A martingale theory of evidence

Aaditya Ramdas

Dept. of Statistics and Data Science (75%)
Machine Learning Dept. (25%)
Carnegie Mellon University

Amazon Research (20%)

Ack support: NSF CAREER, Google Research Award, Adobe Research Award, Sloan Fellowship

Online learning
+ optimization
(Eg: Cover,
Orabona)

Statistics
(Eg: Grunwald,
Shafer, Vovk)

Gambling/betting

Information
Theory
(Eg: Kelly, Cover)

Probability
+ Finance
(Eg: Ville, Shafer, Vovk)

Outline of this lecture series

1. Today: game-theoretic testing
2. Tomorrow morning: game-theoretic estimation
3. Tomorrow afternoon: game-theoretic change detection

Outline of this talk

1. "Sequential, anytime-valid inference (SAVI)"
2. Testing by betting yields SAVI inference (simple example)
3. Core SAVI concepts
4. *Optimal gambling strategies* (*second half of the talk, most interesting!*)

An infamous instance of **"peeking at p-values"** is the power-posing controversy (Amy Cuddy, Dana Carney).

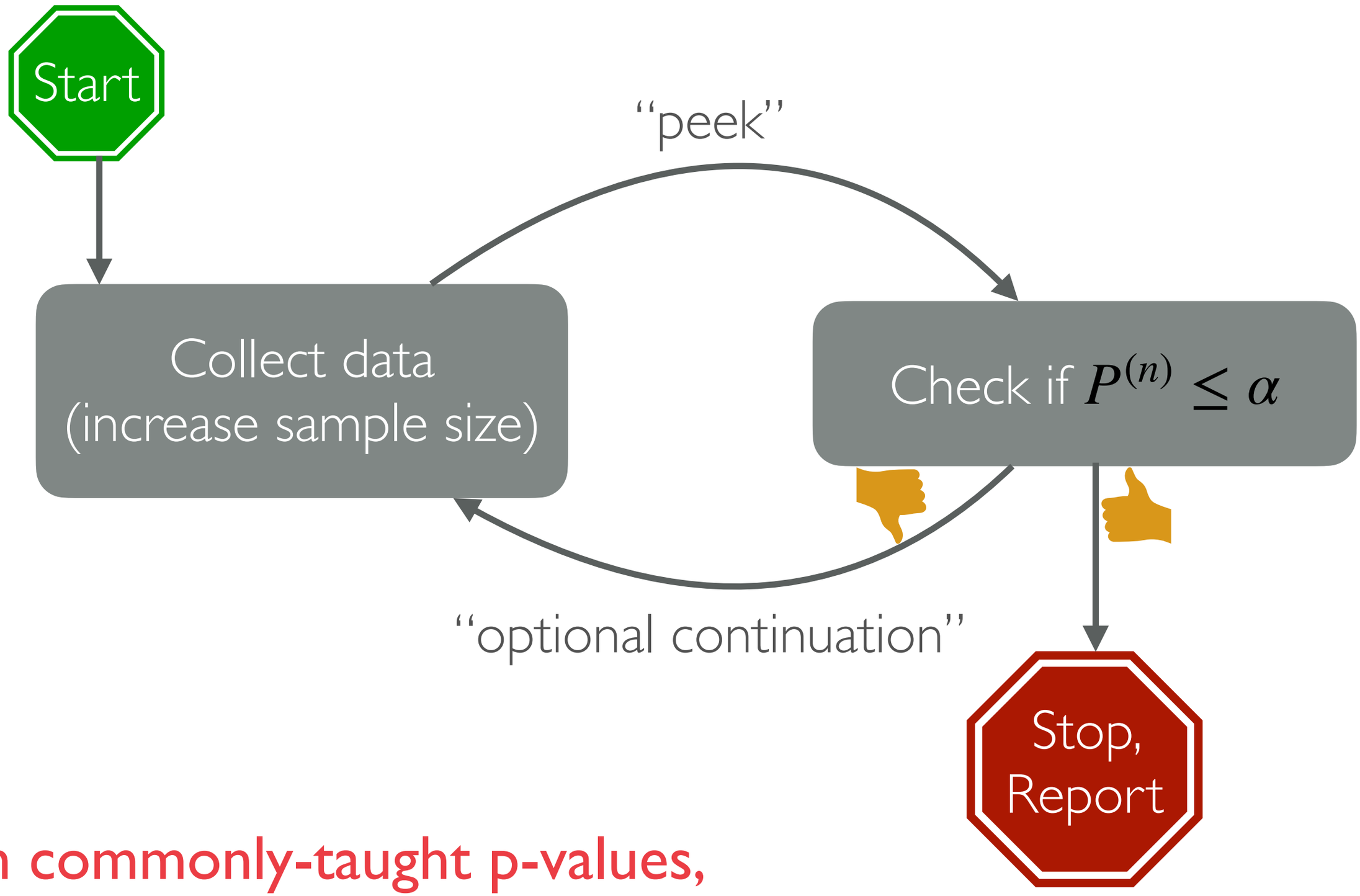
faculty.haas.berkeley.edu/dana_carney/pdf_My%20position%20on%20power%20poses.pdf

Courses | Wshops | UCB | Sports | Vovk | Group Ex | TRL | Jonathan Pon Mem... | Classes | Charities | Podcasts | CausalInf | Other Bookmarks

4. The data are flimsy. The effects are small and barely there in many cases.
5. Initially, the primary DV of interest was risk-taking. We ran subjects in chunks and checked the effect along the way. It was something like 25 subjects run, then 10, then 7, then 5. Back then this did not seem like p-hacking. It seemed like saving money (assuming your effect size was big enough and p-value was the only issue).
6. Some subjects were excluded on bases such as "didn't follow directions." The total number of exclusions was 5. The final sample size was $N = 42$.
7. The cortisol and testosterone data (in saliva at that point) were sent to Salimetrics (which was in State College, PA at that time). The hormone results came back and data were analyzed.
8. For the risk-taking DV: One p-value for a Pearson chi square was .052 and for the Likelihood ratio it was .05. The smaller of the two was reported despite the Pearson being the more ubiquitously used test of significance for a

"Sampling to a foregone conclusion" — Anscombe (1950s)

What is the problem with continuous monitoring?



With commonly-taught p-values, false positive rate $\gg \alpha$.

Let $P^{(n)}$ be a classical p-value (eg: t-test), calculated using the first n samples.

Under the null hypothesis (no treatment effect),

$$\forall n \geq 1, \qquad \underbrace{\Pr(P^{(n)} \leq \alpha)}_{\text{prob. of false positive}} \qquad \leq \alpha \,.$$

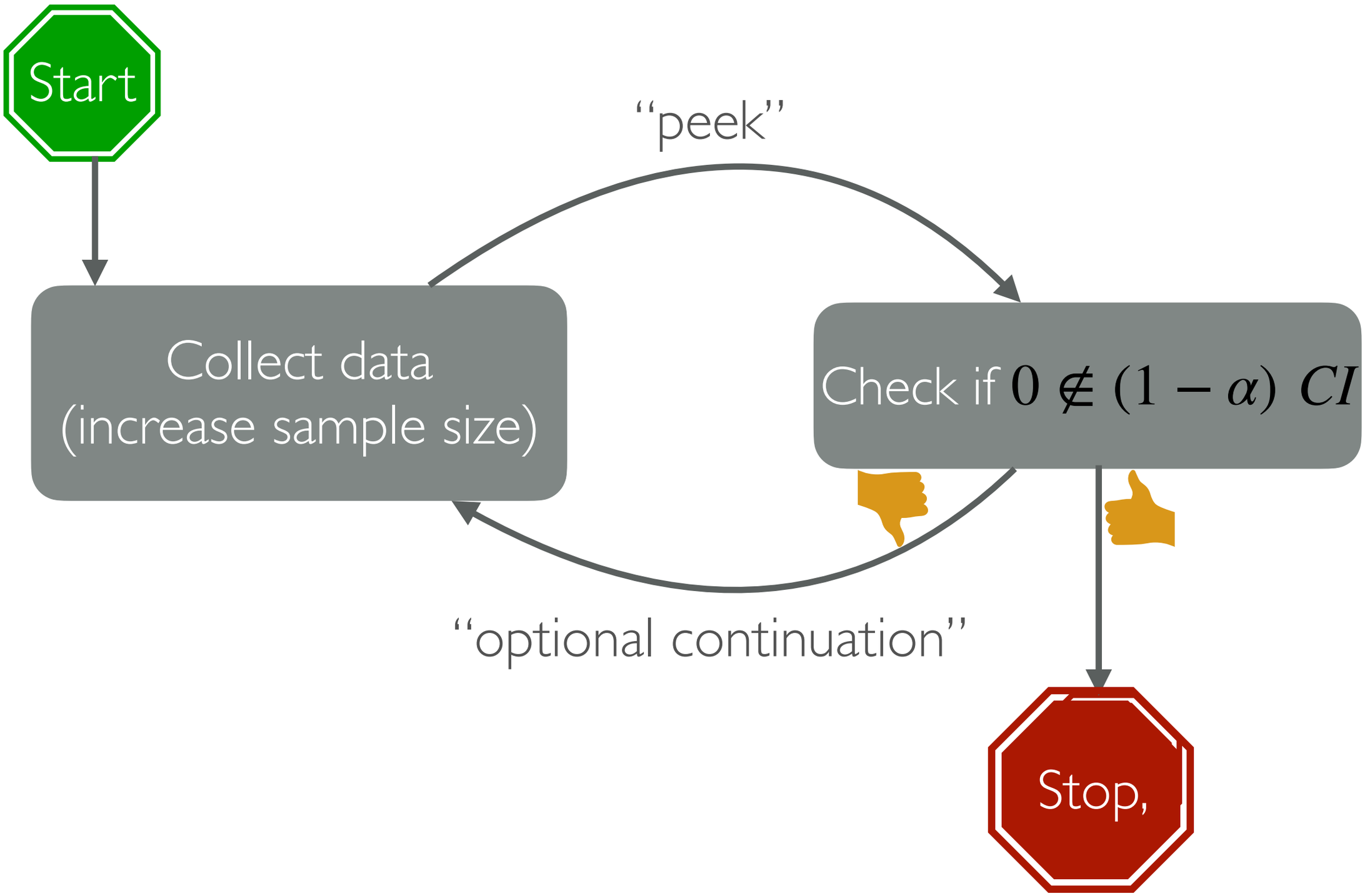
Let τ be the stopping time of the experiment.

Often, τ depends on data, eg: $\tau := \min\{n \in \mathbb{N} : P^{(n)} \leq \alpha\}$.

Unfortunately, $\Pr(P^{(\tau)} \leq \alpha) \not\leq \alpha$. usually $= 1$.

Not special to p-values. Same holds for confidence intervals.

Same issue with confidence intervals



Again, false positive rate $\gg \alpha$.

Let $(L^{(n)}, U^{(n)})$ be any classical $(1-\alpha)$ CI, calculated using the first n samples.

When trying to estimate the treatment effect θ,

$$\forall n \geq 1,\ \underbrace{\Pr(\theta \in (L^{(n)}, U^{(n)}))}_{\text{prob. of coverage}} \geq 1-\alpha\,.$$

Let τ be the stopping time of the experiment.

Again, τ may depend on data, eg: $\tau := \min\{n \in \mathbb{N} : L^{(n)} > 0\}$.

Unfortunately, $\Pr(\theta \in (L^{(\tau)}, U^{(\tau)})) \not\geq 1-\alpha\,.$

usually $= 0$.

We want "safe, anytime-valid inference" (SAVI) methods

SAVI methods are those that yield valid inference at
arbitrary stopping times, possibly not specified or anticipated in advance.

SAVI methods allow for continuous monitoring and analysis of data, adaptive decisions to halt or continue experiments (for any reason), all without violating the validity of the claims.

Provides a lot of flexibility to the statistician ("peeking"), useful for a lot of exploratory settings, or those without oversight (like university labs and tech industry).

Now, forget about statistical inference briefly, let's get to betting.

A New Interpretation of Information Rate

reproduced with permission of AT&T

By J. L. Kelly, jr.

(Manuscript received March 21, 1956)

If the input symbols to a communication channel represent the outcomes of a chance event on which bets are available at odds consistent with their probabilities (i.e., "fair" odds), a gambler can use the knowledge given him by the received symbols to cause his money to grow exponentially. The maximum exponential rate of growth of the gambler's capital is equal to the rate of transmission of information over the channel. This result is generalized to include the case of arbitrary odds.

Suppose we observe iid coin flips B_i of bias $p > 1/2$ (for known p).

We start with one dollar, and can make "double or nothing" bets
(In each round, we bet some fraction λ of our wealth on heads,
if H, we earn that amount, and if T, we lose that amount.)

What fraction λ of our wealth should we bet at each step?

$W_t(\lambda) := \prod_{i=1}^{t}(1 + \lambda(2B_i - 1))$ is wealth after t rounds.

$$W_t = \exp\left(\sum_{i=1}^{t} \log(1 + \lambda B_i)\right) = \exp\left(t\mathbb{E}[\log(1 + \lambda B)] + o(t)\right)$$

Kelly: choose λ to maximize $\lim_{t\to\infty} \frac{\mathbb{E}\log W_t(\lambda)}{t} = \mathbb{E}[\log(1 + \lambda B)].$

Solution: bet $\lambda^* = 2(p - 1/2)$ on heads.

Optimal Wealth $W_t(\lambda^*) = \exp(t \cdot H(p\,|\,0.5) + o(t))$,
where H is the relative entropy (KL divergence)
Equivalently, $\mathbb{E}[\log W_t]/t = H(p\,|\,0.5)$

OPTIMAL GAMBLING SYSTEMS FOR FAVORABLE GAMES

L. BREIMAN

UNIVERSITY OF CALIFORNIA, LOS ANGELES

1. Introduction

Assume that we are hardened and unscrupulous types with an infinitely wealthy friend. We induce him to match any bet we wish to make on the event that a coin biased in our favor will turn up heads. That is, at every toss we have probability $p > 1/2$ of doubling the amount of our bet. If we are clever, as well as unscrupulous, we soon begin to worry about how much of our available fortune to bet at every toss. Betting everything we have on heads on every toss will lead to almost certain bankruptcy. On the other hand, if we bet a small, but fixed, fraction (we assume throughout that money is infinitely divisible) of our available fortune at every toss, then the law of large numbers informs us that our fortune converges almost surely to plus infinity. What to do?

Generalizes Kelly betting to other settings.

Proves that the Kelly criterion also asymptotically optimizes
a) Expected time to reach a threshold wealth
b) Expected wealth at some threshold time

Ok, now let's tie the two together: betting and statistics

Testing by betting

Shafer & Vovk
(+ Robbins, implicitly)

In order to test a hypothesis, one sets up a game such that:
if the null is true, no strategy can systematically make (toy) money,
but if the null is false, then a good betting strategy can make money.

Wealth in the game is directly a measure of evidence against the null.

Each strategy of the statistician = a different estimator or test statistic.
So there are "good" and "bad" strategies for betting,
just as there are good and bad estimators or test statistics.

Testing (and estimation) == game and strategy design.

Kelly's game corresponds to H_0 : fair coin against H_1 : bias p

Outline of this talk

"Sequential anytime-valid inference (SAVI)"

2. Testing by betting yields SAVI inference (parametric)

3. Core SAVI concepts

4. *Optimal gambling strategies*

The lady tasting tea (1920s)

Ronald Fisher

Would you like some tea?

No, T in M ≠ M in T

Can you *really* tell them apart?

Indeed, yes!

Muriel Bristol

T M T T M M M T

What's the probability that a *chance guess* would be *perfect*? 1/70

This is a p-value for H_0 : there is no difference between MT and TM.

Randomization-based causal inference, design of experiments…

The lady tasting tea (1920s)

T M T T M M M T

However, the odds were stacked against Muriel from the start!

The probability that a *chance guess* would yield *at most one error* is 17/70 $\approx \mathbf{0.24}$, which is not so impressive.

With the benefit of 100 years of hindsight, I would have (and did) run the experiment quite differently…

The lady keeps tasting coffee (2020)

(self)

Let's play a game

Umm…sure…?

It involves coffee

Sure!

Leila Wehbe

M C

M C

…

The lady keeps tasting coffee (2020, betting)

Result?

$L_0 = 1$

$R_1 = -1$

$\lambda_1 = 0.2$ (on heads)

$L_1 = L_0 \cdot (1 + \lambda_1 R_1) = 0.8$

$R_2 = +1$

$\lambda_2 = 0.4$ (on heads)

$L_2 = L_1 \cdot (1 + \lambda_2 R_2) = 1.12$

...

$L_t := \prod_{i=1}^{t} (1 + \lambda_i R_i)$, where (λ_i) are "predictable" bets in [-1,1].

Under the null, $(L_t)_{t \in \mathbb{N}}$ is a nonnegative martingale ("fair game").

The lady keeps tasting coffee (2020, betting)

$L_t := \prod_{i=1}^{t} (1 + \lambda_i R_i)$, where (λ_i) are "predictable" bets in [0,1].

Under the null, $(L_t)_{t \in \mathbb{N}}$ is a nonnegative martingale ("fair game").

At any stopping time τ, $\mathbb{E}_{H_0}[L_\tau] \leq 1$ — optional stopping theorem.

Ville's inequality (time-uniform Markov's for nonneg. supermartingales)

$\Pr(\exists t \in \mathbb{N} : L_t \geq 1/\alpha) \leq \alpha$.

If the null holds, then Leila is unlikely to turn one pound into fifty

- L_t directly measures evidence against H_0 ("e-process").
- $\inf_{s \leq t} 1/L_s$ is an "anytime-valid p-value" or "p-process".
- $1\{L_t \geq 1/\alpha\}$ is a level-α sequential test for H_0.

Wald, Robbins, Shafer, Vovk, Grunwald, Ramdas, …

Why is this interesting?

(a) simple and clean approach to sequential experimental design

(b) can express *doubt* naturally

(c) *cooperation between subject and statistician allowed between rounds*

(d) flexible (can design *many* games for each problem)

(e) make up the game (and extend the game) on the fly

(f) evidence only depends on what did occur, not on hypothetical worlds

Addresses 3 issues of p-values: peeking, optional continuation of experiments, reasoning about hypothetical worlds.

Outline of this talk

"Sequential anytime-valid inference (SAVI)"

Testing by betting yields SAVI inference (parametric)

3. Core SAVI concepts

4. *Optimal gambling strategies*

Hypothesis testing in statistical practice

The "null hypothesis" H_0 is a set of distributions $\mathscr{P}$
defined on some filtered measurable space $(\Omega, \mathscr{F})$.
The "alternative hypothesis" H_1 is a set of distributions $\mathscr{Q}$
defined on the same space.

When we are testing H_0 against H_1, we are asking if the data are coming from some distribution in $\mathscr{P}$ or in $\mathscr{Q}$.

Nothing iid is assumed in the above notation,
these distributions could be over sequences of observations.

In statistical practice, the null has a special role (eg: "no effect").
Rejecting the null may correspond to an interesting scientific phenomenon (described by the alternative).
Thus the first goal is to calibrate/control errors under the null.

A p-process (or anytime-valid p-value) for a null $H_0 : P \in \mathscr{P}$ is a sequence $(p_t)_{t \geq 1}$ that satisfies

For any stopping time τ, $P \in \mathscr{P} : P(p_\tau \leq \alpha) \leq \alpha$.

Johari et al. (2015, 2021),
Howard, Ramdas, et al. (2018, 2021)

A p-process (or anytime-valid p-value) for a null $H_0 : P \in \mathscr{P}$ is a sequence $(p_t)_{t\geq 1}$ that satisfies

For any stopping time τ, $P \in \mathscr{P} : P(p_\tau \leq \alpha) \leq \alpha$.

Johari et al. (2015, 2021),
Howard, Ramdas, et al. (2018, 2021)

An e-value for H_0 is a $[0,\infty]$-valued r.v. e s.t.

$\forall P \in \mathscr{P}$, $\mathbb{E}_P(e) \leq 1$. (e for evidence or expectation)

An e-process for H_0 is a sequence of e-values $(e_t)_{t\geq 1}$

s.t. for any stopping time τ, $P \in \mathscr{P} : \mathbb{E}_P(e_\tau) \leq 1$.

Howard, Ramdas, et al. (2018-2021)
Grunwald et al. (2019-2021)
Shafer (2020), Vovk & Wang (2021)

For simple $H_0 = \{P\}$, likelihood ratios are e-processes

$$M_t = \frac{q(X_1, \ldots, X_t)}{p(X_1, \ldots, X_t)}$$

For composite nulls $\mathscr{P}$, *nonnegative martingales* yield e-processes

M is a *test martingale* for $\mathscr{P}$ if $M \geq 0, M_0 = 1$, and
$\mathbb{E}_P[M_t \mid M_1, \ldots, M_{t-1}] = M_{t-1}$ for all $P \in \mathscr{P}, t \geq 1.$

Nonnegative supermartingales are also e-processes

M is a *test supermartingale* for $\mathscr{P}$ if $M \geq 0, M_0 = 1$ and
$\mathbb{E}_P[M_t \mid M_1, \ldots, M_{t-1}] \leq M_{t-1}$ for all $P \in \mathscr{P}, t \geq 1.$

But there are e-processes which are not dominated by any test super martingale.

Averages of dependent e-processes are e-processes (also e-values).
Products of independent e-values are supermartingales (hence e-processes).

Ville's martingale theorem & inequality

Ville'39

If $(L_t)_{t\geq 0}$ is a nonnegative martingale under P, with $L_0 = 1$,

$$P(\exists t \in \mathbb{N} : L_t \geq 1/\alpha) \leq \alpha\,.$$

What he really proved:

For any event A which has measure zero under P, there exists a nonnegative martingale M (under P) such that $M = \infty$ if A occurs.

In fact, both directions hold. An event has measure zero if and only if there exists a nonnegative martingale that blows up on that event.

Further, an event A has probability α iff some nonnegative martingale reaches $1/\alpha$ on A. This is summarized in Ville's inequality.

Finally, if L is an e-process for a set of distributions $\mathscr{P}$, then

$$\sup_{P\in\mathscr{P}} P(\exists t \in \mathbb{N} : L_t \geq 1/\alpha) \leq \alpha\,.$$

A composite generalization of Ville's theorem

For any event A with $P(A) = 0$, there exists a nonnegative martingale M (under P) such that $M = \infty$ if A occurs.

Is there a version of this statement for sets of distributions $\mathscr{P}$?

First try: say $P(A) = 0$ for all $P \in \mathscr{P}$ above. Unfortunately, this is provably incorrect.

Right answer: define a certain "outer measure" $\mathscr{P}(A)$, omitted here.

Theorem: For any event A, we have $\mathscr{P}(A) = 0$ iff there exists an **e-process** M for $\mathscr{P}$ such that $M = \infty$ if A occurs. (+Simplifications fail)

Example application: a distribution-uniform strong law of large numbers has its failure (a "measure zero" event) witnessed by an explicit e-process.

Ruf, Larsson, Koolen, Ramdas (EJP, 2023)

Outline of this talk


Motivation for "safe anytime-valid inference (SAVI)"


Testing by betting yields SAVI inference (parametric)


Core SAVI concepts

4. *Optimal gambling strategies*

Point nulls and alternatives

$$H_0 : X_i \sim P \text{ versus } H_1 : X_i \sim Q \text{ (assuming } P \ll Q, Q \ll P)$$

What is the game?

Initial capital $W_0 = 1$

For each $t = 1,2,\ldots$

Statistician declares "bet" $S_t : \mathcal{X} \to [0,\infty)$ s.t. $\mathbb{E}_P[S_t(X) \mid X_1, \ldots, X_{t-1}] \leq 1$

Reality reveals X_t

Statistician's wealth becomes $W_t = W_{t-1} \cdot S_t(X_t)$

What is the log-optimal betting strategy? Answer: likelihood ratio of Q to P

The **log-optimal bet** is $S_t(x) = \dfrac{q(x)}{p(x)}$.

$W_T^* = \displaystyle\prod_{i=1}^{T} \frac{q(X_i)}{p(X_i)}$ is the log-optimal wealth process:

— it is a positive test martingale under P, $\mathbb{E}_P[W_\tau] \leq 1, \mathbb{E}_P[\log W_\tau] \leq 0.$

— $\mathbb{E}_Q[\log W_T] > 0$ is maximized by this choice of bets, equals $H(Q \mid P)$

Shafer (2021, JRSSA discussion paper)

Proof in two steps (analog of Neyman-Pearson)

Step 1: Dropping the conditioning for simplicity,
statistician declares "bet" $S_t : \mathcal{X} \to [0,\infty)$ s.t. $\mathbb{E}_P[S_t(X)] \leq 1$.

Note that every admissible bet must satisfy the above with equality. Why?
If strict inequality holds, we can use the new bet $\widetilde{S}_t := \dfrac{S_t}{\mathbb{E}_P[S_t(X)]}$, which is still valid. Further, $\mathbb{E}_Q[\log \widetilde{S}_t] > \mathbb{E}_Q[\log S_t]$.

Step 2: Thus, every admissible bet satisfies $\int S_t(x)p(x)dx = 1$.

Define $r_t(x) := S_t(x)p(x)$, and note that the above implies r_t is a density.
Rewriting, we must have $S_t(x) = r_t(x)/p(x)$ for some $r_t(x)$.

Now note that $\mathbb{E}_Q\left[\log \dfrac{r_t(x)}{p(x)}\right] \leq \mathbb{E}_Q\left[\log \dfrac{q(x)}{p(x)}\right]$ because $0 \leq \mathbb{E}_Q\left[\log \dfrac{q(x)}{r_t(x)}\right]$.

E versus P

In the P-world, we judge tests by *probabilities*:

$$\max_{\phi} P_{H_1}(\phi(X) = 1) \text{ power}$$
$$\text{s.t. } P_{H_0}(\phi(X) = 1) \leq \alpha$$

This is a theory of decision making

In the E-world, we judge e-values by *expectations*:

$$\mathbb{E}_{H_0}[W]$$

and

$$\mathbb{E}_{H_1}[\log W]$$

("e-power" or "growth rate")

This is a theory of evidence

We are designing a complementary theory to (say) Neyman-Pearson.
When you see an e-value or e-process, ask about its e-power or growth rate, not its power (a p-concept) — there is some loss in transforming one to other.

Simple null vs. Composite alternative

$$H_0 : X_i \sim P \quad \text{versus} \quad H_1 : X_i \sim \{Q_\theta\}_{\theta \in \Theta}$$

Option 1: Mix (hedge your bets) with "prior" π

$$W_T = \int_\Theta \prod_{i=1}^{T} \frac{q_\theta(X_i)}{p(X_i)} d\pi(\theta)$$

Option 2: Plug-in a representative $\hat{\theta}_i \equiv \theta_i(X_1, \ldots, X_{i-1})$ in each round

$$W_T = \prod_{i=1}^{T} \frac{q_{\hat{\theta}_i}(X_i)}{p(X_i)}$$

Typically, $\lim_{T \to \infty} \mathbb{E}_{Q^*}[\log W_T]/T = \mathbb{E}_{Q^*}\left[\log \frac{q^*(X)}{p(X)}\right]$,
which is the best possible "growth rate", even without knowing Q^*.

To summarize what was known

- For testing a point null P against point alternative Q, likelihood ratios are optimal per-round bets
- The optimal wealth is the likelihood ratio process.
- The optimal rate of growth (exponent) of the wealth is exactly the KL divergence or relative entropy of Q to P.

What about composite nulls?

Significant progress by Peter Grünwald and coauthors in two papers ("Safe Testing" and "Universal Reverse Information Projection and Optimal E-statistics"). We complete the story.

The numeraire e-variable and reverse information projection

arXiv:2402.18810

Martin Larsson (CMU)

Johannes Ruf (LSE)

Our setting: Composite null vs. Simple alternative

- We have a composite null hypothesis $\mathscr{P}$ and a point alternative hypothesis Q. The data is either drawn from some P in $\mathscr{P}$ (the null is true), or from Q (the null is false).

- A *valid* bet is an "e-variable", which is a $X \geq 0$ such that $\mathbb{E}_P[X] \leq 1$ for every $P \in \mathscr{P}$. Think of X as being the multiplier of your wealth in each round of a multi-round game.

- Question: What is the optimal one-round bet X^*? Is it unique? Can we characterize/derive it?

- Answer: It is the likelihood ratio of Q to a special element P^*, which we call the Reverse Information Projection (RIPr).

Our work tells a complete story about (X^*, P^*).

Our setting: rephrased

Ω is a set of possible outcomes

A Forecaster claims that $\mathscr{P}$ describes the world well,
meaning that outcomes/events are well described by some $P \in \mathscr{P}$.

A Skeptic thinks that Forecaster is inaccurate,
and believes that Q is a better model for the world.

Forecaster offers bets based on his forecasts. A valid bet
is an e-variable, which is a random variable $X : \Omega \to \mathbb{R}_0^+$
such that $\mathbb{E}_P[X] \leq 1$ for every $P \in \mathscr{P}$.

After Skeptic picks a particular bet X, we observe the outcome ω.
Skeptic's realized payout is $X(\omega)$.

Which bet should the skeptic pick?

A: The log-optimal bet X^* is the likelihood ratio of Q to the RIPr P^*.

Introducing X^*, the "numeraire" e-variable

Theorem: Under no assumptions on null $\mathscr{P}$ and alternative Q, there *always* exists a special e-variable (bet) X^* which satisfies two properties:

A. First, $X^* \geq 0$ and $\mathbb{E}_P[X^*] \leq 1, \forall P \in \mathscr{P}$ (the e-variable or fair bet property)
B. Second, for any e-variable X, we have $\mathbb{E}_Q[X/X^*] \leq 1$ (the "numeraire property")

Further, X^* is unique up to Q-nullsets. In fact, X^* is the numeraire if and only if it is log-optimal.

Applying Jensen's inequality, we get two other interpretable implications: for any e-variable X, we have $\mathbb{E}_Q[X^*/X] \geq 1$ and $\mathbb{E}_Q[\log(X/X^*)] \leq 0$ (log-optimality!)

Introducing P^*, the reverse information projection

Definition: Define a measure P^* by defining its likelihood ratio (Radon-Nikodym derivative) with respect to Q:

$$dP^*/dQ := 1/X^*$$

- This is understood to be zero on $\{X^* = \infty\}$.
- $P^* \ll Q$ by definition. Also $X^* = dQ/dP^*$ by definition.
- P^* is not a probability measure in general, it is a sub-probability measure, meaning that $\int dP^* \leq 1$.
- P^* lies in the *bipolar* of $\mathscr{P}$, which is defined as follows.
  A. The polar is $\mathscr{P}^\circ := \{X \geq 0 : \mathbb{E}_P[X] \leq 1 \text{ for all } P \in \mathscr{P}\}$, which is simply the set of all e-variables.
  B. The bipolar is $\mathscr{P}^{\circ\circ} := \{P \geq 0 : \mathbb{E}_P[X] \leq 1 \text{ for all } X \in \mathscr{P}^\circ\}$, which we also call "the effective null hypothesis".

Notation

- We making the simplifying assumption in the rest of the talk that $Q \ll \mathscr{P}$, meaning that whenever $P(A) = 0$ for every $P \in \mathscr{P}$, we also have $Q(A) = 0$. Very weak assumption!

- Recall that the KL divergence or relative entropy is defined as $H(Q \mid P) := \mathbb{E}_Q \left[\log \frac{dQ}{dP} \right]$ if $Q \ll P$, ∞ o.w.

- If P^a denotes the absolutely continuous part of P wrt Q, we can rewrite $H(Q \mid P) = \mathbb{E}_Q \left[-\log \frac{dP^a}{dQ} \right]$.

Strong duality of (X^*, P^*)

Theorem: Assume $Q \ll \mathscr{P}$ for simplicity. Let X^* be the numeraire and let P^* be an element of $\mathscr{P}^{\circ\circ}$ that is equivalent to Q.

The following statements are equivalent:

— P^* is the RIPr;

— $\mathbb{E}_Q\left[\dfrac{dP^a}{dP^*}\right] \leq 1$ for all $P \in \mathscr{P}^{\circ\circ}$;

— $\mathbb{E}_Q\left[\log \dfrac{dP^a}{dP^*}\right] \leq 0$ for all $P \in \mathscr{P}^{\circ\circ}$.

If any of these hold, then one has the strong duality:

$$\mathbb{E}_Q[\log X^*] = \sup_{X \in \mathscr{P}^\circ} \mathbb{E}_Q[\log X] = \inf_{P \in \mathscr{P}^{\circ\circ}} H(Q\,|\,P) = H(Q\,|\,P^*),$$

where these quantities may equal $+\infty$.

Estimation of Mixture Models

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Qiang (Jonathan) Li

Dissertation Director: Andrew R. Barron

May 1999

4.2 A New Information Projection Theory

In our theory, we reverse the order of the arguments in the K-L divergence. An analogous information projection theory is obtained. Applications to maximum likelihood estimation require this reversal of the order in the K-L divergence. We build upon a theory of Bell and Cover [1980], who in a portfolio selection contex developed the story under an assumption that a minimizer of $D(P\|Q), Q \in \mathcal{C}$ exists.

Again we consider a convex set $\mathcal{C}$ of probability measures. Let P be a probability measure of our interest. Define

$$D(P\|\mathcal{C}) = \inf_{Q\in\mathcal{C}} D(P\|Q).$$

Similar to T-C theory, we also want to establish existence, uniqueness and characterizing Pythagorean Identity of a projection P^* of P onto $\mathcal{C}$.

DEFINITION 4.2 (Reversed Information Projection) *Given a probability measure P with a density p and a convex set $\mathcal{C}$ of densities q, a function q^* is called the (reversed) information projection if for every q_n with $D(p\|q_n) \to D(p\|\mathcal{C})$, we have $\log q_n \to \log q^*$ in $L^1(P)$.*

THEOREM 4.3 (Properties of the Reversed I-Projection) *Let $\mathcal{C}$ be a convex set of probability measures Q with densities q and let P be a target measure with density p. Then the reversed I-projection q^* of P exists and is unique. Moreover it satisfies the following properties:*

1. $D(p\|q^*) = \inf_{q\in\mathcal{C}} D(p\|q)$,
2. $c_q = \int p\frac{q}{q^*} \le 1, \forall q \in \mathcal{C}$,
3. $D(p\|q) \ge D(p\|q^*) + D(p\|\rho)$ *where* $\rho = \frac{pq/q^*}{c_q}$ *is a density depending on* q.

Our theory recovers these as a special case.

Computer Science > Information Theory

[Submitted on 29 Jun 2023 (v1), last revised 4 Dec 2023 (this version, v2)]

Universal Reverse Information Projections and Optimal E-statistics

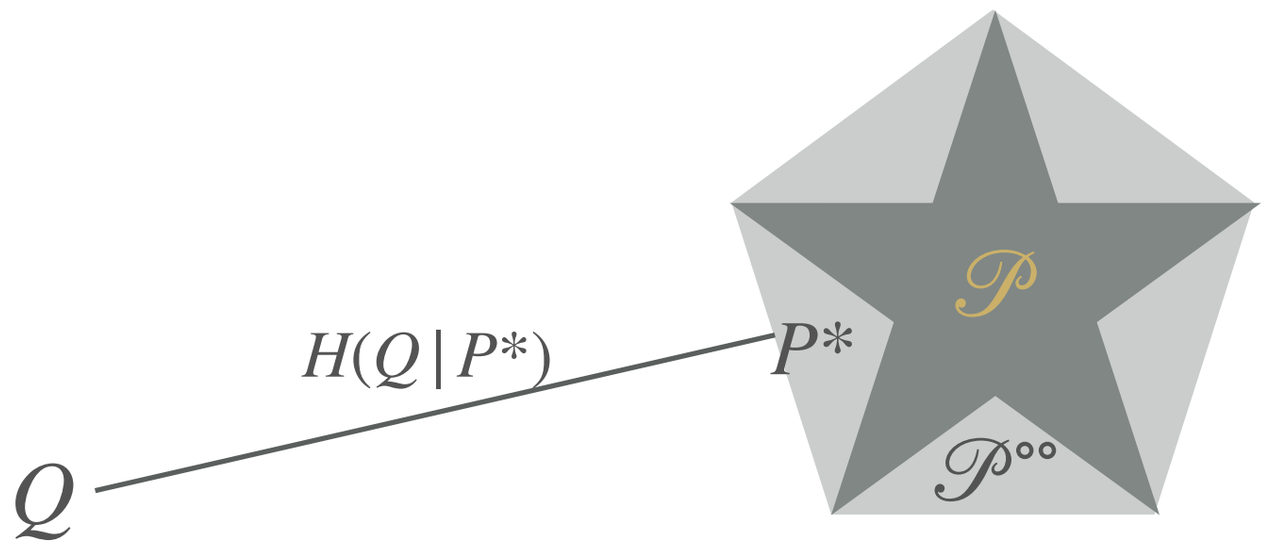
Tyron Lardy, Peter Grünwald, Peter Harremoës

Information projections have found important applications in probability theory, statistics, and related areas. In the field of hypothesis testing in particular, the reverse information projection (RIPr) has recently been shown to lead to so-called growth-rate optimal (GRO) e-statistics for testing simple alternatives against composite null hypotheses. However, the RIPr as well as the GRO criterion are undefined whenever the infimum information divergence between the null and alternative is infinite. We show that in such scenarios there often still exists an element in the alternative that is 'closest' to the null: the universal reverse information projection. The universal reverse information projection and its non-universal counterpart coincide whenever information divergence is finite. Furthermore, the universal RIPr is shown to lead to optimal e-statistics in a sense that is a novel, but natural, extension of the GRO criterion. We also give conditions under which the universal RIPr is a strict sub-probability distribution, as well as conditions under which an approximation of the universal RIPr leads to approximate e-statistics. For this case we provide tight relations between the corresponding approximation rates.

We work with a measurable space $(\Omega, \mathcal{F})$ and, unless specified otherwise, all measures will be defined on this space. Throughout, P will denote a finite measure and $\mathcal{C}$ a set of finite measures, such that P and all $Q \in \mathcal{C}$ have densities w.r.t. a common σ-finite measure μ. These densities will be denoted with lowercase, i.e. p and q respectively. We will assume throughout that $\mathcal{C}$ is σ-convex, i.e. closed under countable mixtures, though we will refer to this simply as 'convex'. Furthermore, we assume that there exists at least one $Q^* \in \mathcal{C}$ such that $P \ll Q^*$. On the one hand, this ensures that $D(P\|Q \rightsquigarrow \mathcal{C})$ is a well-defined number in $[0, \infty]$ for any $Q \in \mathcal{C}$. On the other hand, it aligns with our philosophy when we turn to hypothesis

Our theory avoids all these assumptions

A bit more about the bipolar $\mathscr{P}^{\circ\circ}$



Let M_1 denote the set of all probability measures.

Let M_+ denote the set of all nonnegative measure.

A set $C \subset M_+$ is called "solid" if, for every $P \in C$, we also have $P' \in C$ whenever $P' \leq P$.

If $\mathscr{P}$ is finite, then $\mathscr{P}^{\circ\circ} \cap M_1 = \text{conv}(\mathscr{P})$.

If a reference measure μ exists for $\mathscr{P}$, then every $P \in \mathscr{P}^{\circ\circ}$ is also absolutely continuous wrt μ, and $\mathscr{P}^{\circ\circ}$ is the smallest μ-closed solid convex set that contains $\mathscr{P}$.

A bit more about (X^*, P^*)

Recall that $X^* = \dfrac{dQ}{dP^*}$ by definition of P^*.

We say that P dominates P' if $P(A) \geq P'(A)$ for all A, with strict inequality for some A.

Theorem: Assume $Q \ll \mathscr{P}$. Let X^* be an e-variable, and let P^* be defined by $dP^*/dQ = 1/X^*$. Then X^* is the numeraire iff $P^* \in \mathscr{P}^{\circ\circ}$.

In this case, X^* is Q-a.s. positive, and P^* is equivalent to Q. Further, X^* is the only e-variable which can be written as the likelihood ratio of Q to some element in $\mathscr{P}^{\circ\circ}$. Finally, P^* is also maximal, meaning that it cannot be dominated by any other element of $\mathscr{P}^{\circ\circ}$ that is absolutely continuous wrt Q.

Example 1: Symmetric distributions

Let Z denote the data, in this case real-valued.

$\mathscr{P} := \{P \in M_1 : Z \text{ and } -Z \text{ have the same distribution under } P\}$

Note that $\mathscr{P}$ has no dominating reference measure.

Suppose Q has a Lebesgue density q.

Older theory does not apply in this case. But we can easily show

$p^*(z) = \dfrac{q(z) + q(-z)}{2} 1\{q(z) > 0\}$ is the RIPr density.

It is a probability density iff Q has symmetric support.

We can also show that $X^* = \dfrac{2q(Z)}{q(Z) + q(-Z)}$ is the numeraire.

(In the paper, we generalize beyond Q with Lebesgue density.)

Example 2: 1-Sub-Gaussian distributions

$$\mathscr{P} := \{P \in M_1 : \mathbb{E}_P[e^{\lambda Z - \lambda^2/2}] \leq 1 \text{ for all } \lambda \geq 0\}$$

Above condition implies that $\mathbb{E}_P[Z] \leq 0$ for all $P \in \mathscr{P}$.
Let $Q = N(\mu,1)$ for some known $\mu > 0$.

Once more, $\mathscr{P}$ has no reference measure. So older theory does not apply.

But we can easily show that $\exp(\mu Z - \mu^2/2)$ is the numeraire and $N(0,1)$ is the RIPr.

Example 3: a parametric example from Lardy et al.

$\mathscr{P} := \{P_1, P_2\}$, where $P_1 = N(-1,1)$ and $P_2 = N(1,1)$.

Let Q be a centered Cauchy distribution.

Note that $H(Q \mid P_i) = \infty$.

Nevertheless, the RIPr is $P^* = (P_1 + P_2)/2$
and the numeraire is $X^* = 2q(Z)/(p_1(Z) + p_2(Z))$.

In the paper we generalize this to Q being any symmetric distribution (but note the much smaller null).

Example 4: one-parameter exponential families

Consider an exponential family with density
$p_\theta(z) = \exp(\theta T(z) - A(\theta))$ for $\theta \in \Theta \subset \mathbb{R}$
with respect to a reference measure μ,
and A is convex and differentiable.

The null is given by $\mathscr{P} := \{p_\theta : \theta \in \Theta_0 \subset \Theta\}$,
and we assume Θ_0 is closed, with smallest element θ^*.
The alternative is p_{θ_1} for some $\theta_1 < \theta^*$.

We can show that p_{θ^*} is the RIPr. Thus the numeraire is
the likelihood ratio $p_{\theta_1}/p_{\theta^*}$.

Csiszar and Matus (2003) have an extensive study of the RIPr
for exponential families.

Beyond logarithmic utility

Consider the optimization problem $\sup_{X \in \mathscr{P}^\circ} \mathbb{E}_Q[U(X)]$

for a continuous, increasing, concave, differentiable, bounded U.

For eg, $U(x) = x^{1-\gamma}/(1-\gamma)$ for $\gamma > 1$.

We can show that a maximizer X_γ^* exists and is unique.

Further, define $P_\gamma^* \in \mathscr{P}^{\circ\circ}$ by $\dfrac{(X_\gamma^*)^{-\gamma}}{\mathbb{E}_Q[(X_\gamma^*)^{1-\gamma}]}$.

Define the Renyi divergence of order $1/\gamma$ as

$$D_{1/\gamma}(Q \,|\, P) := \frac{1}{1/\gamma - 1} \log \mathbb{E}_Q \left[\left(\frac{dP^a}{dQ} \right)^{1-1/\gamma} \right].$$

Theorem: X_γ^* and P_γ^* attain the extrema in the strong duality:

$$\sup_{X \in \mathscr{P}^\circ} \mathbb{E}_Q[U(X)] = \frac{1}{1-\gamma} \exp\left((\gamma^{-1} - 1) \inf_{P \in \mathscr{P}^{\circ\circ}} D_{1/\gamma}(Q \,|\, P) \right)^\gamma$$

Summary

We have fully generalized Kelly betting to composite nulls and point alternatives, yielding a strong duality between (X^*, P^*).

We have defined the reverse information projection P^* (RIPr) and the the optimal e-variable X^* (numeraire) without any assumptions.

We showed how to apply this theory to new nonparametric settings that were previously out of reach.

We showed how to generalize this story to general utilities that are continuous, increasing, concave, differentiable, bounded.

Next steps: more general utilities, composite alternatives, and sequential betting strategies (we have ideas for all of these), EM? Covariate shift?

The numeraire e-variable and reverse information projection

arXiv:2402.18810

Martin Larsson (CMU)

Johannes Ruf (LSE)

Outline of this talk

- Motivation for "safe anytime-valid inference (SAVI)"
- Testing by betting yields SAVI inference (parametric)
- Core SAVI concepts
- *Optimal gambling strategies*

E-*processes* for composite null vs. composite alternative?

$$\frac{\text{Likelihood of alternative } P_\nu}{\text{Likelihood of null } P_\theta}$$

"Frequentist"

$$\frac{\text{Maximum (alt) likelihood}}{\text{Maximum (null) Likelihood}}$$

Bayes

$$\frac{\text{Mixture (alt) likelihood}}{\text{Mixture (null) Likelihood}}$$

Game-theoretic
(wealth in a game)

$$\frac{\text{Mixture or plug-in (alt) likelihood}}{\text{Maximum (null) Likelihood}}$$

Can use "prior" information to bet
on the alternative.
But evidence is compared to best null.

Only the last option is an e-process (the "universal inference" e-process).
It has the asymptotically optimal growth rate (Dixit-Martin'23).

Universal Inference (PNAS 2020) and Testing exchangeability (IJAR'22)

(Continued: "universal inference")

$$W_T = \frac{\text{Mixture/Plug-in (alt) likelihood}}{\text{Maximum (null) Likelihood}} = \prod_{i=1}^{T} \frac{q_{\hat{\theta}_i}(X_i)}{p_{\hat{\theta}_T}(X_i)} \text{ is an e-process.}$$

Also if the numerator is nonparametrically chosen smartly, then universal inference (above) is also asymptotically growth rate optimal!

Under mild conditions, $\mathbb{E}[\log W_T]/T \to K(Q^*, \mathscr{P})$

Dixit and Martin (2023, arXiv)

As an e-value, it is always worse than the numeraire,
but the numeraire is an e-value, while universal inference is an e-process.

Open problem: determine when the sequence of numeraires (at increasing sample sizes) does or does not yield an e-process.

Summary

Testing by betting is a simple framework for hypothesis testing that yields sequential, anytime-valid inference.

Optimal gambling strategies are based on likelihood ratios.
Composite alternatives are handled using mixtures (hedging).
Composite nulls are handled using reverse information projections, or via universal inference (maximum-likelihood under the null).

(Composite) Nonnegative (super)martingales are secretly likelihood ratios, even when no reference measure exists.

E-processes exist more generally, even when nonnegative supermartingales do not exist. They are central objects: necessary and sufficient for sequential testing.