Sequential Analysis and Martingales

Liam Donovan

May 2025

Abstract

In certain contexts, such as in clinical trials, making a decision in the least possible samples is paramount. Traditional hypothesis testing considers a fixed sample size, potentially delaying findings. Here, we will examine a type of testing where the test statistic is updated as samples arrive, allowing a decision as soon as the evidence is strong enough. These are called sequential tests. We examine sequential tests which take the form of products of likelihood ratios. These products can be formulated to be martingales, under the null hypothesis, allowing for the use of the optional stopping theorem and Ville's inequality to bound the type-I error rate.

1 Wald's Sequential Ratio Test (SPRT)

We first consider the most straightforward situation-a simple null and alternate hypothesis. Let θ be an unknown parameter:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

and let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from density $f(x;\theta)$. If observations are arriving in time, a natural choice of test is the examine the likelihood test. In particular, by convention, we will consider the reciprocal of the traditional likelihood ratio test statistic, which we denote L_n for n observations.

After 1 observation:

$$L_1 = \frac{f(x_1;\theta_1)}{f(x_1;\theta_0)}$$

and 2 observations:

$$L_2 = \frac{f(x_2;\theta_1)}{f(x_2;\theta_0)}\frac{f(x_1;\theta_1)}{f(x_1;\theta_0)} = \frac{f(x_2;\theta_1)}{f(x_2;\theta_0)}L_1.$$

Proceeding by induction:

$$L_{n+1} = \prod_{i=1}^{n}\frac{f(x_i;\theta_1)}{f(x_i;\theta_0)} = \frac{f(x_{n+1};\theta_1)}{f(x_{n+1};\theta_0)}L_n$$

Observe that this satisfies the Markov property and so $\{L_n\}$ is a Markov chain. Moreover, $\{L_n\}$ is adapted to the natural filtration $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$.

Under H_0 (i.e. when $X_{n+1} \sim f(\cdot;\theta_0)$), take conditional expectations:

$$\mathbb{E}_{\theta_0}\big[L_{n+1} \mid \mathcal{F}_n\big] = L_n\,\mathbb{E}_{\theta_0}\Big[\tfrac{f(X_{n+1};\theta_1)}{f(X_{n+1};\theta_0)}\Big] = L_n \cdot 1 = L_n.$$

Hence $\{L_n, \mathcal{F}_n\}$ is a (nonnegative) martingale under H_0.

Further, notice that the structure under H_0, also applies if we consider the reciprocal of L_n under H_1.

$$\mathbb{E}_{\theta_1}\left[1/L_{n+1} \mid \mathcal{F}_n\right] = \frac{1}{L_n}\,\mathbb{E}_{\theta_1}\left[\tfrac{f(X_{n+1};\theta_0)}{f(X_{n+1};\theta_1)}\right] = \frac{1}{L_n}\cdot 1 = \frac{1}{L_n}.$$

and so $\{1/L_n, \mathcal{F}_n\}$ is a martingale under H_1.

1.1 Decision Rules and Stopping Times

Observe that by the construction of L_n, if the θ_1 is true, then L_n will tend to increase with time, since $f(x_i;\theta_1) > f(x_i;\theta_0)$, on average. Conversely, if the θ_0 is the true parameter, the ratio will tend to decrease with time.

An insight by Wald [1947] was to choose a stopping rule such that we reject the null hypothesis if the ratio grows too large and fail to reject if the ratio becomes too small. That is, let τ be the stopping time:

$$\tau = \inf\{n \geq 1 : L_n \leq A \text{ or } L_n \geq B\}.$$

We must now find suitable A and B. Using the optional stopping theorem on both martingales specified above:

$$\begin{aligned}\mathbb{E}\left(L_\tau\right) &= \mathbb{E}(L_0)\\ \mathbb{E}\left(\frac{1}{L_\tau}\right) &= \mathbb{E}\left(\frac{1}{L_0}\right).\end{aligned}$$

Although, L_0 is not mentioned explicitly, but it must satisfy:

$$L_1 = \frac{f(x_1;\theta_1)}{f(x_1;\theta_0)}L_0$$

which leads to a natural choice of $L_0 = 1$. Thus,

$$\begin{aligned}\mathbb{E}_{\theta_0}\left(L_\tau\right) &= 1\\ \mathbb{E}_{\theta_1}\left(\frac{1}{L_\tau}\right) &= 1\end{aligned}$$

Recall that τ is the time where a decision is made. Therefore, by definition:

$$\begin{aligned}\mathbb{E}_{\theta_0}\left(L_\tau\right) &= A\cdot\mathbb{P}_{\theta_0}(L_\tau = A) + B\cdot\mathbb{P}_{\theta_0}(L_\tau = B) = 1.\\ \mathbb{E}_{\theta_1}\left(\frac{1}{L_\tau}\right) &= \frac{1}{A}\cdot\mathbb{P}_{\theta_1}(L_\tau = A) + \frac{1}{B}\cdot\mathbb{P}_{\theta_1}(L_\tau = B) = 1.\end{aligned}$$

Recall that the $L_\tau = A, B$ corresponds to failing to reject the null and rejecting, respectively. Therefore,

$$\begin{aligned}&\mathbb{P}_{\theta_0}(L_\tau = B) = \alpha \quad \text{(type-I error)}\\ &\mathbb{P}_{\theta_1}(L_\tau = A) = \beta \quad \text{(type-II error)}\\ \implies\ &\mathbb{P}_{\theta_0}(L_\tau = A) = 1-\alpha\\ &\mathbb{P}_{\theta_1}(L_\tau = B) = 1-\beta\end{aligned}$$

which gives:

$$\mathbb{E}_{\theta_0}(L_\tau) = A(1-\alpha) + B\alpha = 1.$$
$$\mathbb{E}_{\theta_1}\left(\frac{1}{L_\tau}\right) = \frac{1}{A}\cdot\beta + \frac{1}{B}\cdot(1-\beta) = 1.$$

solving this system is straightforward:

$$\begin{cases} A = \frac{\beta}{1-\alpha} \\ B = \frac{1-\beta}{\alpha}. \end{cases}$$

Hence your decision rule becomes

$$L_n \leq \frac{\beta}{1-\alpha} \implies \text{accept } H_0, \qquad L_n \geq \frac{1-\beta}{\alpha} \implies \text{reject } H_0.$$

This is called **Wald's Sequential Ratio Test** (SPRT).

1.2 Example: Bernoulli Trials

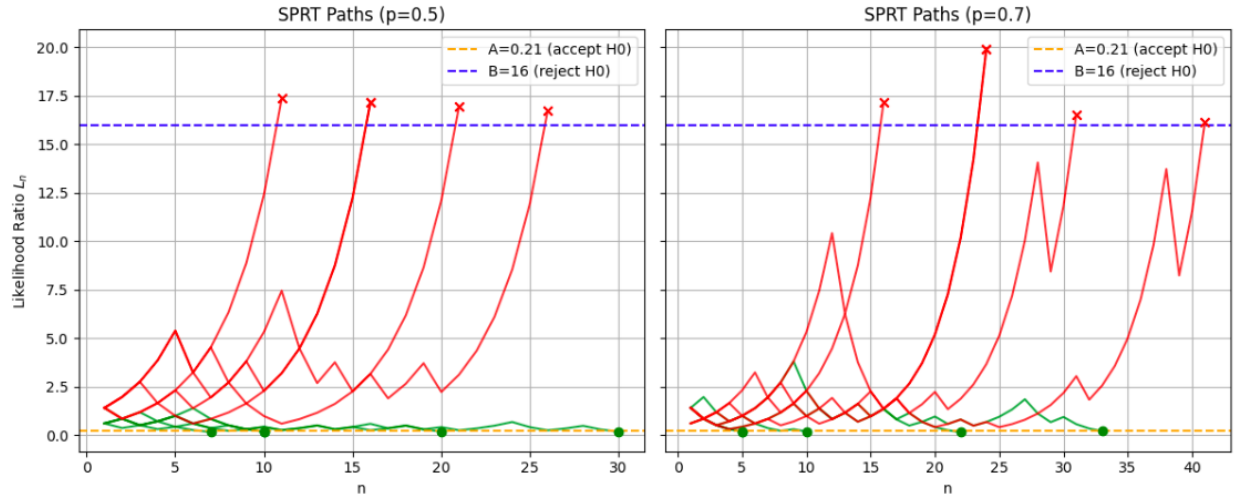
Consider the following test: Let $X_1, X_2 \ldots$ be iid Bernoulli(p). We are interested in testing:

$$H_0 : p = .5$$
$$H_1 : p = .7$$

Further, suppose we chose the standard $\alpha = .05$ and $\beta = .2$ (power of .8). This gives:

$$A = \frac{.2}{.95} \approx .21, \qquad B = \frac{.8}{.05} = 16$$

We run two simulations, one where the null is true ($p = .5$) and one where the alternate is true ($p = .7$). We run 500 trials and track the number of times each rejects and fails to reject, as well as the number of steps to make a decision.



```
p = 0.5: Fail to Reject rate = 0.950, Rejection rate = 0.050,
Avg stop (fail to reject) = 18.4, Avg stop (reject) = 22.7

p = 0.7: Fail to Reject rate = 0.162, Rejection rate = 0.838,
Avg stop (fail to reject) = 17.1, Avg stop (reject) = 28.4
```

We observe that the reject rate under the null is exactly the $\alpha = .05$, as expected. Conversely, the fail to reject rate under the alternate is a bit better than the suggested $\beta = .2$.

Moreover, notice that that, generally, the time to decision is quite small, but is smaller for the null case, since our chosen β was conservative.

2 Generalized Likelihood Ratio

The simple alternate hypothesis is a quite restrictive assumption. If the SPRT is used without a strong prior knowledge about the parameter, it may perform poorly or fail. Therefore, it is natural to ask whether the SPRT can be extended to a composite alternative hypothesis[1]. Let

$$\Theta_1 \subset \mathbb{R}, \text{ with } \theta_0 \notin \Theta_1$$

For instance, Θ_1 may be $\{\theta : \theta \neq \theta_0\}$ (two-sided test), or $\{\theta : \theta > \theta_0\}$ (one-sided test).

Recall that the SPRT was defining by using the reciprocal of the standard likelihood ratio test, iterated in time. Applying the same logic, one defines the **Generalized Likelihood Ratio** (GLR) (Fellouris and Tartakovsky [2012]).

$$G_n \coloneqq \sup_{\theta \in \Theta_1} \prod_{i=1}^{n} \frac{f(x_i;\theta)}{f(x_i;\theta_0)}.$$

One now seeks to derive the same "factorability" that made the SPRT a martingale, but the supremum term spoils this. It can be seen as the supremum of some fixed alternate hypothesis for the SPRT, since it may be rewritten:

$$G_n = \sup_{\theta \in \Theta_1} \prod_{i=1}^{n} M_n(\theta)$$

We can use this connection to bound the expected value of G_n:

$$\mathbb{E}\big[G_{n+1} \mid \mathcal{F}_n\big] = \mathbb{E}\Big[\sup_{\theta} M_{n+1}(\theta) \;\Big|\; \mathcal{F}_n\Big] \;\geq\; \sup_{\theta} \mathbb{E}\big[M_{n+1}(\theta) \mid \mathcal{F}_n\big] = \sup_{\theta} M_n(\theta) = G_n.$$

That is,

$$\mathbb{E}(G_n \mid \mathcal{F}_n) \geq G_n$$

and so $\{G_n, \mathcal{F}_n\}$ is a *submartingale.*.

Thus, it is not easy to create a stopping condition that controls the type-I error.

Also, the supremum often does not have a closed form, so the supremum must be done numerically for each n.

One could define a similar stopping rule as in the SPRT, but it would not directly bound the type-I or type-II error.

[1] One could consider composite nulls also, but the null is often taken to be a fixed reference value, so we will only focus on a simple null.

3 Mixture Martingales

A method of "fixing" the GLR would be to use Bayesian methods in place of the alternate term, to avoid supremum. That is, use a marginal likelihood, and define the **mixture martingale**(Howard et al. [2021]):

$$\widetilde{M}_n := \int_{\Theta_1} \prod_{i=1}^{n} \frac{f(x_i;\theta)}{f(x_i;\theta_0)} \pi(d\theta)$$

for a chosen prior distribution, $\pi(\theta)$. For instance, choosing a beta distribution to estimate a Bernoulli success probability.

We now check the martingale property under the null. Notice that this can be written in terms of M_n:

$$\widetilde{M}_n = \int_{\Theta_1} M_n(\theta)\pi(d\theta)$$

thus,

$$\mathbb{E}_{H_0}[\widetilde{M}_{n+1} \mid \mathcal{F}_n] = \int \mathbb{E}_{H_0}[M_{n+1}(\theta) \mid \mathcal{F}_n]\, \pi(d\theta) = \int M_n(\theta)\, \pi(d\theta) = \widetilde{M}_n.$$

and so $\{\widetilde{M}_{n+1}, \mathcal{F}_n\}$ is a nonnegative martingale. Deriving the expected value at $n = 0$ is straightforward: since $M_0 = 1$:

$$\widetilde{M}_n = \int_{\Theta_1} 1 \cdot \pi(d\theta) = 1 \qquad \text{(integral of pdf over support)}$$

So, $E(\widetilde{M}_0) = 1$.

Unlike the simple-vs-simple SPRT, there is no analogous martingale under H_1 (because we've "mixed" over Θ_1), so you can't write down a pair of equations to get exact A and B in terms of α and β. Instead, you use Ville's inequality Ville [1939] to control the type-I error with a single boundary:

$$\mathbb{P}\left(\sup_n \widetilde{M}_n \geq \frac{1}{x}\right) \leq x$$

Equivalently, for each fixed n,

$$\mathbb{P}\left(\widetilde{M}_n \geq \frac{1}{x}\right) \leq x, \qquad \text{for all } n \geq 1$$

substituting $\alpha = x$, we directly bound the type-I error.

$$\mathbb{P}\left(\widetilde{M}_n \geq \frac{1}{\alpha}\right) \leq \alpha$$

Thus the one-sided stopping rule

$$\tau \;=\; \inf\{n : \widetilde{M}_n \geq 1/\alpha\}$$

Observe that we make no assumptions on the prior distribution, meaning this holds for any choice of prior; although a poor choice would certainly slow convergence.

It is rare that the integral has a closed form solution and must be done numerically. Assuming a "nice" prior and distribution, a method such as quadrature or Laplace's method (i.e., using a second order Taylor expanding $\log \pi(\theta) + \log M_n(\theta)$ about the mode) may be used. These methods have a fairly low cost and have high accuracy if the integral is low dimensional and unimodal. In higher dimensional cases, or when one chooses a non-standard prior, one could use Monte Carlo integration, although this can be very expensive.

3.1 Example: Two Sided Test for Bernoulli Trials

Suppose we are, again, interested in testing the success probability for a Bernoulli(p) distribution. We observe $X_1, X_2, \cdots \sim$ Bernoulli(p) and wish to test

$$H_0 : p = 0.5 \quad \text{vs.} \quad H_1 : p \neq 0.5.$$

A typical choice of prior is a beta distribution, since it is a conjugate prior. We arbitrarily choose an uninformative prior, $\beta(1, 1)$ (a uniform distribution).

For each θ, the SPRT ratio is

$$M_n(\theta) = \prod_{i=1}^{n} \frac{f(X_i; \theta)}{f(X_i; 0.5)} = \big(2\theta\big)^{S_n} \big(2(1-\theta)\big)^{n-S_n},$$

where $S_n = \sum_{i=1}^{n} X_i$. Integrating over the prior yields:

$$\widetilde{M}_n = \int_0^1 M_n(\theta)\, d\theta = \int_0^1 (2\theta)^{S_n} \, (2(1-\theta))^{n-S_n} \, d\theta = 2^n \int_0^1 \theta^{S_n} (1-\theta)^{n-S_n} \, d\theta = 2^n \, B\big(S_n+1, n-S_n+1\big).$$

where $B(\cdot, \cdot)$ denotes the beta function. For our purposes, since S_n is a natural number:

$$B(S_n + 1, n - S_n + 1) = \frac{(S_n + 1)!(n - S_n + 1)!}{(S_n + 1 + n - S_n + 1)!} = \frac{S_n!\,(n - S_n)!}{(n + 1)!}.$$

Next, pick a Type-I level α (say $\alpha = 0.05$), and set

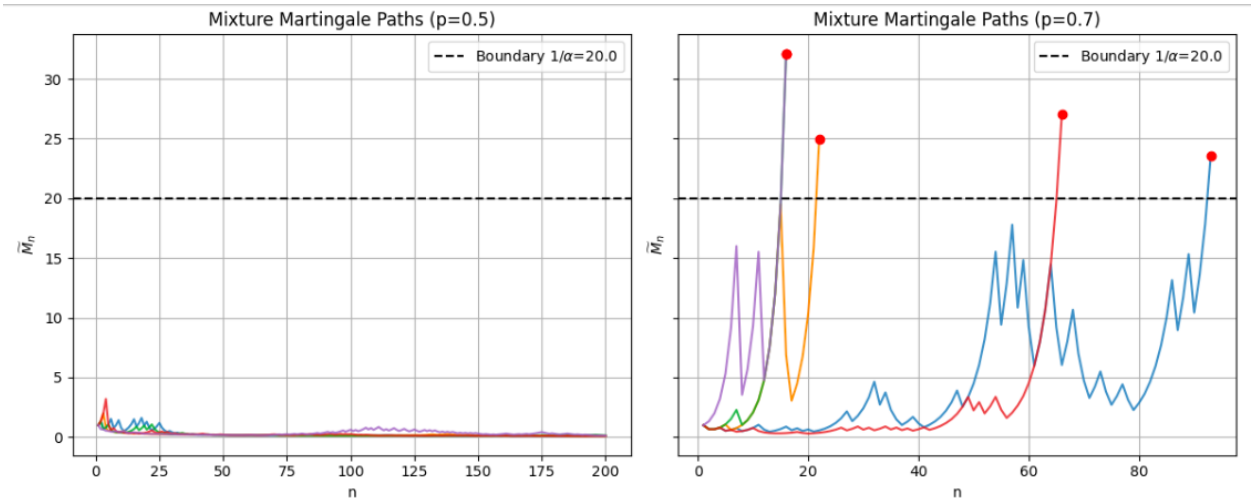
$$\tau = \inf\Big\{ n : \widetilde{M}_n \geq 1/\alpha \Big\} = \inf\big\{ n : \widetilde{M}_n \geq 20 \big\}.$$

Ville's inequality guarantees $\mathbb{P}\left(\tau < \infty\right) \leq \alpha \mid p = .5)$.

Similar to the SPRT test, we simulate 500 trials, one where the null is true ($p = .5$) and one where the null is false ($p = .7$). We report the proportion of the time that the proceses reaches the stopping condition (rejects the null hypothesis) with a max of 200 steps and the average amount of time it took.

```
p = 0.5: Rejection rate = 0.030, Average stop time = 36.3
p = 0.7: Rejection rate = 0.998, Average stop time = 53.9
```

As expected this takes more steps to reject the null, but does not imply that we knew the exact p value under the alternate hypothesis.



4 Cumulant Generating Function-Tilted Martingales

Sometimes it's easier to *reverse-engineer* a martingale by first stating the properties we want and then solving for the function that makes them hold. For example, suppose $X_1, X_2, \ldots$ are i.i.d. observations. The key properties which resulted in the SPRT being a martingale were:

- Multiplicity: $M_n = Z_1 Z_2 \cdots Z_n$, where Z_i only depends on X_i
- $\mathbb{E}(Z_i) = 1 \implies \mathbb{E}(M_n) = 1$

Notice that if these are satisfied:$\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = E(Z_{n+1})M_n = M_n$.

The multiplicity suggests that Z must be either a exponential or logarithmic. Since the expectation of an exponential yields the moment generating function (MGF), we suggest: $Z_i = e^{tX_i}$, but since we want $E(Z_i) = 1$, we normalize:

$$Z_i = \frac{e^{\lambda X_i}}{M_X(\lambda)}$$

$$\rightsquigarrow M_n = \prod_{i=1}^{n} Z_i = \frac{e^{\lambda S_n}}{M_X(\lambda)^n}$$

where $S_n = \sum_{i=1}^{n} X_i$. The MGF is the same for every X_i, since they are iid.

To further show that this has the multiplicity property, one may write this as one exponential function:

$$M_n = \exp\left(\lambda S_N - n \log(M_X(\lambda))\right)$$

The log of the moment generating function is called the *cumulant generating function* (CGF), denoted $\kappa_X(\lambda)$ for random variable X. M_n is then:

$$M_n = \exp\left(\lambda S_N - n\kappa_X(\lambda)\right)$$

thus, under H_0, $\{M_n, \mathcal{F}_n\}$ is a nonnegative martingale.

Observe that t is a free (tuning) parameter—often called the *tilt*—which you can choose to optimize performance or derive sharper concentration bounds. This shift in measure is known as *exponential tilting*, and the resulting process

$$M_n = \exp\big(\lambda S_N - n\,\kappa_X(\lambda)\big)$$

is commonly called the **cumulant-generating-function (CGF) martingale** or simply the **exponential-tilt martingale**. The parameter t is called the *tilt parameter*. Note that one may view the CGF term as a normalization constant, so this is simply an exponential function of X, for some chosen t.

Similar the to the mixture martingale, the CGF martingale does not have a "nice" martingale under H_1. In addition, one must choose a tilt parameter to best bound the martingale.

Remark: Exponential Tilting

Intuitively, exponential tilting constructs a new probability measure where large values of X_i become more likely (for $t > 0$). Under this tilted distribution, the new expected value of X_i becomes $\kappa'_X(\lambda)$, which shifts the center of mass in the direction of the deviation we're trying to measure. This is what gives the martingale its sensitivity to the upper tail.

The CGF term $n\,\kappa_X(\lambda)$ acts as a normalization constant. Thus, the martingale is just a rescaled exponential function of S_n, rescaled such that properties of a pdf are met.

4.1 Choosing the Tilt Variable

We now turn to the question of how to choose the free parameter t, which controls the exponential tilt of the martingale. Suppose we employ the same type of error bounding as we did for mixture martingales. To derive a tight bound for a given significance level α we again apply Ville's inequality to the CGF martingale.

$$\mathbb{P}\left(M_n \geq \frac{1}{\alpha}\right) \leq \alpha$$

$$\iff \mathbb{P}\left(\exp\left(\lambda S_n - n\kappa_X(\lambda)\right) \geq \frac{1}{\alpha}\right) \leq \alpha$$

This gives a valid level-α test:

$$\mathbb{P}\left(S_n \geq a_n(\alpha)\right) \leq \alpha.$$

where

$$a_n(\alpha) = \frac{n\kappa_X(\lambda) - \log(\alpha)}{\lambda}$$

To find the tightest threshold that guarantees a level-α test, we optimize the right-hand side of the bound over all $t > 0$, since the tilt parameter reflects how the probability mass is shifted towards higher values of X. Let this optimal $a_n(\alpha)$ be $a_n^*(\alpha)$ and let λ^* be the minimal t

$$a_n^*(\alpha) = \inf_{\lambda > 0} \frac{n\kappa_X(\lambda) - \log(\alpha)}{\lambda} \quad \implies \quad a_n^*(\alpha) = \frac{n\kappa_X(\lambda^*) - \log(\alpha)}{\lambda^*}.$$

Define the objective function $f(\lambda) = \frac{n\kappa_X(\lambda) - \log(\alpha)}{\lambda}$. The optimality condition is obtained by setting the derivative to zero:

$$f'(\lambda) = \partial_t \left(\frac{n\kappa_X(\lambda) - \log(\alpha)}{\lambda}\right) = 0$$

$$\implies \quad \lambda^*\kappa'_X(\lambda^*) - \kappa_X(\lambda^*) + \frac{\log(\alpha)}{n} = 0.$$

One may check that $f''(\lambda) > 0$ and so this is, indeed, a minimum, noting that $\kappa''(\lambda) = \mathrm{Var}_t(X)$; i.e., the variance of X after exponential tilting.

In practice, one may use a root-finding algorithm, such as Newton's method to compute this.

Notice that this threshold depends on n, so even if α is fixed, the optimal t (and therefore the threshold α_n) must be recomputed at each time step.

4.1.1 Connection between Chernoff Bounds and Tilt Optimization

Optimizing a parameter for a tightest bound is common in concentration inequalities. One may recall that in the derivation of Chernoff bounds (and Hoeffding's bound), the same problem of selecting the optimal coefficient of the MGF appears. We, therefore, claim that these bounds are connected.

Recall: Chernoff bound for S_n

$$\mathbb{P}(S_n \geq a) \leq \frac{M_{S_n}(\lambda)}{e^{ta}}$$

for a chosen constant a. Since X_i are iid, $M_{S_n}(\lambda) = M_X(\lambda)^n$. Hence

$$\mathbb{P}(S_n \geq a) \leq \frac{(M_X(\lambda))^n}{e^{ta}} = \exp\big(-\lambda a + n\,\kappa_X(\lambda)\big)$$

The event inside the probability can be rewritten:

$$S_n \geq a \iff M_n \geq e^{\lambda a - n\kappa_X(\lambda)}$$

Thus, we have:

$$\mathbb{P}\left(M_n \geq e^{\lambda a - n\kappa_X(\lambda)}\right) \leq \exp\big(-\lambda a + n\,\kappa_X(\lambda)\big)$$

One may notice that this is the *same bound as Ville's Inequality*. The key difference is that Chernoff's bound only assumed a fixed sample size, n, but Ville's inequality holds for all n.

Concentration inequalities which only hold for fixed sample sizes are called *static* bounds, in contrast to *time-uniform* bounds Howard et al. [2021]. So, in this case, Ville's inequality can be seen as a time-uniform generalization to the static Chernoff's bound.

4.2 Example: CGF Martingale for Bernoulli Trials

Like before, we will examine the simple case of Bernoulli trials, with unknown parameter, p.

Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli(p)

The MGF is then,

$$M_X(\lambda) = \mathbb{E}[e^{tX_i}] = 1 - p + p\,e^{\lambda},$$

which leads to the CGF of:

$$\kappa_X(\lambda) = \log M_X(\lambda) = \log\big(1 - p + p\,e^{\lambda}\big).$$

Let $S_n = \sum_{i=1}^{n} X_i$. Computing the martingale:

$$M_n(\lambda) = \exp\big(tS_n - n\,\log\big(1 - p + p\,e^{\lambda}\big)\big).$$

To find the optimal tilt parameter λ^* corresponding to a fixed error level $\alpha = 0.05$, we solve:

$$t\,\kappa_X'(\lambda) - \kappa_X(\lambda) + \frac{\log(.05)}{n} = 0.$$

Substituting the expressions for $\kappa_X(\lambda)$ and $\kappa_X'(\lambda)$ yields:

$$t \cdot \frac{p\,e^t}{1 - p + p\,e^t} - \log(1 - p + p\,e^t) + \frac{\log(0.05)}{n} = 0.$$

This equation can be solved numerically (e.g., via Newton's method) to find the optimal λ^*.

4.3 Testing with the CGF Martingale

The natural next question is: how do we derive a test from this martingale. It is apparent that the estimated parameter should be a cumulant (a moment after tilting). For example, for the simple test: for an iid Bernoulli sample with parameter, p

$$H_0 : p = 0.5 \quad \text{vs.} \quad H_1 : p = .7$$

then under H_0

$$\kappa_X(\lambda) = \log\big(1 - .5 + .5\, e^{\lambda}\big).$$

and under H_1:

$$\kappa_X(\lambda) = \log\big(1 - .7 + .7\, e^{\lambda}\big).$$

and similarly, the derivatives of $\kappa_X(\lambda)$ will be some function of the p values under each hypothesis.

In particular, let X_t be a random variable after the tilt is applied to X. It is easy to verify that

$$\kappa_{X_t}(\lambda_t) = \kappa_X(\lambda_t + \lambda) - \kappa_X(\lambda)$$

This identity reveals that exponential tilting corresponds to a *horizontal shift of the CGF*. As a consequence:

A direct consequence is:

$$E(X_t) = \kappa'_X(\lambda)$$

and because $\kappa'_X(0) = E(X)$, the shift simply corresponds to a shift in where we evaluate κ.

Thus, for each given $E(X_t)$, which corresponds to $E(X)$, that we associate to it, a specific λ. That is: suppose we have a test for a mean, as we did for the Bernoulli(p):

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu = \mu_0 + \delta \end{aligned}$$

where δ is the deviation from the mean - the *effect size*.

Then, there exists a unique λ^* such that

$$\kappa'_X(\lambda^*) = \mu_0 + \delta.$$

This value of λ^* is the optimal tilt: it is the parameter that centers the exponential-tilt martingale exactly at the alternative hypothesis. Hence, choosing a tilt parameter corresponds to selecting an effect size. Conversely, if we specify an effect size δ, we can solve for the corresponding λ^* that targets it.

Thus, for testing, if we assume H_0, the chosen λ will not be the "real" optimal one, λ^*. Let us examine how this effects the martingale over many iterations: note that we may approximate $S_n \approx n\mu_0$, under the null

$$M_n \approx \exp\left(n(\lambda\mu_0 - \kappa_X(\lambda))\right)$$

In the case where $\mu_1 > \mu_0$, then for a "reasonable" λ ($\lambda^* \leq \lambda > 0$, where λ^* corresponds to the H_1), $\lambda\mu_1 > \kappa_X(\lambda)$, since κ_X is monotonically increasing until λ^*. That is, M_n drifts upward, as expected.

There are a few very large issues with testing under this martingale:

1. We can only practically "target" one specific effect size at a time; we are restricted to simple alternate hypothesis
2. If the effect size is small, $\lambda\mu_0 - \kappa_X(\lambda)$, under the alternate hypothesis, is very small. Thus, the martingales and the test converges very slow.
3. Martingale only increasing if $\mu_1 > \mu_0$, that is, when the effect size is strictly positive. If $\mu_1 < \mu_0$, then M_n will decrease.
4. We must use Newton's method every iteration to find λ^*
5. We are restricted to simple alternate hypotheses
6. Only practical for testing parameters which are moments (and, therefore, cumulants). Fortunately this situation is far from rare:
    - Bernoulli-derived families (binomial, geometric, negative binomial, Poisson, etc.) all encode their parameters directly in the first (or first two) moments.
    - The Gaussian distribution's two canonical parameters (μ, σ^2) *are* its first two cumulants, making CGF-based tests a natural fit for normal data.

The reason for (3) is: when the effect size is negative, the exponential tilt will move the mass such that the tilted mean is less than the original mean. Thus, the solution is simply to examine the infinium over all $\lambda < 0$, as opposed to the positive λ.

A natural solution to (4) would be to calculate several values of λ and take the one which most closely approximates:

$$\lambda\kappa_X'(\lambda) - \kappa_X(\lambda) + \frac{\log(\alpha)}{n} \approx 0.$$

In fact, this would also capture several possible effects, so it may also be an approach to solving (2) and (5).

5 Stitching Martingales

5.1 Aside: A Naive Approach

To avoid solving for the optimal λ^* at each step (e.g., via Newton's method), a natural idea is to discretize a set of candidate values $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$, and at each time step, take the best-performing one:

$$G_n = \sup_{\lambda \in \Lambda} M_n.$$

While this quantity can grow rapidly under the alternative, it is no longer a martingale under H_0. Like the generalized likelihood ratio (GLR), the supremum introduces a submartingale structure, which complicates analysis. For this reason, G_n is not a desirable test statistic.

5.2 Uniform Stitching Martingale

A more principled approach is to exploit the fact that *convex combinations of martingales are themselves martingales*, due to the linearity of expectation.

Let $\Lambda = \{\lambda_1, \ldots, \lambda_K\}$ be a discrete grid of tilt parameters. Define the *uniform mixture martingale* as:

$$\widetilde{M}_n = \frac{1}{K} \sum_{i=1}^{n} M_n^{(\lambda_i)}$$

This construction maintains the two key properties we imposed in the derivation of the CGF martingale:

- $\widetilde{M}_n = \frac{1}{K} \sum_{i=1}^{K} \prod_{j=1}^{n} Z_i^{(\lambda_j)}$ where $Z_i^{(\lambda_j)} = \frac{e^{\lambda_j X_i}}{M_X(\lambda_j)}$
- and $\mathbb{E}_{H_0}[\widetilde{M}_n] = 1$, so Ville's inequality still applies.

Moreover, while each individual $M_n^{(\lambda_i)}$ targets a specific effect size, the mixture loosely *admits a composite alternative hypothesis*. Recall that $\lambda < 0$ for a negative effect size and $\lambda > 0$ for a positive effect size, so the choice of the grid controls the test.

5.3 General Stitching Martingale

One may select any weights for the above weighted average. Since it may be desirable to place more importance on the troublesome small effect sizes, we consider the weights, w to be functions of the λ. In order to maintain the unit expectation, we further impose that their sum be 1. That is:

$$\widetilde{M}_n = \sum_{i=1}^{n} w(\lambda_i) M_n^{(\lambda_i)}, \qquad \sum_i w(\lambda_i) = 1$$

this is the general case of the *stitching martingale* Ramdas [2018]Kaufmann and Koolen [2021].

For a simple test: consider an iid Bernoulli sample, $X_1, \ldots, X_n$

$$H_0 : p = 0.5 \quad \text{vs.} \quad H_1 : p > .5$$

one may create a grid:

$$\Lambda = \{0.1, 0.2, \ldots, 2.5\}.$$

To test a left-sided alternative ($H_1 : p < 0.5$), one may instead use:

$$\Lambda = \{-0.1, -0.2, \ldots, -2.5\}.$$

For a two-sided test, we simply take the union of both grids. This ensures the test is sensitive to deviations in either direction from the null mean.

If one prefers to calibrate the grid based on desired effect sizes the corresponding λ values can be determined empirically or by solving the CGF-based optimization condition:

$$\lambda \kappa_X'(\lambda) - \kappa_X(\lambda) + \frac{\log(\alpha)}{n} \approx 0.$$

for a chosen α.

Weighting schemes can simply be made by normalized functions with desired properties. For example:

- $w_j \propto 1/|\lambda_j|$ emphasizes sensitivity to small deviations
- A "bell-curve"-shaped function (e.g., $w_j \propto \exp(-\lambda_j^2)$) prioritizes moderate deviations
- - Uniform weights $w_j = 1/|\Lambda|$ offer robustness across the entire range.

One practical difficulty with stitched or mixture martingales is that the designer must choose both the grid of λ values and the weighting scheme $w(\lambda)$ in advance. While certain choices (e.g., a uniform grid with inverse-polynomial weights) are known to yield valid tests and are supported in the literature, there is no universally optimal selection.

For example, the uniform weighting scheme spreads sensitivity evenly across effect sizes, while downward weighting schemes prioritize small deviations and produce more conservative growth. However, selecting between these schemes remains somewhat ad hoc and may require domain knowledge or empirical tuning.

To the author's knowledge, there is no principled method for jointly selecting a grid and weight distribution that guarantees optimal performance across all scenarios, or a method to deduce one a priori, without simply repeating the expensive CGF martingale repeatedly.

This stands in contrast to traditional methods like the likelihood ratio test, which explicitly optimize over the parameter space.

Aside: Amending the GLR One may then try naturally apply the same weighting scheme to the GLR. This is functionally a discretization of the mixture martingale.

$$\widetilde{G}_n = \sum_{\theta \in \Theta'} w(\theta) \cdot \prod_{i=1}^{n} \frac{f(X_i;\theta)}{f(X_i;\theta_0)},$$

where Θ' is a finite grid over the parameter space and $w(\theta)$ is a normalized weight function.

The main challenge is that the parameter space Θ' is typically infinite or continuous, so one must restrict attention to a plausible subset based on prior knowledge or coarse estimation. In practice, this is often reasonable — for instance, if one expects the parameter to fall within a particular range or follow a known trend. In practice this is often possible. However, if a conjugate prior exists, it is more accurate (and often cheaper) to compute the mixture directly.

5.4 Empirical Testing

We run a similar test to before: Bernoulli trials, with $\alpha = .05$. We chose a grid of equally spaces λ from .1 to 2.5 with a 25 points. We compare a downward weighted ($w_j \propto \frac{1}{j^{1.5}}$) scheme against a uniform one.

6 Comparison of Methods

We now compare the performance of these methods against each other. The test will be performed in the following way:

- An iid Bernoulli test
- Maximum of 200 trials
- All tests use a common rejection threshold $1/\alpha$ for ease of visual comparison; consequently the SPRT is *not* tuned for a particular type-II error.

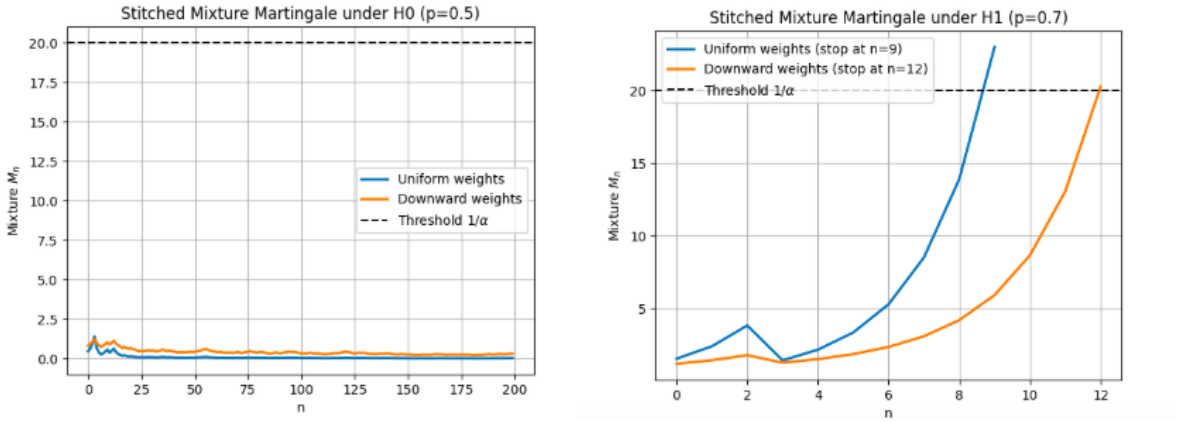


Figure 1: Observe that, even when the weights favor small effects, a relatively large effect is detected early.

- We use a $\beta(1,1)$ distribution as the initial prior for the mixture martingale; this is a uniform(0,1) distribution. This is the most naive prior; no weighting towards larger or smaller values of p.
- We use the same downward weighted scheme for the stitching martingale, $w_j \propto \frac{1}{j^{1.5}}$
- We test:

$$H_0 : p = .5$$
$$H_1 : p = .51 \text{ (Small effect) and } p = .8 \text{ (Large Effect)}$$

- We test the SPRT when it is exact and when it is inaccurate to the true alternate, since, in practice it is unlikely to be exact.

Exact SPRT (well specified)

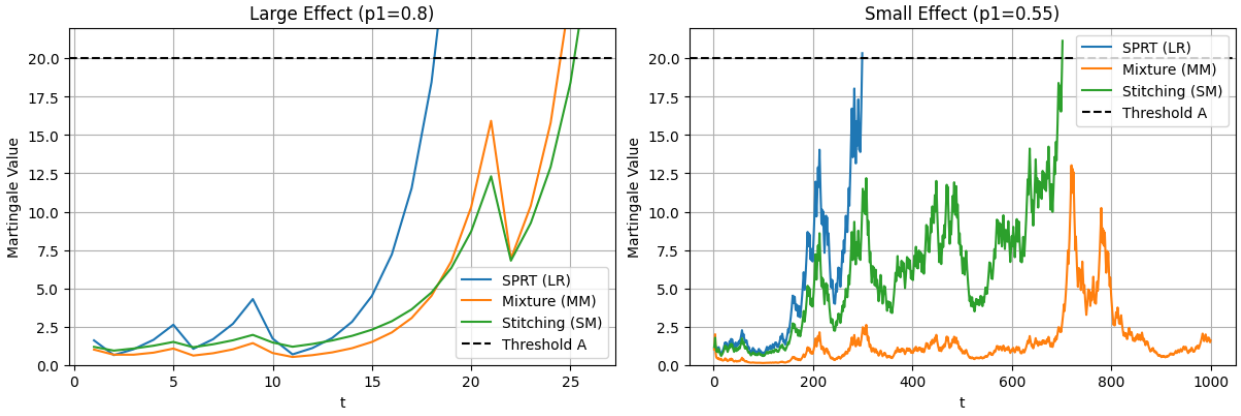


Figure 2: Under both small and large effects the well-specified SPRT converges fastest (note different n-scales).

SPRT with Mis-specified p1 (SPRT = .7)

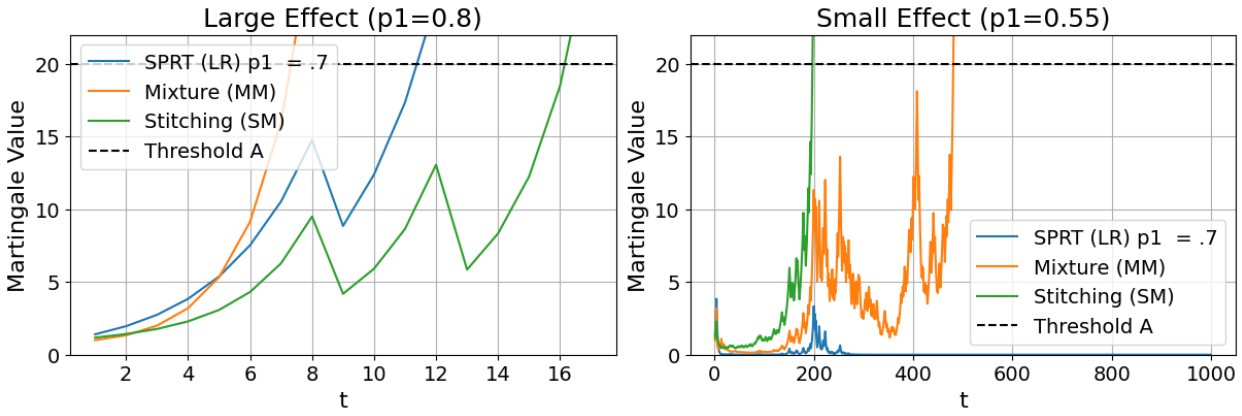


Figure 3: When the design value p_1 is incorrect the SPRT's power suffers: it completely misses the small effect ($p = 0.51$).

One may notice that even when p_1 was inaccurate, under large effect SPRT still performed well. However, an important specification is that this is about the *worst case scenario* for the mixture and stitching martingales, under large effects. A more informative prior (since we are testing $p_1 > .5$, an upward weighted $\beta(2, 1)$) would vastly increase the mixture performance under large effects. Similarly, a uniform or upward weight scheme would improve the performance of the stitching martingale under a large effect. Another observation is that each test rejected the null in under 20 observations, regardless. This should not be surprising: large effects are easier to notice. Thus, these improvements would be marginal and would not improve the practical performance very much.

7 Discussion and Practical Considerations

Each test is designed around the fact that they are martingales under the null hypothesis. Application of the optional stopping theorem then imply that the test statistic should be stable over time. Therefore, one can design time-uniform bounds using Ville's inequality that directly relate to the type-I error. Again, these tests are convenient because they can come to a decision on-the-fly meaning they often notice effects faster than traditional testing.

SPRT. The SPRT has shown to be the most tractable and effective method, if the strong assumption of knowing the exact plausible effect size is known; the circumstance is binary. This is very rarely the case and for more involved trials than ones examined in this document, such as the normal distribution and the Poisson distribution, where the parameters are not on a bounded domain, choosing a simple alternate hypothesis is nearly impossible a priori. It is also worth noting that it is the only test which directly controls its power.

Mixture Martingales. The mixture martingale is a good concept, if a conjugate prior is known. This is common in cases like the normal distribution as well as most Bernoulli-based distributions. The potential problem is its computational cost, but if a conjugate prior is known, the computation is tractable and computationally efficient. Even in slightly more complex settings, low dimensional

cases can be done quickly with numerical integration. More refined mixture models are used in clinical testing.

Stitched CGF Martingales. The stitching martingale is a practical implementation of the CGF martingale and is generally the best option should a conjugate prior not be available. A large discussion in this paper was that deciding the weighting and grid scheme was difficult, but, in practice, a uniform weighting and grid scheme generally perform well enough.

Method	**Type-I Control**	**Type-II Control**	**Requirement**	**Composite H_1**
SPRT	Yes	Yes (Exact)	θ_1 approx.	No
Mixture Martingale	Yes (via Ville)	No	Prior (e.g., Beta)	Yes
Stitched Martingale	Yes (via Ville)	No	Grid + Weights	Yes

Table 1: Comparison of Sequential Testing Methods

Small Effect Sizes. One critique might be that all sequential methods struggle to detect small effects quickly. This is true. But it is not a flaw unique to these methods; all hypothesis tests require a large number of observations to detect subtle deviations from the null. The benefit of sequential methods is that they still preserve valid error control at every step. Like in a typical inference setting, one need not restrict themselves to one test. Just as one could perform the score test, a large sample Z-test, and a likelihood ratio test, one could easily use a fixed sample test repetitively, in addition to using a sequential test. This is, assuming that the setting is not *extremely* time sensitive.

Early Stopping and Burn-in. A practical concern the author could not resolve is the risk of premature stopping. It is possible for an outlier to cause the test statistic to cross the boundary early, leading to a potentially ill informed conclusion. We then assert that one should, perhaps, have a burn-in period (say 10-20 samples) where a rejection cannot be made and perhaps such samples are ignored. Alternatively, early-stage smoothing, penalization, or confidence damping could be introduced to prevent abrupt decisions. Great care should be taken to ensure these methods do not violate the martingale property, however.

References

Georgios Fellouris and Alexander G. Tartakovsky. Nearly minimax mixture-based open-ended sequential tests. *Sequential Analysis*, 31(4):297–325, 2012. doi: 10.1080/07474946.2012.694346. URL https://doi.org/10.1080/07474946.2012.694346.

Alden Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 18:1–29, 2021.

Emilie Kaufmann and Wouter M Koolen. Mixture martingales revisited: Conditional vs. mixture testing and the supermartingale bootstrap. *arXiv preprint arXiv:2107.10417*, 2021.

Aaditya Ramdas. Martingales: Concentration inequalities and sequential analysis (lecture notes). https://www.stat.cmu.edu/~aramdas/martingales18/m18.html, 2018. Carnegie Mellon University.

Jean Ville. *Étude critique de la notion de collectif*. PhD thesis, Université de Paris, Paris, 1939. Doctoral thesis.

Abraham Wald. *Sequential Analysis*. John Wiley & Sons, 1947.