NOTES DURING PHD

BY YONGQI WANG [1,a]

[1] *Centrum Wiskunde & Informatica, Amsterdam, The Netherlands* [a] *yongqi@cwi.nl*.

CONTENTS

$$x = 213123 \tag{0.1}$$

1. TODO.

This is just a playground for me! No need to be too nicely formatted.

1.1. *Questions.*

- What is the regularity condition in universal inference condition?
- Why do we want to have data-dependent significant level α
- What is the difference between p-process and E-process and nonnegative martingale?
- What is a adapted sequence of random sets, random variables
- Wald's sequential likelihood ratio test
- What is Radon-Nikodym derivative
- Try to explain post-hod and ad-hoc in plain English
- What would happen if you just multiply E-variables together? Shouldn't you consider the
- What is the evidence in testing? Is likelihood ratio the best one we have for evidence

- What is the difference between carrying measure and carrying measure?
- Write down the lemma/fact that each element in a regular exponential family is continuous with respect to each other and could be used as carrier. Is it a one-to-one mapping?
- Mean-value Parameterization Convexity of mean-value spaces and canonical spaces
- Loewner ordering

1.2. *Filtrations and sigma-fields.*

DEFINITION 1.1 (Filtration). Let $(\Omega, \mathcal{A}, P)$ be a probability space. An increasing sequence $\mathcal{F}_n$ of sub-σ algebras of $\mathcal{A}$ (i.e. $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \mathcal{F}_n \subseteq \cdots \subseteq \mathcal{A}$) is called a filtration.

The sub-σ-algebra is just the sigma algebra of the X_0, X_1, X_n.
What is the difference between the stopping time and random times that can take possibly infinity.
Filtration sigma-field at time t, filtration is a n increasing sequence of sigma-fields.
What is the lim sup here? A_t is an adapted sequences of events in some filtered probability space.

$$A_\infty := \limsup_{t\to\infty} A_t := \bigcap_{t\in\mathbb{N}} \bigcup_{s\geq t} A_s$$

DEFINITION 1.2 (Absolute continuous). Let p, q be two probability distributions. p is called *absolutely continuous* with respect to q if the
More explicitly, this reads ($p \ll q$) that

$$q(A) = 0 \quad \Rightarrow \quad p(A) = 0, \quad \text{for any } A \in \mathcal{F}_t \text{ and } t \in \mathbb{N}.$$

1.3. *All the inequalities.*

Hoeffding, Bernstein, McDiarmid, Talagrand's
Evidence lower bound
line crossing inequality average treatment effect mixture adaptive design

1.4. *Papers, talks, textbook, and more topics.*

Clarke Barron 1990 1994 about

- Stein's Methods
- Ito's process
- Levy's process
- Gaussian process

1.5. *Basic Concepts.*

DEFINITION 1.3 (Admissibility). In $\mathtt{E}$-process, a process is inadmissible if there exists a process that is strictly better than it at certain time points.

Calibration

Understanding the Fourier transform in terms of transform Placing a restriction on the Fourier transform == smaller function space Minimax setup

Integration / Measure

Topology crash course

The topologist Stephen Smale stunned the mathematical community in 1958 [Sma58] when he proved it was possible to turn the sphere inside out without introducing any creases. Several ways to do this are beautifully illustrated in video recordings [Max77, LMM94, SFL98].

Tower properties

2. Meetings.

26.09.2025. There are quite a lot of things that I need to check after this. Nick presented on conformal prediction where my take-away about is that one could relate the

Another theme is still the contingency table. There is something strange about the NML sequential in which the estimation is on the entire sequence.

Another problem that I need to write down is that for a given prior π and Bayesian updating on the blocks, is the multiplication equal to the global one.

Simpler proof: 1) write something in closed form, e.g. exponential 2) compare it to the proof of seq-RIPr and RIPr in general case. 3) but I think the seq-COND is not there yet?

I need to refresh on NP-lemma, product measure, outer measure, push-forward measure and etc.

I also wanna do a step-by-step guide on plotting and figure

05.09.2025.

What I did this week. In terms of coding, what I did is transform the big genotype G and phenotype P data into a small subset. There are mainly two reasons: 1) 455k SNPs exceeds our scope of multiple testing, and we really need a handful of interesting targets 2) the 44 SNPs in the small dataset are not "significant"

1) Run Fisher exact test with `PLINK -assoc fisher` 2) Rank them by p-value, FDR control (0.05) 3) Obtain about 125 SNPs with very small p

Noted that this is for the 4688 sample points where we are simply testing with Fisher exact test for association. Furthermore, I synthesise sequential dataset by shuffling into blocks of 50 whose ratio of case between control is close to global.

TABLE 1
rs17375018: Fisher's Exact Test p-value of 9.08×10^{-9}.

	SNP+ (a)	SNP- (b)
case (1)	1779	866
ctrl (0)	4089	2626

One might think that in a global significant SNPs, the pseudo-sequential data blocks would likely yield significant results. However, we observed multiple $p_i = 1$, e.g., on tables as below.

	a	b
1	46	29
0	16	9

Furthermore, we are trying to calculate the probability of $p_{\text{fisher}} = 1$ for the pseudo blocks. Start with the simple one and use union bound.

$$\mathrm{P}\left(\frac{n'_{a1}}{n'_{a0}} \approx \frac{n_1}{n_0} \,\middle|\, n'_a, n'_b, \frac{n_a}{n_b} \approx \frac{n'_a}{n'_b}\right)$$

For now, we consider the cases, $\approx$ is $=$ avoiding situations where non-integers appears and ceiling/flooring functions are applied.

- Check out this https://scholar.google.com/citations?user=UnrY-40AAAAJ&hl=en
- https://en.wikipedia.org/wiki/Boole%27s_inequality

29.08.2025. I mainly talked about the gap between the test supermartingale and $\mathtt{E}$-process. I was confused but Peter corrected me, saying even though you can find a supermartingale upper bounding the $\mathtt{E}$-process.

In the simple null case, they are the same. However, in the composite settings, it is not, for each $P \in \mathcal{P}$, there is a test supermartingale but there is not a single one for all $P \in \mathcal{P}$. In fact, it's very easy to give examples where a random process is $\mathtt{E}$-process but not supermartingale.

He mentioned examples from UI, which is very straightforward to show it's $\mathtt{E}$and Alex also said something about the canonical filtrations, by Ramdas, in the t-test.

Seb talks about Gaussian channelling, a paper on 2004 with 3rd derivative and information theory. Interesting, worth a look.

25.08.2025. Practical goal:

- Compare speed of conditional E in BiasedUrn and R: about 3x performance, not worth it to import
- Add save E into the package, not done yet but the styles are already similar to the safestats pacage

This week's goal theory-wise

- Just write down again the KL for Gaussian family for repetition.
- Restate the connection between integral and sequential product of UI: haven't found the literature yet
- Try to re-state the simple and anti-simple case
- What is the seq-RIPr and seq-COND?
- Why do we need to have a general KL measure in general paper
- Read carefully how the maximum is found in the log-optimal paper
- Reproduce with safestats packages for the
- UMP in math. stat. course lecture notes.
- I wanna write down $\mathrm{d}\nu$ and $\mathrm{d}X$ and all that.

The difference between the simple and anti-simple case In short, the simple case is where $\Sigma_q - \Sigma_p$ is negative which mean we can find a RIPr via a prior (or a element) of P.

15.08.2025. Sebastian did a great presentation regarding testing quantile given filtrations. I would formulate here and also add a picture.

The question in mind is to test whether data X is from a hypothesis $\mathcal{H}$ that is non-parametric:

$$\mathcal{H} = \{P \in \mathcal{P}(X) \mid \mathbb{E}_p\left[\phi_i(X) = 0\right], ß = 1, 2, \ldots, d\}$$

where $\mathcal{P}(X)$ is all distribution on X.

The simplest instance where testing X with the same mean μ where $\phi_1(X) = X - \mu$. Larson has proven that the 'optimal' $\mathtt{E}$-variable must be in the form:

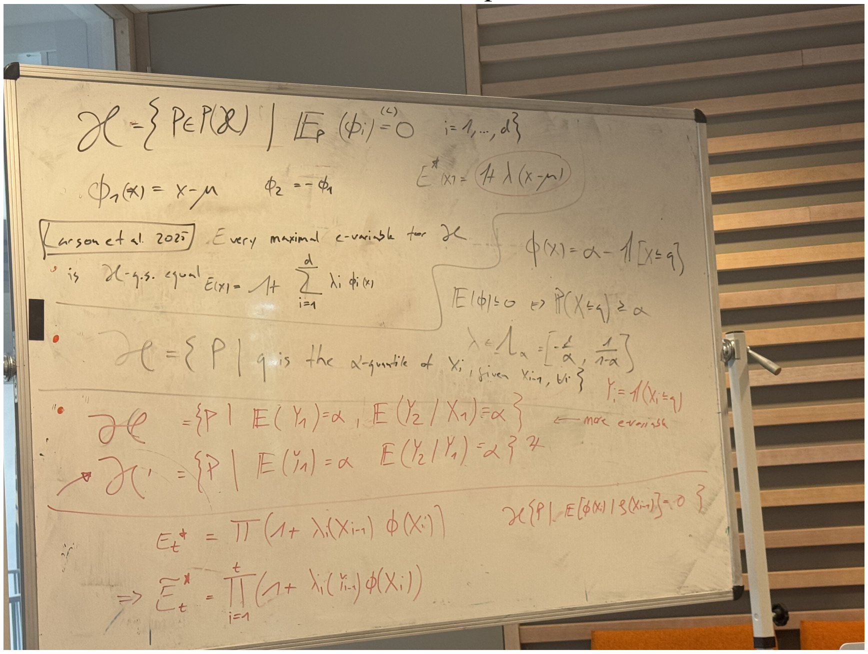
$$S := 1 + \sum_{i=1}^{d} \lambda_i \phi_i(X)$$

Sebastian is mainly interested in the cases if there is any gain in $\mathtt{E}$-variables compared to a coraser filtrations. Imagine we have conditioned on the original data $X_i, i = 1, 2, \ldots,$ you would think that the hypothesis case constructed as below:

$$\mathcal{H} = \{P \in \mathcal{P}(X) \mid \mathbb{E}\left[Y_2 \mid X_1\right] = \alpha\} \tag{2.1}$$

$$\mathcal{H}' = \{P \in \mathcal{P}(X) \mid \mathbb{E}\left[Y_2 \mid Y_1\right] = \alpha\} \tag{2.2}$$

where $Y_i = 1_{X_i \leq q}$. He showed that both $\mathcal{H}$ and $\mathcal{H}'$ are convex and $\mathcal{H}$'s closed convex hull is not the same, otherwise it would be kind of pointless.



14.08.2025. This is a short discussion regarding UI, specifically about the difference between prequential and integral representation.

Supposed iid data $\mathcal{D} = \{X_1, X_2, \ldots, X_n\}$ and we denote the data up to time i by $X^{(i)} = \{X_1, X_2 \ldots, X_i\}$.

One way to instantiate UI by

$$\frac{\prod_{i=1}^{n} P_{\tilde{\theta}_{alt|i-1}}(X_i)}{P_{\hat{\theta}_0}(X^{(n)})}$$

where $\tilde{\theta}_{alt|i-1}$ is any estimator in alternative based on the first $i-1$ data points $X^{(i-1)}$ and $\hat{\theta}_0 = \arg\max_{\theta \in \Theta_0} \prod_i p_\theta(X_i)$ is the MLE estimator under the null. See more details in section 7 of UI paper.

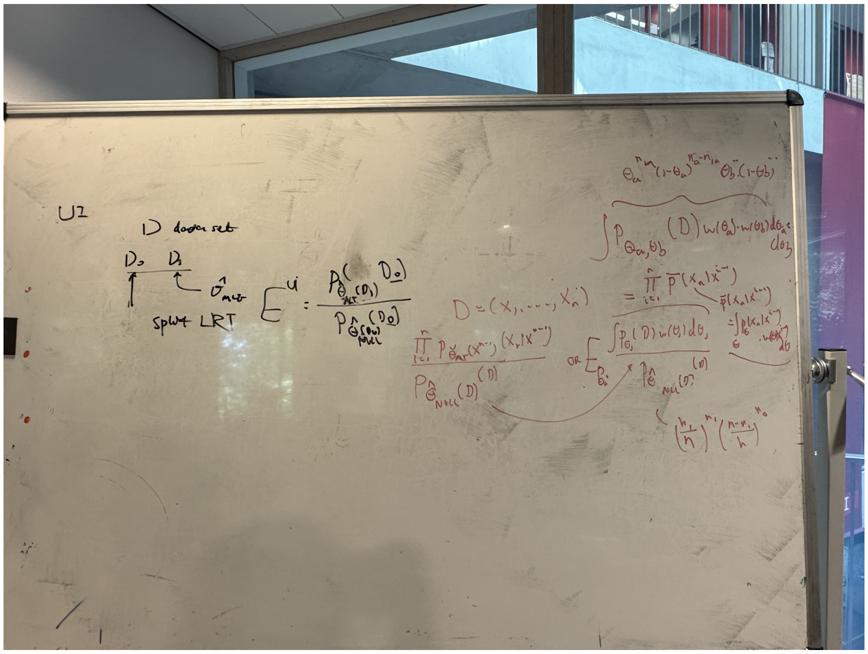
Notice that the sequence of the data $\mathcal{D} = \{X_1, \ldots X_n\}$ really matters. Imagine you obtian $\mathcal{D}'$ with a rearranged sequence and thus slightly different $\tilde{\theta}$ and hence slightly different value at the denominator.

Revisit factorization of probability, it's also called chain rule or general product rule:

$$\begin{aligned} \mathrm{P}\left(X_1, X_2, \ldots, X_n\right) &= \mathrm{P}\left(X_1\right) \mathrm{P}\left(X_2 \mid X_1\right) \cdots \mathrm{P}\left(X_n \mid X_1, X_2, \ldots, X_{n-1}\right) \\ &= \prod_{i=1}^{n} \mathrm{P}\left(X_i \,\middle|\, X^{(i-1)}\right) \end{aligned}$$

QUESTION 2.1. Below is a mixture of all nonnegative test martingale/e-process?

$$\frac{\int p_\theta(X^{(n)}) w(\theta) \, \mathrm{d}\theta}{P_{\hat{\theta}_{\text{null}}}(X^{(n)})}$$



25.07.2025. I only briefly went over the conformal prediction and fisher's noncentral hypergeometric distribution. There were some discussions on what the conformal prediction is.

Imagine you have a classifier for images (dogs, cat, etc.) and is trained via N datasets. For the next prediction, we need something to quantify the uncententy to say that

> The label for X_{N+1} I gave being X (here could be any label), has

Peter gave a algorithmic explanantion where the predicted labels are gaven for each label, then run through against the previous training datasets. Rank them, cut off the tailing α percent then we can say we are confident about our prediction with $1-\alpha$.

However, this is awfully similar to the permutation test, by Sebastian. Yeah it does look a lot like just ranking the prediction and give a p-value.

Alexander also suggested using Gaussian for conformal prediction might be too confusing as the parameter and the prediction kind of just are the same thing. Maybe try Poisson example where parameter is in real number while the prediction is in $\mathbb{Z}$.

What I do not follow is the output of the classifier is a weighted matrix over all the labels. What is the ranking being done over? Is it

21.05.2025. Papers discussed: PNAS and general case

For the simple case (the Gaussian location family in 3.2.1 in Hao (2025)), my interpretation is as followed: If negative semidefnite $\Sigma_q - \Sigma_p$, then we reduced to the simple case where the RIPr $\mathtt{E}$-variable is not only in $\operatorname{conv}(\mathcal{P})$ (the convex hull of the null dist.) but rather a element of $\mathcal{P}$ itself.

Otherwise, we can find a prior on $\mathcal{P}$ with sharp variance $\mathcal{N}(\boldsymbol{\mu}^*, (\Sigma_q - \Sigma_p)/n)$. The results in turn suggests that we cast a sharp prior on $\mathcal{P}$ with the same mean. However, extra argument about the alternative Q, I think we move to the distributions from Q that shares the same sufficient statistics. This is still the part where I am having some doubts. I am still sure is that what does it mean to operate on a 'enlarged' or 'modified' alternative even if the alternative $\mathcal{Q} = \{Q\}$ being just Gaussian with same mean and a different variance.

Another point is in the Anti-Simple case, then $S^{(n)}_{Q,\text{SEQ,RIP}} = 1$. First of all, the superscript $^{(n)}$ says it is considering a sequential settings but $X^{(i)}$ is *singular* following either null dist.

or alternative dist. This $S = 1$ breaks simply when we are considering two X following the same distribution, we then test whether $X' = (X, Y) \overset{\text{i.i.d.}}{\sim} P \in \mathcal{P}$ or alternative.

Data spilting in UI In a sense, the spilting is done sequentially or done across n data points in contrast to the D_0 D_1 in the original paper. Now we are talking about the UI in (3.2.14) I think! The classical on would be $L = L_0(\hat{\theta}_1)/L_0(\hat{\theta}_0)$

$$S^{(n)}_{\breve{\boldsymbol{\mu}},\mathrm{UI}} = \frac{\prod_{i=1}^{n} q_{\breve{\boldsymbol{\mu}}|_{i-1}(U_{(i)})}}{p_{\breve{\boldsymbol{\mu}}|_{n}(U^{(n)})}}$$

We have a regular ML estimator likelihood in the denominator after looking at the whole sequence n. For the nominator, we have a product of the ML likelihood for each time i. In the end, we have effectively losing just $U_{(1)}$ in calculating likelihood.

Beyond NP We discussed about the similarities among the loss function in the ERM scheme and in this paper. The ERM loss is always with respect to some data (empirical) while here we are concerning about a data-dependent loss?

The problem or rather common difficulty with p-value in NP paradigm is that why do we report p-value at all? What does it mean for a small p-value when $p \ll \alpha$?

Peter proposed a alternative as in we should report E-value instead. I think also the notation of seeing α as the error probability is somewhat challenged.

Inspiration In vaccination study, we see a extremely small p-value on null but we are not allowed to make *new* decision without setting up new hypothesises and new studies. In other words, if we looked at a promising data *post-hoc*, there is really not much we can do based on the original data. But with $\mathtt{E}$-value, he argued that new decision can be based/conditioned on the data.

Is l in Eq. 1 Grünwald (2024) just a new α?

In this loss function $L(\kappa, a)$, κ is the true state of nature that we don't know? We just assumed if the data is coming from null or alternative.

In all, I think the main idea is to change from Type-I error safe to Type-I risk safe. I think section 2's main take home message is that any maximally compatible decision rule must be a $\mathtt{E}$-variable

This is cited from Micheal I Jordan's notes: https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter8.pdf

3. TLDR for papers. Here I really hope that I can somehow summarise in my words of the talks or presentation I have been to or literature I have read into a *short* paragraph.

There are only two requirement:

- short: ideally in one sentence, but it should definitely shorter than the abstract
- memorable/impressionable: for example, instead of precisely stating the LLN, it is better to write "keep tossing a fair coin, we are sure about the average being 0.5".

Admissible Anytime-Valid Sequential Inference Must Rely on Nonnegative Martingale (Ramdas et al., 2022a):

> p-process, e-process, stopping times are sort of the same: we can always find a nonnegative martingale for them. However, there is a gap.

The numeraire e-variable and reverse information projection (**?**):

> Shows that there is always a numeraire $\mathtt{E}$-variable X^* in the form of $\mathbb{E}_Q\left[X/X^*\right] \leq 1$ for every $\mathtt{E}$-variable X in the simple alternative Q vs composite null $\mathcal{P}$ setting. It is unique (up to Q-nullset), log-optimal, and connection between the effective null !

Reverse Information Projections and Optimal E-statistics (Lardy, Grünwald and Harremoës, 2024):

> There is always a RIPr (?) even when the KL is infinite in the simple alternative Q vs composite null $\mathcal{P}$ setting. In simple words, if the KL is infinite (bad at describing compared), there is still a relatively better version to describe that!.

4. Reverse Inverse Projections (RIPr).

4.1. *Li's Algorithm.* Originally developed by Li (1999), the following algorithm is written based on the notes from Grünwald, de Heide and Koolen (2024a); Hao et al. (2024). To quote Brinda (2018), Li's inequality requires the family $\mathcal{Q}$ to have a uniformly bounded density ratio.

Li obtains the RIPr in a greedy manner where the KL divergence between distribution Q onto the convex hull of a set of distributions $\mathcal{Q}$ (composite null) is minimized. It is assumed that the KL divergence between Q and any distribution $Q \in \mathcal{Q}$ is finite [1] (often call "nondegenerate" condition).

Algorithm 1: Li's Algorithm

$Q_{(1)} = \arg\min_{Q \in \mathcal{Q}} D(P \| Q)$
for $m = 2, 3, \ldots, K$ **do**
  $Q := \alpha Q_{(m-1)} + (1-\alpha) Q'$
  $\alpha, Q' \leftarrow \arg\min_{\alpha, Q'} D(P \| Q)$

Here, the distribution Q' and α is chosen (coupled) such that the divergence is minimized. The minimizer need not be unique.

Regularity condition on alternative

Li's algorithm is apparently greedy with high fluctuation in initial steps. Additionally, this task is computationally expensive, and it is not clear of the convexity.

Is the returned mixture in the convex hull?

The returned $Q_{(m)}$ is in the convex hull. The first step returned a single element in $\mathcal{Q}$ with the smallest KL divergence. Iteratively, the linear combination is still in the convex hull, i.e.

$$
\begin{aligned}
Q_{(2)} &= \alpha_1 \cdot Q_{(1)} + (1-\alpha_1) \cdot Q'_{(1)} \\
Q_{(2)} &= \alpha_2 \cdot Q_{(2)} + (1-\alpha_2) \cdot Q'_{(2)} \\
&= \alpha_2 \cdot \alpha_1 \cdot Q_{(1)} + \alpha_2 \cdot (1-\alpha_1) \cdot Q'_{(1)} + (1-\alpha_2) \cdot Q'_{(2)}.
\end{aligned}
$$

It is clear that $Q_{(2)}$ or $Q_{(m)}$ would still be a convex combination of elements in $\mathcal{Q}$.

4.2. *Cisszar Algorithm.*

---

[1] Which should I use? $\mathcal{Q}$ or $\mathbb{Q}$

5. Nonnegative martingales and E-process.

DEFINITION 5.1 (inadmissible). For some $t \in \mathbb{N}_0$, and some distribution $P \in \mathcal{P}$, there exist $P(p'_t < p_t) > 0$.

The argument is the same for $\mathtt{E}$-process with 'greater' rather than 'smaller'
Gap between necessary and sufficient conditions

EXAMPLE 5.2. Let U be a uniform random variable in $[0, 1]$ and $\mathcal{Q}$ is the set of probability measures where X_1 is Bernoulli and $X_2, X_3, \dots = 0$. Let thje

5.1. *Stuff.* Being integrable or not, I think, is the main contribution from Wang and Ramdas (2024).

QUESTION 5.3. Which class of random processes is bigger? $\mathtt{E}$-processes or martingales?

If the e-process is defined by the stuff upper bounded NNSM, then martingales is a bigger? Or should I say for a set of distribution $\mathcal{Q}$ we can always find some NNSM upper bounding the e_t??

5.2. *Notation.* We limit ourself to discrete time only $n \in \mathbb{N}_0 = \{0, 1, 2 \dots\}$, usually written as $n \geq 0$. Fix a probability triplet $(\Omega, \mathcal{A}, \mathcal{P})$ and $\mathcal{F}_t$, a sigma-field at time t, defines a filtration, $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \mathcal{A}$. We call $\mathcal{F}_t$ the canonical filtration and $X = \{X_n\}_{n \geq 0}$ a stochastic process adapted to $\mathcal{F}$ if X_n is $\mathcal{F}_n$-measurable.

DEFINITION 5.4. An $\mathtt{E}$-process for a family $\mathcal{P}$ of probability measures, namely $\mathcal{P}$-$\mathtt{E}$-process, is a nonnegative process E, adapted to some filtrations such that

$$\mathbb{E}_{\mathtt{P}}\left[E_\tau\right] \leq 1 \quad \text{for every stopping time} \quad \tau$$

It can also be defined as any nonnegative process that is upper bounded, for every $P \in \mathcal{P}$, by a P-martingale starting at one.
Focus on the section 5-8 in Ramdas et al. (2022a).

DEFINITION 5.5 (Admissibiliy).

Have to check Prop. 36 first where the set Q is reduced to simpler subclass: 1) null with Lebesgue densities up to finite time 2) discrete distributions with iid.
Lemma 38: subGaussian with one atom. We can always create a subGaussian RV that has point mass in some ball (a, b) while behaving like Gaussian outside

6. Integrals. What is it other than the area under curve?

Interesting examples where the Riemann's integration fails is that $f(x) = 1$ if x is rational and 0 otherwise over the interval $[0, 1]$. Why this function's integral is problematic in Riemann's definition.

Lebesgue used measure theory to define integral.

Many times, when Riemann fails, Lebesgue works.

Set being countable and uncountable rely on if we can arrange them in a 'readable' sequence, could be infinite like rational numbers.

Proving that $\mathbb{R}$ is uncountable is not so trivial: decimal expansion.

No matter how hard you try to write the numbers in $[0, 1)$, there is always some values left out.

Continuum hypothesis: How do you check which ∞ is bigger? Or which set has the higher cardinality?

Definition 1.2.2 ()

EXAMPLE 6.1 (Set of Lebesgue measure zero, 1.2.2). A set S is zero in Lebesgue measure when it can be covered with a sequence of open intervals $(I_1, I_2, \dots)$. And the sum $\sum_0^\infty m(I_n)$ can be made arbitrarily small

In other words, for any $\epsilon > 0$, we can always find a sequence of I_n that covers S.

THEOREM 6.2. *Any countable infinite set of S has Lebesgue measure zero.*

REMARK 6.3. The set of measure zero would not change a damn thing on f

Before writing this section, I sometimes came across expressions such as $p(x) = A\,\mathrm{d}x$ or $p(x) = A'm(\mathrm{d}x)$ in defining exponential family. It relates to the fundamental concept of measure and a density in measure theory. When I see $\mathrm{d}x$ in evaluating integral, e.g. $f(x) = \int_{\mathbb{R}} x\,\mathrm{d}x = x^2$, it usually refers to the Lebesgue measure. In the case of $m(\mathrm{d}x)$, it generally emphasize that you are integrating with respect to a general measure m.

We skipped defining the carrier density and its measure for now.

- Correct carrying density
- Check if cumulant function can be differentiated infinitely

7. Exponential Family. The exponential family is a collection of parametric models with very elegant properties. Peter and Hao utilised it to arrive some nice results about the optimality of $\mathtt{E}$-variables, mainly relying on the duality. We called a model belonging to the exponential family if the underlying distributions can be written below.

DEFINITION 7.1 (Exponential Family).

$$p_\beta(U) := \exp\left(\beta^T t(U) - \psi(\beta)\right) r(U) \tag{7.1}$$

$$= \frac{1}{Z(\beta)} \exp\left(\beta^T X\right) r(U). \tag{7.2}$$

- U: *random variable*
- $X = t(U)$: *sufficient statistics*
- β: *canonical parameter*
- $\psi(\beta)$: *cumulant function*
- $r(y)$: the *carrying density*
- $Z(\beta)$: the *partition function*, $\psi(\beta) = \log Z(\beta)$

7.1. *Basic Properties.* Cumulant generating function $\psi(\beta)$: The first and second *cumulant* is mean and variance. $\psi(\beta)$ is differentiable infinitely.

DEFINITION 7.2 (Minimal). We say an exponential family representation is *minimal*. If not minimal, there is t_k that can be represented by the linear combination of t_j for $j \neq k$, requiring 1 less parameter.

We usually assume minimal if not mentioned.

DEFINITION 7.3 (Canonical Parameter Space). Canonical parameter space $\mathtt{B}$ is the set of parameter β where the following integration is finite

$$\mathtt{B} := \left\{ \beta : \int_{\mathcal{U}} \exp\left(\beta^T t(U)\right) r(U) < \infty \right\}.$$

REMARK 7.4. Every exponential family has a canonical parameterization with carrier density p_0 such that the canonical parameter space contains the origin, i.e. $\mathbf{0} \in \mathtt{B}$. (Section 18.4.3 in MDL)

PROOF. If $\mathbf{0} \notin \mathtt{B}$ or $r(U)$ is not a carrier density, we can pick any $\beta_0 \in \mathtt{B}$ and set

$$r'(U) = \frac{r(U)\exp(\beta t(U))}{\sum_U r(U)\exp(\beta t(U))}$$

Now the family corresponding to r' with parameter $\beta' = \beta - \beta_0$ is identical to the original family containing $\mathbf{0}$. $\square$

Definition 7.5 (Regular Exponential Family). If the canonical parameter space of exponential families are nonempty open set, we call such families *regular*.

Definition 7.6 (Full Exponential Family). The family is called *full* if the dimension of β equals the dimension of $\mathtt{B}$.

Definition 7.7 (Order). The order of an exponential family is the minimal dimension of $t(u)$ such that we can express the family using Eq. (7.1).

Definition 7.8 (Minimal Exponential Family). An exponential family is referred to as *minimal* if: a) there are no linear constraints among the components of the parameter vector β; b) there are no linear constraints among the components of the sufficient statistic $t(u)$ (in the latter case, with probability one under the measure m).

Example 7.9 (Non-minimal distribution). The simplest would be a multinomial distribution with parameter $()$. The PMF can be written as bit

$$s$$

We can reparameterize the probability distribution even in the case of minimal distribution. Given a (?arbitrary) set Θ and a mapping $\Phi : \Theta \to \mathtt{B}$, we consider the densities with canonical parameters replaced by $\Phi(\theta)$

add example

$$p_\theta(u) := \exp\big(\Phi(\theta)^T t(u) - \phi(\Psi(\theta))\big).$$

Here Φ is a one-to-one mapping whose image is all of $\mathtt{B}$.

If Φ's image is a strict subset of $\mathtt{B}$ ($\Phi(\Theta) \subset \mathtt{B}$), it is then OK to reparameterize on that subset. If it can not be reducible, we refer this exponential family as curved.

We are also interested in cases in which the image of ψ is a strict subset of N. If this subset is a linear subset, then it is possible to transform the representation into an exponential family on that subset. When the representation is not reducible in this way, we refer to the exponential family as a curved exponential family.

Definition 7.10 (Curved Exponential Family). TBH

Example 7.11 (Normal distribution with variance equalling to mean). Let $u \in \mathbb{R} \sim \mathcal{N}(\mu, \mu^2)$, $\mu \neq 0$. The density of u is

$$\begin{aligned} p_\mu(u) &= |\mu|^{-1}(2\pi)^{-1/2} \exp\left(-\frac{1}{2}(\frac{u-\mu}{\mu})^2\right) \\ &= (2\pi)^{-1/2} \exp\left(-\frac{u^2}{2\mu^2} + \frac{u}{\mu} - \frac{1}{2} + \log(|\mu|)\right) \\ &= \exp\big(\beta^T t(u) - \psi(\beta)\big). \end{aligned}$$

The $\beta^T = (-1/2\mu^2, 1/\mu)$, the sufficient statistics $t(u) = (u^2, u)$. The dimension of the sufficient statistic is more than the dimension of β for curved exp. family.

I have found some other versions of the same statement regarding minimal exponential family:

- A minimal exponential family is where the $t(u)$ are linearly independent
- A minimal exponential family is one where representation reaches the *order*

All share the same statement regarding the linear dependency of sufficient statistics $t(u)$, only Michael I. Jordan's Chapter 8 stated on β. I wonder if linear dependency in β implies linear dependency in $t(u)$, i.e.

$$a^T \beta = C \Leftrightarrow b^T t(u) = C'$$

He also claimed that non-minimal families can always be reduced to minimal families via a suitable transformation and reparameterization.

If an exponential family is not minimal, it is called *overcomplete*. Both minimal and overcomplete representations are useful

7.2. *Mean-value parameterization.* We

DEFINITION 7.12 (Matching pair). Let $\mathcal{P}, \mathcal{Q}$ be exponential family for random variable U. We can they are *matching pair* if

- they are mutually absolutely continuous wrt each other
- they have the same sufficient statistics X
- their mean-value space, $\mathtt{M}_q \subseteq \mathtt{M}_p$
- for every matching canonical parameterisation of $\mathcal{P}$ and $\mathcal{Q}$, we have $\mathtt{B}_p \subseteq \mathtt{B}_q$

7.3. *Simple Case.*

- Why I think this is the carrier probability with some measure v $p_{\boldsymbol{\mu}^*} = p_{\mathbf{0},\boldsymbol{\mu}^*} \quad q_{\mathbf{0},\boldsymbol{\mu}^*} = q = q_{\boldsymbol{\mu}^*}$:

  It is rather simple when you write down $\boldsymbol{\beta} = \mathbf{0}$

$$\begin{aligned}
p_{\mathbf{0};\boldsymbol{\mu}^*}(u) &:= \frac{1}{Z_p(\mathbf{0};\boldsymbol{\mu}^*)} \exp\big(\mathbf{0}^T t(u)\big) \cdot p_{\boldsymbol{\mu}^*}(u) \\
&= \frac{p_{\boldsymbol{\mu}^*}(u)}{Z_p(\mathbf{0};\boldsymbol{\mu}^*)} \\
&= \frac{p_{\boldsymbol{\mu}^*}(u)}{\int \exp(\mathbf{0}^T t(u)) p_{\boldsymbol{\mu}^*}(u)\, \mathrm{d}\nu} = p_{\boldsymbol{\mu}^*}(u)
\end{aligned}$$

- A naive question follows: what is the distribution $p_{\boldsymbol{\mu}^*}(u)$? Is it in the mean-value parameterization or canonical form?

7.4. *General Case.*

- Why Hao explicitly denote the mean of the carrier density in Grünwald, de Heide and Koolen (2024a):

We first discuss the Gaussian location family where the mean $\boldsymbol{\mu}^* \in \mathtt{M}_p = \mathbb{R}^d$. The null distribution will be family with $\boldsymbol{\mu}^*$ and a positive semidefinite covariance matrix Σ_p and the alternative $\mathcal{Q} = \{Q\}$ would be Gaussian distribution with the same $\boldsymbol{mu}^*$ and a covariance matrix $\Sigma_p \neq \Sigma_q$.

The notation D_{GAUSS} in Eq. (3.2.1) in Hao (2025) is just the KL divergence. Here $X = U = (X_1, \ldots, X_d)$ is the d-dimensional random vector with distribution p or q

$$\begin{aligned}
&D_{\text{GAUSS}}(B) = \mathrm{D_{KL}}(P \,\|\, Q) \\
&\qquad := \int_{\mathcal{U}} p(X) \log \frac{p(X)}{q(X)} = \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right]
\end{aligned}$$

$$= \mathbb{E}_p\left[\log \frac{(2\pi)^{-d/2}\det(\Sigma_p)^{-1/2}\exp\left(-\frac{1}{2}(X-\boldsymbol{\mu}^*)^T\Sigma_p^{-1}(X-\boldsymbol{\mu}^*)\right)}{(2\pi)^{-d/2}\det(\Sigma_q)^{-1/2}\exp\left(-\frac{1}{2}(X-\boldsymbol{\mu}^*)^\top\Sigma_q^{-1}(X-\boldsymbol{\mu}^*)\right)}\right]$$

$$= \mathbb{E}_p\left[\log \frac{\det(\Sigma_q)^{-1/2}\exp\left(\Sigma_p^{-1}\right)}{\det(\Sigma_q)^{-1/2}\exp\left(\Sigma_q^{-1}\right)}\right] \quad \text{Wrong!}$$

For more details, check https://statproofbook.github.io/P/mvn-kl.html for cases where the means are different.

Why Eq. (3.4.2) holds? The part I don't understand is that why putting a prior will equal to $p_{\boldsymbol{\mu}^*}$

$$S_{\text{COND}} := \frac{q_{W_1}(U^{(n)} \mid Z)}{p_{W_0}(U^{(n)} \mid Z)} \overset{?}{=} \frac{q_{\boldsymbol{\mu}^*}(U^{(n)} \mid Z)}{p_{\boldsymbol{\mu}^*}(U^{(n)} \mid Z)}$$

The key is just simple derivation with the fact that $\hat{X}_{|n} := \sum X_i/n \sim \mathcal{N}(n\boldsymbol{\mu}^*, n\Sigma_q)$. Setting $Z = \hat{X}_{|n}/\sqrt{n} \sim \mathcal{N}(\sqrt{n}\boldsymbol{\mu}^*, \Sigma_q)$, we have

$$\begin{aligned}
q_{W_1}(X^{(n)} \mid Z) &:= \frac{q_{W_1}(X^{(n)}, Z)}{q_{W_1}(Z)} = \frac{q_{W_1}(X^{(n)})}{q_{W_1}(Z)} \overset{iid}{=} \frac{\prod_{i=1}^n q_{W_1}(X_i)}{q_{W_1}(Z)} \\
&= \frac{\prod (2\pi)^{-d/2}(\det\Sigma_q)^{-1}\exp\left(-1/2(X_i-\boldsymbol{\mu}^*)^\top\Sigma_q^{-1}(X_i-\boldsymbol{\mu}^*)\right)}{(2\pi)^{-d/2}(\det\Sigma_q)^{-1}\exp\left(-1/2(Z-\sqrt{n}\boldsymbol{\mu}^*)^\top\Sigma_q^{-1}(Z-\sqrt{n}\boldsymbol{\mu}^*)\right)} \\
&= \frac{\exp\left(\sum_{i=1}^n X_i^\top\Sigma_q^{-1}X_i - 2\sqrt{n}\boldsymbol{\mu}^{*\top}\Sigma_q^{-1}Z + n\boldsymbol{\mu}^{*\top}\Sigma_q^{-1}\boldsymbol{\mu}^*\right)}{\exp\left(Z^\top\Sigma_q^{-1}Z - 2\sqrt{n}\boldsymbol{\mu}^{*\top}\Sigma_q^{-1}Z + n\boldsymbol{\mu}^{*\top}\Sigma_q^{-1}\boldsymbol{\mu}^*\right)} \\
&= \exp\left(\sum_{i=1}^n (X_i - Z)^\top\Sigma_q^{-1}(X_i - Z)\right)
\end{aligned}$$

The equality follows due to the joint distribution $X^{(n)}$ and Z is simply $X^{(n)}$.
For UI case, I don't see any sign of splitting? It looks like just ML prequential settings
What makes the E-process-ness? th
Why the $S_{\text{SEQ}} = S_{\text{SEQ,RIP}}$ in Thm. 2?

7.5. *One-dimensional special case.*

$$\mathcal{G} := \{g_\eta(y), \eta \in A, \in \mathcal{Y}\}, \quad A \text{ and } \mathcal{Y} \in \mathbb{R}^p$$
$$g_\eta(y) := \exp\left(\eta y - \psi(\eta)\right) g_0(y) m(\mathrm{d}y)$$

REMARK 7.13 (Carrying density in p.3 efron). Any $g_{\eta_0}(y) \in \mathcal{G}$ could be the carrier density and the members of $\mathcal{G}$ are absolutely continuos with respect to each other, i.e. their null measure sets agree.

REMARK 7.14. *Cumulant generating function* for $\psi(\eta)$ originates from the old techniques for finding expectations, variances and higher-order moments.

7.6. *Moment Relationships.* We can differentiate $\exp(\psi(\eta)) = \int_{\mathcal{Y}} \exp(\eta y) g_0(y) m(\mathrm{d}y)$

$$\dot{\psi}(\eta)\exp(\psi(\eta)) = \int_{\mathcal{Y}} y\exp(\eta y)g_0(y)m(\mathrm{d}y)$$

$$\mathbb{E}_{g_0}[y] = \dot{\psi}(\eta) = \int_{\mathcal{Y}} y\exp(\eta y - \psi(\eta))g_0(y)m(\mathrm{d}y)$$

Differentiating $\exp(\psi(\eta))$ twice gives:

$$(\ddot{\psi}(\eta) + \dot{\psi}(\eta))\exp(\psi(\eta)) = \int_{\mathcal{Y}} y^2\exp(\eta y)g_0(y)m(\mathrm{d}y)$$

$$\mathrm{Var}(y) = \ddot{\psi}(\eta) = \int_{\mathcal{Y}} (y^2 - y)\exp(\eta y - \psi(\eta))g_0(y)m(\mathrm{d}y)$$

Efron's book (Chap. 2)

$$\mathcal{G} := \{g_\eta(y), \eta \in A, y \in \mathcal{Y}\}, \quad A \text{ and } \mathcal{Y} \in \mathbb{R}^p$$

$$g_\eta(y) := \exp\left(\eta^T y - \psi(\eta)\right) g_0(y)m(\mathrm{d}y)$$

- $g_0(y)$: *carrying density* w.r.t. some *carrying measure* $m(\mathrm{d}y)$ on $\mathcal{Y}$
- A is the *canonical parameter space*:
- $\psi(\eta)$ is the *normalizing function* or *cumulant generating function* (CGF)

REFERENCES

ADAMS, R. J. (2020). Safe Hypothesis Tests for the 2×2 Contingency Table, Master's thesis, TU delft.

ALTHAM, P. M. E. (1969). Exact Bayesian Analysis of a 2 x 2 Contingency Table, and Fisher's "Exact" Significance Test. *Journal of the Royal Statistical Society. Series B (Methodological)* **31** 261–269.

ARNOLD, S., HENZI, A. and ZIEGEL, J. F. (2023). Sequentially Valid Tests for Forecast Calibration. *The Annals of Applied Statistics* **17**. https://doi.org/10.1214/22-AOAS1697

BERGER, J. O. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statistical Science* **18**. https://doi.org/10.1214/ss/1056397485

BRANNATH, W. and SCHACHERMAYER, W. (1999). A Bipolar Theorem for Subsets of $L^0_+(\omega, \mathcal{F}, P)$. In *Séminaire de Probabilités XXXIII* 349–354. Springer, Berlin, Heidelberg. https://doi.org/10.1007/BFb0096525

BRINDA, W. D. (2018). Adaptive Estimation with Gaussian Radial Basis Mixtures, PhD thesis, Yale University.

CASGRAIN, P., LARSSON, M. and ZIEGEL, J. (2023). Anytime-Valid Sequential Testing for Elicitable Functionals via Supermartingales. https://doi.org/10.48550/arXiv.2204.05680

CHOE, Y. J. (2023). Comparing Forecasters and Abstaining Classifiers, PhD thesis, Carnegie Mellon University. https://doi.org/10.1184/R1/23576451.v1

CHOE, Y. J. and RAMDAS, A. (2024). Comparing Sequential Forecasters. *Operations Research* **72** 1368–1387. https://doi.org/10.1287/opre.2021.0792

CLARKE, B. S. and BARRON, A. R. (1994). Jeffreys' Prior Is Asymptotically Least Favorable under Entropy Risk. *Journal of Statistical Planning and Inference* **41** 37–60. https://doi.org/10.1016/0378-3758(94)90153-8

CLERICO, E. (2025). Optimal E-Value Testing for Properly Constrained Hypotheses. https://doi.org/10.48550/arXiv.2412.21125

CSISZÁR, I., KATONA, G. O. H. and TARDOS, G. (2007). *Entropy, Search, Complexity*. *Bolyai Society Mathematical Studies* **16**. Springer, Berlin.

CSISZÁR, I. and TUSNÁDY, G. (1984). Information Geometry and Alternating Minimization Procedures. *Statistics and Decisions, Supplement Issue* **1** 205–237.

DE HEIDE, R. and GRÜNWALD, P. D. (2021). Why Optional Stopping Can Be a Problem for Bayesians. *Psychonomic Bulletin & Review* **28** 795–812. https://doi.org/10.3758/s13423-020-01803-x

DE ROOS, D. (2025). Uniformly Growth-Rate Optimal e-Variables for Composite Hypotheses Sharing a Sufficient σ-Algebra, PhD thesis, Leiden University.

DIACONIS, P. and YLVISAKER, D. (1979). Conjugate Priors for Exponential Families. *The Annals of Statistics* **7** 269–281. https://doi.org/10.1214/aos/1176344611

FISCHER, L. and RAMDAS, A. (2024). Improving the (Approximate) Sequential Probability Ratio Test by Avoiding Overshoot. https://doi.org/10.48550/arXiv.2410.16076

GRUNWALD, P. A Picture That Says A Lot.

GRUNWALD, P. (2004). A Tutorial Introduction to the Minimum Description Length Principle. https://doi.org/10.48550/arXiv.math/0406077

GRÜNWALD, P. (2023). The E-Posterior. https://doi.org/10.1098/rsta.2022.146

GRÜNWALD, P. D. (2024). Beyond Neyman–Pearson: E-values Enable Hypothesis Testing with a Data-Driven Alpha. *Proceedings of the National Academy of Sciences* **121** e2302098121. https://doi.org/10.1073/pnas.2302098121

GRUNWALD, P. D. and DAWID, A. P. (2004). Game Theory, Maximum Entropy, Minimum Discrepancy and Robust Bayesian Decision Theory. *The Annals of Statistics* **32**. https://doi.org/10.1214/009053604000000553

GRÜNWALD, P., DE HEIDE, R. and KOOLEN, W. (2024a). Safe Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86** 1091–1128. https://doi.org/10.1093/jrsssb/qkae011

GRÜNWALD, P., DE HEIDE, R. and KOOLEN, W. (2024b). Authors' Reply to the Discussion of 'Safe Testing'. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86** 1163–1171. https://doi.org/10.1093/jrsssb/qkae069

GRÜNWALD, P., HAO, Y. and BALSUBRAMANI, A. (2024). Growth-Optimal E-variables and an Extension to the Multivariate Csiszár-Sanov-Chernoff Theorem. https://doi.org/10.48550/arXiv.2412.17554

GRÜNWALD, P. D. and MEHTA, N. A. (2020). Fast Rates for General Unbounded Loss Functions: From ERM to Generalized Bayes. *Journal of Machine Learning Research* **21** 1–80.

GRÜNWALD, P., LARDY, T., HAO, Y., BAR-LEV, S. K. and DE JONG, M. (2024). Optimal E-values for Exponential Families: The Simple Case.

HAO, Y. (2025). E-Values for Anytime-Valid Inference with Exponential Families, PhD thesis, Leiden University, Leiden.

HAO, Y. and GRÜNWALD, P. (2024). E-Values for Exponential Families: The General Case. https://doi.org/10.48550/arXiv.2409.11134

HAO, Y., GRÜNWALD, P., LARDY, T., LONG, L. and ADAMS, R. (2024). E-Values for k-Sample Tests with Exponential Families. https://doi.org/10.48550/arXiv.2303.00471

HARTOG, W. and LEI, L. (2025). Family-Wise Error Rate Control with E-values. https://doi.org/10.48550/arXiv.2501.09015

HASHEMI, L., NANDRAM, B. and GOLDBERG, R. (1997). Bayesian Analysis for a Single 2×2 Table. *Statistics in Medicine* **16** 1311–1328. https://doi.org/10.1002/(SICI)1097-0258(19970630)16:12<1311::AID-SIM568>3.0.CO;2-3

HENDRIKSEN, A., DE HEIDE, R. and GRÜNWALD, P. (2021). Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations. *Bayesian Analysis* **16**. https://doi.org/10.1214/20-BA1234

HENZI, A. and ZIEGEL, J. F. (2022). Valid Sequential Inference on Probability Forecast Performance. https://doi.org/10.48550/arXiv.2103.08402

HONDA, J. and TAKEMURA, A. (2010). An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In *Annual Conference Computational Learning Theory*.

HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. (2020). Time-Uniform Chernoff Bounds via Nonnegative Supermartingales. https://doi.org/10.48550/arXiv.1808.03204

HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. (2021). Time-Uniform, Nonparametric, Nonasymptotic Confidence Sequences. *The Annals of Statistics* **49**. https://doi.org/10.1214/20-AOS1991

JOHNSON, V. E. (2013). Uniformly Most Powerful Bayesian Tests. *The Annals of Statistics* **41** 1716–1741. https://doi.org/10.1214/13-AOS1123

KIRSCH, W. (2018). An Elementary Proof of de Finetti's Theorem. https://doi.org/10.48550/arXiv.1809.00882

KOOLEN, W. M. and GRÜNWALD, P. (2022). Log-Optimal Anytime-Valid E-values. *International Journal of Approximate Reasoning* **141** 69–82. https://doi.org/10.1016/j.ijar.2021.09.010

KOOLEN, W. M., PÉREZ-ORTIZ, M. F. and LARDY, T. (2025). A Generalisation of Ville's Inequality to Monotonic Lower Bounds and Thresholds. https://doi.org/10.48550/arXiv.2502.16019

LARDY, T., GRÜNWALD, P. and HARREMOËS, P. (2024). Reverse Information Projections and Optimal E-statistics. *IEEE Transactions on Information Theory* **70** 7616–7631. https://doi.org/10.1109/TIT.2024.3444458

LARSSON, M., RAMDAS, A. and RUF, J. (2025a). The Numeraire E-Variable and Reverse Information Projection. https://doi.org/10.48550/arXiv.2402.18810

LARSSON, M., RAMDAS, A. and RUF, J. (2025b). The Numeraire E-Variable and Reverse Information Projection. *The Annals of Statistics* **53** 1015–1043. https://doi.org/10.1214/24-AOS2487

LARSSON, M., RAMDAS, A. and RUF, J. (2025c). Testing Hypotheses Generated by Constraints. https://doi.org/10.48550/arXiv.2504.02974

LI, Q. J. (1999). Estimation of Mixture Models, PhD thesis, Yale University, USA.

LIEB, E. H., OSHERSON, D. and WEINSTEIN, S. (2006). Elementary Proof of a Theorem of Jean Ville. https://doi.org/10.48550/arXiv.cs/0607054

LY, A., BOEHM, U., GRÜNWALD, P., RAMDAS, A. and VAN RAVENZWAAIJ, D. (2024). Safe Anytime-Valid Inference: Practical Maximally Flexible Sampling Designs for Experiments Based on e-Values. https://doi.org/10.31234/osf.io/h5vae

MARSHALL, R. J. (1988). Bayesian Analysis of Case-control Studies. *Statistics in Medicine* **7** 1223–1230. https://doi.org/10.1002/sim.4780071203

MONIRI, B. (2022). Universal Inference.

MYUNG, J. I., NAVARRO, D. J. and PITT, M. A. (2006). Model Selection by Normalized Maximum Likelihood. *Journal of Mathematical Psychology* **50** 167–179. https://doi.org/10.1016/j.jmp.2005.06.008

PÉREZ, M.-E. (1994). An Automatic and Proper Bayesian Estimation Analysis of 2×2 Contingency Tables with One and Two Fixed Margins. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* **3** 101–112.

PÉREZ-ORTIZ, M. F., LARDY, T., DE HEIDE, R. and GRÜNWALD, P. (2023). E-Statistics, Group Invariance and Anytime Valid Testing.

PÉREZ-ORTIZ, M. F., LARDY, T., DE HEIDE, R. and GRÜNWALD, P. D. (2024). E-Statistics, Group Invariance and Anytime-Valid Testing. *The Annals of Statistics* **52** 1410–1432. https://doi.org/10.1214/24-AOS2394

POSNER, E. (1975). Random Coding Strategies for Minimum Entropy. *IEEE Transactions on Information Theory* **21** 388–391. https://doi.org/10.1109/TIT.1975.1055416

RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. (2022a). Admissible Anytime-Valid Sequential Inference Must Rely on Nonnegative Martingales. https://doi.org/10.48550/arXiv.2009.03167

RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. M. (2022b). Testing Exchangeability: Fork-Convexity, Supermartingales and e-Processes. *International Journal of Approximate Reasoning* **141** 83–109. https://doi.org/10.1016/j.ijar.2021.06.017

RAMDAS, A., GRÜNWALD, P., VOVK, V. and SHAFER, G. (2023). Game-Theoretic Statistics and Safe Anytime-Valid Inference.

Ruf, J., Larsson, M., Koolen, W. M. and Ramdas, A. (2023). A Composite Generalization of Ville's Martingale Theorem Using e-Processes. *Electronic Journal of Probability* **28**. https://doi.org/10.1214/23-EJP1019

Shafer, G. What Is Game-Theoretic Statistics?

Shafer, G., Shen, A., Vereshchagin, N. and Vovk, V. (2011). Test Martingales, Bayes Factors and p-Values. *Statistical Science* **26**. https://doi.org/10.1214/10-STS347

Topsøe, F. (2007). Information Theory at the Service of Science. In *Entropy, Search, Complexity*, **16** 179–207. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-32777-6_8

Rosanne J Turner (2019). Safe Tests for 2 x 2 Contingency Tables and the Cochran-Mantel-Haenszel Test, PhD thesis, Leiden University.

Rosanne J. Turner and Peter Grünwald (2022). Anytime-Valid Testing and Confidence Intervals in Contingency Tables and Beyond.

Turner, R. and Grunwald, P. (2023). Safe Sequential Testing and Effect Estimation in Stratified Count Data. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research* **206** 4880–4893. PMLR.

Turner, R. J. and Grünwald, P. D. (2023). Exact Anytime-Valid Confidence Intervals for Contingency Tables and Beyond. *Statistics & Probability Letters* **198** 109835. https://doi.org/10.1016/j.spl.2023.109835

Turner, R. J., Ly, A. and Grünwald, P. D. (2024). Generic E-variables for Exact Sequential k-Sample Tests That Allow for Optional Stopping. *Journal of Statistical Planning and Inference* **230** 106116. https://doi.org/10.1016/j.jspi.2023.106116

Ville, J. (1939). *Étude critique de la notion de collectif*. *Monographies des probabilités* **3**. Gauthier-Villars, Paris.

Vovk, V. and Wang, R. (2021). E-Values: Calibration, Combination, and Applications. *The Annals of Statistics* **49**. https://doi.org/10.1214/20-AOS2020

Vovk, V. and Wang, R. (2023). Confidence and Discoveries with E-Values. *Statistical Science* **38**. https://doi.org/10.1214/22-STS874

Wagenmakers, E.-J. and Ly, A. (2020). Bayesian Scepsis about SWEPIS: Quantifying the Evidence That Early Induction of Labour Prevents Perinatal Deaths. https://doi.org/10.31234/osf.io/5ydpb

Wang, R. (2022). Testing with P*-Values: Between p-Values, Mid p-Values, and e-Values.

Wang, R. (2023). A Tiny Review on E-Values and e-Processes.

Wang, R. and Ramdas, A. (2021). False Discovery Rate Control with E-Values.

Wang, H. and Ramdas, A. (2024). The Extended Ville's Inequality for Nonintegrable Nonnegative Supermartingales. https://doi.org/10.48550/arXiv.2304.01163

Wasserman, L., Ramdas, A. and Balakrishnan, S. (2020). Universal Inference. *Proceedings of the National Academy of Sciences* **117** 16880–16890. https://doi.org/10.1073/pnas.1922664117

Xu, Z. and Ramdas, A. (2024). Online Multiple Testing with E-Values. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics* 3997–4005. PMLR.

Xu, Z., Wang, R. and Ramdas, A. (2021). A Unified Framework for Bandit Multiple Testing. In *Advances in Neural Information Processing Systems* **34** 16833–16845. Curran Associates, Inc.

A Note on Universal Inference - Tse - 2022 - Stat - Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1002/sta4.501.

A Statistical Analysis of Adverse Impact of Employer Decisions. https://www.periodicos.capes.gov.br/index.php/acervo/buscador.html?task=de

A Case Study on the Bayesian Analysis of 2 × 2 Tables with All Margins Fixed.