Reliable sequential testing for statistical model checking

Daniël Reijsbergen Pieter-Tjerk de Boer
Werner Scheinhardt Boudewijn Haverkort

University of Twente, The Netherlands

ANZAP workshop, July 8–11, 2013
University of Queensland, Brisbane, Australia

Outline

Outline

Context

Consider

- Some transition system, and
- Some path-property, e.g. path ends in deadlock before termination.

Model checking gives answer to:

- Do such paths exist?

  $\rightarrow$ (Non-probabilistic) Model Checking

Context

Consider

- Some transition system, and
- Some path-property, e.g. path ends in deadlock before termination.

Model checking gives answer to:

- Do such paths exist?

  $\rightarrow$ (Non-probabilistic) Model Checking
- Is probability p of such paths smaller/larger than some p_0?

  $\rightarrow$ Probabilistic Model Checking

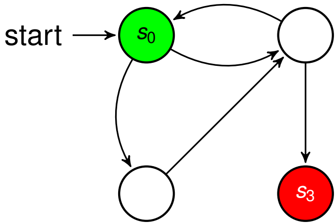
E.g. is $p = P(\text{path ends in deadlock}) < 0.05$?

Concrete example

For instance:

- Transition system: DTMC
- Property (event): "reach state s_3 before returning to s_0"
- Is $P(\text{event}) < 0.05$ or > 0.05?

So is probability of reaching s_3 before s_0 smaller than 5%?



How to do it?

Traditional approach: numerical analysis
But state spaces are huge...

Alternative approach: Stochastic Model Checking (SMC)
Based on (discrete event) simulation:

- Run n independent random samples
- Count S= # runs that satisfies path-property
- Compare estimate $\hat{p} = S/n$ to p_0

Advantage: No need to store and compute large system
$\Rightarrow$ Currently implemented in UPPAAL and PRISM

Concrete example (cont'd)

Computer program:

- In *i*-th run, simulate the DTMC until
  - reach $s_3 \Rightarrow$ return $X_i = 1$; quit;
  - reach $s_0 \Rightarrow$ return $X_i = 0$; quit;
- Repeat this N times (how to choose N ??)
- Accept or reject

  $H_0 : p = p_0$

  $H_{+1} : p > p_0$

  $H_{-1} : p < p_0$

  Such that $P(\text{accept } H_{+1} | H_0) < 0.05$, etc.

Approaches in literature

Used so far:

- Confidence intervals (Gauss)
- Sequential Probability Ratio Test (SPRT)
- Approximate Model Checking (Chernoff)
- Bayesian

All have (dis)advantages. In particular:

- Gauss: solid, but no outcome guaranteed
- SPRT: efficient: no need for many simulations,
  but validity of outcome depends on a-priori parameter δ

Gaus

Gauss:

- Fixed sample size N
- Test statistic $S_N = \sum_{i=1}^{N} X_i$
- Based on Central Limit Theorem
- Optimize N, based on guess γ for $p - p_0$

SPRT

Sequential Probability Ratio Test:

- Sequential test
- Based on Wald (1945)
- Test statistic $\frac{p_{+1}^{S_N}(1-p_{+1})^{N-S_N}}{p_{-1}^{S_N}(1-p_{-1})^{N-S_N}}$
- Indifference level δ: take

  $p_{+1} = p_0 + \delta$

  $p_{-1} = p_0 - \delta$
- Always draws conclusion
- Don't care what conclusion is when $p \in (p_0 - \delta, p_0 + \delta)$

Outline

General framework

All methods:

- Perform N consecutive simulation runs, leading to i.i.d. sequence of $X_i \sim \text{Bernoulli}(p)$
- Classical test statistic $S_N = \sum_{i=1}^{N} X_i \sim \text{Binom}(N, p)$
- Need to identify in which direction S_N deviates from $p_0 N$...

General framework

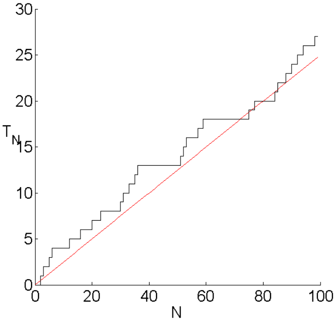
All methods:

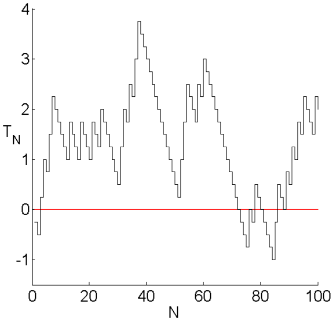
- Perform N consecutive simulation runs, leading to i.i.d. sequence of $X_i \sim \text{Bernoulli}(p)$
- Classical test statistic $S_N = \sum_{i=1}^{N} X_i \ \sim \text{Binom}(N, p)$
- Need to identify in which direction S_N deviates from $p_0 N$...
- ... in a statistically sound way, i.e. with guaranteed upper bounds on probability of wrong conclusion

Only difference between methods:

- when to stop, and
- what to conclude?

General framework



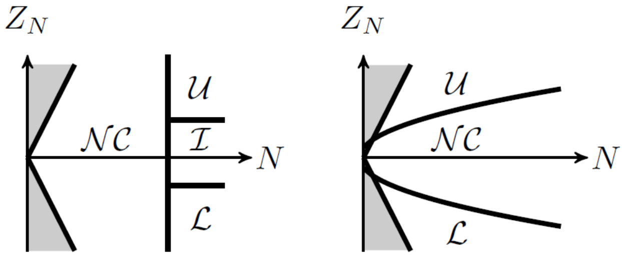


Sample path of: S_N $\qquad$ $Z_N = S_N - p_0 N$

Z_N has positive drift ($p > p_0$)
or negative drift ($p < p_0$)

General framework

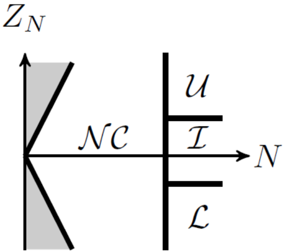
When to stop and what to conclude?



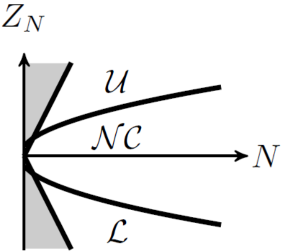
$\mathcal{NC}$	Non-critical:	no conclusion yet, continue
$\mathcal{U}$	Upper:	stop, conclude $H_{+1} : p > p_0$
$\mathcal{L}$	Lower:	stop, conclude $H_{-1} : p < p_0$
$\mathcal{I}$	Inconclusive:	stop, no conclusion (keep $H_0 : p = p_0$)
Grey	unreachable (slopes $1 - p_0$ and $-p_0$)	

General framework

When to stop and what to conclude?



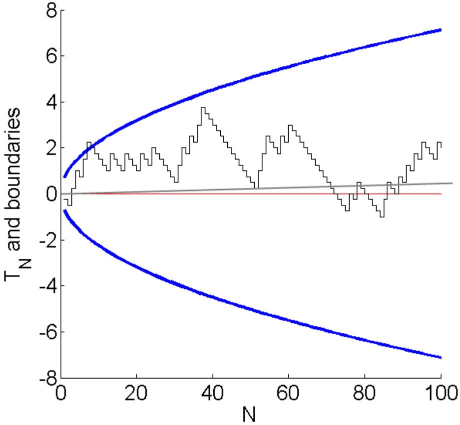
Fixed sample size test



Sequential test

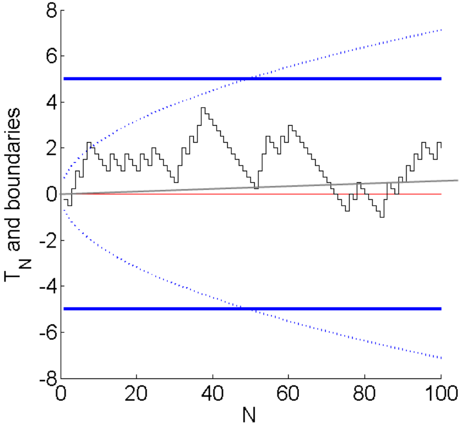
- Typical shape depends on type of test
- Specifics depend on parameters and confidence level

Fixed sample size test (Gauss)



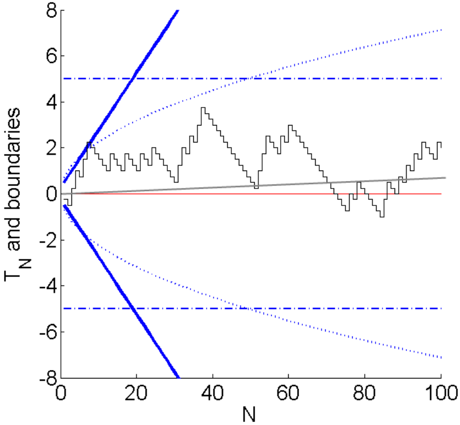
Boundaries as a function of (predetermined) N behave $\sim \sqrt{N}$

Sequential test (SRPT)



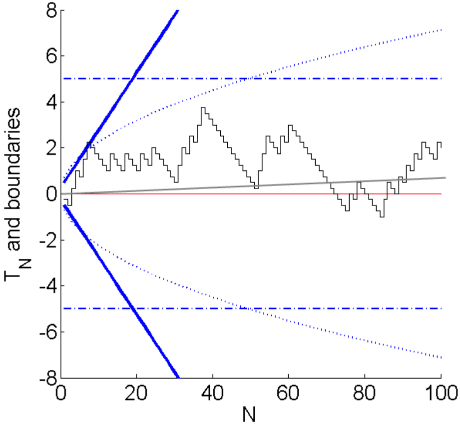
Boundaries almost constant

Sequential test (Linear)



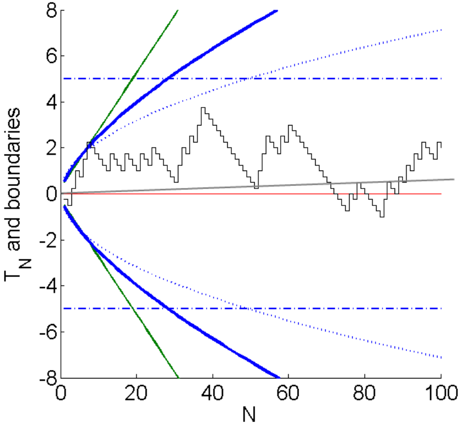
Linearly diverging boundaries better?

Sequential test (Linear)



Linearly diverging boundaries better? No

Sequential test (new)



Boundaries ‘in between’ square root and linear

Outline

New sequential techniques

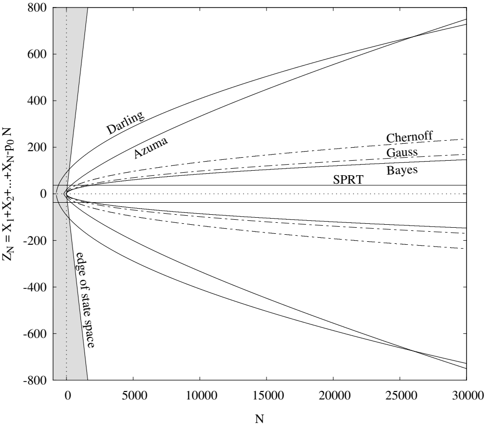
Boundaries of $\mathcal{NC}$ should not be

- Too wide (like linear)
  $\rightarrow$ may never terminate
- Too narrow (like square root)
  $\rightarrow$ too easy to draw wrong conclusion when $|p_0|$ small

Propose:

- 'Azuma' $\sim a(N+k)^b$, with $b \in (\frac{2}{3}, 1)$
- 'Darling' $\sim a\sqrt{(N+k)\log(N+k)}$

New sequential techniques



Azuma and Darling compared to earlier tests

Azuma, bounding P(wrong conclusion)

Bound on $P(\text{accept } H_{+1}|H_0)$

$= P(Z_N \text{ ends up in } \mathcal{U}|Z_N \text{ has drift } 0)$

Based on Generalized Azuma-Hoeffding inequality (writing n for N):

- $f_n = a(n+k)^b$, with $b \in (\frac{2}{3}, 1)$, $k, a > 0$
- Let Z_n have drift 0, be stopped at $-f_n$
- Let $W_n = e^{c_n(Z_n - f_n)}$ for some sequence c_n

Lemma

W_n is a supermartingale, i.e.

$$\mathbb{E}(W_n|W_{n-1}, \ldots, W_1) \le W_{n-1},$$

if we take $c_n = 8(3 - \frac{2}{b})\frac{d}{dn}f_n$

Azuma, bounding P(wrong conclusion)

Theorem

$$\mathbb{P}(\exists n \geq 0 : Z_n > f_n) \leq e^{-8(3b-2)a^2 k^{2b-1}}$$

Proof.

Define bounded stopping time

$$N(m) = \min\{n : |Z_n| \geq f_n \text{ or } n = m\}$$

for supermartingale $W_n = e^{c_n(Z_n - f_n)}$. Then

$$\begin{aligned}
\mathbb{P}(Z_{N(m)} \geq f_{N(m)}) &= \mathbb{P}(W_{N(m)} \geq 1) \\
&\leq \mathbb{E}(W_{N(m)}) \\
&\leq \mathbb{E}(W_0) = e^{-f(0)c(0)} \\
&= e^{-8(3b-2)a^2 k^{2b-1}}.
\end{aligned}$$

Azuma, bounding P(wrong conclusion)

Corollary

Azuma test with boundaries $+a(N+k)^b$ *and* $-a(N+k)^b$ *satisfies*

$$\mathbb{P}(Accept\ H_{+1} \mid \neg H_{+1}) \leq \alpha$$
$$\mathbb{P}(Accept\ H_{-1} \mid \neg H_{-1}) \leq \alpha$$

and

$$\mathbb{P}(Reject\ H_{+1} \mid H_{+1}) \leq \beta$$
$$\mathbb{P}(Reject\ H_{-1} \mid H_{-1}) \leq \beta$$

with $\alpha = \beta = e^{-8(3b-2)a^2k^{2b-1}}$

Darling

- Boundary of $\mathcal{NC}$ is $f_n = a\sqrt{(n+k)\log(n+k)}$
- Darling and Robbins (1968) on iterated logarithm:

If $\epsilon > 0$ exists such that

$$\sum_{n=1}^{\infty} e^{-f_n^2/(n+1)} \leq \epsilon$$

then P(wrong conclusion) $\leq 2\sqrt{2}\epsilon$.

Optimize parameters

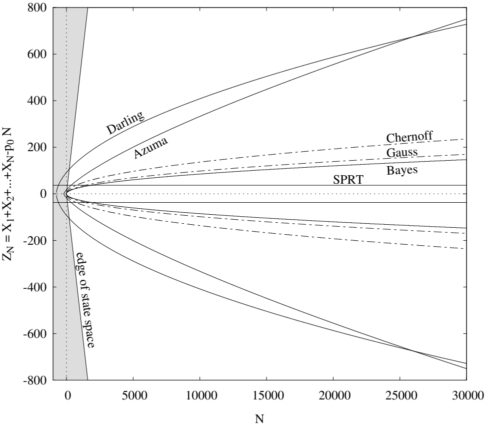
- Azuma: $k(a, \alpha, b) = \left(\frac{\log\left(\frac{\alpha}{2}\right)}{8a^2(2-3b)} \right)^{\frac{1}{2b-1}}$
- Darling: $k(a, \alpha) = \left(\frac{\alpha(a-1)}{2\sqrt{2}} \right)^{-\frac{1}{a-1}} - 1$
- Then, minimize (approximate) expected hitting time over *a*, i.e. solve

  $$f_n = |p - p_0| n,$$

  using guess γ for $|p - p_0|$
- Azuma: take $b = \frac{3}{4}$

Outline

Shape of $\mathcal{NC}$



Azuma and Darling compared to earlier tests

Experimental results

Test	γ (or δ)	probability of correct conclusion	probability of no conclusion	average number of samples
Gauss	0.1	0.036 ± 0.012	0.953 ± 0.013	$1.64 \cdot 10^2$
	0.01	$\mathbf{0.946 \pm 0.014}$	$\mathbf{0.054 \pm 0.014}$	$\mathbf{2.04 \cdot 10^4}$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$2.39 \cdot 10^6$
SPRT	0.1	0.489 ± 0.031	0	$(3.70 \pm 0.17) \cdot 10^1$
	0.01	$\mathbf{0.949 \pm 0.014}$	**0**	$\mathbf{(2.19 \pm 0.10) \cdot 10^3}$
	0.001	1.0 ± 0.0	0	$(2.39 \pm 0.03) \cdot 10^4$
Chernoff	0.1	0.007 ± 0.005	0.993 ± 0.005	$6.67 \cdot 10^2$
	0.01	$\mathbf{1.0 \pm 0.0}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{6.67 \cdot 10^4}$
	0.001	1.0 ± 0.0	0.0 ± 0.0	$6.67 \cdot 10^6$
Bayes	uniform	0.599 ± 0.030	0	$(5.64 \pm 0.56) \cdot 10^2$
Azuma	0.1	1.0 ± 0.0	0	$(1.41 \pm 0.01) \cdot 10^6$
	0.01	$\mathbf{1.0 \pm 0.0}$	**0**	$\mathbf{(4.79 \pm 0.10) \cdot 10^4}$
	0.001	1.0 ± 0.0	0	$(2.24 \pm 0.01) \cdot 10^5$
Darling	0.1	1.0 ± 0.0	0	$(2.04 \pm 0.02) \cdot 10^5$
	0.01	$\mathbf{1.0 \pm 0.0}$	**0**	$\mathbf{(1.78 \pm 0.02) \cdot 10^5}$
	0.001	1.0 ± 0.0	0	$(2.10 \pm 0.02) \cdot 10^5$

$p = 0.19$, $p_0 = 0.20$, **bold:** $\gamma = |p - p_0|$ (guess correct)

Outline

Conclusions

- Existing tests have shortcomings:
  - Gauss: depends on N, possibly no conclusion
  - SRPT: depends on indifference level δ
- New tests do not have these shortcomings
- ... at the expense of longer simulation times

Future Work:

- Improve bounds.
- Generalize results: importance sampling??

Thanks mates!