E-Readers @ CWI Amsterdam
24 April 2026

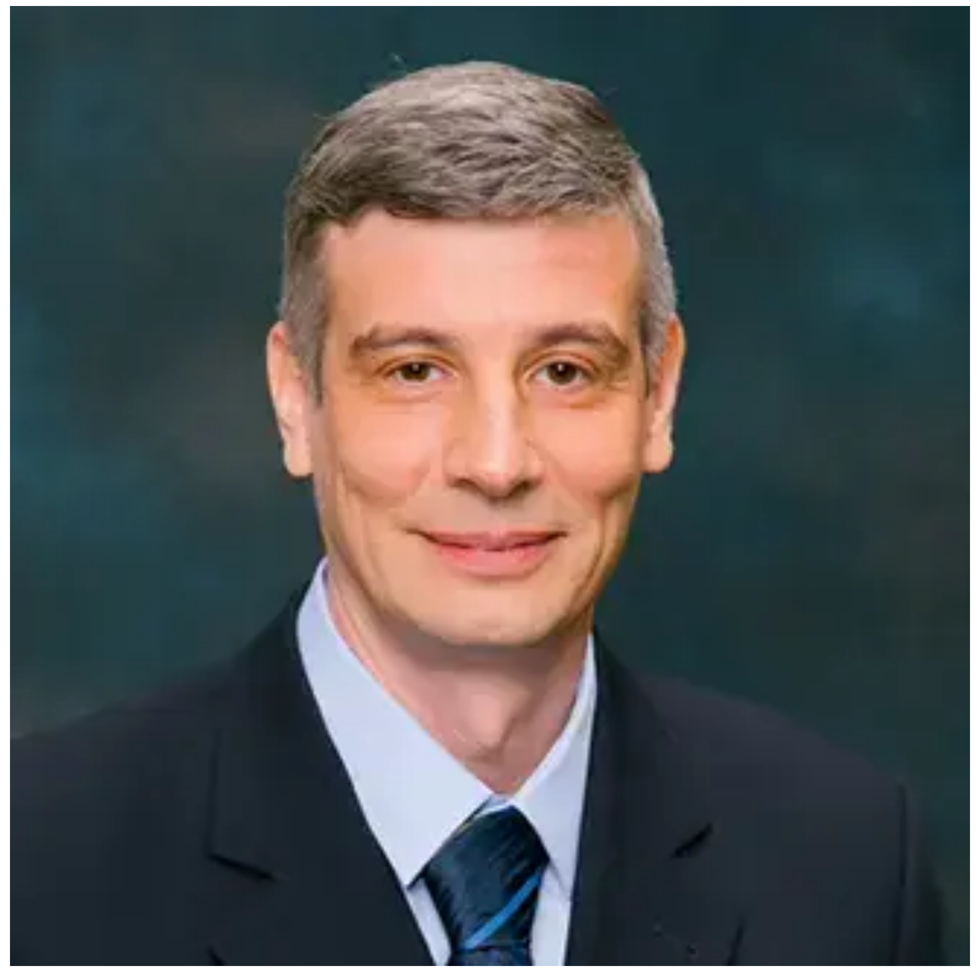
Betting on Bets:
Anytime-Valid Tests for Stochastic Dominance

Yo Joong "YJ" Choe, INSEAD Decision Sciences
Sebastian Arnold, CWI Amsterdam

Sebastian Arnold*
CWI Amsterdam

YJ Choe*
INSEAD

Marco Scarsini
LUISS



Ilia Tsetlin
INSEAD

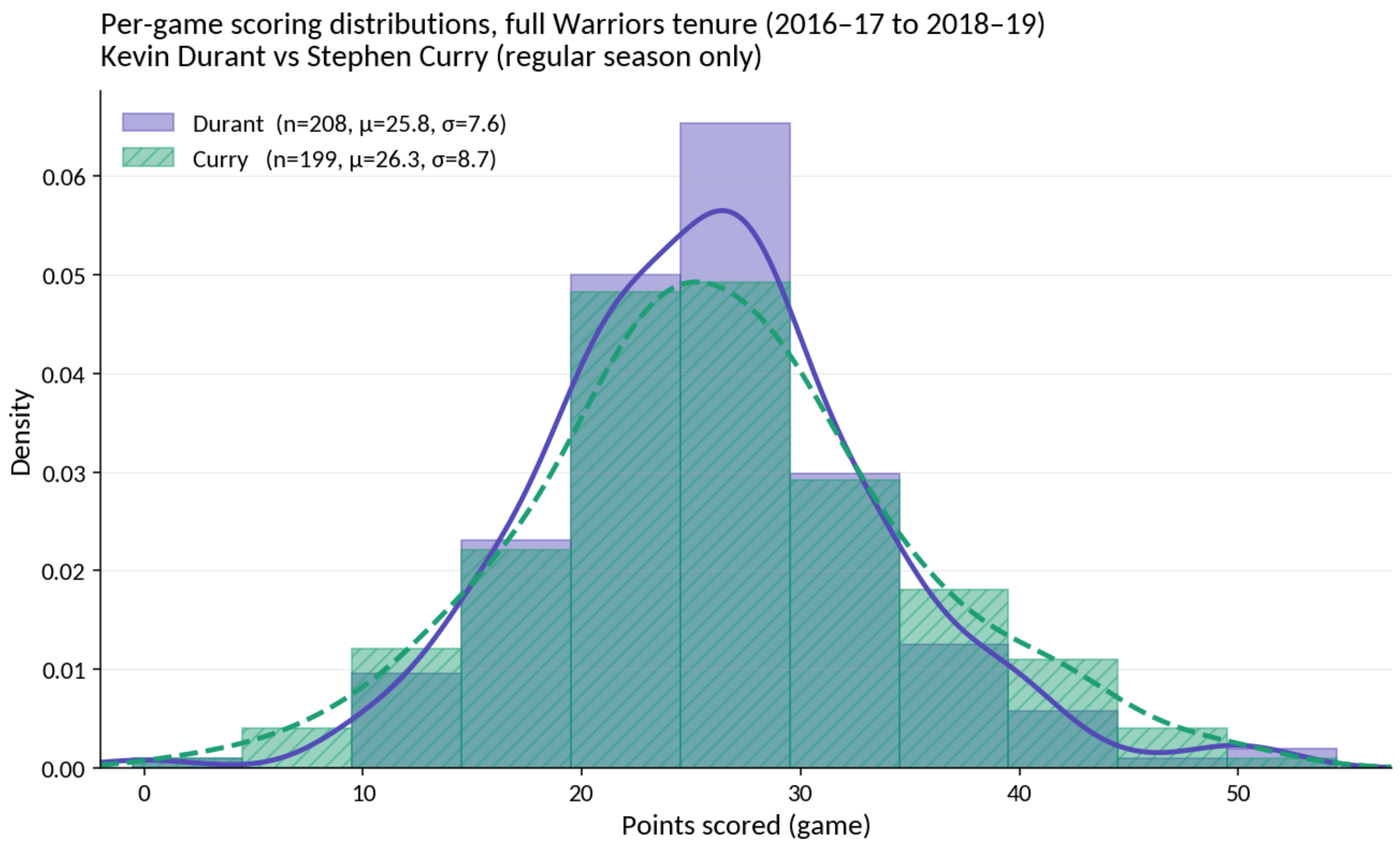
**Equal contribution.*

Motivation

Motivation #1: Does one prospect have an **upside** over another?

🏀 Comparing player performances:

Is there an upside to playing Curry over Durant for the final game of the season?


Per-game scoring distributions, full Warriors tenure (2016–17 to 2018–19)
Kevin Durant vs Stephen Curry (regular season only)

💵 Testing distributional effects:

Is there any upside in future incomes to veteran status? (Plot: Abadie, 2002)

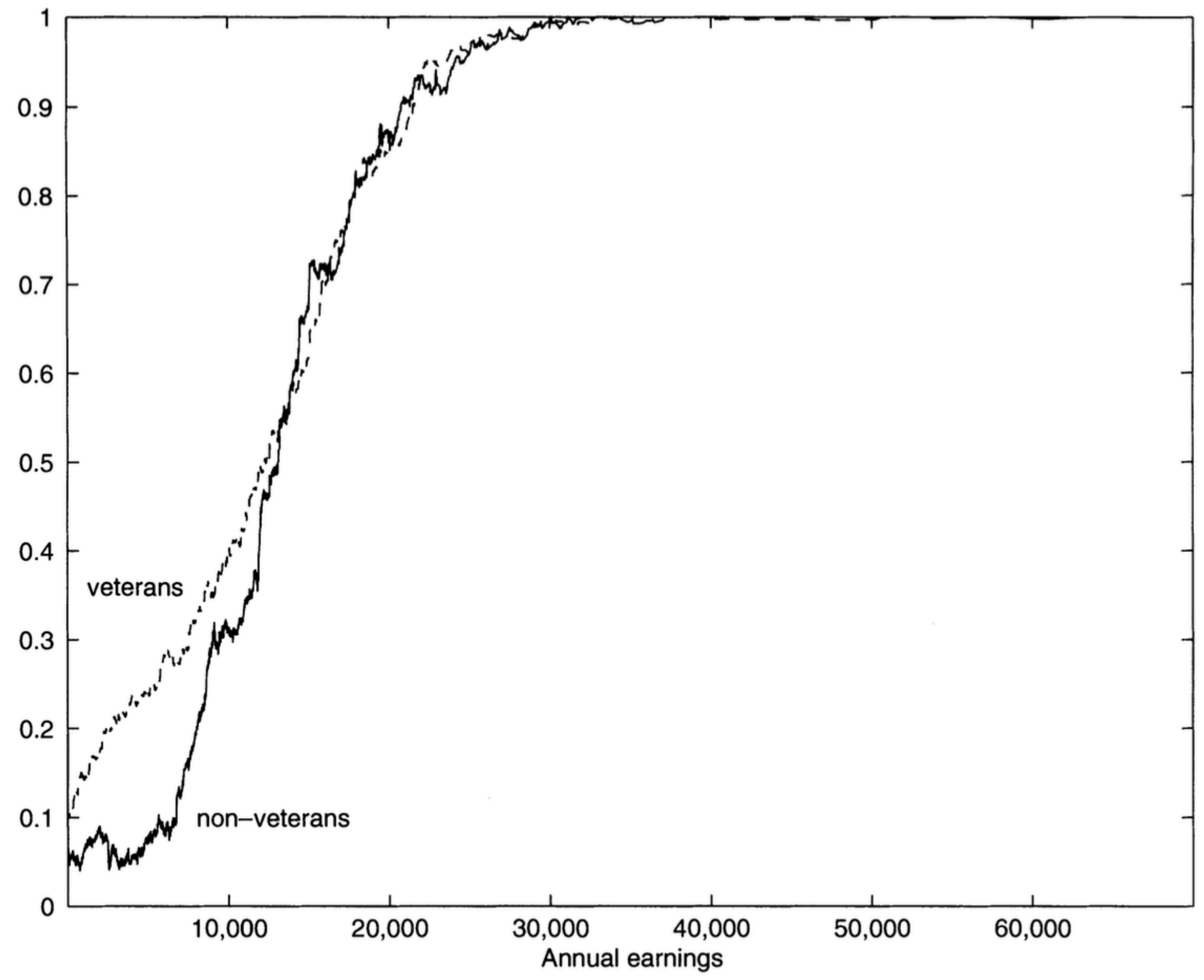


Figure 2. Estimated Distributions of Potential Earnings for Compliers.

These examples motivate **distributional** comparisons of random variables, beyond just their means, in different regions of the support.

Motivation #2: Can I monitor this test in real time?

In reality, these are *sequential* inference problems (literally, game time decisions!)

Can we compare distributions & upsides **over sequentially observed data,** such that the inference remains valid (say, under continuous monitoring)?

The Goal

Develop **“anytime-valid”** methods for testing **stochastic dominance** that can be monitored for upsides at adaptive sample sizes.

…without sacrificing statistical power.

Stochastic Dominance

(a.k.a. Stochastic Ordering)

First-order stochastic dominance (1-SD)

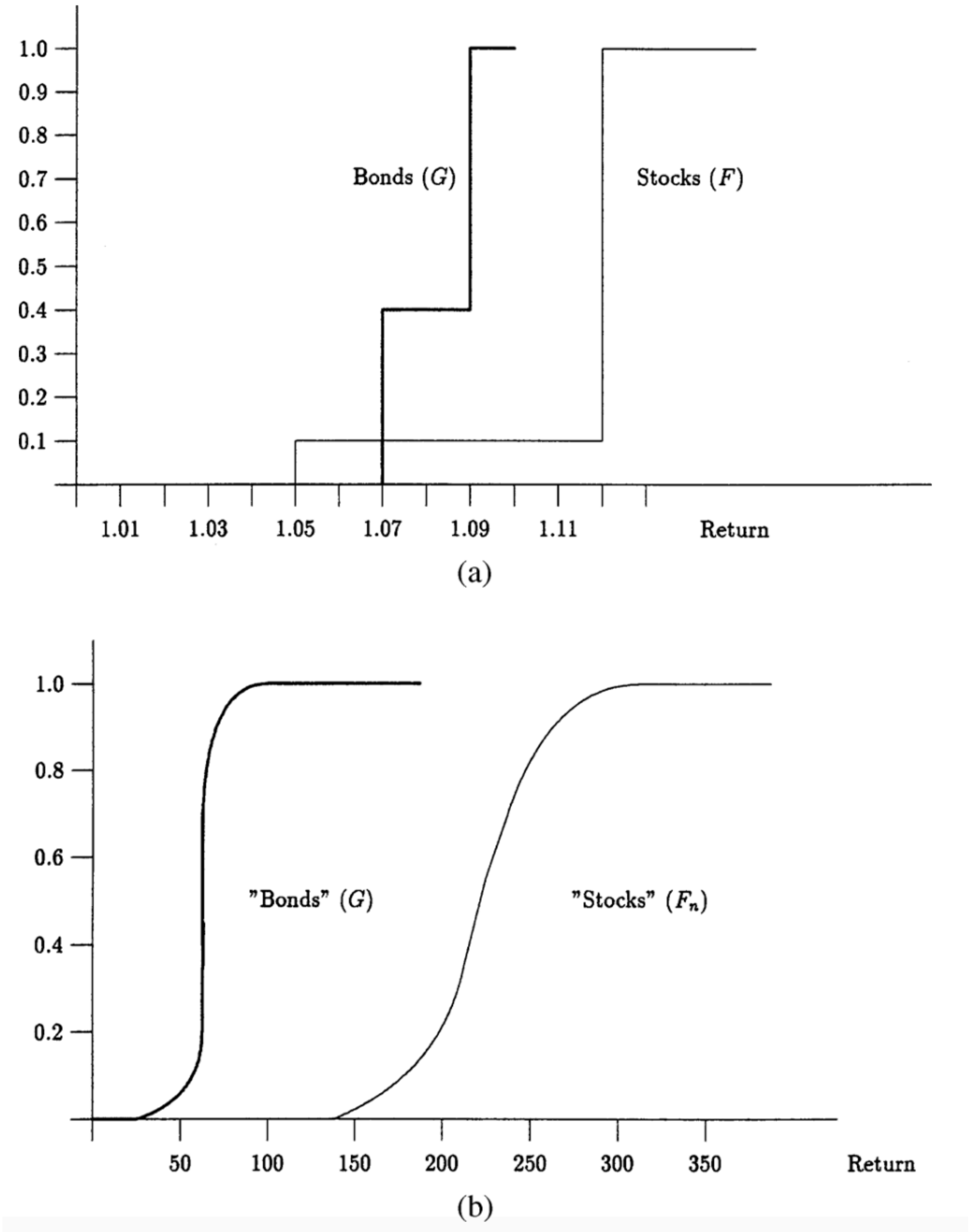
- Let X and Y be random variables with CDFs F_X and F_Y.
- **Defn:** X *stochastically dominates* Y *in the first order* if the CDF of X is entirely below the CDF of Y:

$$Y \preceq_1 X \quad \Longleftrightarrow \quad F_X(z) \leq F_Y(z), \forall z$$

- **Expected utility view:** $Y \preceq_1 X$ *if and only if*

$$\mathbb{E}[u(Y)] \leq \mathbb{E}[u(X)]$$

for every increasing function u.



SD visualized in CDFs:

(a) Bond $\not\preceq_1$ Stock; (b) Bond $\preceq_1$ Stock.

Plot: Leshno & Levy (2002).

Testing for upside = Testing the 1-SD null

- Suppose we observe pairs of observations $(X_1, Y_1), (X_2, Y_2), \ldots \sim \mathbb{P}_{XY}$.

- Define the (nonparametric and composite) null hypothesis that X dominates Y:

$$H_0 = \{\mathbb{P} : Y \preceq_1 X \text{ under } \mathbb{P}\} \quad \text{vs.} \quad H_1 = \{\mathbb{P} : Y \npreceq_1 X \text{ under } \mathbb{P}\}$$

- The null hypothesis H_0 says that Y **has no upside over** X over the entire support ($z \in \mathbb{R}$).

- **Rejecting $H_0 \equiv$ there is an upside with which Y has an advantage over X.**

E-values &
Sequential, anytime-valid inference (SAVI)

Ville, Wald, Kelly, Robbins, Cover, Vovk, Shafer, Grünwald, Ramdas, Wang, …

E-value: "E is the new P"

- Given **n** data points $X_1, \ldots, X_n$ and a null hypothesis H_0 (possibly composite), an **e-value** $E = E_n(X_1, \ldots, X_n)$ is any non-negative random variable satisfying:

$$\mathbb{E}_{H_0}[E] \le 1.$$

e.g., $E = \prod_{i=1}^{n} \frac{q(X_i)}{p(X_i)}$

- **E-values can be used for testing H_0:** for any $\alpha \in (0, 1)$, by Markov's inequality,

$$P(E \ge 1/\alpha) \le \alpha, \quad \forall P \in H_0.$$

- **For intersection nulls, e-values can be combined easily under arbitrary dependence:** Given e-values $E^{(z)}$ for $H_0^{(z)}$, their **(weighted) average** is an e-value for $H_0 = \cap_{z \in Z} H_0^{(z)}$:

$$\mathbb{E}_{H_0}\left[\sum_{z \in Z} w^{(z)} E^{(z)}\right] = \sum_{z \in Z} w^{(z)} \mathbb{E}_{H_0}\left[E^{(z)}\right] \le 1.$$

A key benefit of using e-values over p-values!

Ramdas et al. (2020; 2022)

E-processes quantify evidence in sequential experiments

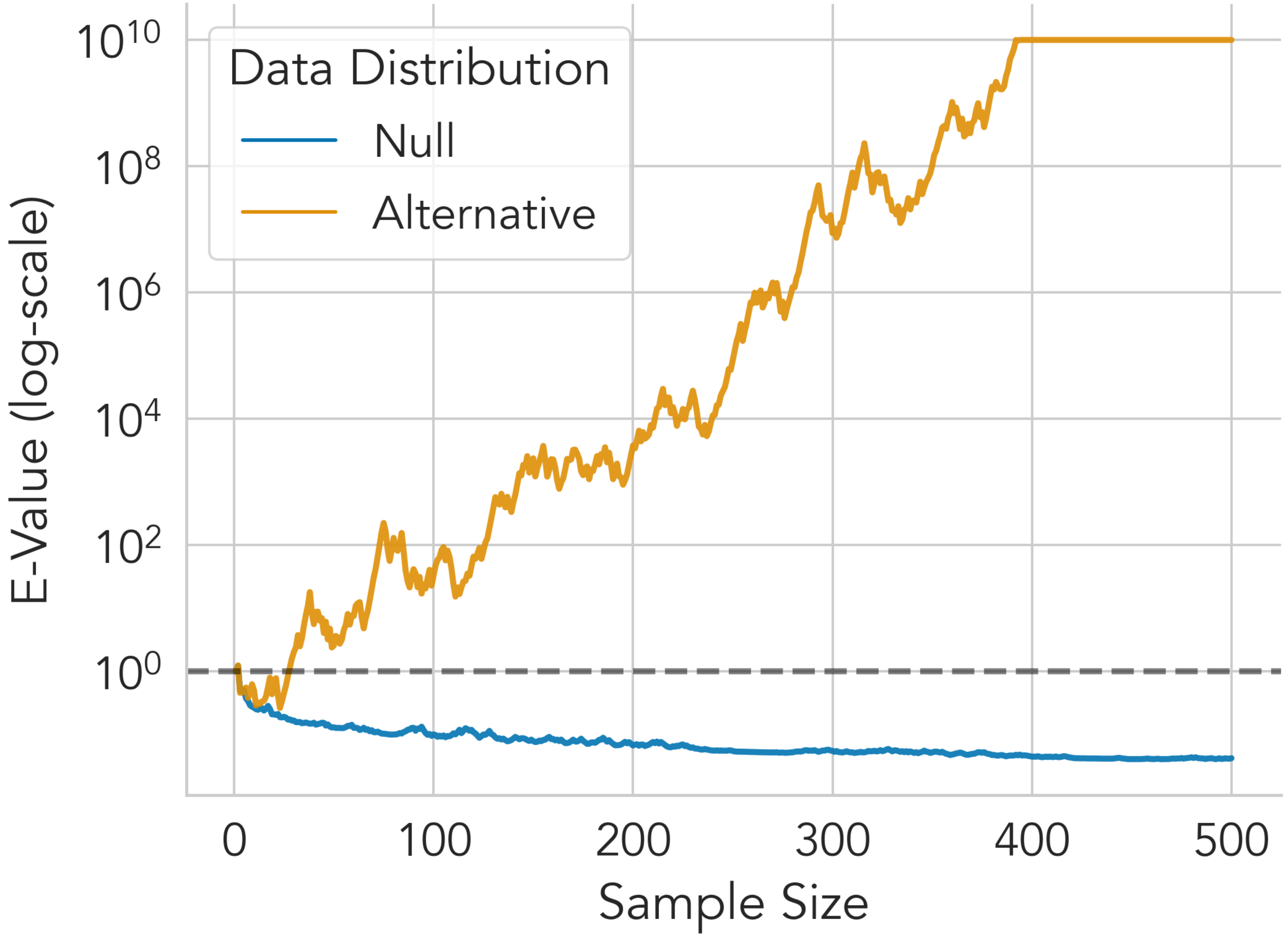
E-process $(E_t)_{t \geq 0}$

A non-negative process for H_0

For any stopping time τ,

$$\mathbb{E}_{H_0}[E_\tau] \leq 1.$$

“ANYTIME-VALIDITY”



An e-process is expected to be small under the null.

We want it to grow large under the alternative.

Ville's inequality: From e-processes to sequential tests

Jean Ville

- Let $\alpha \in (0, 1)$ be any significance level.
- **Ville's inequality** for test martingales & e-processes:

$$P\left(\exists t \geq 1 : E_t \geq 1/\alpha\right) \leq \alpha, \ \forall \alpha \in (0,1).$$

- This is **equivalent** to a time-uniform guarantee for sequential testing:

$$P\left(\exists t \geq 1 : E_t \geq 1/\alpha\right) \leq \alpha, \ \forall \alpha \in (0,1).$$

Betting on bets:
An optimal e-variable for testing SD nulls

First step: Decompose the null hypothesis

$$H_0 = \{\mathbb{P} : \mathbb{P}(X \leq z) \leq \mathbb{P}(Y \leq z), \forall z\} = \{\mathbb{P} : Y \preceq_1 X \text{ under } \mathbb{P}\}.$$

- This is a highly composite family of distributions. However, notice it is an intersection null:

$$H_0 = \bigcap_z H_0(z), \quad \text{where} \quad H_0(z) = \{\mathbb{P} : \mathbb{P}(X \leq z) \leq \mathbb{P}(Y \leq z)\}.$$

- Generally, if we have an e-value $E(z)$ for each null $\mathscr{H}_0(z)$, then **any mixture** (i.e., weighted average) of the e-values over $\mathbf{z}$ is an e-value for the intersection null:

$$E = \int E(z) d\psi(z) \text{ is an e-value for } H_0.$$

The building-block GRO e-value

- For each "threshold" $\mathbf{z} \in \mathbb{R}$, consider the individual hypothesis

$$\mathcal{H}_0(z) = \{\mathbb{P} : \mathbb{P}(X \leq z) \leq \mathbb{P}(Y \leq z)\} = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[D(z)] \leq 0\}, \text{ for } D(z) = \mathbf{1}(X \leq z) - \mathbf{1}(Y \leq z).$$

- **Growth-rate:** $\gamma_z(\lambda) = \mathbb{E}_{\mathbb{Q}}[\log S(\lambda, z)] = \mathbb{Q}(D(z) = 1)\log(1 + \lambda) + \mathbb{Q}(D(z) = -1)\log(1 - \lambda)$.

Lemma (The building-block GRO e-value: single round). Fix any threshold $\mathbf{z} \in \mathbb{R}$.

(a) For any bet $\lambda \in [0, 1]$, $S(\lambda, z) = 1 + \lambda[\mathbf{1}(X \leq z) - \mathbf{1}(Y \leq z)]$ is an **e-value** for $\mathcal{H}_0(z)$.

(b) For each alternative $\mathbb{Q} \notin \mathcal{H}_0(z)$, there is a **growth-rate optimal (GRO)** bet:

$$\lambda^\star(z) = \left[\frac{\mathbb{Q}(X \leq z < Y) - \mathbb{Q}(Y \leq z < X)}{\mathbb{Q}(X \leq z < Y) + \mathbb{Q}(Y \leq z < X)}\right]_+ = \frac{\left[\mathbb{Q}(X \leq z) - \mathbb{Q}(Y \leq z)\right]_+}{\mathbb{Q}(X \leq z < Y) + \mathbb{Q}(Y \leq z < X)}.$$

Handling the intersection hypothesis

- Now that we have an optimal e-value for each threshold $\mathbf{z} \in Z$, we can take any weighted average to obtain a valid e-value for the SD null $\mathscr{H}_0$ (intersection over z's).

$$E = \int_Z E(z) d\psi(z),$$ for any mixture distribution ψ on Z.

- Then, across multiple rounds of data (sequentially observed), we can simply **multiply** these mixtures of GRO e-values!

Main result: A powerful e-process & a test of power one for SD

- In a sequential setup: given $(t-1)$ observations, we can estimate the *GRO bet* uniformly closely with the joint empirical CDF $\hat{\mathbb{Q}}_{t-1}$. Then, calculate the (mixture of) GRO e-value for the $\mathbf{t}$-th data points. Define the resulting GRO e-process with predictable mixture $(\psi_t)_{t\in\mathbb{N}}$:

$$E_t = \prod_{\ell=1}^{t} S_\ell, \text{ where } S_\ell = \int_Z S(\hat{\lambda}^\star_\ell, z)\, d\psi_\ell(z).$$

Theorem (Anytime-validity and power of the GRO e-process).

(a) $(E_t)_{t\in\mathbb{N}}$ is an e-process for the 1-SD null H_0.

(b) For reasonable* choices of $(\psi_t)_{t\in\mathbb{N}}$, the e-process is powerful under any alternative $\mathbb{Q}$:

$$\mathbb{Q}\left(\liminf_{t\to\infty} E_t = \infty\right) = 1,$$ and the resulting sequential test has asymptotic power one.

**Eventually assign enough mass on non-dominance regions:* $\psi_t(\{z : \mathbb{Q}_X(z) - \mathbb{Q}_Y(z) > \varepsilon\}) > \delta$ *for large* t*, for some* $\varepsilon, \delta > 0$.

Choosing the predictable mixture weights

- For each threshold, we already have a GRO e-value. So, we want to make sure that the mixture weights is adaptive and do not "miss" any important (non-dominance) regions.

- The simplest choice,

$$\psi_t = \psi = \frac{1}{m}\sum_{i=1}^{m} \delta_{z_i},$$

  can work well if $\{z_1, \ldots, z_m\}$ "sufficiently cover" the non-dominance regions of interest.

- When the support is unbounded, we recommend adjusting the **m** thresholds to the empirical quantiles of observed data, and use *exponential weights* adapted to those quantiles:

$$\psi_t = \sum_{i=1}^{n_t} w_t(z_i^t)\delta_{z_i^t}, \text{ where } w_t(z_i^t) \propto \exp\left(\eta \cdot \frac{\hat{F}_X^{t-1}(z_i^t) - \hat{F}_Y^{t-1}(z_i^t)}{\hat{\sigma}_{t-1}}\right)$$

*What do you mean by "betting on bets"?

- Consider two uncertain prospects (really, bets) X and Y, say the returns of S&P 500 and Bitcoin.

- Given a threshold (say, $\mathbf{z} = 0\,\%$), Forecaster claims that

  "the probability of X *falling below* $\mathbf{z}$ *is smaller than that of* Y *falling below* $\mathbf{z}$*"*

- Forecaster then proposes a **double-or-nothing-or-push bet** with the following rules:

  - If $X \leq z$ but $Y > z$ (i.e., $D(z) = +1$), then you **double** your money;

  - If $X \leq z$ but $Y > z$ (i.e., $D(z) = -1$), then you **lose** your money;

  - Otherwise, nothing happens **("push")**.

- Skeptic places λ-fraction of her money on this bet, where $\lambda \in [0, 1]$.

Skeptic makes precisely the building block e-value: $S(\lambda, z) = 1 + \lambda D(z)$.

Testing higher-order SD

Higher-order SD

- More generally, **X** *stochastically dominates* **Y** *in the* **k**-*th order* if

$$\mathbb{E}[u(Y)] \le \mathbb{E}[u(X)] \text{ for every “utility” function } \mathbf{u} \in \mathscr{U}^{[k]}.$$

- *E.g.*, 2nd order utility class $\mathscr{U}^{[2]}$ consists of all **increasing & concave** functions (risk-averse DMs).
- Generally, the **k**-th order utility class consists of functions that alternate signs in their first **k** derivatives.
  - 3rd order models “prudent” DMs (marginal utility is convex): implies positive skew
  - 4th order captures “temperate” DMs (reluctance to accept additional risk): implies thinner tails
- **Testing for higher-order SD = Testing for upside by a DM with a partially specified utility function.**

Characterizing k-th order SD with generators

- Formally, X *stochastically dominates* Y *in the* k*-th order* **(k-SD)**, or $Y \preceq_k X$, if

$$F_X^{[k]}(z) \le F_Y^{[k]}(z), \quad \forall z \in \mathbb{R}, \quad \text{where} \quad F^{[k]}(z) = \int_{-\infty}^{z} F^{[k-1]}(u)du.$$

(we set $F^{[1]} \equiv F$).

Lemma (Characterizations of k-SD). The following statements are equivalent:

(a) $Y \preceq_k X$, or $F_X^{[k]}(z) \le F_Y^{[k]}(z)$ for all $z \in \mathbb{R}$.

(b) [Utility View] $\mathbb{E}[u(Y)] \le \mathbb{E}[u(X)]$ for any k-th order utility function $u \in \mathscr{U}^{[k]}$.

(c) [Generator View] $\mathbb{E}[(z-X)_+^{k-1}] \le \mathbb{E}[(z-Y)_+^{k-1}]$ **for all** $z \in \mathbb{R}$.

Extension via the integral identity

- Proof of the generator characterization is based on the following identity (via Fubini):

$$F^{[k]}(z) = \frac{1}{(k-1)!}\mathbb{E}\left[(z-X)_+^{k-1}\right].$$

- Thus, we can rewrite **the k-SD null hypothesis** as another intersection null over **z**!

$$\mathscr{H}_0^{[k]} = \bigcap_{z\in\mathbb{R}} \mathscr{H}_0^{[k]}(z), \quad \text{where} \quad \mathscr{H}_0^{[k]}(z) = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[(z-X)_+^{k-1}] \le \mathbb{E}_{\mathbb{P}}[(z-Y)_+^{k-1}]\}.$$

Main result #2: An asymptotically powerful e-process using generators

Assumption. The joint support of X and Y are bounded from below: $Z = [a, \infty)$.

- Define the k-th order, *normalized* utility generator for each $z \in \mathbb{R}$ (always within $[-1, 0]$):

$$u_z(x) = -\left[\frac{(z-x)_+}{z-a}\right]^{k-1}, \; x \in Z.$$

- Define the *generalized* bet $S^{[k]}(\lambda, z) = 1 + \lambda[u_z(Y) - u_z(X)]$.

Theorem (Validity of the generator-based e-variable/process for k-SD nulls).

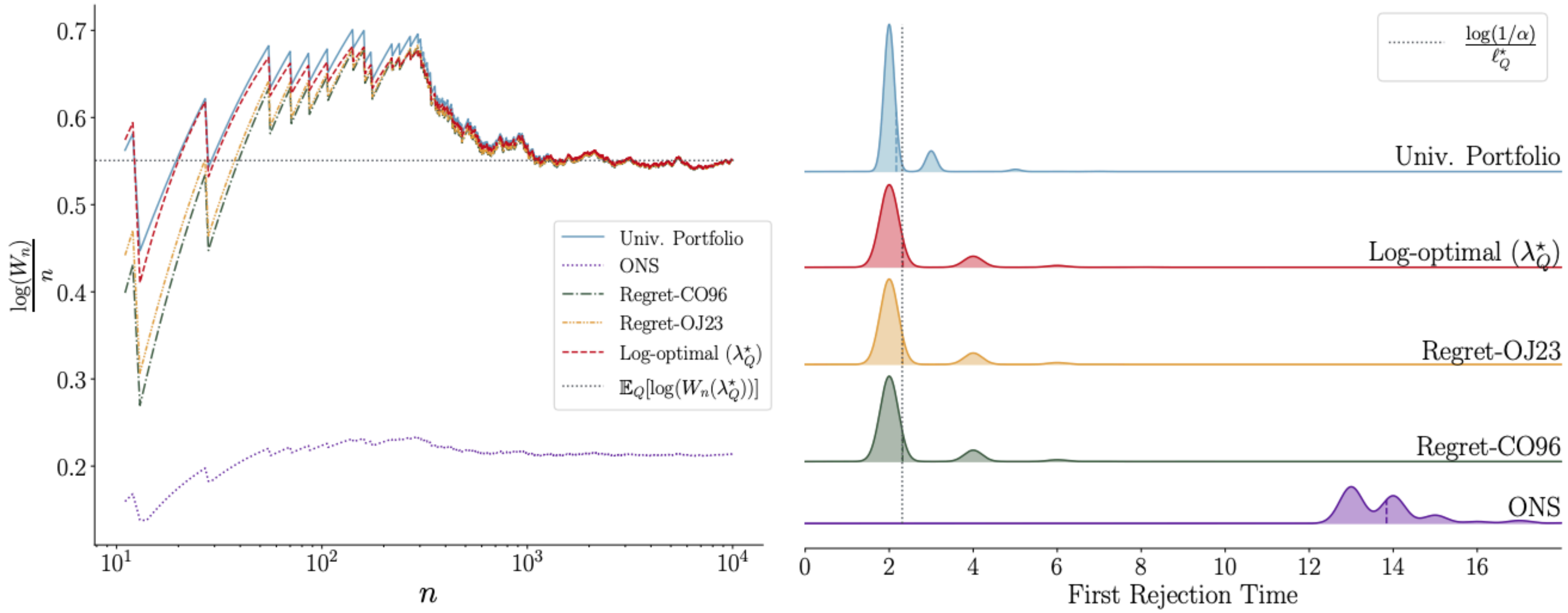
(a) For each $z \in Z$ and $\lambda \in [0, 1]$, $S^{[k]}(\lambda, z)$ is an e-variable for $\mathscr{H}_0^{[k]}$.

(b) For any predictable mixture $(\psi_t)_{t\in\mathbb{N}}$ and bets $(\lambda_t(z))_{t\in\mathbb{N}}$, we have an e-process for $\mathscr{H}_0^{[k]}$:

$$E_t^{[k]} = \prod_{\ell=1}^{t} S_\ell^{[k]}, \text{ where } S_\ell^{[k]} = \int_Z S^{[k]}(\lambda_\ell(z), z) d\psi_\ell(z).$$

*How do we choose the bets?

- For k-SD, we don't have a growth-rate optimal (GRO) bet; bet outcome is no longer ternary.
- But we have a *bounded* outcome, and many betting strategies that we know would work.
- For each threshold **z**, we default to **universal portfolio bets $\lambda_t^{\mathrm{UP}}(\mathbf{z})$**, which are asymptotically log-optimal and achieves sublinear portfolio regret (Waudby-Smith et al., 2025, figure below).



Simulations

Baselines

Denote $\Delta_{t-1}(z) = \sum_{\ell=1}^{t-1} D_{\ell}(z) = t\left[\hat{F}_X^{t-1}(z) - \hat{F}_Y^{t-1}(z)\right]$, and $V_{t-1}(z) = \sum_{\ell=1}^{t-1} D_{\ell}^2(z)$.

- **AdaGRO-Exp:** exponential weights with self-normalization: $w_t(z) \propto \exp\left\{\eta \cdot \Delta_{t-1}(z)/\sqrt{V_{t-1}(z)}\right\}$
- **AdaGRO-Hedge:** exponential weights without self-normalization: $w_t(z) \propto \exp\left\{\eta \cdot \Delta_{t-1}(z)\right\}$
- **AdaGRO-Linear:** linear weights using GRO bets themselves: $w_t(z) \propto \hat{\lambda}_t^{\star}(z)$
- **GRO:** non-adaptive GRO baseline where $\psi_t = \psi = \frac{1}{m}\sum_{i=1}^{m} \delta_{z_i}$
- **Constant:** non-adaptive, constant-bet (non-GRO) baseline with $\lambda_t(z) = 0.1$

Simulation #1: Finite support with anti-monotonicity

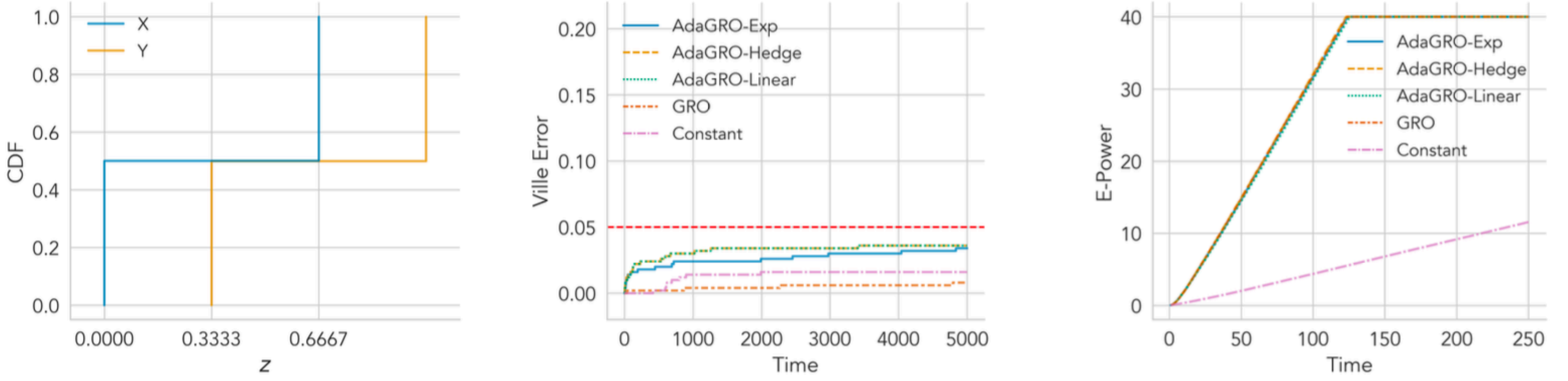
$$\mathbb{P}(X = 0, Y = 1) = 1/2$$

$$\mathbb{P}(X = 2/3, Y = 1/3) = 1/2$$

$$(\rho(X, Y) = -1)$$

*Ville Error: $\hat{\mathbb{P}}(\exists t \geq 1 : E_t \geq 1/\alpha)$

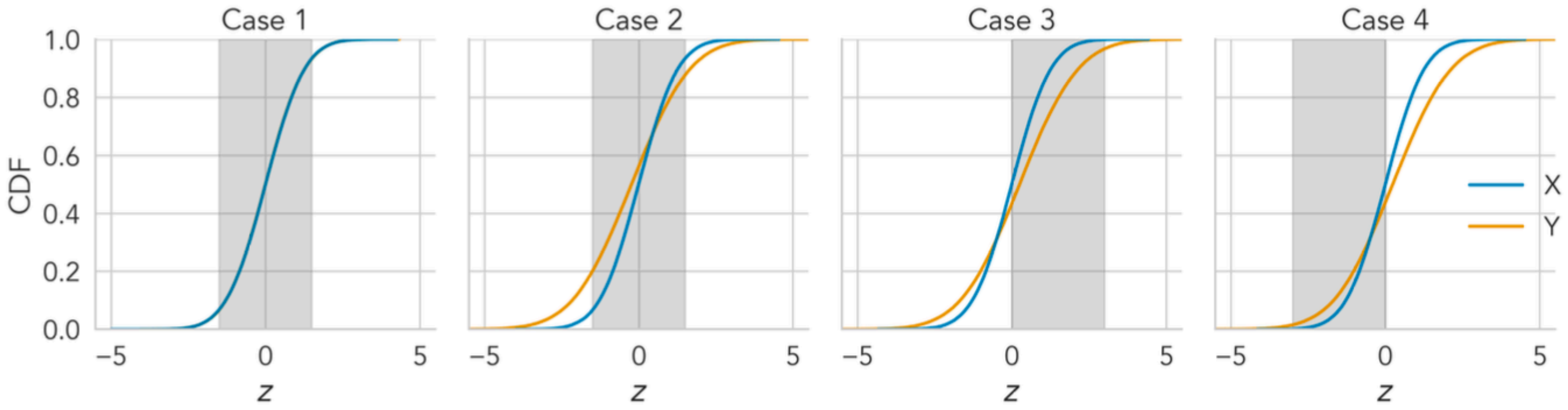
*E-Power: $\hat{\mathbb{E}}[\log E_t]$



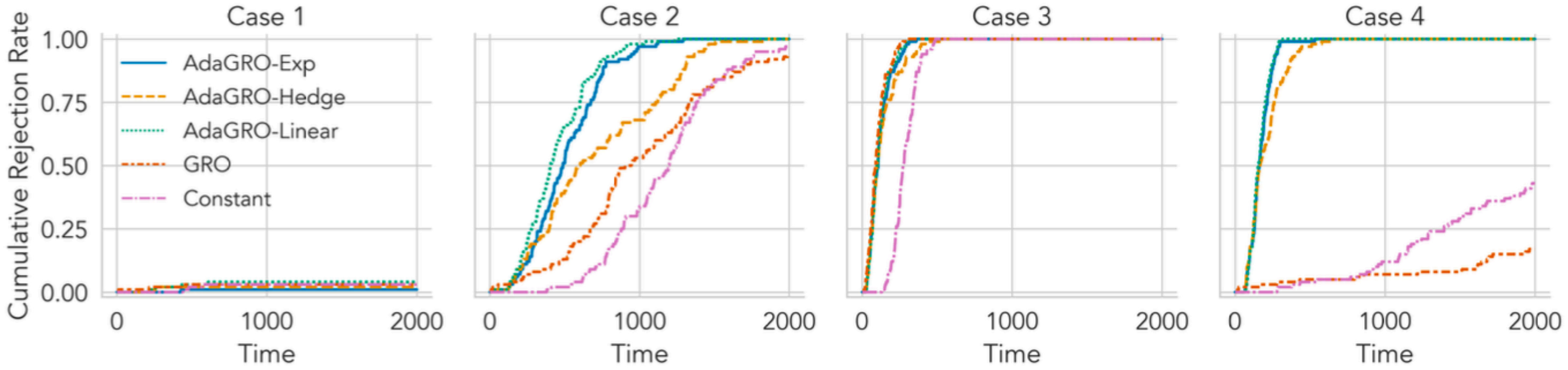
(a) Marginal CDFs, F_X and F_Y. (b) Ville error for $\mathcal{H}_0 : X \preceq_1 Y$. (c) E-power against $\mathcal{H}_0 : Y \preceq_1 X$.

Simulation #2: Adaptivity to non-dominance regions

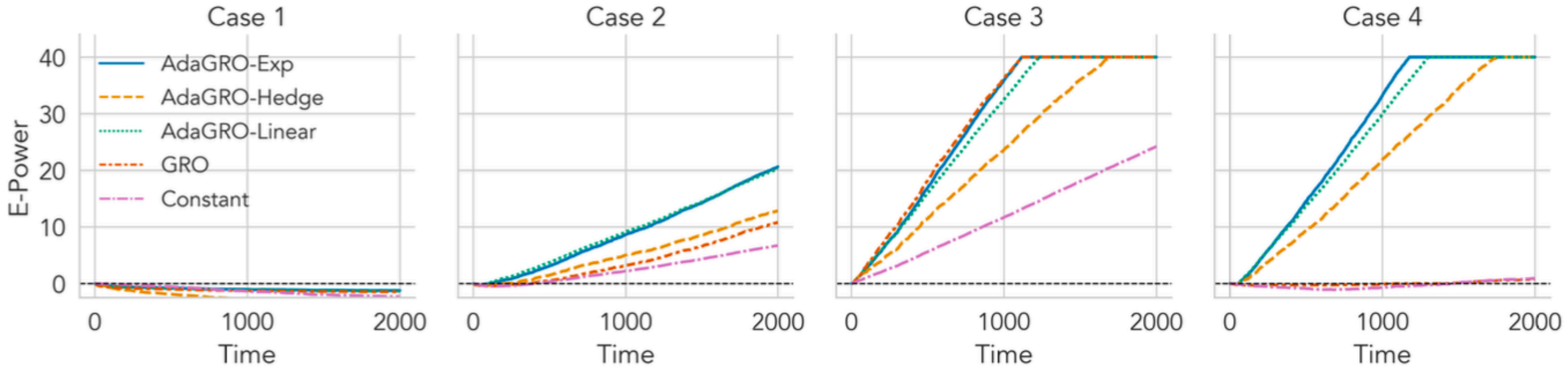
	Case 1	Case 2	Case 3	Case 4
F_X F_Y	$\mathcal{N}(0,1)$ $\mathcal{N}(0,1)$	$\mathcal{N}(0,1)$ $\mathcal{N}(-0.25,1.5^2)$	$\mathcal{N}(0,1)$ $\mathcal{N}(0.25,1.5^2)$	$\mathcal{N}(0,1)$ $\mathcal{N}(0.25,1.5^2)$
$\tilde{\mathcal{Z}}_0$	$[-1.5,1.5]$	$[-1.5,1.5]$	$[0.0,3.0]$	$[-3.0,0.0]$
$Y \preceq_1 X$?	Yes	No	No	No



(a) CDFs of X and Y. The shaded area indicates the initial search interval $\tilde{\mathcal{Z}}_0$.

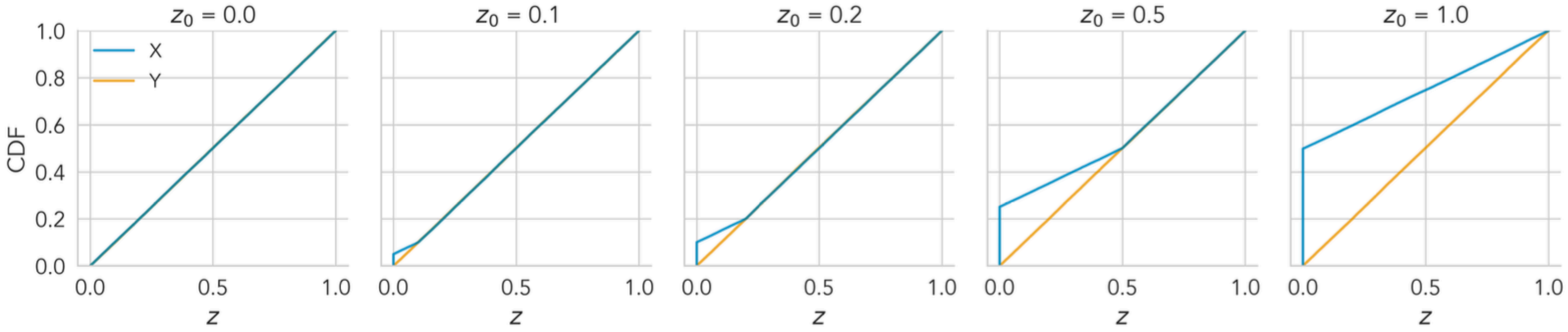


(b) Cumulative rejection rates for testing $\mathcal{H}_0 : Y \preceq_1 X$ (FSD) at $\alpha = 0.05$. For Case 1 (null case), this is the Ville error, which should be controlled at level α at all times.

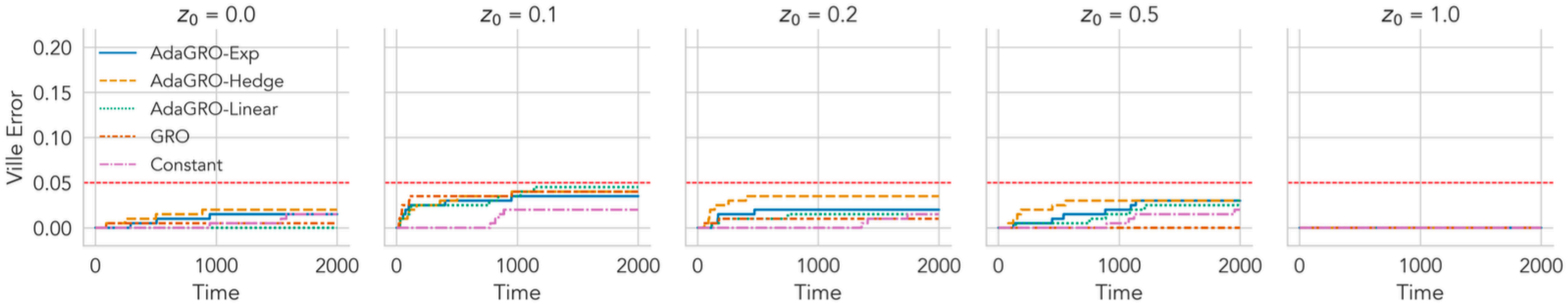


(c) E-powers against $\mathcal{H}_0 : Y \preceq_1 X$ (FSD).

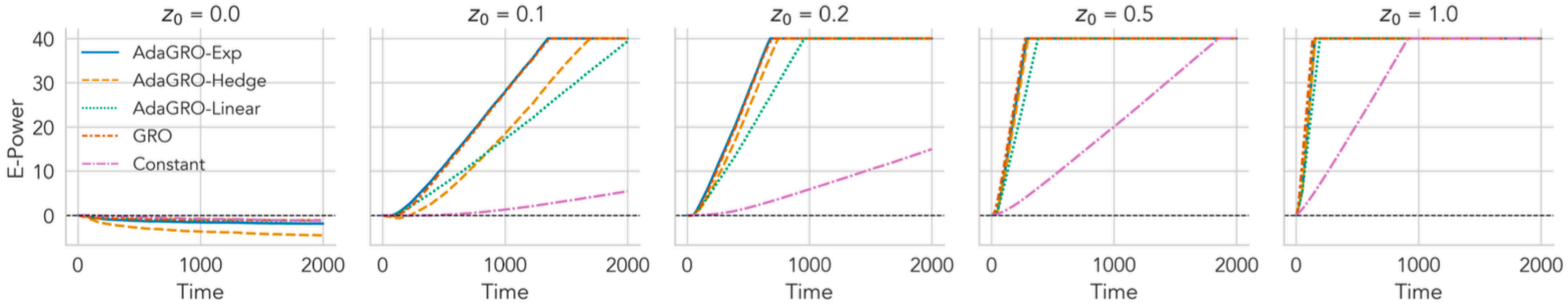
Simulation #3a: Robustness to “contact sets” between CDFs



(a) CDFs of X and Y for $z_0 \in \{0.0, 0.1, 0.2, 0.5, 1.0\}$.

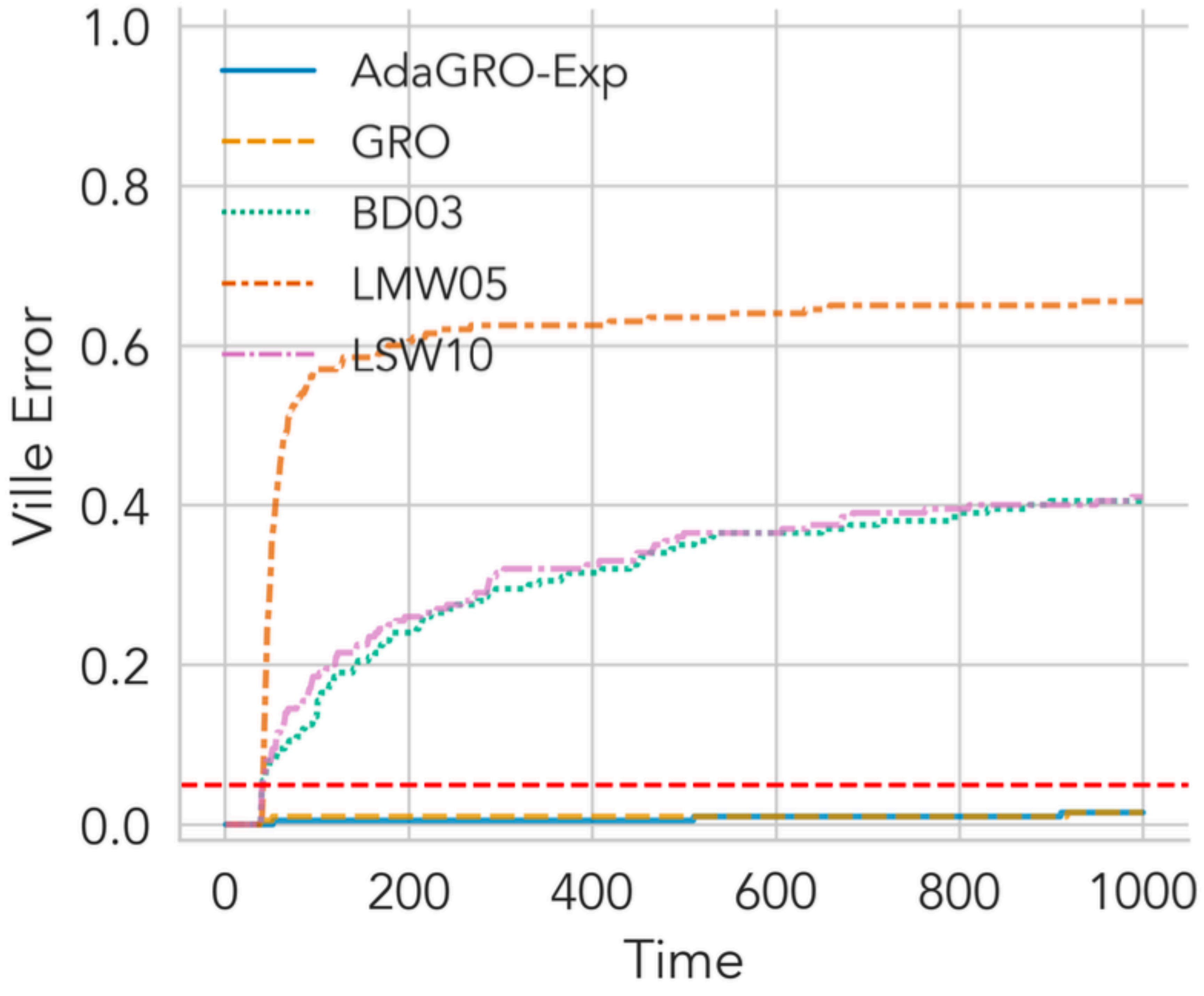


(b) Ville error for testing $\mathcal{H}_0 : X \preceq_1 Y$.

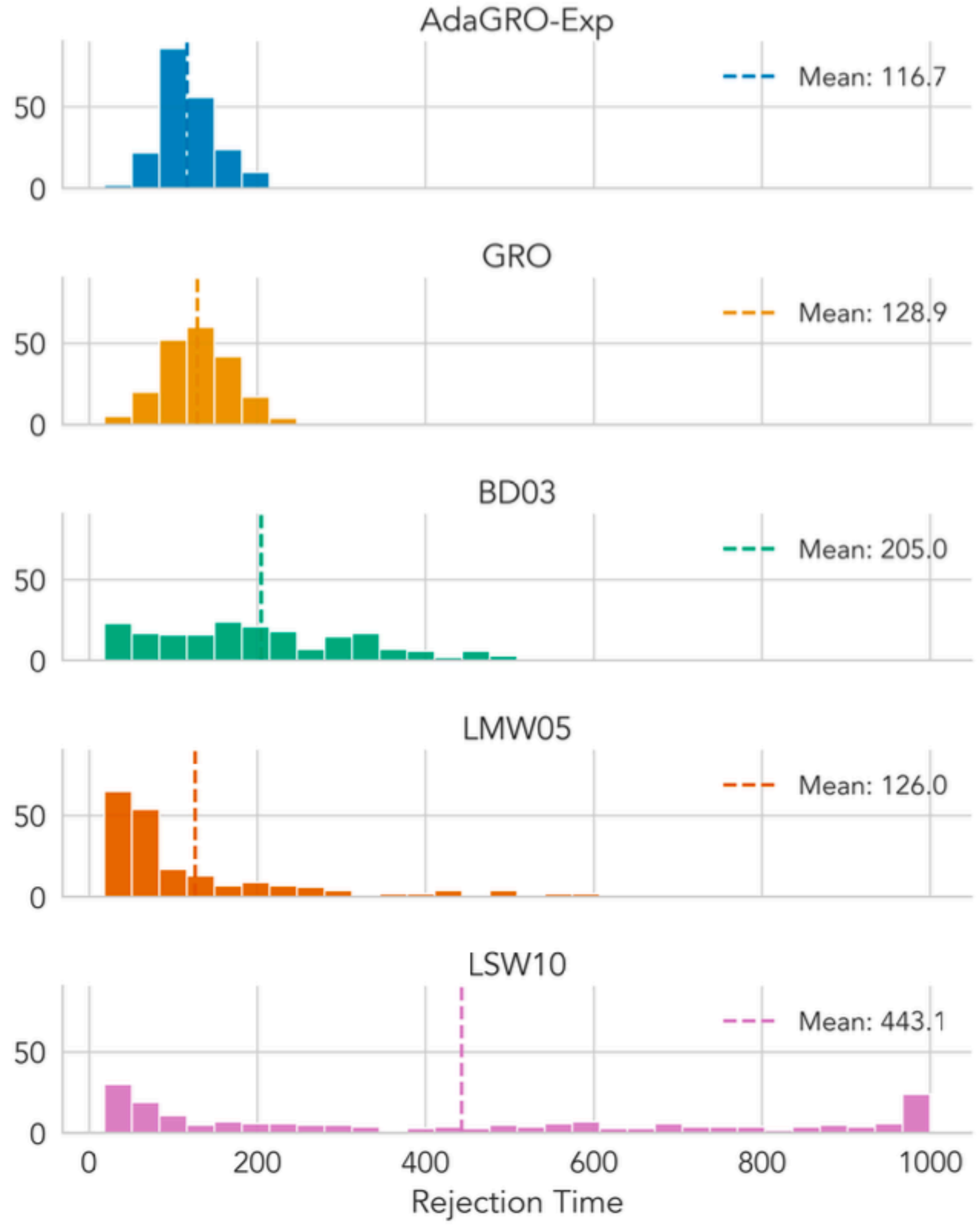


(c) E-power against $\mathcal{H}_0 : Y \preceq_1 X$.

Simulation #3b: Comparison with classical, non-SAVI tests

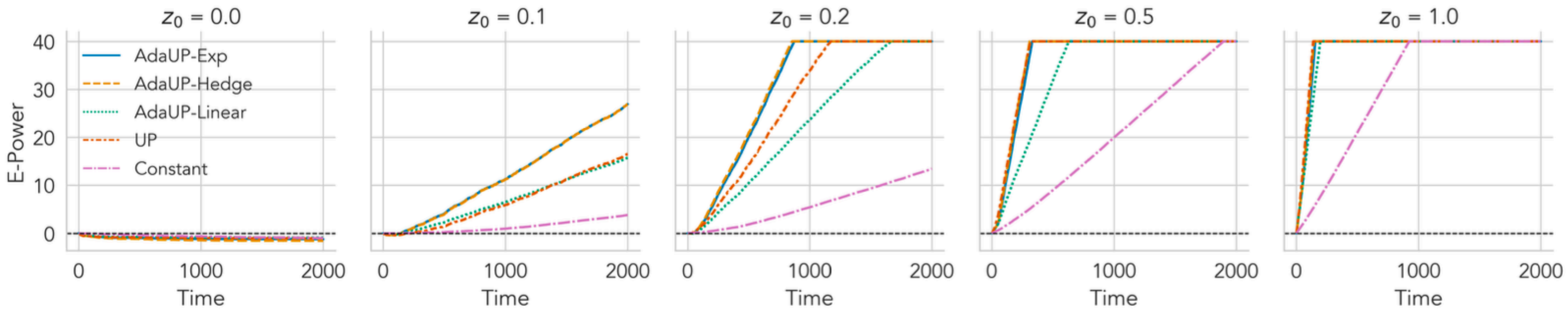


(a) Ville error under $\mathcal{H}_0$.

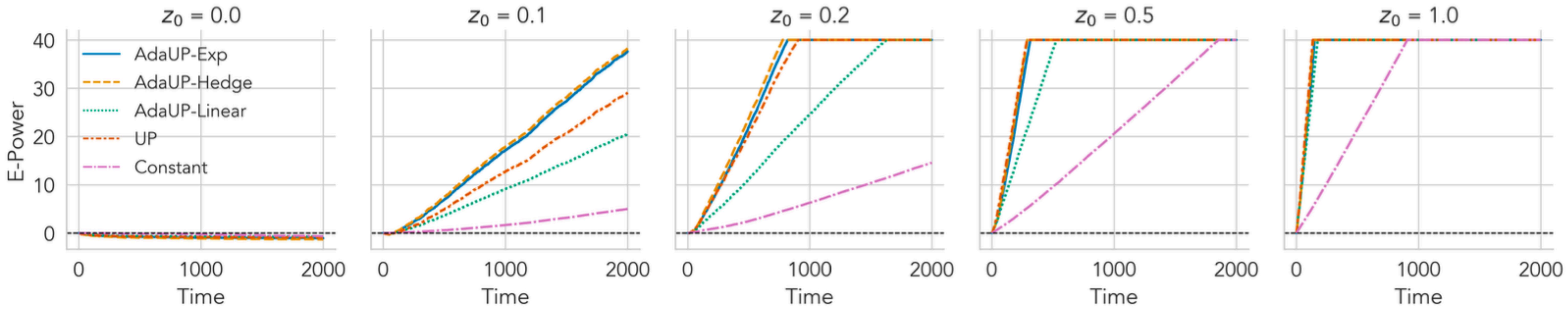


(b) Rejection times under $\mathcal{H}_1$.

Simulation #3c: Testing 2-SD and 3-SD nulls

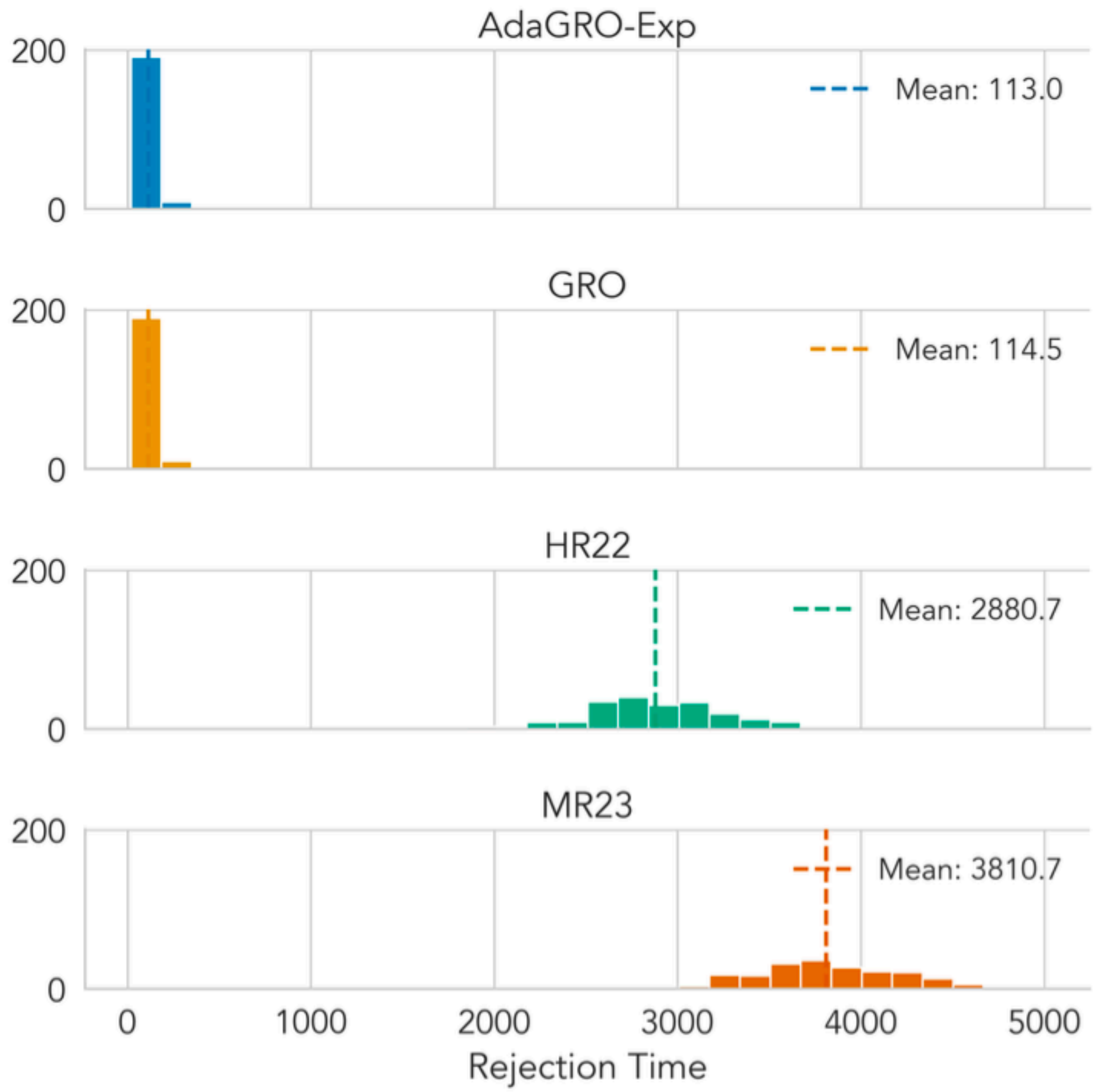


(a) E-power against $\mathcal{H}_0 : Y \preceq_2 X$ (2-SD).

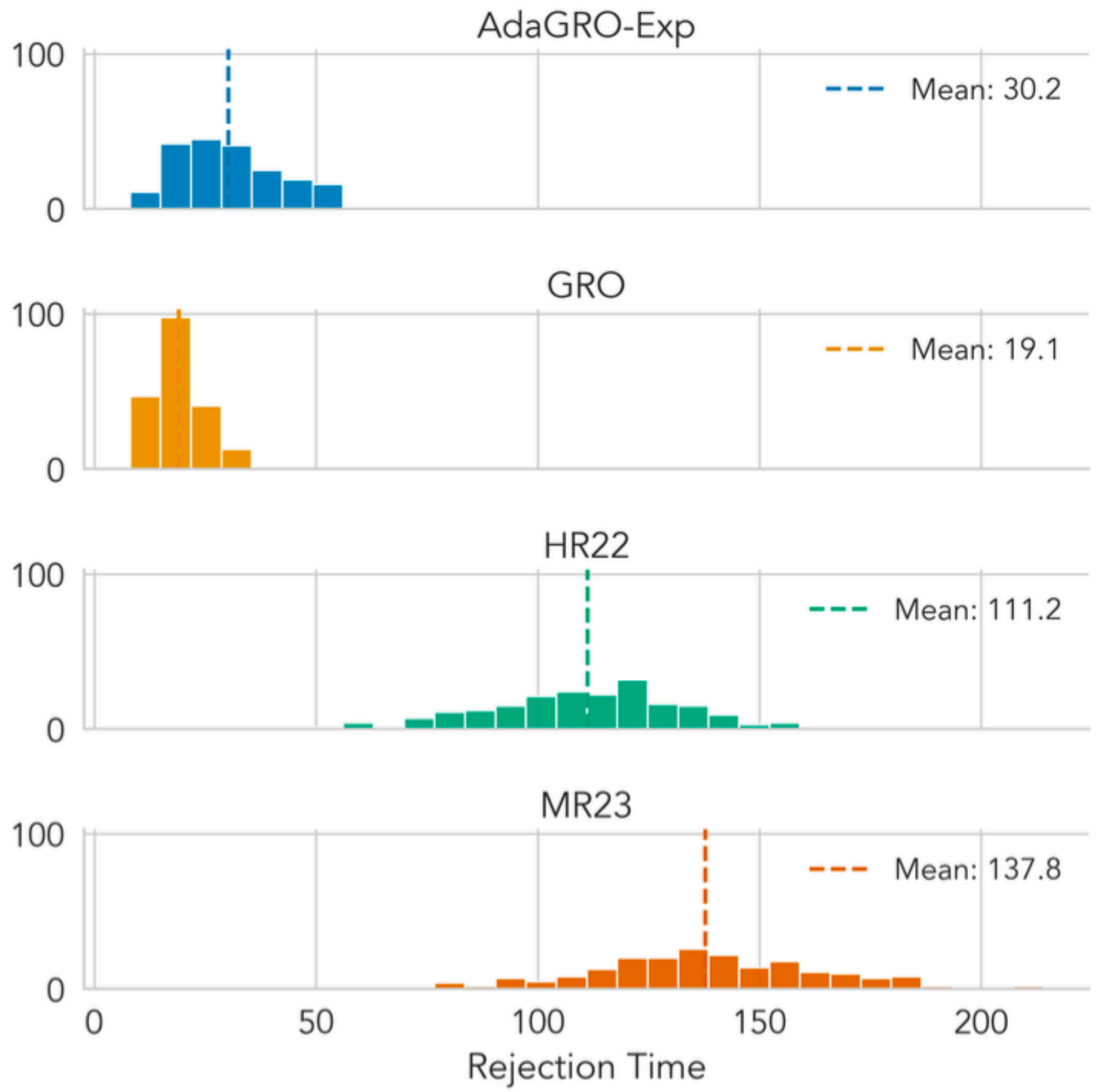


(b) E-power against $\mathcal{H}_0 : Y \preceq_3 X$ (3-SD).

Simulation #3d: Comparison with time-uniform CDF bands



(a) Substantial contact between CDFs ($z_0 = 0.2$).



(b) No contact between CDFs ($z_0 = 1.0$).

Affirming SD (i.e., *definite* upside): Testing the non-dominance null

An impossibility result

- Assume we want to show the stronger claim that Y has a *definite upside* over X.
- Im this case, we want to reject the the *non-SD null (in first order)*

$$\mathscr{H}_0' = \{\mathbb{P} : \mathbb{P}(X \leq z) < \mathbb{P}(Y \leq z) \text{ for some } z\},$$

a ***non-convex*** union-hypothesis.

Corollary (No nontrivial e-variable exists for the non-SD null). There exists no e-variable E for $\mathscr{H}_0'$ which is nontrivial for all $\mathbb{Q} \in (\mathscr{H}_0')^c$ (that is, $\mathbb{E}_{\mathbb{Q}} \log E > 0$) simultaneously.

follows from Theorem 6.2 by Zhang et al. (2024)

The minimum GRO e-value (finite support)

- Assume **finite** (and known) support and add separation from the null:

$$\mathcal{Q}(\varepsilon) = \{\mathbb{Q} : \mathbb{Q}(X \leq z_i) > \mathbb{Q}(Y \leq z_i) + \varepsilon, \text{ for all } z = z_1, \ldots, z_{m-1}\} .$$

Proposition (Validity and asymptotic optimality of the min-approach)

(a) $S = \min_{i=1,\ldots,m-1} S(\lambda_i, z_i)$ is an e-value for $\mathcal{H}'_0$, for any bets $\lambda_i \in [0,1]$,

(b) For the predictable GRO plug-in, $\mathfrak{e}_t = \min_{i=1,\ldots,m-1} \left(\prod_{\ell=1}^{t} S_\ell \right) \to \infty$, for any $\mathbb{Q} \in \mathcal{Q}(\varepsilon)$,

and we have the asymptotically optimal growth-rate: $\liminf_{t\to\infty} \frac{1}{t}(\log \mathfrak{e}_t - \log E_t) \geq 0$.

Affirming FSD for continuous outcomes

- Continuous CDFs F_X and F_Y on Z will *always* cross at the boundaries/tails.

- We have to restrict the domain of interest to $\tilde{\mathsf{Z}} \subseteq \mathsf{Z}$ (Davidson and Duclos, 2013).

- Again, consider *separated* (and *restricted)* alternative

$$\mathcal{Q}(\varepsilon, \tilde{\mathsf{Z}}) = \{\mathbb{Q} : \mathbb{Q}(X \leq z_i) > \mathbb{Q}(Y \leq z_i) + \varepsilon, \text{ for all } z \in \tilde{\mathsf{Z}}\}.$$

Proposition. We can construct sequential tests for the non-SD null that achieve *asymptotic power one* against alternatives in $\mathbb{Q} \in \mathcal{Q}(\varepsilon, \tilde{\mathsf{Z}})$.

- These tests are based on *time-uniform CDF bands*, see e.g. Howard and Ramdas (2022), Mineiro and Howard (2023), or Clerico et al (2026).

Discussion

Discussion

- **We developed a novel family of e-processes & sequential tests for various SD notions.**
  - *These methods are fully nonparametric and robust to dependence between X and Y.*
  - *The e-based approach performs competitively in power with classical, non-SAVI methods.*
- **We discuss potential extensions to non-i.i.d., unpaired, and multivariate data.**
- **We also discuss extensions to general, integral stochastic orders.**
  - Notable examples: increasing convex orders (i.e., risk-seeking DMs) & infinite-order SD.
- **Non-SD null testing remains challenging, and there are various extensions for future work.**
  - In particular, affirming weaker notions SD (higher-order SD; almost SD; …) remains hard.

Thank you!

Any questions?

https://arxiv.org/abs/2604.21851