2. Techniken zur Varianzreduktion

Grundsätzliche Beobachtung bei der Auswertung stochastischer Simulationen:
zufällige Eingaben erzeugen zufällige Ausgaben

$\Rightarrow$ Statistische Auswertung ist notwendig

$\Rightarrow$ Ergebnisse in Form von Mittelwertschätzern und Konfidenzintervallen zu einem gegebenem Konfidenzniveau α

Simulation komplexer Systeme erfordert hohen Aufwand, für genaue Resultate werden viele Replikationen oder lange Beobachtungsreihen benötigt

$\Rightarrow$ Beschränkung von Zeit und Rechenkapazität verhindert oft genügend genaue Analysen

Techniken zur Vergrößerung der Genauigkeit bei vorgegebenem Aufwand sind von großer praktischer Bedeutung!

Aber gibt es solche Techniken überhaupt?

Literatur zu diesem Kapitel:

- Law/Kelton 2000, Kap. 11
- P. Heidelberger. Fast Simulation of Rare Events in Queueing and Reliability Models. ACM Trans. on Model. Comp. Simul. 5 (1), 1995, 43-85.

Zur Erinnerung: Berechnung des Konfidenzintervalls erfolgt über

$$\widetilde{Y} \pm \varepsilon_{\alpha} \cdot \widetilde{S} / \sqrt{n}$$

Beispiel unfaire Münze:

- mit Wahrscheinlichkeit *p* fällt die Münze auf Kopf (mit W. *1-p* auf Zahl)

Ziel:

Ermittlung von *p* mittels Experimenten/Simulation

$\tilde{p}$ soll mit einer relativen Genauigkeit von 10% des ermittelten Mittelwertes zum Signifikanzniveau α=0.01 (99%) geschätzt werden

Da in diesem Fall die Varianz als p(1-p) bekannt ist, erhalten wir nach n Replikationen (Münzwürfen):

$$\tilde{p}_n \pm \varepsilon_{\alpha/2} \cdot \sqrt{p(1-p)/n}$$

wobei $\varepsilon_{\alpha/2}$ das $\alpha/2$-Quantil von N(0,1) ist (hier 2.576), ist gefordert

$$2.576 \cdot \left(\sqrt{p(1-p)/n} \right) / \tilde{p}_n \leq 0.1$$

Und da $\tilde{p}_n$ für $n \to \infty$ gegen p konvergiert gilt $n \approx 100 \cdot 2.576^2 \cdot (1-p)/p$

für kleine p gilt $n \in O(1/p)$

Einige Beispiele:

$p=0.5 \Rightarrow n\approx 664$, $p=0.1 \Rightarrow n\approx 6000$, $p=10^{-6} \Rightarrow n\approx 6.64\cdot 10^{8}$

Beobachtung: Wenn die Wahrscheinlichkeit des Ereignisses kleiner wird, so werden mehr Beobachtungen benötigt, um den Erwartungswert mit gleicher relativer Genauigkeit zu schätzen!

Sind Ereignisse mit kleiner Wahrscheinlichkeit interessant?

In vielen Fällen (leider) ja!

Beispiele:

- Ausfallwahrscheinlichkeit eines Systems in einem vorgegebenem Intervall soll $<10^{-6}$ sein
- Wahrscheinlichkeit des Paketverlusts in einem Rechnernetz soll $< 10^{-9}$ sein
  - Um eine Verlustwahscheinlichkeit von mit 10% Genauigkeit (bei 99% Signifikanzniveau) zu schätzen sind 664 Milliarden Beobachtungen (d.h. oft Simulationsläufe/Replikationen) notwendig!
  - Bei 10 Sekunden Dauer einer Replikation wären dies mehr als 200 000 Jahre!

Wie könnte es schneller gehen?

- Schnellere Rechner (durch Technologie begrenzt, teuer)
- Effizienterer Code (oft nur geringe Effekte, aufwendig)
- Parallelisierung (siehe nächstes Kapitel)
- Reduktion der Varianz (in diesem Kapitel behandelt)

Wenn es gelingt, die Varianz des Schätzers zu verkleinern, ohne den Mittelwert zu verändern, dann reduziert sich die Breite der Konfidenzintervalle proportional zur Reduktion der Varianz!
Zugehörige Techniken werden als **Varianzreduktionstechniken (VRTs)** bezeichnet

Es existieren zahlreiche VRTs, aber

- Normalerweise hängen die Techniken vom konkreten Modell ab (und sind nicht blind anwendbar)
- Ist eine Abschätzung der Varianzreduktion vor Simulationsdurchführung nicht möglich (oft kann nicht einmal eine Reduktion garantiert werden!)
- Sind Pilotläufe zur Abschätzung des Effekts notwendig
- Erfordert der Einsatz von VRTs zusätzlichen Aufwand, der zur erreichten Varianzreduktion in Beziehung gesetzt werden muss

Unterschiedliche Techniken der Varianzreduktion existieren, wobei zwei Ansätze Verwendung finden

1. Veränderung des beobachteten stochastischen Prozesses, so dass der Erwartungswert unverändert bliebt oder rekonstruiert werden kann, die Varianz aber kleiner wird.
2. Beobachtung des Modells primär in dem Bereich, in dem seltene Ereignisse auftreten und anschließendes Zurückrechnen des Erwartungswertes

Wir behandeln

2.1 Gemeinsame Zufallsvariablen

2.2 Antithetische Variablen

2.3 Kontrollvariationen

2.4 Konditionierung

2.5 Importance Sampling

2.1 Gemeinsame Zufallszahlen

Im Gegensatz zu den folgenden Verfahren werdem gemeinsame ZZs zum **Vergleich von zwei Modellen** eingesetzt (ähnliches Vorgehen existiert, wenn ein Teil der ZZs durch Werte aus Traces ersetzt werden)

Annahme: Benutzung identischer ZZs in beiden Modellen

Dies bedeutet, dass

- eine eindeutige Zuordnung von ZZs zu Prozessen im System möglich ist
  z.B. Ankunft zu Ankunft, Bedienung zu Bedienung
  Dies kann problematisch sein, da
  - unterschiedliche Systemkonfigurationen verglichen werden
  - ZZs aus unterschiedlichen Verteilungen kommen können (und damit mit unterschiedlicher Zahl von [0,1)-verteilten Werten generiert werden)
- Eingriffe in die Struktur des Simulators notwendig sind
  Problematisch, da
  - viele Simulationsumgebungen nur beschränkte Eingriffsmöglichkeiten zulassen
  - Änderbarkeit und Übersichtlichkeit komplexer Modelle eingeschränkt wird
- Generatoren reproduzierbare ZZs erzeugen (heute i.d.R. gegeben)

Annahme: Zwei Modellvarianten, die mit gemeinsamen ZZ simuliert werden

Seien $v_1=(v_{11},\ldots,v_{1n})$ und $v_2=(v_{21},\ldots,v_{2n})$ die beobachteten Stichproben und $w_i= v_{i1} - v_{i2}$ die beobachtete Differenz

Stichproben beschreiben Realisierungen von ZVs V_1, V_2 und W

$\tilde{V}_1$, $\tilde{V}_2$, $\tilde{S}_1$ und $\tilde{S}_2$ sind die Mittelwert- und Varianzschätzer für V_1 und V_2

Ziel: Schätzung von $E(W) = E(V_1) - E(V_2)$ unter Verwendung von $\sigma^2(W) = VAR(W)$

Schätzer $\tilde{W} = \sum_{i=1}^{n} w_i \Big/ n$

und der Varianz $\sigma^2(\tilde{W}) = \frac{\sigma^2(W)}{n} = \frac{\sigma^2(V_1)+\sigma^2(V_2)-2\cdot COV(V_1,V_2)}{n}$

Bei unabhängigen ZVs gilt $COV(V_1,V_2) = 0$ (wie bisher angenommen)

Wenn V_1 und V_2 positiv korreliert sind, so gilt $COV(V_1,V_2) > 0$ wodurch $\sigma^2(\tilde{W})$ reduziert wird!

Bisheriger Schätzer für $\sigma^2(W)$ (bei unkorrelierten ZVs V_1 und V_2):

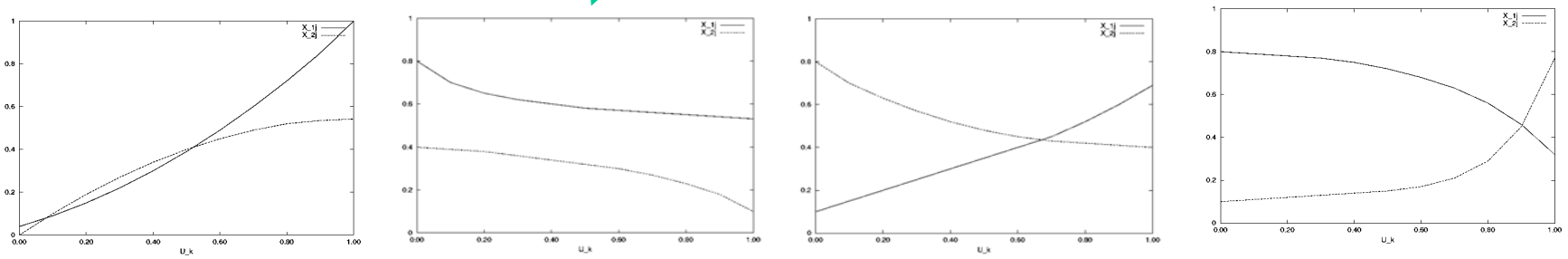
$$\tilde{S}_{12}^2 = \tilde{S}_1^2 + \tilde{S}_1^2 \text{ mit } \tilde{S}_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(v_{ij} - \tilde{V}_i)^2 \ (i = 1,2)$$

Nun Schätzung direkt aus $\tilde{S}_{12}^2 = \frac{1}{n-1}\sum_{j=1}^{n}((v_{1j} - v_{2j}) - \tilde{W})^2$

$\Rightarrow$ Die einzelnen Werte v_{ij} müssen gespeichert und offline ausgewertet werden!

Vorgehen klingt logisch und einfach, aber gemeinsame ZZs führen nicht zwangsläufig zu positive Korrelation. Es können auch negative Korrelationen auftreten und die Varianz vergrößert werden!

Beispiele für mögliche Abhängigkeiten: Positive Korrelation



Positive Korrelation bei Verwendung von gemeinsamen ZZs ist nachzuweisen:
Möglichkeiten:

- Intuitive Erklärung (z.B. lange Bedienzeiten führen zu langen Wartezeiten)
- Nachweis der Verhaltensmonotonie (z.B. durch Analyse der Transformationen)
- Schätzung von $\tilde{S}_1$, $\tilde{S}_2$ und $\tilde{S}_{12}$ in Pilotläufen

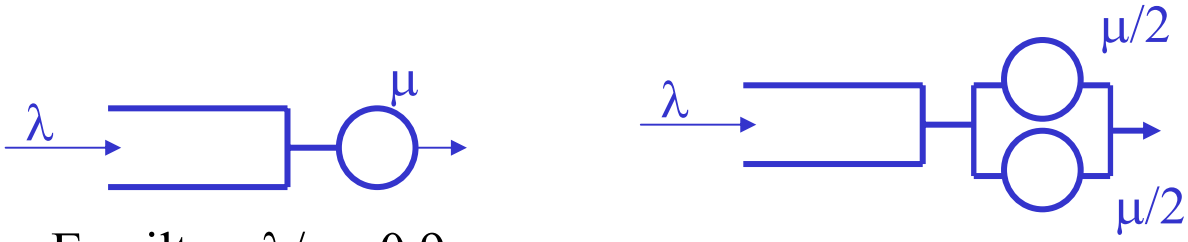
  Falls nicht $\tilde{S}_1 + \tilde{S}_2 > \tilde{S}_{12}$ eindeutig gilt, so bringt die Verwendung gemeinsamer ZZs keine Vorteile

Weiterer Nachteil: Durch gemeinsame ZZs eingeführt Korrelation verhindert die Anwendung mancher statistischer Testverfahren

Verwendung gemeinsamer ZZs in modernen Simulationssystemen:

- Definition von ZZ-Strömen und Verwendung von identischen Strömen für korrespondieren Ereignissen
  (Bedienzeiten an einer Ressource, Ankunftszeiten eines Prozesstyps, …)
  - In den meisten Simulationssystemen wird die Definition von unabhängigen Teilströmen ermöglicht (Startpunkte der Ströme müssen möglichst weit auseinander liegen um Abhängigkeiten zu vermeiden)
  - Die Verwendung gleicher Saaten reicht i.d.R. nicht aus!

Beispiel: Vergleich eines M/M/1 und M/M/2 Systems



- Es gilt $\rho=\lambda/\mu = 0.9$
- Die durchschnittliche Verweilzeit der ersten 100 Aufträge ist zu ermitteln, wenn mit dem leeren System gestartet wird.
- Es werden 100 Replikationen durchgeführt

Ergebnisse

	I	A	S	A+S
$\hat{S}_{12}^2$	18.0	9.02	8.80	0.07
90% Konf. Int.	0.70	0.49	0.49	0.04
Corr(V_1,V_2)	-0.17	0.33	0.40	0.995

I unabhängige Beobachtungen
A Zwischenankunftszeiten aus einen ZZ-Strom
S Bedienzeiten aus einem ZZ-Strom
A+S Zwischenankunfts- und Bedienzeiten aus jeweils einem ZZ-Strom

Ansatz ist interessant für die trace-getriebene Simulation!

Beispiel Caching-Algorithmen für Web-Server

Zugriffssequenz auf Web-Seiten
(aus einer Messung)

Algorithmus 1

Algorithmus 2

Stochastische Zugriffszeiten (evtl. gemeinsame ZZs)

Vergleich der Algorithmen auf Basis der Simulationsergebnisse

2.2 Antithetische Variablen

Ziel: Verkleinerung der Varianz eines Schätzer bei der Analyse eines Modells

Bestimmung von Konfidenzintervallen über Replikationen!

Idee der antithetischen Variablen:
Gepaarte Replikationen:

Kleine Beobachtungswerte im ersten Lauf korrespondieren zu großen Beobachtungswerten im zweiten Lauf und umgekehrt
(d.h. Einführung negativer Korrelation)

$\Rightarrow$ Mittelwerte der zwei werden im Durchschnitt näher am Erwartungswert liegen, als der Durchschnitt der Mittelwerte zweier unabhängiger Replikationen

Realisierung:

- Verwendung komplementärer ZZs (d.h. u_i im ersten und $1-u_i$ im zweiten Lauf)
- Es ist sicherzustellen, dass u_i und $1-u_i$ für identische Ereignisse verwendet werden (d.h. Definition von ZZ-Strömen für Ereignisse)

Mathematische Grundlagen:

- Seien $v_1 = (v_{11}, \ldots, v_{1n})$ und $v_2 = (v_{21}, \ldots, v_{2n})$ die beiden beobachteten Sequenzen

Es gelten folgende Annahmen:

- v_1 und v_2 entstehen durch Verwendung antithetischer Variablen
- Die Werte von v_{ij} und v_{kl} sind unabhängig, falls $j \neq l$ (auch wenn $i=k$)
- $\mu = E(v_{ij}) = E(w_j)$ mit $w_j = (v_{1j} + v_{2j})/2$
- v_{ij} sind Realisierungen der ZV V_i,
  w_j sind Realisierungen der ZV W

Mittelwertschätzer $\tilde{W} = \frac{1}{n}\sum_{j=1}^{n} w$

Varianz des Schätzers $\sigma^2(\tilde{W}) = \frac{\sigma^2(W)}{n} = \frac{\sigma^2(V_1) + \sigma^2(V_2) + COV(V_1, V_2)}{n}$

- Bei unabhängigen Beobachtungen gilt $COV(V_1, V_2)=0$,
- Falls V_1 und V_2 negativ korreliert sind, gilt $COV(V_1, V_2)<0 \Rightarrow \sigma^2(\tilde{W})$ wird kleiner

Wie schon bei gemeinsamen ZZs ist nicht klar, ob antithetische Variablen wirklich zu einer Varianzreduktion führen

- Voraussetzung dafür ist ein monotones Modellverhalten in den antithetischen Variablen
- Zur Anwendung von antithetischen Variablen muss durch Probeläufe oder Interpretation des Systemverhaltens sichergestellt sein, dass Varianzreduktion eintritt

Beispiel: Mittlere Wartezeit der ersten 100 Aufträge in einem M/M/1-System mit ρ=0.9 und Start bei leerem System

Ergebnisse:

	I	AV
$\hat{S}_{12}^2$	4.84	1.94
90% Konf. Int.	0.36	0.23
Corr(V_1,V_2)	-0.07	-0.52

- I unabhängige Beobachtungen
- AV antithetische Variablen für Ankunft und Bedienung

Varianzreduktion wurde erreicht, aber deutlich geringer als bei gemeinsamen ZZs.

2.3 Kontrollvariationen

Idee auch hier: Ausnutzung der Korrelation zwischen verschiedenen Größen zur Reduktion der Varianz des Schätzers der gesuchten Größe.

Geschätzt werden soll μ, der Erwartungswert der ZV V aus der Stichprobe $(v_1, \ldots, v_n)$

Wir nehmen an, dass V positiv oder negativ mit einer ZV X korreliert ist und der Erwartungswert ν von X bekannt ist

Beispiele:

- In einem einfachen Bediensystem wird die Verweilzeit steigen, wenn die Bedienzeiten steigen
  (positive Korrelation zwischen Bedienzeit und Verweilzeit)
- In einem einfachen Bediensystem wird die Verweilzeit sinken, wenn die Ankunftsabstände größer werden
  (negative Korrelation zwischen Ankunftsabstand und Verweilzeit)
- In einem technischen System wird die Systemverfügbarkeit steigen, wenn die Zuverlässigkeit der Komponenten steigt

Definiere „kontrollierte“ ZV: $Z = V - a \cdot (X - \nu)$

Es gilt $E(Z) = E(V) = \mu$

und $\sigma^2(Z) = \sigma^2(V) + \underbrace{a^2 \cdot \sigma^2(X) - 2a \cdot COV(V,X)}_{\text{Wähle a so, dass } <0}$

Zur Wahl von a:

- $a > 0$, falls V und X positiv korreliert sind
- $a < 0$, falls V und X negativ korreliert sind

Die Varianz von Z ist kleiner als die Varianz von V, falls

$$2a \cdot COV(V,X) > a^2 \cdot \sigma^2(X)$$

und das Optimum wird erreicht, falls die erste Ableitung 0 wird

$$2a \cdot \sigma^2(X) - 2 \cdot COV(V,X) = 0$$

mit der Lösung $a_{opt} = COV(V,X) / \sigma^2(X)$

Falls a = a_{opt}, so gilt:

$$\sigma^2(Z) = \sigma^2(V) - \frac{(COV(V,X))^2}{\sigma^2(X)} = (1 - \rho_{VX}^2) \cdot \sigma^2(V)$$

$$\text{mit } 0 \le \rho_{VX}^2 = \frac{(COV(V,X))^2}{\sigma^2(V) \cdot \sigma^2(X)} \le 1 \quad \Rightarrow \text{Varianz wird nicht größer!}$$

Probleme bei der praktischen Umsetzung:

- σ^2(X) könnte unbekannt sein
- COV(V,X) wird in fast allen Fällen unbekannt sein

$\Rightarrow$ a_{opt} kann nicht exakt ermittelt werden

$\Rightarrow$ a_{opt} muss aus Simulationsergebnissen geschätzt werden

Mögliche Schätzung von a_{opt}:

Seien $(x_1, \ldots, x_n)$ Realisierungen von X und $(v_1, \ldots, v_n)$ Realisierungen von V in der Simulation und $\tilde{V}$ und $\tilde{X}$ die Mittelwertschätzer von V und X

Schätzer für die Kovarianz $\widetilde{COV}(V,X) = \dfrac{\sum_{j=1}^{n}(v_j - \tilde{V})(x_j - \tilde{X})}{n-1}$

und für die Varianz $\tilde{S}^2(X) = \dfrac{\sum_{j=1}^{n}(x_j - \tilde{X})^2}{n-1}$ womit gilt $\tilde{a}_{opt} = \dfrac{\widetilde{COV}(V,X)}{\tilde{S}^2(X)}$

und schließlich als Schätzer der gesuchten Größe $\tilde{Z} = \tilde{V} - \tilde{a}_{opt} \cdot (\tilde{X} - \upsilon)$

Problem bei diesem Vorgehen: Die einzelnen Schätzer sind abhängig, wenn sie aus den Daten eines Simulationslauf resultieren

$\Rightarrow$ Es ist nicht klar, ob $\tilde{Z}$ ein erwartungstreuer Schätzer ist

$\Rightarrow$ Es ist nicht klar, ob die Varianz wirklich reduziert wird

Beispiel: Mittelere Verweilzeit der ersten 100 Kunden in einem M/M/1 System mit mittlerer Zwischenankunftszeit 1 Min., mittlerer Bedienzeit 0.9 Minuten

x_i wird als Mittelwert aus 99 Bedienungen berechnet

- Wert fest, da Simulation mit der 100. Bedienung endet
- Annahme Bedienzeit und Verweilzeit sind positiv korreliert

Resultat aus 10 unabhängigen Replikationen:

j	1	2	3	4	5	6	7	8	9	10	MW	Var
v_j	13.8	3.18	2.26	2.76	4.33	1.35	1.82	3.01	1.68	3.60	3.78	13.33
x_j	0.92	0.95	0.88	0.89	0.93	0.81	0.84	0.92	0.85	0.88	0.89	0.002

$\Rightarrow \tilde{C}OV(V,X)=0.07$, $\tilde{a}_{opt}=35$ und $\tilde{Z}=4.13$ (wahrer Wert bekannt und ist ebenfalls 4.13)

100 Wiederholungen dieses Experimentswerden werden mit diesen Parametern durchgeführt, wobei $\tilde{Y}$ und $\tilde{Z}$ sowie deren Varianzen geschätzt werden

Resultate dieser Experimente:

$\tilde{S}^2(\tilde{Y})=0.99$ und $\tilde{S}^2(\tilde{Z})=0.66$ sowie 4.18 ± 0.16 als Konfidenzint. für die Verweilzeit

$\Rightarrow$ Reduktion der Varianz um 1/3 und keine Verzerrung des Schätzers (in diesem Beispiel)

2.4 Konditionierung

Beobachtung der „interessierenden“ ZV V unter der Bedingung, dass eine andere ZV X im Modell einen bestimmten Wert annimmt

E(V|X=x) ist der Erwartungswert von V, wenn X den Wert hat

Sei X eine diskrete ZVs, dann gilt: $E_X(E(V \mid X)) = \sum_{x_i} E(V \mid X = x_i) \cdot p(x_i)$

mit p(x) W., dass der Wert von X gleich x ist
(normalerweise ist p(x) unbekannt)

Für die Varianz gilt: $\sigma_X^2(E(V \mid X)) = \sigma^2(X) - E_X(\sigma^2(V \mid X)) \leq \sigma^2(X)$

<u>Vorgehen</u>:

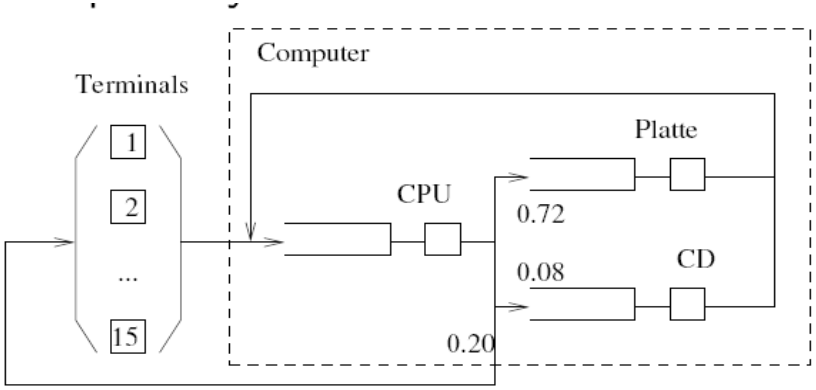
- Simuliere Realisierungen von X und
- Für jedes beobachtete x simuliere Realisierungen von V|X=x

Vorgehen bringt Vorteile, falls

- Realisierungen von X effizient realisiert werden können
- E(V|X) analytisch berechnet werden kann
- $E_X(\sigma^2(V|X))$ groß ist

Methode ist sehr stark modellabhängig, deshalb Beschreibung am Beispiel

Typisches Beispiel: Modell eines Computersystems



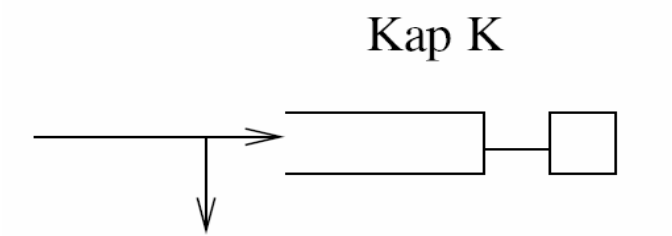
Ziel: Bestimmung der mittleren Verweilzeit an CPU, Platte und CD

Problem: nur 8% der Jobs greifen in einem Zyklus auf die Platte zu
⇒ Simulation liefert wenige Beobachtungen der Zugriffszeit
⇒ große Varianz des Schätzers

Mögliche Konditionierung

- Annahme, jeder Job, der die CPU verlässt, würde zur CD wechseln
  - Berechnung der „potenziellen“ Verweilzeit (analytisch, falls Bedienzeit exponentiell, auf Basis des Systemzustandes sonst)
- Falls sich mehrere Jobs in der CD-Warteschlange befinden, wird der Systemzustand festgehalten und mehrere Replikationen gestartet

Weiteres Beispiel: Bestimmung der Überlaufwahrscheinlichkeit eines Puffers in einem Kommunikationssystem



Bestimmung der Wahrscheinlichkeit, dass der Puffer bei Ankunft eines Paketes voll ist

- Falls K groß ist und die Auslastung moderat ist, dann ist die gesuchte Wahrscheinlichkeit sehr klein
- Trotzdem ist ihre Bestimmung von großer praktischer Bedeutung (z.B. Dimensionierung von Puffern in Rechnernetzen zur Vermeidung von Datenverlust mit typischen Anforderungen von 10^{-8} – 10^{-9} für die Verlustwahrscheinlichkeit)

Vorgehen bei der Simulation zur Bestimmung der Wahrscheinlichkeit:

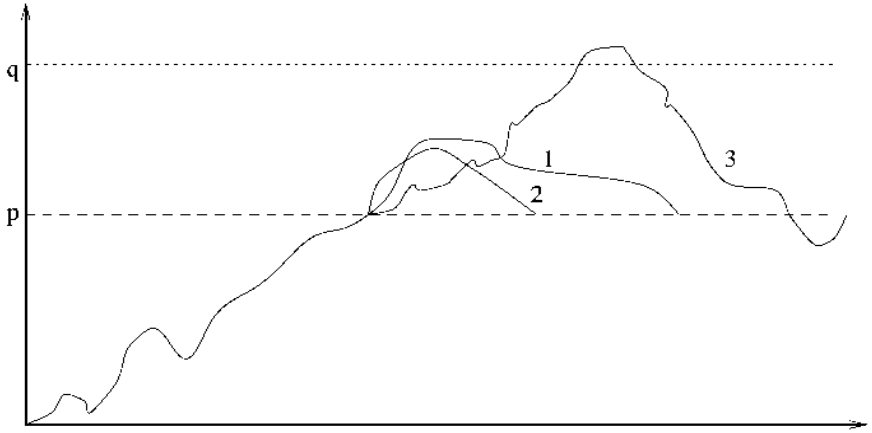
- Definiere Regenerationszustand (d.h. Zustand, bei dem das zukünftige Verhalten nicht von der Vergangenheit abhängt, z.B. leeres System)
- Simuliere R Zyklen, die jeweils im Regenerationszustand starten und enden
- Sei b_r (= Anzahl verlorene Pakete/Anzahl Pakete) Schätzer für die Verlustwahrscheinlichkeit im Zyklus r (b_r sind unabhängig identisch verteilt)

Problem: Für die meisten Zyklen ist $b_r=0$

Idee eines Konditionierungsansatzes:

- Wähle L < K
- Simuliere bis Pufferfüllung L erreicht und speichere Zustand
- Beginne vom gespeicherten Zustand mit r Replikationen, die jeweils enden, wenn Warteschlangenlänge L wieder erreicht wird
- Nach der r-ten Replikation fahre mit der Simulation fort, bis die Replikation endet

Skizze des Vorgehens:



Aus der Simulation kann geschätzt werden:

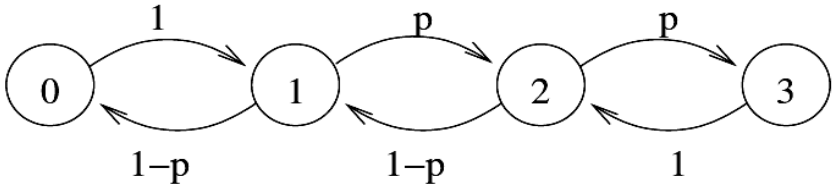
- P(n=L) Wahrscheinlichkeit, dass Population L in einer Replikation erreicht wird
- $P(n=K|n_0=L)$ Wahrscheinlichkeit, dass Population K erreicht wird, falls mit Population L begonnen wird

$\Rightarrow P(n=K) = P(n=K|n_0=L)\cdot P(n=L)$

Vorgehen kann auf mehrere Zwischenstufen L_1, L_2, .. erweitert werden

Beispiel: M/M/1/3-System (analytisch analysierbar):
Ankunftsrate λ, Bedienrate μ und Übergangswahrscheinlichkeit $p = \lambda / (\lambda + \mu)$

Systemverhalten kann durch folgende Markov-Kette beschrieben werden



Regenerationszustand 0

<u>Analyseziel</u>: Bestimmung der Wahrscheinlichkeit, dass ausgehend von Zustand 1, Zustand 3 erreicht wurde, ohne dass Zustand 0 oder 1 vorher erreicht wird

X ZV mit X=1 falls 3 erreicht wird, 0 sonst, E(X) ist der gesuchte Wert

Vorgehen bei der Simulation:

- Simuliere R Replikationen, die in Zustand 1 starten
- Bestimme x_r = Wert von X in Replikation r
- Berechne $\tilde{X} = \frac{1}{R}\sum_{r=1}^{R} x_r$ und $\tilde{S}^2(X) = \frac{1}{R-1}\sum_{r=1}^{R}(x_r - \tilde{X})^2$
- Konfidenzintervall $\tilde{X} \pm \varepsilon_\alpha \cdot \tilde{S}(X)/\sqrt{R}$

Erwartungswert $E(\tilde{X})=E(X)=p^2$ und Varianz $\sigma^2(X) = p^2 - p^4 = p^2 \cdot (1 - p^2)$

Für $p << 1$ gilt $\sigma^2(X) \approx p^2 \Rightarrow \sigma^2(\tilde{X}) \approx p^2 / R$

Konfidenzintervall nach R Replikationen $\tilde{X} \pm \varepsilon_\alpha \cdot \tilde{S}(X) / \sqrt{R}$

mit Erwartungswert: $p \pm \varepsilon_\alpha \cdot O(p / \sqrt{R})$

Simulationsaufwand, damit die Breite des 99% Konfidenzintervalls $\approx 0.1 \cdot \tilde{X}$

Es muss gelten $2.576 \cdot p / \sqrt{R} \leq 0.1 \cdot p^2 \Rightarrow 100 \cdot 2.576^2 / p^2 \leq R$

Beispiele

- Für $p = 0.1 \Rightarrow R \geq 66\,358$
- Für $p = 0.01 \Rightarrow R \geq 6\,635\,776$

$\Rightarrow$ Extrem hoher Simulationsaufwand für kleine Werte von p

Aufwandsreduktion durch Konditionierung:

- Simuliere K Replikationen bis zum Erreichen von Zustand 2 oder bis zur Rückkehr in Zustand 0 $\Rightarrow$
  Wahrscheinlichkeit Zustand 2 zu erreichen wird beschrieben durch ZV Y
- Jedes mal, wenn Zustand 2 erreicht wird, simuliere L Replikationen bis Zustand 3 oder 1 erreicht wird $\Rightarrow$
  Wahrscheinlichkeit Zustand 3 zu erreichen, unter der Bedingung, dass Zustand 2 erreicht wurde, wird beschrieben durch ZV X|Y

Es gilt $E(X) = E(X|Y) \cdot E(Y)$

Für die einzelnen ZVs gilt:

- $E(\tilde{Y})=E(Y)=p$ und $\sigma^2(Y)= p - p^2$
- $E(\tilde{X}|\tilde{Y}) = E(X|Y) = p$ und $\sigma^2(Y) = p - p^2$

Zur Vereinfachung für $p << 1$:

$E(\tilde{Y}) \approx E(\tilde{X}|\tilde{Y}) \approx p$ und $\sigma^2(Y) \approx \sigma^2(X|Y) \approx p$

$\Rightarrow \sigma^2(\tilde{Y}) \approx p / K$ und $\sigma^2(\tilde{X}|\tilde{Y}) \approx p / (p \cdot K \cdot L)$

Zur Auswahl von K und L:

Wähle K und L so, dass zur Schätzung von Y und X|Y gleich viele Replikationen verwendet werden (ZVs mit identischer Varianz!)

$\Rightarrow$ R/2 = K = p ·K ·L $\Rightarrow$ L = 1 / p

Für die Varianz gilt dann
(siehe Originalliteratur z.B. P. Glasserman et. al. Multilevel Splitting for estimating rare event probabilities, Operations Research 47, 1999.):

$$\sigma^2((\tilde{X} \mid \tilde{Y}) \cdot \tilde{X}) = \frac{p^2 \cdot (1-p) \cdot (p^2 R^2 - 1)}{R^2 \cdot (pR - 1)} \text{ für } pR \neq 1$$

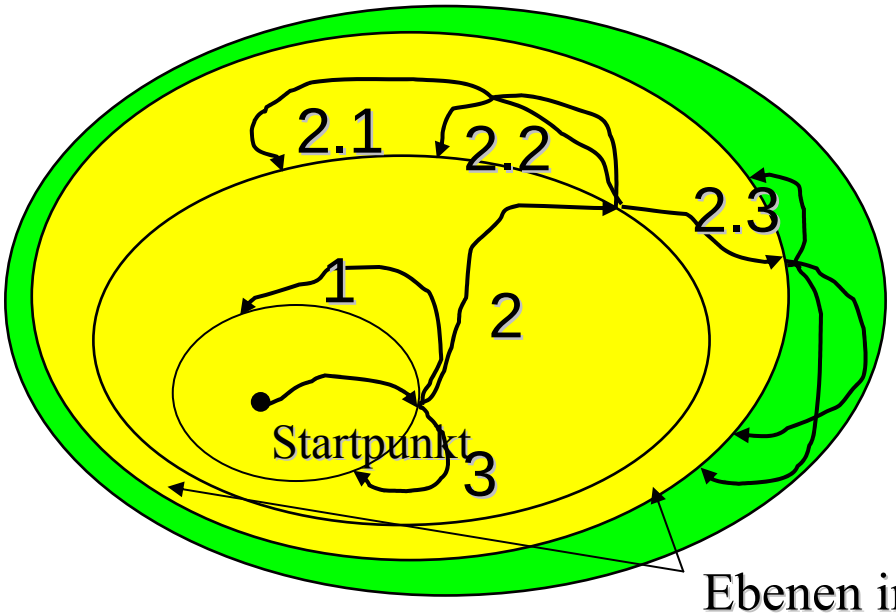
Erwartungswert für das Konfidenzintervall $p^2 \pm \varepsilon_\alpha \cdot \sigma((\tilde{X} \mid \tilde{Y}) \cdot \tilde{X})$

Simulationsaufwand, damit die Breite des 99% Konfidenzintervalls $\approx 0.1 \cdot \tilde{X}$

- Für p = 0.1 $\Rightarrow$ R $\approx$ 6 500 (statt $\geq$ 66 358)
- Für p = 0.01 $\Rightarrow$ R $\approx$ 65 000 (statt $\geq$ 6 635 776)

$\Rightarrow$ Deutliche Aufwandsredutkion ist möglich

Natürliche Verallgemeinerung



Ebenen im Zustandsraum

- Mehrere Punkte, an denen neue Replikationen ansetzen
- Unterschiedliche Punkte für ab- und aufsteigende Kurven
- Dynamische Bestimmung der jeweiligen Replikationszahlen

Probleme bei der konkreten Anwendung:

- Festlegung der Punkte, wo neue Replikationen ansetzen
- Festlegung der Zahl der Replikationen
- Integration in Modellierungswerkzeuge

2.5 Importance Sampling

Sei f(x) Dichtefunktion einer ZV X, die den Ausgang eines Simulationsexperiments beschreibt und g(x) eine Funktion

Ermittelt werden soll $\gamma = E_f(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f(x) dx$

Sei ferner g(x) = 0 für x < y

Falls $P(y < X) = \int_{y}^{\infty} f(x) dx \ll 1$ ist die Ermittlung von $E_f(g(X))$ schwierig, da die meisten Experimente x < y liefern

Sei X‘ eine andere ZV mit (beliebiger) Dichtefunktion f‘(x), so dass

$P(y < X') = \int_{y}^{\infty} f'(x) dx \gg 0$ und $f(x) > 0 \Rightarrow f'(x) > 0$

Dann gilt $\gamma = E_f(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot \frac{f(x)}{f'(x)} \cdot f'(x) dx = E_{f'}(g(x)L(x))$

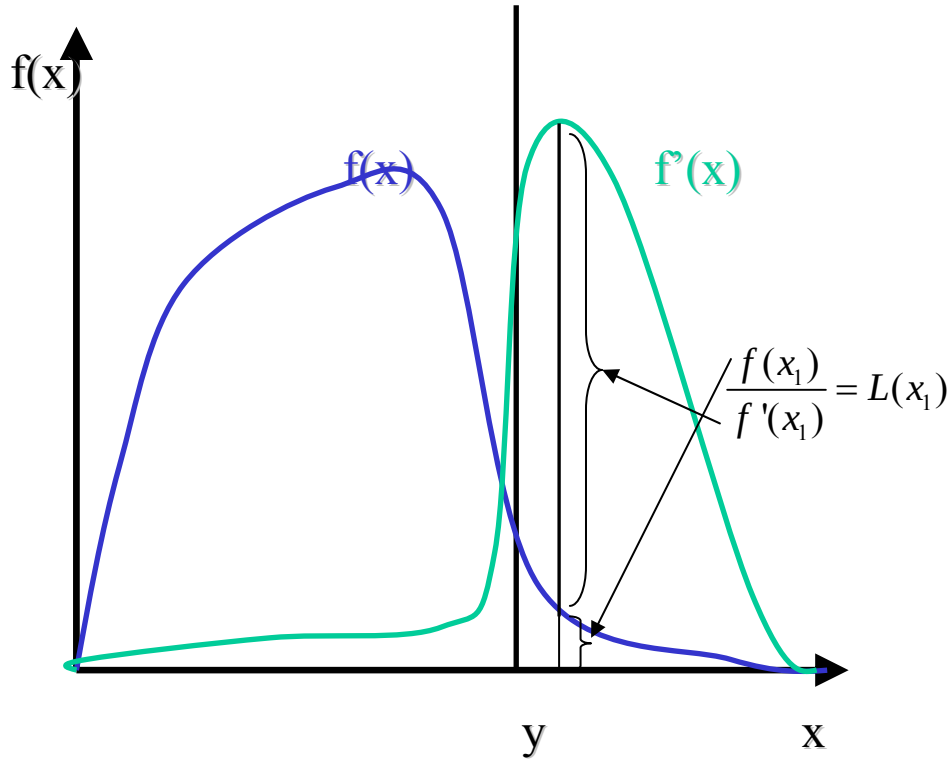
Mit der Likelihood-Funktion $L(x) = f(x) / f'(x)$

Skizze des Vorgehens

Voraussetzung für die Verbesserung der Schätzung:

- f‘(x) muss im „interessanten“ Bereich groß sein!

⇒ Viele Beobachtungen führen zu relevanten Resultaten (X>y)



Formale Betrachtung der Situation:
Sei zur Vereinfachung g(x) = 1 falls x > y und 0 sonst

$$\gamma = E_f(g(X)) = \int_{-\infty}^{\infty} \delta(x > y) \cdot \frac{f(x)}{f'(x)} \cdot f'(x)dx = E_{f'}(\delta(x > y) \cdot L(x))$$

Vorgehen in der Simulation zur Bestimmung von γ:

- Erzeuge R Stichproben $x_1,\ldots,x_R$ gemäß Dfkt. f(x)
- Berechne $\hat{\gamma}_R = 1/R \sum_{r=1}^{R} \delta(x_r > y)$ (offensichtlich $E_f(\hat{\gamma}) = \gamma$)
- Konfidenzintervall $\hat{\gamma}_R \pm \varepsilon_\alpha \cdot \sqrt{\gamma \cdot (1-\gamma)/R}$
- Für eine Konfidenzintervallbreite, die proportional zu γ sein soll, sind $R \in O(1/\gamma)$ Stichproben notwendig
- Simulationsaufwand wächst proportional zum Kehrwert von γ über alle Grenzen (unbounded relative error)

Vorgehen beim Einsatz von „importance sampling“ zur Bestimmung von γ:

- Erzeuge R Stichproben $y_1,\ldots,y_R$ gemäß Dfkt. f‘(x) und bestimme dabei jeweils die Werte der Likelihood-Funktion $L(x_r) = f(x_r)/f'(x_r)$
- Berechne $\hat{\gamma}'_R = 1/R \sum_{r=1}^{R} \delta(y_r > y) \cdot L(y_r)$ (offensichtlich $E_f(\hat{\gamma}) = \gamma$)

  Da f‘ mit höherer Wahrscheinlichkeit Werte größer y liefert, werden mehr Beobachtungswerte ≠ 0 sein, dies wird durch $L(x_r)$ ausgeglichen!

Sei $Z = \delta(Y > y) \cdot L(Y)$ und $\hat{\gamma}'_R = 1/R \sum_{r=1}^{R} z_r$

(z_r unabhängige Realisierungen von Z)

Da $(\delta(x>y))^2=\delta(x>y)$

Für das zweite Moment gilt:

$$E_{f'}(Z^2) = E_{f'}((\delta(Y > y) \cdot L(Y))^2) = \int \delta(x > y) \cdot \left(\frac{f(x)}{f'(x)} \right)^2 \cdot f'(x)dx$$

$$= \int \delta(x > y) \cdot \frac{f(x)}{f'(x)} \cdot f(x)dx = E_f(\delta(X > y) \cdot L(Y))$$

- Zur Reduktion der Varianz muss f(x)/f‘(x) klein sein, wenn x > y
- Da f(x) für x > y klein ist (seltenes Ereignis), sollte f‘(x) in diesem Fall groß sein

Angestrebt wird, dass die Größe der Stichprobe, die zum Erzielen einer bestimmten relativen Genauigkeit notwendig ist, unabhängig von der Größe von γ ist (bounded relative error)

Formale Formulierung:

- Sei ε ein Seltenheitsparameter (d.h. $\lim_{\varepsilon \to \infty} \gamma(\varepsilon) = 0$)
- γ(ε) ist mit vorgegebener relativer Genauigkeit α mit O(R) Replikationen für alle ε > 0 bestimmbar

Beispiele für die Wahl von ε:

- Zur Bestimmung der Pufferüberlaufswahrscheinlichkeit eines Systems mit Puffergröße K wähle $\varepsilon = 1/K$
- Zur Bestimmung der Verfügbarkeit von Systemen sei die Ausfallrate $\lambda_i(\varepsilon) = a_i\varepsilon^{bi}$ für Konstanten $a_i, b_i \geq 1$

Zur Herleitung einer Dichtefunktion, die „relative bounded errors“ erreicht:

- Gelte $\gamma(\varepsilon) \sim c \cdot f(\varepsilon)$ (d.h. $\lim_{\varepsilon\to\infty}(\gamma(\varepsilon) / (c \cdot f(\varepsilon)) = 1)$)
- Falls $E_{f'}(Z^2) \sim d \cdot (f(\varepsilon))^2 \Rightarrow \sigma_{f'}(Z) \sim k \cdot f(\varepsilon)$, wobei
  - $\sigma_{f'}(Z)$ die Standardabweichung von Z ist und
  - $k = (d - c^2)^{1/2}$
- Dann gilt für den relativen Fehler des mit importance sampling ermittelten Schätzers

$$\lim_{\varepsilon\to 0} \frac{\sigma(\hat{\gamma}'_R)}{\gamma(\varepsilon)} = \lim_{\varepsilon\to 0} \frac{\sigma_{f'}(Z)}{\gamma(\varepsilon)\cdot\sqrt{R}} = \frac{k \cdot f(\varepsilon)}{c \cdot f(\varepsilon)\cdot\sqrt{R}} = \frac{k}{c\cdot\sqrt{R}} < \infty$$

Bisherige Beschreibung beruht auf der Kenntnis von f(ε)

- In der Simulation ist aber gerade $E_f(g(X))$ zu ermitteln und damit f in der Regel unbekannt!

Was bleibt!

- Heuristiken für verschieden Modellklasse (gleich mehr dazu) oder
- Abschätzung von f

Reicht intuitives Verständnis von f aus, um importance sampling effektiv einzusetzen? (z.B. vergrößere Ankunftsrate, verkleinere Bedienrate etc)

Leider nein!! (Wie viele Beispiel und naive Ansätze zeigen)

Ein einfaches Beispiel:

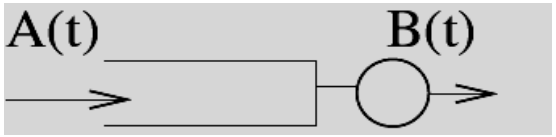
- Sei $f(x) = \lambda \cdot e^{-\lambda x}$ gesucht $P(x>t)$ für festes $t>0$
- Falls $t >> \lambda^{-1}$ seltenes Ereignis

Einfache Idee: ersetze λ durch λ‘ und simuliere mit Hilfe von importance sampling

$$E(Z^2) = \int_t^\infty \frac{f(x)}{f'(x)} \cdot f(x)dx = \frac{\lambda^2}{\lambda'} \cdot \int_t^\infty e^{(\lambda'-2\lambda)x} dx$$

Für $\lambda' > 2\lambda$ unendliche Varianz
$\Rightarrow$ Situation kann beliebig verschlechtert werden

Heuristiken für Warteschlangensysteme



Untersuchung der Wahrscheinlichkeit, dass ausgehend von einer Ankunft im leeren System die Population im Puffer einen festen (großen) Wert K überschreitet, bevor das System wieder leer läuft

- Für allgemeine Systeme vom Typ GI/GI/1 können asymptotisch optimale Werte aus den momentgenerierenden Funktionen der Zwischenankunfts- und Bedienzeitverteilung gewonnen werden
- Für M/M/1-Systeme resultiert das asymptotisch optimale Modell aus der Vertauschung der Ankunfts- und Bedienrate (System wird instabil!)
- Mehrbedienersystem GI/GI/N lassen sich ähnlich wie GI/GI/1 behandeln
- Kaum Resultate für Warteschlangennetze (außer Tandemnetzen) oder System mit korrelierten Ankunftsströmen

Heuristiken funktionieren nur

- für die angegebenen Systeme
- für das angegebene Leistungsmaß
- für große K und potenziell unendliche Zeitintervalle

Heuristiken für Zuverlässigkeitsuntersuchungen

Modelle können aus mehreren Komponenten bestehen, die ausfallen können und repariert werden

In der Regel sind Ausfallraten wesentlich kleiner als Reparaturraten

$\Rightarrow$ Wahrscheinlichkeit eines Systemausfalls ist sehr klein (seltenes Ereignis)

Zahlreiche Heuristiken existieren, insbesondere für exponentiell verteilte Ausfall- und Reparaturzeiten

Beispiel: Failure biasing

- Fehlerwahrscheinlichkeit wird gleichmäßig für alle Fehler erhöht (Fehlerwahrscheinlichkeit = Wahrscheinlichkeit, dass innerhalb eines vorgegebenen Intervalls mindestens ein Ausfall stattfindet)

Verfahren funktioniert gut, solange alle Fehler ungefähr gleich wahrscheinlich sind und ähnliche Effekte bzgl. des Systemverhaltens haben

Falls aber unterschiedliche Fehlerraten vorliegen, kann es zu Verfälschungen im Systemverhalten kommen, so dass Schätzer verzerrt sind

Beispiel:

System mit

- 3 Komponenten vom Typ 1 mit Fehlerwahrscheinlichkeit $O(\varepsilon)$
- 1 Komponente vom Typ 2 mit Fehlerwahrscheinlichkeit $O(\varepsilon^2)$
- Systemausfall bei Ausfall von allen Type 1 Komponenten oder der Typ 2 Komponente
- Defekte Komponenten werden nicht repariert

$\Rightarrow$ Wahrscheinlichkeit Systemausfall durch Ausfall der Typ 1 Komponenten $O(\varepsilon^3)$
Wahrscheinlichkeit Systemausfall durch Ausfall der Typ 2 Komponente $O(\varepsilon^2)$

Failure biasing modifiziert die Fehlerwahrscheinlichkeiten zu

- $O(1)$ für Typ 1 Komponenten
- $O(\varepsilon)$ für Typ 2 Komponenten

$\Rightarrow$Wahrscheinlichkeit Systemausfall durch Ausfall der Typ 1 Komponenten $O(1)$
Wahrscheinlichkeit Systemausfall durch Ausfall der Typ 2 Komponente $O(\varepsilon)$

Systemverhalten ändert sich, da der wahrscheinlichste Pfad zum Fehler unwahrscheinlicher wird !

Insgesamt

- bietet importance sampling theoretisch die Möglichkeit mit festem Aufwand beliebig kleine Wahrscheinlichkeiten zu schätzen
- ist in der Praxis aber eine optimale (oder auch nur relativ gute) Realisierung nur für sehr einfache Beispiele und Maße bekannt
- gibt es Modelle, bei denen nachweislich keine statische Parameteränderung zu guten Ergebnissen führt (bounded relative error)
- fehlt der Durchbruch zur praktischen Anwendbarkeit
- ist eine automatische Anwendung auf komplexere Modelle und damit auch eine Integration in Simulationswerkzeuge noch nicht absehbar