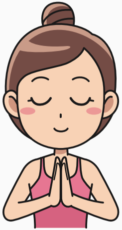
Log-Optimal Anytime-Valid E-Values

Wouter M. Koolen
CWI and University of Twente

SAVI Workshop Eindhoven, June 2, 2022

Warm Thanks



Peter Grünwald

Francesco Gili

Aaditya Ramdas

Johannes Ruf

Martin Larsson

Dmitri Ostrovskii

Manifesto

We are at SAVI

Manifesto

We are at SAVI

Safety
(no evidence against true null)

Power
(large evidence against false null)

Main Message

We are at SAVI

Main Message

We are at SAVI

I like e-values
You like e-values
We like e-values
⋱

Main Message

We are at SAVI

I like e-values
You like e-values
We like e-values
⋱

I especially like optimal e-values, e.g. GRO

Main Message

We are at SAVI

I like e-values

You like e-values

We like e-values

$\ddots$

I especially like optimal e-values, e.g. GRO

But e-values are a **batch** concept.

Main Message

We are at SAVI

I like e-values
You like e-values
We like e-values
⋱

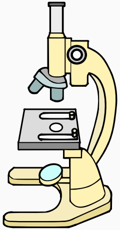
I especially like optimal e-values, e.g. GRO

But e-values are a **batch** concept.

We need an anytime analogue of GRO!

Starting Point

Setup



Goal: Keep it simple, focus on essence.

We fix

- Binary alphabet $\mathcal{X} = \{0, 1\}$
- Final time T

  $T = 2$ is interesting, and so is $T = 3$.
- Set $\mathcal{H}$ of distributions on $\mathcal{X}^T$ ("the null")

  i.i.d. case is interesting. Simple (point null) is boring
- Single distribution Q on $\mathcal{X}^T$ ("point alternative")

  i.i.d. case is interesting. Simple is interesting.

Running Example: Bernoulli Twins

Let P_θ denote i.i.d. Bernoulli(θ).

Running Example: Bernoulli Twins

Let P_θ denote i.i.d. Bernoulli(θ).

Throughout this talk we will focus on testing null

$$\mathcal{H} = \left\{P_{\frac{1}{3}}, P_{\frac{2}{3}}\right\}$$

with help of point alternative

$$Q = P_{\frac{1}{2}}$$

based on T observations.

E-value

Definition

A random variable $E : \mathcal{X}^T \to \mathbb{R}_+$ is an e-value for $\mathcal{H}$ if

$$\underset{X^T \sim P}{\mathbb{E}} \left[E(X^T) \right] \leq 1 \qquad \text{for all } P \in \mathcal{H}.$$

Example E-value

For the Bernoulli Twins case, one possible e-value is

$$E(X^T) = \frac{P_{\frac{1}{2}}(X^T)}{\max\left\{P_{\frac{1}{3}}(X^T), P_{\frac{2}{3}}(X^T)\right\}}$$

Aaditya's Universal Inference e-value

Example E-value

For the Bernoulli Twins case, one possible e-value is

$$E(X^T) = \frac{P_{\frac{1}{2}}(X^T)}{\max\left\{P_{\frac{1}{3}}(X^T), P_{\frac{2}{3}}(X^T)\right\}}$$

Aaditya's Universal Inference e-value

Does it have any power?

Optimal E-values

Definition

The GRO e-value is the maximiser of

$$\max_{E \text{ an e-value for } \mathcal{H}} \; \mathop{\mathbb{E}}_{X^T \sim Q} \left[\ln E(X^T) \right]$$

Example of GRO E-value: Bernoulli Twins case

Recall $Q = P_{\frac{1}{2}}$ and $\mathcal{H} = \left\{P_{\frac{1}{3}}, P_{\frac{2}{3}}\right\}$.

Example of GRO E-value: Bernoulli Twins case

Recall $Q = P_{\frac{1}{2}}$ and $\mathcal{H} = \left\{ P_{\frac{1}{3}}, P_{\frac{2}{3}} \right\}$.

Then the GRO e-value is

$$E^{\mathrm{GRO}}(X^T) \; = \; \frac{P_{\frac{1}{2}}(X^T)}{\frac{1}{2} P_{\frac{1}{3}}(X^T) + \frac{1}{2} P_{\frac{2}{3}}(X^T)}$$

Example of GRO E-value: Bernoulli Twins case

Recall $Q = P_{\frac{1}{2}}$ and $\mathcal{H} = \left\{P_{\frac{1}{3}}, P_{\frac{2}{3}}\right\}$.

Then the GRO e-value is

$$E^{\mathrm{GRO}}(X^T) \;=\; \frac{P_{\frac{1}{2}}(X^T)}{\frac{1}{2}P_{\frac{1}{3}}(X^T) + \frac{1}{2}P_{\frac{2}{3}}(X^T)}$$

“the RIPR prior is 50-50”

Example of GRO E-value: Bernoulli Twins case

Recall $Q = P_{\frac{1}{2}}$ and $\mathcal{H} = \left\{P_{\frac{1}{3}}, P_{\frac{2}{3}}\right\}$.

Then the GRO e-value is

$$E^{\text{GRO}}(X^T) \;=\; \frac{P_{\frac{1}{2}}(X^T)}{\frac{1}{2}P_{\frac{1}{3}}(X^T) + \frac{1}{2}P_{\frac{2}{3}}(X^T)}$$

"the RIPR prior is 50-50"

Instance of Muriel's Group Invariance (Haar prior)

Example of GRO E-value: Bernoulli Twins case

Recall $Q = P_{\frac{1}{2}}$ and $\mathcal{H} = \left\{P_{\frac{1}{3}}, P_{\frac{2}{3}}\right\}$.

Then the GRO e-value is

$$E^{\mathrm{GRO}}(X^T) = \frac{P_{\frac{1}{2}}(X^T)}{\frac{1}{2}P_{\frac{1}{3}}(X^T) + \frac{1}{2}P_{\frac{2}{3}}(X^T)}$$

"the RIPR prior is 50-50"

Instance of Muriel's Group Invariance (Haar prior)

Theorem (Grünwald, de Heide, and Koolen, 2019)

$E^{GRO}(X^T) = \frac{Q(X^T)}{P_W(X^T)}$ *for* $P_W \in \mathrm{conv}(\mathcal{H})$ *minimising* $\mathrm{KL}(Q \| P_W)$.

The good

- "Type-I" error control under $\mathcal{H}$ (Markov)
- For any e-value E, on replications

$$X_1^T, X_2^T, \ldots \qquad \text{i.i.d. from } Q$$

the LLN gives

$$\frac{1}{n}\sum_{i=1}^{n} \ln E(X_i^T) \quad \xrightarrow{\text{a.s.}} \quad \underbrace{\mathop{\mathbb{E}}_{X^T \sim Q}\left[\ln E(X^T)\right]}_{\text{maximised by GRO } E}$$

The bad

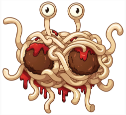
At SAVI, we are interested in optional stopping!

Problem

Recall e-value $E : \mathcal{X}^T \to \mathbb{R}_+$ *was only defined on batch* $\mathcal{X}^T$.

So $E(X^t)$ *for* $t < T$ *nonsensical.*

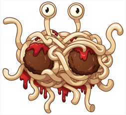
Wait, wait, wait ... Perhaps a miracle occurs?



For simple $\mathcal{H} = \{P\}$, $E^{\text{GRO}}(X^T) = \frac{Q(X^T)}{P(X^T)}$ is a P-martingale.

So $E^{\text{GRO}}(X^\tau)$ is safe (i.e. an e-value) for any stopping rule τ after all.

Wait, wait, wait . . . Perhaps a miracle occurs?

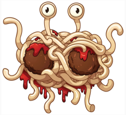


For simple $\mathcal{H} = \{P\}$, $E^{\text{GRO}}(X^T) = \frac{Q(X^T)}{P(X^T)}$ is a P-martingale.

So $E^{\text{GRO}}(X^\tau)$ is safe (i.e. an e-value) for any stopping rule τ after all.

What about composite $\mathcal{H}$?

Wait, wait, wait . . . Perhaps a miracle occurs?



For simple $\mathcal{H} = \{P\}$, $E^{\text{GRO}}(X^T) = \frac{Q(X^T)}{P(X^T)}$ is a P-martingale.

So $E^{\text{GRO}}(X^\tau)$ is safe (i.e. an e-value) for any stopping rule τ after all.

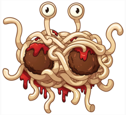
What about composite $\mathcal{H}$?

The Bernoulli Twins GRO e-value

$$E^{\text{GRO}}(X^T) \;=\; \frac{P_{\frac{1}{2}}(X^T)}{\frac{1}{2}P_{\frac{1}{3}}(X^T) + \frac{1}{2}P_{\frac{2}{3}}(X^T)}$$

had 50-50 RIPR prior independent of T. So maybe . . .

Wait, wait, wait . . . Perhaps a miracle occurs?



For simple $\mathcal{H} = \{P\}$, $E^{\text{GRO}}(X^T) = \frac{Q(X^T)}{P(X^T)}$ is a P-martingale.

So $E^{\text{GRO}}(X^\tau)$ is safe (i.e. an e-value) for any stopping rule τ after all.

What about composite $\mathcal{H}$?

The Bernoulli Twins GRO e-value

$$E^{\text{GRO}}(X^T) \;=\; \frac{P_{\frac{1}{2}}(X^T)}{\frac{1}{2}P_{\frac{1}{3}}(X^T) + \frac{1}{2}P_{\frac{2}{3}}(X^T)}$$

had 50-50 RIPR prior independent of T. So maybe . . .

. . . but **NO!** $(E^{\text{GRO}}(X^t))_t$ is not a (super)-martingale under either element of the null, and $E^{\text{GRO}}(X^\tau)$ is unsafe for some τ.

Still, are we already satisfied?

A product of sequential E-variables is an e-process.

$$E(X^m) = \prod_{i=1}^{m} E(X_i)$$

Still, are we already satisfied?

A product of sequential E-variables is an e-process.

$$E(X^m) = \prod_{i=1}^{m} E(X_i)$$

But is it a powerful E-process?

Still, are we already satisfied?

A product of sequential E-variables is an e-process.

$$E(X^m) = \prod_{i=1}^{m} E(X_i)$$

But is it a powerful E-process?

NO, it is too safe!

Still, are we already satisfied?

A product of sequential E-variables is an e-process.

$$E(X^m) = \prod_{i=1}^{m} E(X_i)$$

But is it a powerful E-process?

NO, it is too safe!

The expected stopped value is ≤ 1 even under a switching sequence of elements from null $\mathcal{H}$. "Imprecise Probability"

Still, are we already satisfied?

A product of sequential E-variables is an e-process.

$$E(X^m) = \prod_{i=1}^{m} E(X_i)$$

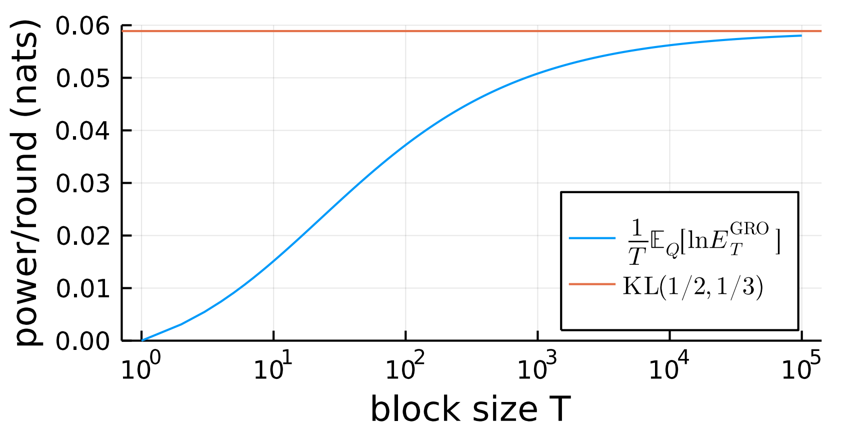
But is it a powerful E-process?

NO, it is too safe!

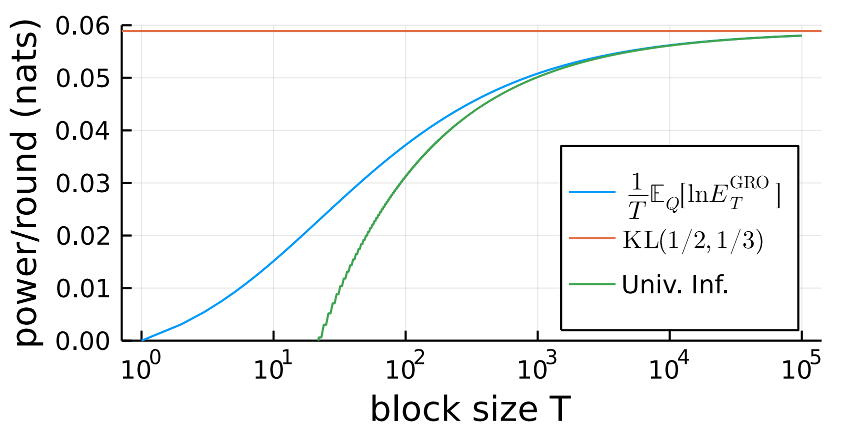
The expected stopped value is ≤ 1 even under a switching sequence of elements from null $\mathcal{H}$. “Imprecise Probability”

No Free Lunch: superfluous safety is hampering our power.

Hampering Power?



Hampering Power?



The plan

Roll up sleeves and

1. Define E on $\mathcal{X}^{\leq T}$ ("random variable" $\Rightarrow$ "random process")
2. Upgrade safety criterion ("e-value" $\Rightarrow$ "e-process")
3. Upgrade power criterion (GRO $\Rightarrow$ LOAVEV)
4. Profit

Anytime-Validity

Setup with stopping

Distribution P on $\mathcal{X}^T$ and stopping rule τ give rise to stopped distribution P^τ on $\mathcal{X}^{\leq T}$.

Setup with stopping

Distribution P on $\mathcal{X}^T$ and stopping rule τ give rise to stopped distribution P^τ on $\mathcal{X}^{\leq T}$.

Example: if τ stops after seeing a one, then $P^\tau_{\frac{1}{3}}$ can output

X^τ	$P^\tau_{\frac{1}{3}}(X^\tau)$
1	$\frac{1}{3}$
01	$\frac{2}{3}\frac{1}{3}$
001	$\left(\frac{2}{3}\right)^2 \frac{1}{3}$
0001	the remainder

Setup with stopping

To keep a point alternative yet make optional stopping interesting, we will employ randomised stopping times.

Setup with stopping

To keep a point alternative yet make optional stopping interesting, we will employ randomised stopping times.

We encode these by $\tau(X^m) \in [0, 1]$ indicating the conditional probability of stopping:

$$P^\tau(X^m) \;=\; P(X^m)\left(\prod_{i=0}^{m-1} \big(1 - \tau(X^i)\big)\right)\tau(X^m) \qquad \text{for any } X^m \in \mathcal{X}^{\leq T}\ .$$

Setup with stopping

To keep a point alternative yet make optional stopping interesting, we will employ randomised stopping times.

We encode these by $\tau(X^m) \in [0,1]$ indicating the conditional probability of stopping:

$$P^\tau(X^m) \;=\; P(X^m)\left(\prod_{i=0}^{m-1}\left(1-\tau(X^i)\right)\right)\tau(X^m) \qquad \text{for any } X^m \in \mathcal{X}^{\leq T}\ .$$

Example: if τ stops with probability $1/2$, then $P^\tau_{\frac{1}{3}}$ outputs

X^τ	$P^\tau_{\frac{1}{3}}(X^\tau)$
ϵ	$\frac{1}{2}$
0	$\frac{2}{3}\left(\frac{1}{2}\right)^2$
1	$\frac{1}{3}\left(\frac{1}{2}\right)^2$
00	$\left(\frac{2}{3}\right)^2\left(\frac{1}{2}\right)^3$
01	[illegible]

E-processes

Recall that the null $\mathcal{H}$ is a set of distributions on $\mathcal{X}^T$.

Definition (AVEV aka e-process)

A process $E : \mathcal{X}^{\leq T} \to \mathbb{R}_+$ is an anytime-valid e-value for $\mathcal{H}$ if

$$\mathop{\mathbb{E}}_{X^m \sim P^\tau} [E(X^m)] \leq 1 \qquad \text{for all } P \in \mathcal{H}, \text{ any stopping time } \tau$$

Interpretation: $E(X^\tau)$ is an e-value regardless of stopping rule imposed.

Upgrading the GRO criterion

To upgrade GRO, we need an upgraded alternative.

Upgrading the GRO criterion

To upgrade GRO, we need an upgraded alternative.

We will pick a distribution Q on $\mathcal{X}^T$ and a (randomised) stopping time σ.

Upgrading the GRO criterion

To upgrade GRO, we need an upgraded alternative.

We will pick a distribution Q on $\mathcal{X}^T$ and a (randomised) stopping time σ.

As before, this encodes our hope for what happens when null $\mathcal{H}$ is false.

Upgrading the GRO criterion

Fix $\mathcal{H}$, alternative Q on $\mathcal{X}^T$ and randomized stopping time σ.

Definition (LOAVEV)

The log-optimal anytime-valid e-value is the maximiser of

$$\max_{E \text{ an AVEV for } \mathcal{H}} \mathbb{E}_{X^m \sim Q^\sigma} [\ln E(X^m)]$$

Reduction to batch case

LOAVEV is actually standard GRO for the "arbitrarily stopped" null

$$\mathcal{H}' = \{P^\tau | P \in \mathcal{H}, \tau \text{ stopping time}\}$$

and the "σ-stopped" alternative

$$Q' = Q^\sigma$$

both defined on $\mathcal{X}^{\leq T}$.

Reduction to batch case

LOAVEV is actually standard GRO for the "arbitrarily stopped" null

$$\mathcal{H}' = \{P^\tau | P \in \mathcal{H}, \tau \text{ stopping time}\}$$

and the "σ-stopped" alternative

$$Q' = Q^\sigma$$

both defined on $\mathcal{X}^{\leq T}$.

So what do LOAVEVs look like?

Representation (1)

LOAVEVs are admissible e-processes, and hence

Theorem (Ramdas, Ruf, Larsson, and Koolen, 2020)

$$E^{LOAVEV}(X^m) \;=\; \min_{P \in \mathcal{H}} M^P(X^m)$$

where M^P is a P-martingale for each P.

Representation (2)

$$E^{\mathrm{LOAVEV}}(X^m) \;=\; \frac{Q^{\sigma}(X^m)}{\int P^{\tau}(X^m) w^*(P,\tau)\,\mathrm{d}(P,\tau)}$$

"RIPR form"

In fact, $E^{\mathrm{LOAVEV}}(X^m) = \frac{Q^{\sigma}(X^m)}{P_W(X^m)}$ where $P_W \in \mathrm{conv}\,(\mathcal{H}')$ minimises $\mathrm{KL}(Q^{\sigma} \| P_W)$

Simple special cases

Theorem (Koolen and Grünwald, 2021)

If $\mathcal{H}$ and Q are i.i.d. and the 1-outcome RIPR $\tilde{P}$ is in $\mathcal{H}$ (not a mixture) then

$$E^{LOAVEV}(X^m) = \frac{Q(X^m)}{\tilde{P}(X^m)}$$

for any σ.

The generic case

What can we hope to see?

- Perhaps

$$E^{\mathrm{LOAVEV}}(X^m) \; = \; \frac{P^{\sigma}_{\frac{1}{2}}(X^m)}{\frac{1}{2}P^{\sigma}_{\frac{1}{3}}(X^m) + \frac{1}{2}P^{\sigma}_{\frac{2}{2}}(X^m)}?$$

- Martingale-like behaviour? Betting interpretation? Explainable strategies?
- E^{LOAVEV} should retain invariances common to Q, σ and $\mathcal{H}$.
  Cases in point
  - label flips
  - exchangeability (if nobody cares about the order, why should LOAVEV)

Numerical example

Bernoulli Twins case with uniform stopping time $\sigma(x^m) = \frac{1}{T+1-m}$.

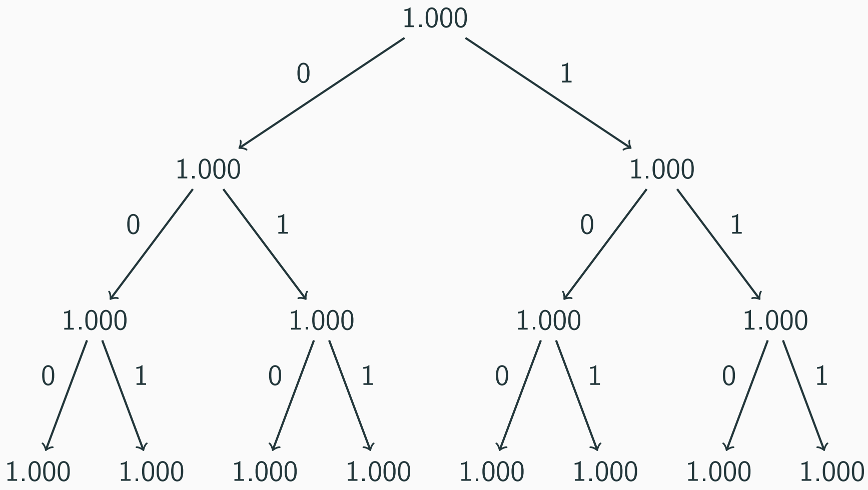
LOAVEV problem is strictly concave (solution unique)

Not fun to enumerate all deterministic stopping rules. Instead we use use rewrite based on flow representation (see paper)

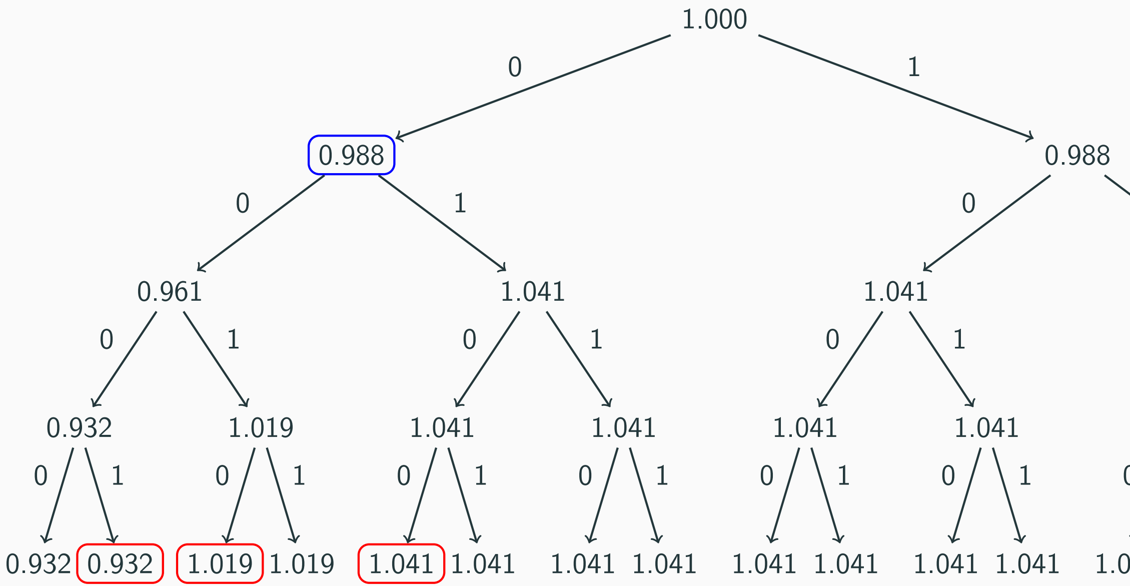
Computation with CVX convex optimisation library.

I'll show you E^{LOAVEV} for $T = 3, 4, 6$.

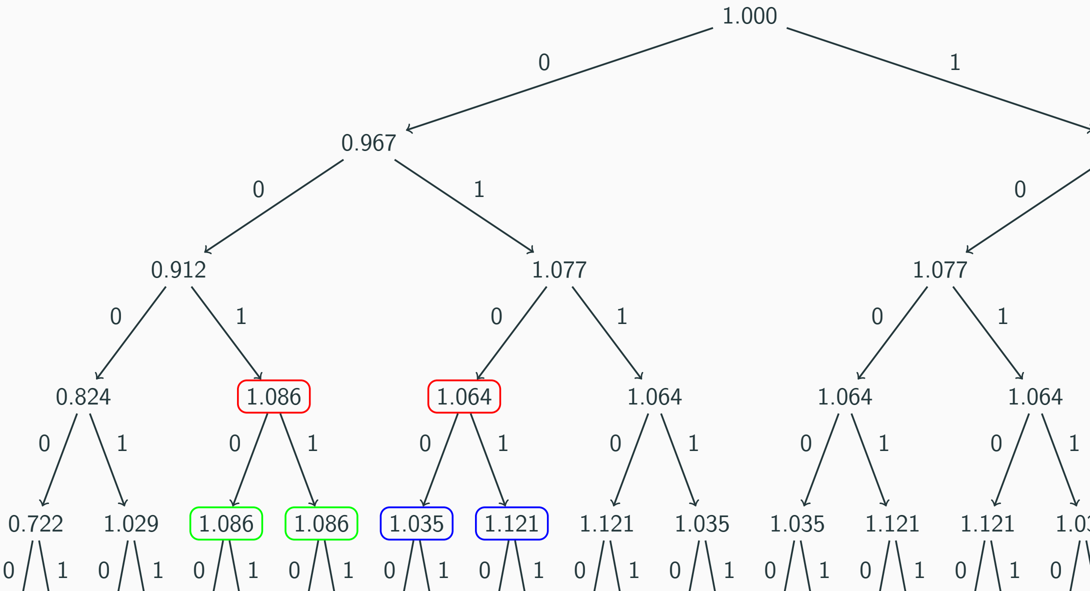
LOAVEV for Bernoulli Twins, $T = 3$ rounds



LOAVEV for Bernoulli Twins, $T = 4$ rounds



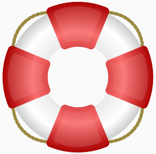
LOAVEV for Bernoulli Twins, $T = 6$ rounds



Weird stuff

- Rebounds
- Stopping
- Not based on sufficient statistic

Open Problem



Only just started

Examples/remarks/ideas/conjectures welcome!

We are very likely not smart enough to see the elegant structure of the LOAVEV.

Can you?

Conclusion

Conclusion

- Natural upgrade of GRO criterion to anytime validity
- First results
  - Trivialises in special cases
  - As-of-yet ill understood general case
  - As-of-yet computational nightmare

Thanks!

References

- Grünwald, P., R. de Heide, and W. M. Koolen (June 2019). “Safe Testing”. In: *ArXiv*.
- Koolen, W. M. and P. Grünwald (Sept. 2021). “Log-optimal anytime-valid E-values”. In: *International Journal of Approximate Reasoning*.
- Ramdas, A., J. Ruf, M. Larsson, and W. M. Koolen (Sept. 2020). “Admissible anytime-valid sequential inference must rely on nonnegative martingales”. In: *ArXiv*.