With the two-group model, we have

$$\mathbb{E}[\theta_t \mid \mathbf{X}_t] = 1 - \Pr(\theta_t = 0 \mid W_t) = 1 - \frac{(1-q)f_0(W_t)}{f(W_t)} := 1 - L_t$$

where $f = (1-\pi)f_0 + \pi f_1$. So θ_t is equivalent to $1 - L_t$ in the sense of taking expectation.

- Individual constraint $C_1(\boldsymbol{\delta}^t)$ can be exactly satisfied if

$$\frac{V_t}{R_t} := \frac{\sum_{i\leq t}\{L_i G_0(\mathbf{X}_i) + (1-L_i)G_1(\mathbf{X}_i)\}\delta_i}{(\sum_{i\leq t}\delta_i)\vee 1} \leq \alpha$$

holds, where V_t is the expected cost of summation until time t and R_t is the number of selection.

- Accordingly, interactive constraint $C_2(\boldsymbol{\delta}^t) \leq K$ can be achieved if

$$\frac{\mathrm{TS}_t}{\mathrm{NS}_t} := \frac{\sum\sum_{1\leq i<j\leq t} g(\mathbf{X}_i,\mathbf{X}_j)(1-L_i)(1-L_j)\delta_i\delta_j}{\sum\sum_{1\leq i<j\leq t}(1-L_i)(1-L_j)\delta_i\delta_j} \leq K,$$

where the expected total mutual effects conditional on $\{\mathbf{X}_i\}_{i\leq t}$ and the expected number are denoted as TS_t and NS_t, respectively.

Oracle II-COS procedure

Oracle II-COS (Individual and Interactive Constrained Online Selection)

If L_t is known, when t comes before the first selection (i.e, $R_{t-1} = 0$), the decision rule is $\delta_t = 1$ if

$$\frac{V_{t-1} + L_t G_0(\mathbf{X}_t) + (1 - L_t) G_1(\mathbf{X}_t)}{R_{t-1} + 1} \leq \alpha, \tag{2.3}$$

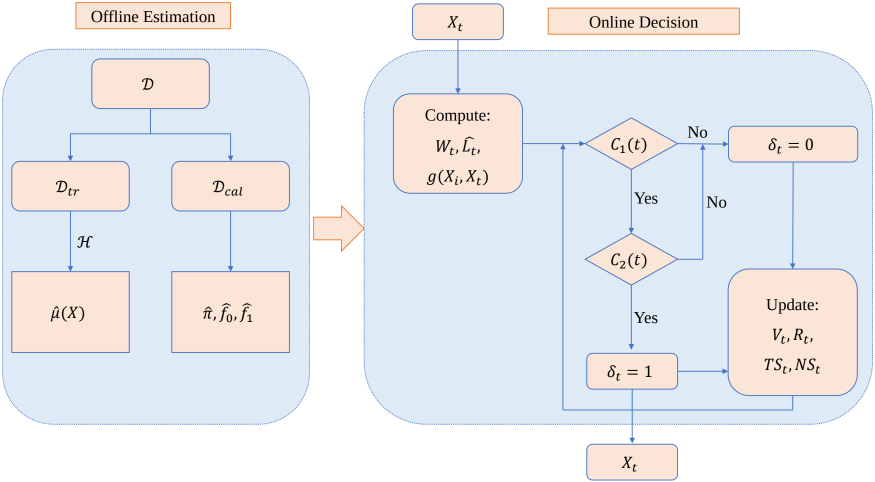
holds; otherwise, $\delta_t = 0$ which means $\mathbf{X}_t$ is not selected. When $\mathbf{X}_t$ arrives with $R_{t-1} \geq 1$, then $\delta_t = 1$ if (2.3) and

$$\frac{\mathsf{TS}_{t-1} + \left[\sum_{i \leq t-1} g(\mathbf{X}_i, \mathbf{X}_t)(1 - L_i)\delta_i\right](1 - L_t)}{\mathsf{NS}_{t-1} + \left[\sum_{i \leq t-1} (1 - L_i)\delta_i\right](1 - L_t)} \leq K \tag{2.4}$$

hold simultaneously; otherwise, $\delta_t = 0$.

- Assume L_t values are known. Then the oracle II-COS selection rule controls both constraints at any time t, i.e., $C_1(\boldsymbol{\delta}^t) \leq \alpha$ and $C_2(\boldsymbol{\delta}^t) \leq K$.
- In practice, since L_t is unknown, we adopt a data-splitting strategy to estimate L_t and propose a data-driven II-COS procedure.

Workflow of data-driven II-COS procedure



Individual and interactive constraints guarantee

Theorem

Suppose Assumptions 1-2 hold and take the bandwidths for estimating f and f_0 in the order of $n^{-1/(2\beta+1)}$. Let $T_m = \inf\{t : \sum_{i=1}^{t} \delta_i = m\}$ for $m > 2$. Then for any given time $t \geq T_m$, the data-driven II-COS procedure satisfies

- *(Bound for individual constraint) Denote $\Delta_n = Dn^{\frac{-\beta}{2\beta+1}}\sqrt{\log n}$ and D is a constant depending on M, ℓ, c_β, β, π, c_G and $K(\cdot)$, then*

$$C_1(\boldsymbol{\delta}^t) \leq \alpha + \Delta_n.$$

- *(Bound for interactive constraint) Assume there exists a constant $\alpha' \in (0,1)$ such that $\sum_{i\leq t} \widehat{L}_i \delta_i/(1 \vee R_t) \leq \alpha'$, then*

$$C_2(\boldsymbol{\delta}^t) \leq K + \frac{(K + c_g)\Delta_n}{0.5 - \frac{m\alpha'}{m-1} - \Delta_n}.$$

The II-COS procedure controls the two constraints asymptotically at any $t \leq T$:

$$\lim_{n\to\infty} C_1(\boldsymbol{\delta}^t) \leq \alpha \quad \text{and} \quad \lim_{n\to\infty} C_2(\boldsymbol{\delta}^t) \leq K.$$

Basic simulation setting

Two constraints of interest:

- Individual constraint: False selection rate (FSR), computed by the average value of FSP among 500 replications.
$$\mathrm{FSP}(t) = \frac{\sum_{i=1}^{t} \delta_i (1-\theta_i)}{(\sum_{i=1}^{t} \delta_i) \vee 1}.$$
- Interactive constraint: Empirical similarity (ES), the average of ES_0
$$\mathrm{ES}_0(t) = \frac{\sum\sum_{1 \le i < j \le t} g(\mathbf{X}_i, \mathbf{X}_j)\theta_i\theta_j\delta_i\delta_j}{\sum\sum_{1 \le i < j \le t} \theta_i\theta_j\delta_i\delta_j},$$
where g is taken as the RBF kernel.

Stopping rule:

- Stop when selecting $m = 100$ samples, $T_m = \inf_t\{t : \sum_{i=1}^{t} \delta_i = m\}$.

Benchmarks:

- Compare the **II-COS** procedure with four benchmarks from online multiple testing.
- **SAST**: implemented with the same lFDR estimator $\widehat{L}_t$
- **LOND, SAFFARON, ADDIS**: implemented with the conformal *p*-values suggested by Bates et al. [2023]

Model setting:

- A classification model:
  - $\mathbf{X} \mid Y = 0 \sim \mathcal{N}_4(\boldsymbol{\mu}_1, \mathbf{I}_4)$ and $\mathbf{X} \mid Y = 1 \sim \mathcal{N}_4(\boldsymbol{\mu}_2, \mathbf{I}_4)$
  - $\boldsymbol{\mu}_1 = (5, 0, 0, 0)^\top, \boldsymbol{\mu}_2 = (0, 0, -3, -2)^\top$ and $\Pr(Y = 1) = 0.2$.
- The target region is $\mathcal{A} = \{1\}$.
- Random forest with defaulted parameters is trained to give prediction.

Simulation: results in real time

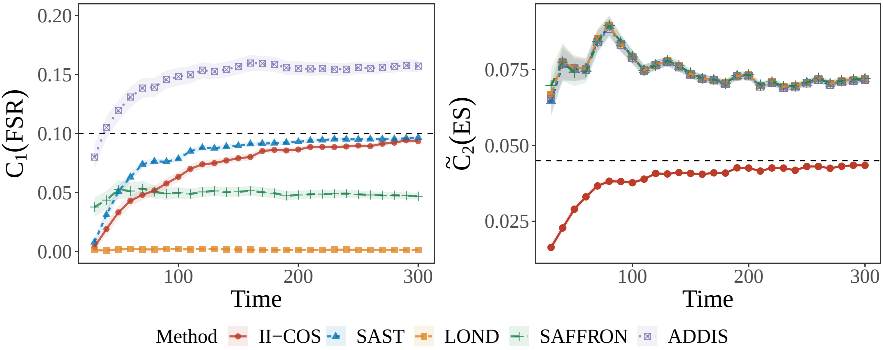


图: The real-time plot of FSR and ES for II-COS, SAST, LOND, SAFFRON and ADDIS. The black dashed lines denote the FSR level $\alpha = 0.1$ and the ES level $K = 0.045$. We use training data size $n_{\text{tr}} = 1,000$ and calibration data size $n_{\text{cal}} = 4,000$.

Prediction-assisted Inference

Summary:

- Predictive inference plays a central role to guarantee the validity of the decision-making.
- A real-time selection rule under general constraints to extract informative samples from target space. (II-COS)

Discussion:

- More general interactive constraints: uniform/space-filling design criterion.
- Incorporation of auxiliary information: neighbor information or the feedback information.
- More general requirements: deterministic sample selection, obtaining a better training model.

Thank you!

See more details and experiments results in our paper:

- Real-Time Selection Under General Constraints via Predictive Inference, *NeurIPS*, 2024.

References I

E. Aharoni and S. 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.

R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

S. Bates, E. Candès, L. Lei, Y. Romano, and M. Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

V. Chernozhukov, K. Wüthrich, and Y. Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

D. P. Foster and R. 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.

B. Gang, W. Sun, and W. Wang. Structure–adaptive sequential testing for online false discovery rate control. *Journal of the American Statistical Association*, 118(541):732–745, 2023.

A. Javanmard and A. Montanari. On online control of false discovery rate. *arXiv preprint arXiv:1502.06197*, 2015.

Y. Jin and E. J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

A. Ramdas, T. Zrnic, M. Wainwright, and M. Jordan. Saffron: an adaptive algorithm for online control of the false discovery rate. In *International Conference on Machine Learning*, pages 4286–4294. PMLR, 2018.

Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. In *Advances in neural information processing systems*, pages 3538–3548, 2019.

J. Tian and A. Ramdas. Addis: an adaptive discarding algorithm for online fdr control with conservative nulls. *Advances in Neural Information Processing Systems*, 32:9388–9396, 2019.

V. Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28, 2015.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. New York: Springer, 2005.

X. Wu, Y. Huo, H. Ren, and C. Zou. Optimal subsampling via predictive inference. *Journal of the American Statistical Association*, pages 1–29, 2023.