Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

Optimale importance sampling in Markovketens: efficiënte simulatie zeldzame gebeurtenissen.

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

Friso J.M. Scholten

Delft, Nederland
augustus 2015

BSc verslag TECHNISCHE WISKUNDE

"Optimale importance sampling in Markovketens: efficiënte simulatie zeldzame gebeurtenissen."

Friso J.M. Scholten

Technische Universiteit Delft

Begeleider

Dr.ir. L.E. Meester

Overige commissieleden

Dr. J.L.A. Dubbeldam

Dr.ir. M. Keijzer

Augustus, 2015 Delft

Inhoudsopgave

Hoofdstuk 1

Introductie

1.1 Monte Carlo-simulatie voor zeldzame gebeurtenissen

De impact van zeldzame gebeurtenissen kunnen soms zo groot zijn dat men geïnteresseerd is in de precieze kans dat deze gebeurtenissen optreden. Een aansprekend voorbeeld van deze zeldzame gebeurtenissen zijn storingen in de veiligheidssystemen in een vliegtuig. Hoewel de veiligheidssystemen zeer betrouwbaar zijn, is de impact van het falen van de combinatie van veiligheidssystemen (namelijk het neerstorten van een vliegtuig) zo groot dat men de kans van het neerstorten van een vliegtuig wil schatten. Andere voorbeelden van zeldzame gebeurtenissen die onderzocht worden zijn het vollopen van wachtrijen in telecommunicatieapparatuur en het schatten van het risico van extreem grote verliezen bij financiële producten.

De kans van dit soort zeldzame gebeurtenissen kan worden geschat door middel van Monte-Carlosimulatie. Monte-Carlosimulatie is een simulatietechniek waarbij een model niet één, maar een groot aantal keer wordt doorgerekend, waarbij de invoerwaarden van iedere replicatie realisaties zijn van vooraf bekende kansverdelingen. Hiermee kan de verdeling van de uitvoervariabele worden bepaald. Monte-Carlosimulatie is voornamelijk handig als de kans van zeldzame gebeurtenis moeilijk analytisch te berekenen is.

Voor het nauwkeurig schatten van de kans van een zeldzame gebeurtenis door middel van Monte-Carlosimulatie is echter een groot aantal replicaties nodig. Als de gebeurtenis optreedt met kans 10^{-6}, treedt de gebeurtenis bij een miljoen replicaties gemiddeld maar één keer op. Voor het nauwkeurig bepalen van de kans van de zeldzame gebeurtenis zijn dus nog veel meer replicaties nodig. In de volgende sectie wordt de efficiëntie van Monte-Carlosimulatie verder uitgewerkt. Hierbij wordt de relatie tussen het aantal benodigde replicaties en de gewenste nauwkeurigheid bepaald.

1.2 Efficiëntieanalyse Monte-Carlosimulatie

We beschouwen het volgende probleem: de kans $\gamma = P(A)$ dat gebeurtenis A optreedt moet worden geschat. We definiëren de stochast Y als de indicatorfunctie van A: $Y = \mathbb{1}(A)$. Dan geldt dat de verwachting van Y leidt tot $\mathrm{E}[Y] = 0 \cdot P(\neg A) + 1 \cdot P(A) = \gamma$. De variantie van Y, genoteerd door σ^2, wordt gegeven door $\sigma^2 = \gamma(1-\gamma)$.

Vervolgens passen we Monte-Carlosimulatie toe door een model N maal te simuleren. De output

van het model in run i wordt gegeven door de stochast $Y_i = \mathbb{1}(A \text{ treedt op in run } i)$. We definiëren $\hat{\gamma}$, een schatter van γ, door

$$\hat{\gamma} = \frac{1}{N} \sum_{k=1}^{N} Y_i$$

De verwachting van $\hat{\gamma}$ is

$$\mathrm{E}\left[\hat{\gamma}\right] = \mathrm{E}\left[\frac{1}{N} \sum_{k=1}^{N} Y_i\right] = \frac{1}{N} \sum_{k=1}^{N} \mathrm{E}\left[Y_i\right] = \frac{1}{N} \sum_{k=1}^{N} \gamma = \gamma,$$

dus $\hat{\gamma}$ is een zuivere schatter. De variantie van $\hat{\gamma}$ is

$$\mathrm{Var}\left[\hat{\gamma}\right] = \mathrm{Var}\left[\frac{1}{N} \sum_{k=1}^{n} Y_i\right] = \frac{1}{N^2} \sum_{k=1}^{n} \mathrm{Var}\left[Y_i\right] = \frac{1}{N}\gamma(1-\gamma) = \frac{1}{N}\sigma^2$$

waarbij de variantie uit de som kan worden genomen doordat de Y_i's onafhankelijk zijn.

Voor grote N kan de verdeling van $\hat{\gamma}$ worden benaderd door een normale verdeling als gevolg van de centrale limietstelling. Het betrouwbaarheidsinterval voor γ wordt dan gegeven door $\left(\hat{\gamma} \mp z\frac{\sigma}{\sqrt{N}}\right)$, waarbij z wordt gegeven door de mate van gewenste betrouwbaarheid. Een typische waarde van z is $z = 1.96$ voor een 95% betrouwbaarheidsinterval.

De grootte van het betrouwbaarheidsinterval $z\frac{\sigma}{\sqrt{N}}$, ook wel de absolute fout genoemd, geeft in het algemeen aan hoe accuraat de simulatie is: hoe kleiner de absolute fout, hoe waarschijnlijker dat de schatting dichtbij de gezochte waarde zit. Voor zeldzame gebeurtenissen, waarbij dus $\gamma \ll 1$, is de absolute fout echter op zichzelf niet erg interessant. Een voorbeeld illustreert dit goed: het 95%-betrouwbaarheidsinterval $[0.499, 0.501]$ zegt veel meer over de waarde van γ dan het 95%-betrouwbaarheidsinterval $[0, 0.002]$, terwijl de absolute fouten van de twee betrouwbaarheidsintervallen gelijk zijn. Het tweede betrouwbaarheidsinterval geeft immers niet aan of bijvoorbeeld $\gamma \approx 0.001$ of $\gamma \approx 0.000001$.

De relevantie van een absolute fout is dus afhankelijk van de waarde die je zoekt. Daarom kan er in plaats van de absolute fout de relatieve fout gebruikt, d.w.z. de absolute fout gedeeld door de te bepalen waarde: $RE = \frac{z\sigma}{\sqrt{N}\gamma}$. Er geldt voor de relatieve fout:

$$RE = \frac{z\sigma}{\sqrt{N}\gamma} = \frac{z\sqrt{\gamma(1-\gamma)}}{\sqrt{N}\gamma} \approx \frac{z\sqrt{\gamma}}{\sqrt{N}\gamma} = \frac{z}{\sqrt{N}\sqrt{\gamma}}.$$

Bij een gegeven gewenste relatieve fout en betrouwbaarheid is het aantal runs dat uitgevoerd moet worden omgekeerd evenredig met de grootte van de te schatten kans. Voor zeldzame gebeurtenissen, waarbij $\gamma \ll 1$, geldt dus dat het aantal replicaties N erg groot moet zijn voor een gegeven relatieve fout.

Er kan dus geconcludeerd worden dat de standaard Monte-Carlo-aanpak ongeschikt is voor het simuleren van zeldzame gebeurtenissen, omdat de variantie van de schatter relatief groot is, en er dus een groot aantal replicaties nodig is om een redelijke nauwkeurigheid te krijgen. Een van de mogelijke technieken om de variantie te reduceren is het toepassen van *importance sampling*. Importance sampling is een techniek waarbij, kort gezegd, een steekproef uit een andere verdeling wordt getrokken dan de verdeling waarin men geinteresseerd is.

1.3 Onderzoeksvraag

Het doel van dit onderzoek is het onderzoeken van de toepassing van importance sampling in het schatten van de kans van zeldzame gebeurtenissen in een credit risk model. Dit credit risk model behelst een bank die aan m schuldenaren geld heeft uitgeleend, waarbij schuldenaar i een schuld heeft van c_i en kans p_i heeft in gebreke te blijven en de lening dus niet kan terugbetalen. Het doel is om de kans in te schatten dat de bank meer dan een bepaald bedrag x niet zal terugkrijgen.

De hoofdvraag van dit onderzoek is: *Hoe goed werkt importance sampling in een credit risk model met onafhankelijke schuldenaren?*

De theorie achter importance sampling wordt besproken in hoofdstukken 2 en 3. In deze hoofdstukken wordt ook onderzocht hoe de importance sampling optimaal gekozen moet worden. Deze optimale keuze van de importance sampling verdeling wordt de *zero-variance verdeling* genoemd. In hoofdstuk 4 wordt een simulatie-experiment uitgevoerd waarbij importance sampling wordt gebruikt in een credit risk model met onafhankelijke schuldenaren, waarbij de schuldenaren gelijke eigenschappen hebben. In hoofdstuk 5 wordt een simulatie-experiment uitgevoerd waarbij gevarieerd wordt met de schuld en kans op niet-terugbetaling van de schuldenaren.

Hoofdstuk 2

Importance sampling

2.1 Principe achter importance sampling

In het vorige hoofdstuk is laten zien dat standaard Monte Carlo (MC) niet geschikt is voor het schatten van de kans van zeldzame gebeurtenissen, omdat de MC-schatter in dat geval een relatief grote variantie heeft, en er dus een zeer groot aantal replicaties uitgevoerd moet worden om een redelijke (relatieve) nauwkeurigheid te verkrijgen.

Het basisidee van importance sampling (IS) is dat de kansen in het model zodanig wordt aangepast, dat de gebeurtenissen die "belangrijk" zijn vaker zullen optreden. Stel dat we geïnteresseerd zijn in de stochast Y, waarbij we de kans γ willen schatten dat de zeldzame gebeurtenis $Y \in A$ optreedt. Er geldt dan $\gamma = P(Y \in A) = \mathrm{E}\left[\mathbb{1}(Y \in A)\right]$. De standaard MC-methode schat dan $\mathrm{E}\left[\mathbb{1}(Y \in A)\right]$ door $\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i \in A)$, waarbij $Y_1, \ldots, Y_n$ onderling onafhankelijke en gelijkverdeelde stochasten zijn. We nemen aan aan dat Y een continue stochast is met kansdichtheid f; de situatie voor een discrete stochast is analoog.

Het idee van importance sampling is als volgt: we kiezen een kansdichtheid $\tilde{f}$ zodanig dat $\tilde{f}(y) > 0$ als $\mathbb{1}(y \in A)f(y) \neq 0$. Er geldt dan:

$$\begin{aligned}
\mathrm{E}\left[\mathbb{1}(Y \in A)\right] &= \int \mathbb{1}(y \in A) f(y) dy && (2.1)\\
&= \int \mathbb{1}(y \in A) \frac{f(y)}{\tilde{f}(y)} \tilde{f}(y) dy && (2.2)\\
&= \int \mathbb{1}(y \in A) L(y) \tilde{f}(y) dy && (2.3)\\
&= \tilde{\mathrm{E}}\left[\mathbb{1}(Y \in A) L(Y)\right], && (2.4)
\end{aligned}$$

waarbij $L(y) = \frac{f(y)}{\tilde{f}(y)}$, de *likelihood ratio*, en $\tilde{\mathrm{E}}\left[\cdot\right]$ de verwachting onder kansdichtheid $\tilde{f}$ is.

De voorwaarde dat als $\mathbb{1}(y \in A)f(y) \neq 0$, dan $\tilde{f}(y) > 0$ is benodigd omdat anders er wordt gedeeld door 0 in de integraal.

$\tilde{\mathrm{E}}\left[\mathbb{1}(Y \in A)L(Y)\right]$ kan worden geschat door $\tilde{\gamma} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i \in A)L(Y_i)$. In het geval dat Y een discrete stochast is, geldt hetzelfde principe, waarbij de integralen door sommen en de kansdichtheden door kansfuncties vervangen worden.

Maar lost het kiezen van een andere kansverdeling ook het probleem van de inefficiëntie van de MC-schatter op? Het antwoord is ja, mits de dichtheid $\tilde{f}$ zo gekozen wordt dat $\tilde{f} \gg f$ op het goede gebied, namelijk $\tilde{f}(y) \gg f(y)$ voor $y \in A$. Er geldt nu voor de variantie van de IS-schatter:

$$\begin{aligned}
\text{Var}\,[\tilde{\gamma}] &= \frac{1}{n}\tilde{\text{Var}}\,[\mathbb{1}(Y \in A)L(Y)] && (2.5)\\
&= \frac{1}{n}\left(\tilde{\text{E}}\,[\mathbb{1}(Y \in A)L^2(Y)] - \gamma^2\right) && (2.6)\\
&\ll \frac{1}{n}\left(\tilde{\text{E}}\,[\mathbb{1}(Y \in A)L(Y)] - \gamma^2\right) && (2.7)\\
&= \frac{1}{n}\left(\text{E}\,[\mathbb{1}(Y \in A)] - \gamma^2\right) && (2.8)\\
&= \text{Var}\,[\hat{\gamma}] && (2.9)
\end{aligned}$$

De ongelijkheid volgt uit de keuze van $\tilde{f}$: $\tilde{f} \gg f$ op A impliceert dat $L = f/\tilde{f} \ll 1$ en dus dat $L^2 \ll L$.

We moeten dus $\tilde{f}$ zo kiezen dat $\tilde{f}(y) \gg f(y)$ voor $y \in A$. Theoretisch gezien is er een optimale keuze van de IS-kansdichtheid $\tilde{f}$. Stel namelijk dat je $\tilde{f}$ zodanig kiest $\tilde{f} = f\frac{\mathbb{1}(y\in A)}{\gamma}$, dan geldt dat:

$$\begin{aligned}
\text{Var}\,[\tilde{\gamma}] &= \frac{1}{n}\tilde{\text{Var}}\,[\mathbb{1}(Y \in A)L(Y)] && (2.10)\\
&= \frac{1}{n}\tilde{\text{Var}}\left[\mathbb{1}(Y \in A)\frac{f(Y)}{f(Y)\frac{\mathbb{1}(Y\in A)}{\gamma}}\right] && (2.11)\\
&= \frac{1}{n}\tilde{\text{Var}}\,[\gamma] && (2.12)\\
&= 0 && (2.13)
\end{aligned}$$

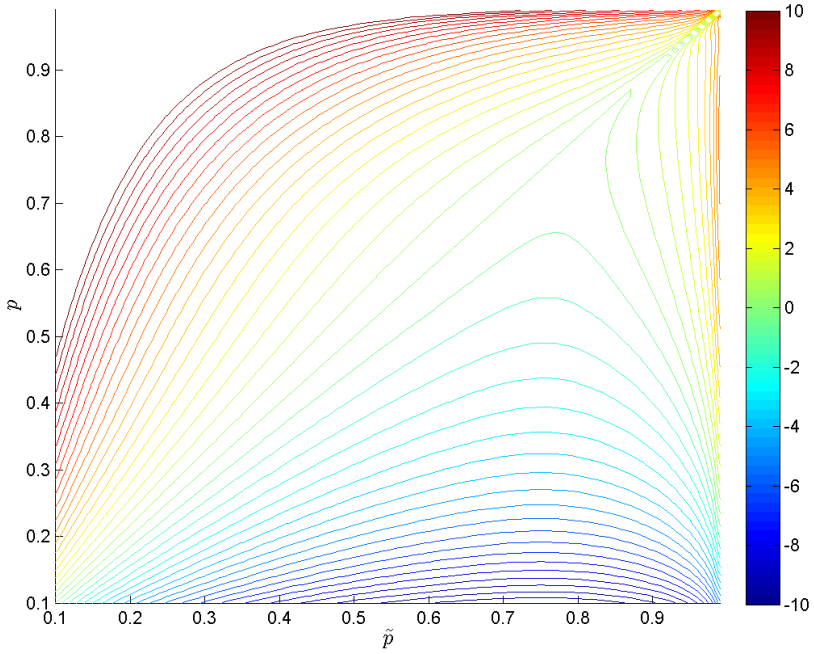
De laatste stap volgt door het feit dat γ een constante is. Er is dus een keuze van $\tilde{f}$ die variantie gelijk aan nul geeft! Deze $\tilde{f}$ noemen we de *zero-variance*-dichtheid. Met deze $\tilde{f}$ krijg je dus altijd de exacte waarde. Er zit echter een addertje onder het gras: deze keuze van $\tilde{f}$ hangt af van de waarde γ, wat juist de onbekende waarde is die geschat wordt. In de praktijk is deze optimale keuze dus niet haalbaar, maar er kan wel geprobeerd worden deze te benaderen.

2.2 Importance sampling met een binomiale verdeling

Om meer gevoel te krijgen voor de werking en de mogelijke variantiereductie door het toepassen van importance sampling wordt nu de binomiale verdeling als expliciet voorbeeld genomen.

We zijn nu geïnteresseerd in de rechterstaartkans van de binomiale verdeling. Dit betekent dat, als $Y \sim \text{Bin}(n, p)$, de waarde $\gamma = P(Y \geq k)$ geschat moet worden voor een zekere k. In de notatie van hiervoor definiëren we het interval $A = [k, \infty]$ en $\gamma = P(Y \geq k) = \text{E}\,[\mathbb{1}(Y \in A)]$.

Stel we willen een andere binomiale verdeling, namelijk $\text{Bin}(n, \tilde{p})$, gebruiken om de waarde van γ te schatten door middel van importance sampling (IS). Voor de variantie onder de IS-verdeling



Figuur 2.1: Isolijnen van de logaritme van de variantieratio van importance sampling verdeling van een $\text{Bin}(n, \tilde{p})$ ten opzichte van een $\text{Bin}(n, p)$, afhankelijk van de verschillende $\tilde{p}$ en p. Overige parameters: $n = 20$ en $k = 15$.

geldt dan het volgende:

$$\tilde{\text{Var}}\left[\mathbb{1}(Y \in A)L(Y)\right] = \tilde{\text{E}}\left[\mathbb{1}(Y \in A)L^2(Y)\right] - \gamma^2 \tag{2.14}$$

$$= \sum_{i=k}^{n} \left(\frac{\binom{n}{i} p^i (1-p)^{n-i}}{\binom{n}{i} \tilde{p}^i (1-\tilde{p})^{n-i}} \right)^2 \binom{n}{i} \tilde{p}^i (1-\tilde{p})^{n-i} - \gamma^2 \tag{2.15}$$

$$= \sum_{i=k}^{n} \binom{n}{i} \left(\frac{p^2}{\tilde{p}} \right)^i \left(\frac{(1-p)^2}{1-\tilde{p}} \right)^{n-i} - \gamma^2 \tag{2.16}$$

Welke $\tilde{p}$ moet dan gekozen worden om de variantie te verkleinen? De variantieratio tussen de IS-schatter en de MC-schatter is:

$$\text{VR} = \frac{\tilde{\text{Var}}\left[\mathbb{1}_A(Y)L(Y)\right]}{\text{Var}\left[\mathbb{1}_A(Y)\right]} = \frac{\sum_{i=k}^{n} \binom{n}{i} \left(\frac{p^2}{\tilde{p}} \right)^i \left(\frac{(1-p)^2}{1-\tilde{p}} \right)^{n-i} - \gamma^2}{\gamma(1-\gamma)}. \tag{2.17}$$

Hoe kleiner de variantieratio, hoe efficiënter de IS-schatter is.

We nemen nu het voorbeeld voor $n = 20$, $k = 15$ en verschillende p. In figuur 2.1 worden isolijnen van de logaritme van de variantieratio weergegeven voor verschillende combinaties van p en $\tilde{p}$. De blauwe lijnen geven aan waar importance sampling voordelig is. In het gebied van de rode lijnen is het gebruik van importance sampling juist contraproductief.

Figuur 2.1 illustreert dat het in het algemeen beter is om $\tilde{p} > p$ te kiezen, hoewel een te grote $\tilde{p}$ de variantieratio weer laat toenemen. Het optimimum voor de kleinere waarden van p lijkt te liggen rond $\tilde{p} = \frac{k}{n} = 0.75$. Opvallend is dat in dit binomiale voorbeeld de variantieratio niet daalt

tot 0. Dit laat zien dat de optimale keuze van de IS-verdeling niet in de klasse van $\text{Bin}(n, \tilde{p})$-verdelingen zit. In het volgende hoofdstuk wordt de theorie uitgebreid naar importance sampling bij Markovketens, waar de optimale keuze van de IS-verdeling wel gevonden kan worden.

Hoofdstuk 3

Importance sampling in Markov-ketens

3.1 Inleiding Markov-ketens

De modellen die moeten worden doorgerekend door middel van Monte-Carlosimulatie bestaan vaak niet uit een enkele stochast met een bekende kansverdeling, maar bestaan vaak uit gecompliceerde stochastische processen zoals Markov-ketens. In dit hoofdstuk wordt het gebruik van importance sampling in zogenaamde discrete tijdshomogene Markov-ketens (DTMC) bekeken. Een DTMC is een stochastich proces $\{Y_j, j \geq 0\}$ met een aftelbare toestandsruimte, waarbij de Markoveigenschap geldt en de overgangskansen onafhankelijk zijn van de tijdsparameter j.

De Markoveigenschap houdt in dat, gegeven dat de huidige toestand bekend is, de kans van het toekomstig gedrag van het proces niet veranderd wordt door het toevoegen van kennis uit het verleden van het stochastisch proces:

$$P(Y_{n+1} = y_{n+1}|Y_n = y_n, Y_{n-1} = y_{n-1}, \ldots, Y_0 = y_0) = P(Y_{n+1} = y_{n+1}|Y_n = y_n). \tag{3.1}$$

De tijdshomogeniteit betekent dat voor alle $n \geq 0$:

$$P(Y_{n+1} = y_{n+1}|Y_n = y_n) = P(Y_1 = y_1|Y_0 = y_0) \tag{3.2}$$

In de volgende behandeling van importance sampling bij Markov-ketens wordt er een aantal eigenschappen gebruikt van conditionele verwachtingen en varianties, gegeven in de volgende stellingen.

Stelling 3.1.1. $\mathrm{E}\left[h((Y)X|Y)\right] = h(Y)\mathrm{E}\left[X|Y\right]$

Bewijs. De stochast $\mathrm{E}\left[h(Y)X|Y\right]$ is een functie van Y. We noemen $r(y) = \mathrm{E}\left[h(Y)X|Y = y\right]$, dan geldt:

$$r(y) = \sum_x xh(y)P(X = x|Y = y) \tag{3.3}$$

$$= h(y)\sum_x xP(X = x|Y = y) \tag{3.4}$$

$$= h(y)\mathrm{E}\left[X|Y = y\right] \tag{3.5}$$

Daarom geldt:

$$\mathrm{E}\left[h(Y)X|Y\right] = r(Y) = h(Y)\mathrm{E}\left[X|Y\right] \tag{3.6}$$

□

Stelling 3.1.2. $\mathrm{E}\left[\mathrm{E}\left[Y|X\right]\right] = \mathrm{E}\left[Y\right]$
Stelling 3.1.3. $\mathrm{Var}\left[X\right] = \mathrm{E}\left[\mathrm{Var}\left[X|Y\right]\right] + \mathrm{Var}\left[\mathrm{E}\left[X|Y\right]\right]$

De bewijzen van stellingen 3.1.2 en 3.1.3 kunnen gevonden worden op respectievelijk pagina 149 en 151 [1].

3.2 Importance sampling bij Markov-ketens

In deze sectie wordt eerst het basisraamwerk van Rubino en Tuffin [2] besproken, waarbij de notatie voor de verschillende elementen wordt geïntroduceerd. In de subsectie daarna wordt er dieper ingegaan op de keuze van Rubino en Tuffin om de standaard IS-schatter XL_τ aan te passen naar $\tilde{X}$. In het onderdeel daarna worden de optimale overgangskansen, de zogenaamde zero-variance kansen, afgeleid.

3.2.1 Raamwerk Rubino en Tuffin

Rubino en Tuffin [2] beschouwen een vrij algemeen raamwerk voor het toepassen van importance sampling op Markov-ketens. Zij motiveren hun beweringen echter wat summier. Daarom wordt in deze paragraaf het raamwerk van Rubino en Tuffin uitgebreid besproken.

We beschouwen een Markov-keten met discrete tijdsparameter, $\{Y_j, j \geq 0\}$, met toestandsruimte $\mathcal{Y}$. We noemen $\tau = \inf\{j \geq 0 : Y_j \in \Delta\}$ het stoptijdstip, oftewel het tijdstip dat de keten voor de eerste keer in de verzameling toestanden $\Delta \subset \mathcal{Y}$ komt. Er wordt aangenomen dat $\mathrm{E}\left[\tau\right] < \infty$. De overgangskansen van de keten worden genoteerd als $P(y,z) = P\left[Y_j = z|Y_{j-1} = y\right]$ voor $y, z \in \mathcal{Y}$, en de beginkansen als $\pi_0(y) = P[Y_0 = y]$ voor $y \in \mathcal{Y}$. Vervolgens definiëren we de functie $c : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$, die de kosten voorstelt van elke overgang van een toestand naar de volgende toestand. De totale kosten van het hele pad tot aan het stoptijdstip wordt aangegeven door de stochast $X = \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i)$. We nemen $\gamma(y) = \mathrm{E}\left[X|Y_0 = y\right]$, en we definiëren vervolgens $\beta = \mathrm{E}\left[X\right] = \sum_{y\in\mathcal{Y}} \pi_0(y)\gamma_0(y)$.

Men is hier dus geïnteresseerd in de verwachting van X. X kan van alles voorstellen: bijvoorbeeld het aantal stappen voordat de keten in een bepaalde toestand komt, of de kosten die gemaakt moeten worden als de keten een bepaald traject doorloopt. In onze context van het schatten van de kans van het optreden van zeldzame gebeurtenissen wordt de functie c gegeven als $c(y,z) = \mathbb{1}(y \notin A, z \in A)$, waarbij $A \subset \Delta$ de verzameling toestanden is waarvan we de kans dat de keten in die verzameling komt willen weten.

Het basisidee van importance sampling is om nu de kansen van de mogelijke paden aan te passen. Zo wordt de kans op het pad $(y_0, \ldots, y_n)$, waarbij $y_j \notin \Delta$ voor $j < n$, en $y_n \in \Delta$, gegeven door

$$P[(Y_0, \ldots, Y_\tau) = (y_0, \ldots, y_n)] = \pi_0(y_0) \prod_{j=1}^{n} P(y_{j-1}, y_j).$$

De overgangskansen $P(y_{i-1}, y_i)$ worden vervangen door de nieuwe kansen $\tilde{P}(y_{i-1}, y_i)$. Deze IS-kansen moeten zo worden gekozen dat na de vervanging van de overgangskansen de paden in verwachting een eindige lengte moeten hebben. Bovendien moeten de mogelijke paden die tot

kosten leiden in de originele situatie ook bestaan bij gebruik van de nieuwe overgangskansen. Wiskundig genoteerd betekent dit dat $\tilde{\mathrm{E}}[\tau] < \infty$ en $\tilde{P}[(Y_0, \ldots, Y_\tau) = (y_0, \ldots, y_n)] > 0$ waar $\sum_{i=1}^{n} c(y_{i-1}, y_i) P[(Y_0, \ldots, Y_\tau) = (y_0, \ldots, y_n)] > 0$.

We kunnen de likelihood ratio van een overgang definiëren als:

$$L(y_{j-1}, y_j) = \frac{P(y_{j-1}, y_j)}{\tilde{P}(y_{j-1}, y_j)}$$

De likelihood ratio van een mogelijk pad definiëren we dan als:

$$L_n(Y_0, \ldots, Y_n) = \prod_{j=1}^{n} L(Y_{j-1}, Y_j)$$

De IS-schatter voor $\gamma(y)$ wordt dan gegeven als XL_τ.

3.2.2 Keuze importance sampling schatter

Rubino en Tuffin gebruiken echter in plaats van XL_τ de schatter

$$\tilde{X} = \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) L_i \tag{3.7}$$

Bij de schatter XL_τ loopt in de i-de term van de som het product van de likelihood ratio's tot en met τ, terwijl bij de $\tilde{X}$ het product slechts loopt tot en met i. In deze sectie wordt bewezen dat $\tilde{X}$ dezelfde verwachting heeft als XL_τ. Daarna wordt kort gerefereerd naar de literatuur die de reden achter de keuze van $\tilde{X}$ voorschrijven.

Lemma 3.2.1. *Er geldt voor vaste $t \geq 1$:*
als $\tilde{P}[(Y_0, \ldots, Y_t) = (y_0, \ldots, y_t)] > 0$ waar $P[(Y_0, \ldots, Y_t) = (y_0, \ldots, y_t)] > 0$, dan geldt:

$$\tilde{\mathrm{E}}[L_t | Y_0 = y] = 1$$

.

Bewijs. Dit lemma wordt bewezen met behulp van volledige inductie.

De begininductie voor $t = 1$:

$$\begin{aligned}
\tilde{\mathrm{E}}[L(Y_0, Y_1) | Y_0 = y] &= \tilde{\mathrm{E}}\left[\frac{P(y, Y_1)}{\tilde{P}(y, Y_1)}\right] && (3.8)\\
&= \sum_{z \in \{\tilde{P}(y,z) > 0\}} \tilde{P}(y, z) \frac{P(y, z)}{\tilde{P}(y, z)} && (3.9)\\
&= \sum_{z \in \{\tilde{P}(y,z) > 0\}} P(y, z) && (3.10)\\
&= \sum_{z \in \{P(y,z) > 0\}} P(y, z) && (3.11)\\
&= 1 && (3.12)
\end{aligned}$$

De inductieveronderstelling is nu dat voor een $n \geq 1$ geldt dat:

$$\tilde{\mathrm{E}}\left[L_n | Y_0 = y\right] = 1$$

De inductiestap gaat als volgt:

$$\begin{aligned}
\tilde{\mathrm{E}}\left[L_{n+1} | Y_0 = y\right] &= \tilde{\mathrm{E}}\left[\tilde{\mathrm{E}}\left[L_{n+1} | Y_1, \ldots, Y_n\right] | Y_0 = y\right] && (3.13)\\
&= \tilde{\mathrm{E}}\left[L_n \tilde{\mathrm{E}}\left[L(Y_n, Y_{n+1}) | Y_1, \ldots, Y_n\right] | Y_0 = y\right] && (3.14)\\
&= \tilde{\mathrm{E}}\left[L_n \tilde{\mathrm{E}}\left[L(Y_n, Y_{n+1}) | Y_n\right] | Y_0 = y\right] && (3.15)\\
&= \tilde{\mathrm{E}}\left[L_n \tilde{\mathrm{E}}\left[L(Y_0, Y_1) | Y_0\right] | Y_0 = y\right] && (3.16)\\
&= \tilde{\mathrm{E}}\left[L_n | Y_0 = y\right] && (3.17)\\
&\overset{ind}{=} 1 && (3.18)
\end{aligned}$$

De eerste stap gebruikt stelling 3.1.2. De tweede stap gebruikt stelling 3.1.1. De derde stap gebruikt de Markoveigenschap. De vierde stap gebruikt de tijdshomogeniteit. De vijfde stap gebruikt de hiervoor bewezen vergelijking $\tilde{\mathrm{E}}\left[L(Y_0, Y_1) | Y_0\right] = 1$. □

Eerst bewijzen we dat XL dezelfde verwachting heeft als $\tilde{X}$.

Stelling 3.2.2.

$$\tilde{\mathrm{E}}\left[XL_\tau \;\middle|\; Y_0\right] = \tilde{\mathrm{E}}\left[\tilde{X} \;\middle|\; Y_0\right],$$

Bewijs. We beschouwen $\tilde{\mathrm{E}}\left[XL_\tau \;\middle|\; Y_0, \tau = t\right]$ voor willekeurige t. Er geldt:

$$\begin{aligned}
\tilde{\mathrm{E}}\left[XL_\tau \;\middle|\; Y_0, \tau = t\right] &= \tilde{\mathrm{E}}\left[\sum_{i=1}^{t} c(Y_{i-1}, Y_i) L_t \;\middle|\; Y_0, \tau = t\right] && (3.19)\\
&= \sum_{i=1}^{t} \tilde{\mathrm{E}}\left[c(Y_{i-1}, Y_i) L_t \;\middle|\; Y_0, \tau = t\right] && (3.20)
\end{aligned}$$

$\tilde{\mathrm{E}}\left[\tilde{X} \;\middle|\; Y_0, \tau = t\right]$ kan op een analoge manier worden omgeschreven. Het is dus voldoende om te bewijzen dat geldt

$$\tilde{\mathrm{E}}\left[c(Y_{i-1}, Y_i) L_t\right] = \tilde{\mathrm{E}}\left[c(Y_{i-1}, Y_i) L_i\right]$$

voor alle vaste t en i met $i \leq t$, aangezien de sommen over deze verwachtingen ook gelijk zijn.

Als $c(y_{i-1}, y_i) = 0$, dan is de gelijkheid triviaal. We kijken daarom verder naar het geval dat $c(y_{i-1}, y_i) > 0$, en dus $\tilde{P}[y_{i-1}, y_i] > 0$ waar $P[y_{i-1}, y_i] > 0$. Volgens lemma 3.2.1 is de verwachting van de likelihood ratio dus gelijk aan 1.

Er geldt dan:

$$\begin{aligned}
\tilde{\mathrm{E}}\left[c(Y_{i-1}, Y_i) L_t\right] &= \tilde{\mathrm{E}}\left[\tilde{\mathrm{E}}\left[c(Y_{i-1}, Y_i) L_t \;\middle|\; Y_i, Y_{i-1}, \ldots, Y_0\right]\right] && (3.21)\\
&= \tilde{\mathrm{E}}\left[c(Y_{i-1}, Y_i) L_i \tilde{\mathrm{E}}\left[\prod_{j=i+1}^{t} L(Y_{j-1}, Y_j) \;\middle|\; Y_i\right]\right] && (3.22)\\
&= \tilde{\mathrm{E}}\left[c(Y_{i-1}, Y_i) L_i\right] && (3.23)
\end{aligned}$$

Dit geldt onafhankelijk van t voor $t \geq i$, dus er geldt $\tilde{\mathrm{E}}\left[XL_\tau \;\middle|\; Y_0, \tau = t\right]$. □

Nu bewezen is dat XL_τ dezelfde verwachting heeft als $\tilde{X}$ rest nog de vraag waarom $\tilde{X}$ wordt gebruikt in plaats van XL_τ. Uit de literatuur volgen kortweg twee redenen:

1. Het kiezen van de (zero-variance) IS-verdeling laat de Markov-eigenschap verdwijnen uit het stochastisch proces. Een voorbeeld hiervan wordt beschreven in Example 2.4 van [3].
2. De variantie van $\tilde{X}$ is kleiner dat de variantie van XL_τ, wanneer wordt voldaan aan bepaalde voorwaarden. Een voorbeeld van een voldoende voorwaarde is positieve covariantie tussen de termen van de som in $\tilde{X}$, zoals beschreven in [4].

Daarom gebruiken we vanaf nu $\tilde{X}$ als IS-schatter.

3.2.3 Zero-variance overgangskansen: optimale keuze

Ook hier is weer de vraag: wat is de optimale keuze van $\tilde{P}(y,z)$?

Stelling 3.2.3 (Zero-variance in Markov-model)**.** *Als $\tilde{P}(y,z)$, voor alle $y,z \in \mathcal{Y}$, zodanig wordt gekozen dat*

$$\tilde{P}(y,z) = \frac{P(y,z)\,(c(y,z)+\gamma(z))}{\gamma(y)}, \tag{3.24}$$

dan is $\tilde{\mathrm{Var}}\left[\tilde{X}\right] = 0$ *(en* $\tilde{\mathrm{E}}\left[\tilde{X}\right] = \gamma(Y_0)$*).*

Merk op dat deze keuze van $\tilde{P}(y,z)$ geldige kansen oplevert, aangezien er geldt:

$$\begin{aligned}
\gamma(y) &= \mathrm{E}\left[X \;\middle|\; Y_0 = y\right] && (3.25)\\
&= \mathrm{E}\left[\mathrm{E}\left[X \;\middle|\; Y_1\right] \;\middle|\; Y_0 = y\right] && (3.26)\\
&= \mathrm{E}\left[\mathrm{E}\left[\sum_{i=1}^{\tau} c(Y_{i-1},Y_i) \;\middle|\; Y_1\right] \;\middle|\; Y_0 = y\right] && (3.27)\\
&= \mathrm{E}\left[c(Y_0,Y_1) + \mathrm{E}\left[\sum_{i=2}^{\tau} c(Y_{i-1},Y_i) \;\middle|\; Y_1\right] \;\middle|\; Y_0 = y\right] && (3.28)\\
&= \mathrm{E}\left[c(Y_0,Y_1) + \gamma(Y_1) \;\middle|\; Y_0 = y\right] && (3.29)\\
&= \sum_{z\in\mathcal{Y}} P(y,z)\,(c(y,z)+\gamma(z)) && (3.30)
\end{aligned}$$

Hierdoor geldt dat $\sum_{z\in\mathcal{Y}} \tilde{P}(y,z) = 1$.

Bewijs. Voor het bewijs van deze stelling vullen we de keuze van $\tilde{P}(y,z)$ rechtstreeks in de schatter $\tilde{X}$. De schatter valt dan te versimpelen tot $\tilde{X} = \gamma(Y_0)$, en is dus een constante bij een gegeven begintoestand Y_0 en heeft dus variantie 0.

$$\begin{aligned}
\tilde{X} &= \sum_{t=0}^{\infty} \mathbb{1}(\tau = t) \sum_{i=1}^{t} c(Y_{i-1},Y_i) \prod_{j=1}^{i} \frac{P(Y_{j-1},Y_j)}{\tilde{P}(Y_{j-1},Y_j)} && (3.31)\\
&= \sum_{t=0}^{\infty} \mathbb{1}(\tau = t) \sum_{i=1}^{t} c(Y_{i-1},Y_i) \prod_{j=1}^{i} \frac{\gamma(Y_{j-1})}{c(Y_{j-1},Y_j)+\gamma(Y_j)} && (3.32)
\end{aligned}$$

We kunnen nu alle waarden die τ aan kan nemen afgaan. Het geval van $t = 0$ is triviaal aangezien het proces zich dan al in de stoptoestand bevindt.

Voor $t = 1$, dan is $\gamma(Y_1) = 0$, en volgt uit bovenstaande vergelijking dat

$$\tilde{X} = c(Y_0, Y_1)\frac{\gamma(Y_0)}{c(Y_0, Y_1) + \gamma(Y_1)} = \gamma(Y_0)$$

.

Voor $t \geq 2$ geldt door de laatste term van de som af te splitsen:

$$\begin{aligned}\tilde{X} = &\sum_{i=1}^{t-1} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \\ &+ c(Y_{t-1}, Y_t) \prod_{j=1}^{t} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \qquad (3.33)\\ = &\sum_{i=1}^{t-1} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \\ &+ c(Y_{t-1}, Y_t) \left(\frac{\gamma(Y_{t-1})}{c(Y_{t-1}, Y_t) + \gamma(Y_t)}\right) \prod_{j=1}^{t-1} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \qquad (3.34)\end{aligned}$$

Het invullen van $\gamma(Y_t) = 0$ geeft

$$\begin{aligned}= &\sum_{i=1}^{t-1} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \\ &+ c(Y_{t-1}, Y_t) \left(\frac{\gamma(Y_{t-1})}{c(Y_{t-1}, Y_t)}\right) \prod_{j=1}^{t-1} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \qquad (3.35)\\ = &\sum_{i=1}^{t-1} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \\ &+ \gamma(Y_{t-1}) \prod_{j=1}^{t-1} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \qquad (3.36)\end{aligned}$$

We kunnen deze som door neerwaartse inductie op de waarde van t terugbrengen naar een

gemakkelijkere vergelijking. Er geldt namelijk voor elke $n \geq 2$ dat:

$$\begin{aligned} &\sum_{i=1}^{n} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \\ &\qquad + \gamma(Y_n) \prod_{j=1}^{n} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \end{aligned} \tag{3.37}$$

$$\begin{aligned} &= \sum_{i=1}^{n-1} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \\ &\qquad + (c(Y_{n-1}, Y_n) + \gamma(Y_n)) \prod_{j=1}^{n+1} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \end{aligned} \tag{3.38}$$

$$\begin{aligned} &= \sum_{i=1}^{n-1} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \\ &\qquad + \gamma(Y_{n-1}) \prod_{j=1}^{n-1} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \end{aligned} \tag{3.39}$$

Voor elke waarde van $t \geq 2$ kan de vergelijking voor $\tilde{X}$ dus worden teruggebracht naar de vergelijking

$$\tilde{X} = c(Y_0, Y_1) \frac{\gamma(Y_0)}{c(Y_0, Y_1) + \gamma(Y_1)} + \gamma(Y_1) \frac{\gamma(Y_0)}{c(Y_0, Y_1) + \gamma(Y_1)} \tag{3.40}$$

$$= \gamma(Y_0) \tag{3.41}$$

De schatter $\tilde{X}$ heeft dus een constante waarde bij deze keuze van $\tilde{P}(y, z)$ en gegeven Y_0, dus de schatter heeft in dat geval variantie 0. □

3.3 Importance sampling in een credit risk model

3.3.1 Credit risk model als Markovketen

We passen nu een credit risk model in de context die hiervoor beschreven is.

Het credit risk model is als volgt: een bank heeft aan m schuldenaren geld uitgeleend. Schuldenaar i heeft een schuld van waarde c_i met een kans van p_i dat hij in gebreke blijft. We beschouwen hier de schuldenaren als onafhankelijk: de kans dat een schuldenaar in gebreke blijft hangt niet af van de andere schuldenaren. Het totale verlies is gegeven door

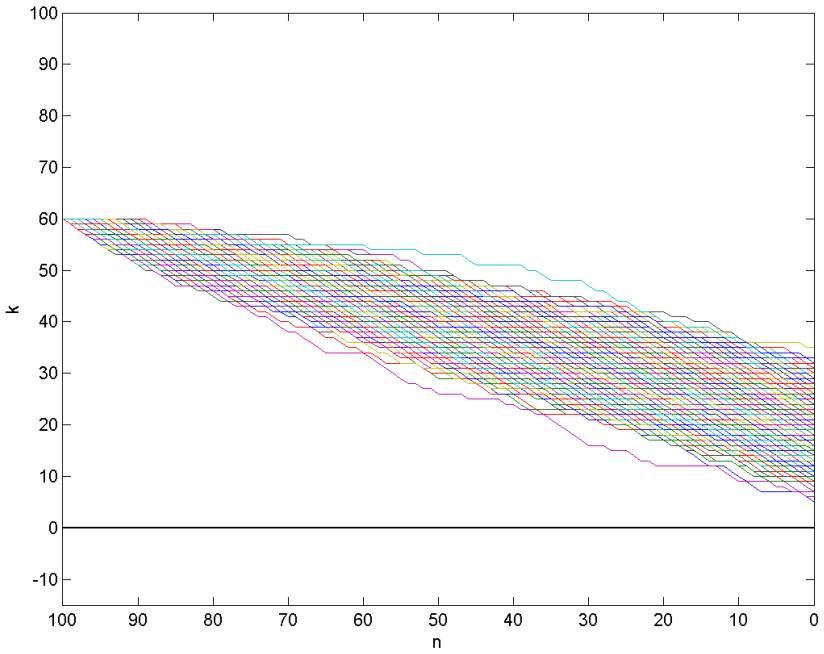
$$V = \sum_{i=1}^{m} c_i Y_i, \tag{3.42}$$

met $Y_i \sim \text{Ber}(p_i)$ een Bernoulli-stochast. De bank probeert de kans uit te rekenen dat de bank een verlies V draait van meer dan x:

$$\gamma = P(V \geq x). \tag{3.43}$$

Bij de simpele situatie dat alle klanten gelijk zijn, dat wil zeggen $p_i = p$ en $c_i = c$ voor alle i, reduceert dit model naar een binomiale verdeling $V' \sim \text{Bin}(m, p)$. De te schatten kans wordt dan gegeven door $\gamma = P(V' \geq \frac{x}{c})$.

We gaan bovenstaand probleem beschrijven door een Markovketen, zodat we de theorie over importance sampling uit de vorige secties kunnen toepassen. De Markovketen heeft de vector (n, v) als toestand, waarbij (n, v) correspondeert met n schuldenaren, waarbij er nog voor tenminste v euro in gebreke moet blijven. Elke schuldenaar correspondeert met een overgang van de toestand: een schuldenaar kan in gebreke blijven of niet. We lopen "achteruit" door de rij schuldenaren: bij het in gebreke blijven van schuldenaar n (met een schuld van c_n) springt de keten dus van toestand (n, v) naar $(n-1, v-c_n)$. Bij het in gebreke blijven van schuldenaar n met schuld c_n hoeft er immers door de resterende schuldenaren nog een totale schuld van tenminste $v - c_n$ euro in gebreke te blijven. Mocht schuldenaar n niet in gebreke blijven, dan springt de keten van (n, v) naar $(n-1, v)$.



Figuur 3.1: Mogelijke paden door de toestandsruimte.

Om meer gevoel te krijgen bij de verschillende paden zijn in figuur 3.1 zijn verschillende paden weergegeven vanaf begintoestand $(100, 60)$. De paden lopen vanaf $m = 100$ totdat alle schuldenaren zijn gesimuleerd. Bij elke schuldenaar is er de mogelijkheid dat er naar beneden wordt gesprongen, wat correspondeert met een niet-terugbetalende schuldenaar. De kans dat het pad onder de $v = 0$-lijn uitkomt wordt geschat.

We gieten bovenstaand verhaal in de hiervoor beschreven notatie van het raamwerk van Rubino en Tuffin. We bekijken het stochastisch proces $\{Y_t, t \geq 0\}$ met $Y_t = (n, v) \in \mathcal{Y} = \mathbb{N} \times \mathbb{R}$. De overgangskansen worden gegeven door

$$P(y, z) = \begin{cases} p_n & \text{als } y = (n, v) \ \& \ z = (n-1, v-c_n) \\ 1 - p_n & \text{als } y = (n, v) \ \& \ z = (n-1, v) \end{cases}. \tag{3.44}$$

Het stopgebied Δ wordt gedefinieerd als $\Delta = \{(n,v)|n=0\}$. We definiëren de kostenfunctie $c(y_{i-1},y_i) = \mathbb{1}(y_i \in A \;\&\; y_{i-1} \notin A)$. Het gebied A, waarvan we de kans willen schatten dat de keten daar eindigt is $A = \{(n,v) \;\middle|\; n=0, v \leq 0\}$. We zijn geïnteresseerd in de (additieve) stochast $X = \sum_{i=1}^{\tau} c(Y_{i-1},Y_i) = \mathbb{1}(Y_\tau \in A \;\&\; Y_{\tau-1} \notin A)$. We nemen begintoestand $Y_0 = (m,x)$. Met deze keuze geldt voor de stoptijd $\tau = m$.

In deze Markovketen kunnen we importance sampling toepassen door nieuwe overgangskansen $p_{n,v}$ te geven:

$$\tilde{P}(y,z) = \begin{cases} p_{n,v} & \text{als } y=(n,v) \;\&\; z=(n-1,v-c_n) \\ 1-p_{n,v} & \text{als } y=(n,v) \;\&\; z=(n-1,v) \end{cases}. \tag{3.45}$$

In principe zijn $p_{n,v}$ redelijk vrij te kiezen. Er moet worden voldaan aan $\tilde{\mathrm{E}}[\tau] < \infty$, hetgeen gegarandeerd is doordat τ is vastgesteld op $\tau = m$. De andere voorwaarde is dat als $\mathbb{1}(y_m \in A \;\&\; y_{m-1} \notin A)\prod_{j=1}^{m} P(y_{j-1},y_j) > 0$, dan $\prod_{j=1}^{m} \tilde{P}(y_{j-1},y_j) > 0$. Deze voorwaarde kan geïnterpreteerd worden als dat er geen $p_{n,v}$ mag worden gekozen die een pad afsluit dat naar het gebied A leidt.

3.3.2 Recursieve betrekking van variantie

Om de opbouw van de variantie gedurende de simulatie en de invloed van de overgangskansen te verduidelijken, wordt een recursieve vergelijking van de variantie afgeleid.

De recursieve betrekking wordt afgeleid met behulp van de variantiedecompositie van stelling 3.1.3. We beginnen bij (n,v), met $n > 1$. De variantie van $\tilde{X}_{n,v}$ wordt geconditioneerd naar $\tilde{B}_{n,v}$, de Bernoulli-stap op (n,v). Deze Bernoulli-stap is klant n die zijn schuld niet kan terugbetalen (en dus naar toestand $(n-1,v-c_n)$ wordt gesprongen) of wel kan terugbetalen (en dus naar toestand $(n-1,v)$ wordt gesprongen). We krijgen dan:

$$\tilde{\mathrm{Var}}\left[\tilde{X}_{n,v}\right] = \tilde{\mathrm{Var}}\left[\tilde{\mathrm{E}}\left[\tilde{X}_{n,v} \mid B_{n,v}\right]\right] + \tilde{\mathrm{E}}\left[\tilde{\mathrm{Var}}\left[\tilde{X}_{n,v} \mid B_{n,v}\right]\right]. \tag{3.46}$$

We werken de componenten van deze vergelijking verder uit.

$$\tilde{\mathrm{E}}\left[\tilde{X}_{n,v} \mid B_{n,v}\right] = \begin{cases} \frac{p_n}{p_{n,v}}\tilde{\mathrm{E}}\left[\tilde{X}_{n-1,v-c_n}\right] & \text{met kans } p_{n,v} \\ \frac{1-p_n}{1-p_{n,v}}\tilde{\mathrm{E}}\left[\tilde{X}_{n-1,v}\right] & \text{met kans } 1-p_{n,v} \end{cases} \tag{3.47}$$

Het nemen van de variantie hiervan geeft:

$$\tilde{\mathrm{Var}}\left[\tilde{\mathrm{E}}\left[\tilde{X}_{n,v} \mid B_{n,v}\right]\right] = \left(\frac{p_n}{p_{n,v}}\tilde{\mathrm{E}}\left[\tilde{X}_{n-1,v-c_n}\right] - \frac{1-p_n}{1-p_{n,v}}\tilde{\mathrm{E}}\left[\tilde{X}_{n-1,v}\right]\right)^2 p_{n,v}(1-p_{n,v}). \tag{3.48}$$

Het eerste deel van de decompositie is dus uitgewerkt. De uitwerking van het tweede deel gaat soortgelijk:

$$\tilde{\mathrm{Var}}\left[\tilde{X}_{n,v} \mid B_{n,v}\right] = \begin{cases} \frac{p_n^2}{p_{n,v}^2}\tilde{\mathrm{Var}}\left[\tilde{X}_{n-1,v-c_n}\right] & \text{met kans } p_{n,v} \\ \frac{(1-p_n)^2}{(1-p_{n,v})^2}\tilde{\mathrm{Var}}\left[\tilde{X}_{n-1,v}\right] & \text{met kans } 1-p_{n,v} \end{cases} \tag{3.49}$$

De vewachting hiervan nemen geeft:

$$\tilde{\mathrm{E}}\left[\tilde{\mathrm{Var}}\left[\tilde{X}_{n,v} \mid B_{n,v}\right]\right] = \frac{p_n^2}{p_{n,v}}\tilde{\mathrm{Var}}\left[\tilde{X}_{n-1,v-c_n}\right] + \frac{(1-p_n)^2}{1-p_{n,v}}\tilde{\mathrm{Var}}\left[\tilde{X}_{n-1,v}\right] \tag{3.50}$$

Vergelijking (3.46) kan nu worden ingevuld door middel van (3.48) en (3.50):

$$\tilde{\text{Var}}\left[\tilde{X}_{n,v}\right] = \left(\frac{p_n}{p_{n,v}}\tilde{\text{E}}\left[\tilde{X}_{n-1,v-c_n}\right] - \frac{1-p_n}{1-p_{n,v}}\tilde{\text{E}}\left[\tilde{X}_{n-1,v}\right]\right)^2 p_{n,v}(1-\tilde{p}_{n,k}) + \frac{p_n^2}{p_{n,v}}\tilde{\text{Var}}\left[\tilde{X}_{n-1,v-c_n}\right] + \frac{(1-p_n)^2}{1-p_{n,v}}\tilde{\text{Var}}\left[\tilde{X}_{n-1,v}\right] \tag{3.51}$$

Om de vergelijking wat overzichtelijker te maken schrijven we de vergelijking anders, door $\sigma^2_{n,v} = \tilde{\text{Var}}\left[\tilde{X}_{n,v}\right]$, $\gamma_{n,v} = \tilde{\text{E}}\left[\tilde{X}_{n,v}\right]$, en $\overline{p} = 1-p$. De recurrente betrekking wordt dus gegeven door:

$$\sigma^2_{n,v} = \frac{p_n^2}{p_{n,v}}\sigma^2_{n-1,v-c_n} + \frac{\overline{p}_n^2}{\overline{p}_{n,k}}\sigma^2_{n-1,v} + p_{n,v}\overline{p}_{n,v}\left(\frac{p_n}{p_{n,v}}\gamma_{n-1,v-c_n} - \frac{\overline{p}_n}{\overline{p}_{n,v}}\gamma_{n-1,v}\right)^2 \tag{3.52}$$

De recurrente betrekking geeft aan dat de variantie op (n,v) afhangt van de varianties van eerdere stappen, plus extra variantie die wordt toegevoegd per stap.

We proberen nu $p_{n,v}$ zo te kiezen dat de $\sigma^2_{n,v}$ geminimaliseerd wordt. Voor de overzichtelijkheid substitueren we:

$$a = p_n\sigma_{n-1,v-1} \tag{3.53}$$
$$b = \overline{p}_n\sigma_{n-1,v} \tag{3.54}$$
$$c = p_n\gamma_{n-1,v-1} \tag{3.55}$$
$$d = \overline{p}_n\gamma_{n-1,v}. \tag{3.56}$$
$$x = p_{n,v} \tag{3.57}$$

Dan geldt er vervolgens:

$$\begin{aligned}
\arg\min_{0\le x\le 1}\sigma^2_{n,v} &= \arg\min_{0\le x\le 1}\frac{a^2}{x} + \frac{b^2}{1-x} + x(1-x)\left(\frac{c}{x} - \frac{d}{1-x}\right)^2 && (3.58)\\
&= \arg\min_{0\le x\le 1}\frac{a^2}{x} + \frac{b^2}{1-x} + \left(\frac{c^2(1-x)^2 + d^2x^2 - cdx(1-x)}{x(1-x)}\right) && (3.59)\\
&= \arg\min_{0\le x\le 1}\frac{a^2}{x} + \frac{b^2}{1-x} + \left(\frac{c^2(1-x) + d^2x - x(1-x)(c^2+cd+d^2)}{x(1-x)}\right) && (3.60)\\
&= \arg\min_{0\le x\le 1}\frac{a^2}{x} + \frac{b^2}{1-x} + \frac{c^2}{x} + \frac{d^2}{(1-x)} - c^2 - cd - d^2 && (3.61)\\
&= \arg\min_{0\le x\le 1}\frac{a^2+c^2}{x} + \frac{b^2+d^2}{1-x} && (3.62)
\end{aligned}$$

Hierbij gebruikten we in de derde stap de identiteiten $(1-x)^2 = (1-x) - x(1-x)$ en $x^2 = x - x(1-x)$.

Door deze laatste uitdrukking te differentiëren naar x en gelijk aan 0 te stellen krijg je dat de optimale keuze $\frac{p_{n,v}}{\overline{p}_{n,v}}$ wordt gegeven door:

$$\frac{p_{n,v}}{\overline{p}_{n,v}} = \frac{p_n\sqrt{\sigma^2_{n-1,v-c_n} + \gamma^2_{n-1,v-c_n}}}{\overline{p}_n\sqrt{\sigma^2_{n-1,v} + \gamma^2_{n-1,v}}}. \tag{3.63}$$

In het algemeen zijn de varianties klein bij het gebruik van importance sampling, dus als we veronderstellen dat in bovenstaande vergelijking $\sigma^2_{n-1,v-c_n} = \sigma^2_{n-1,v} = 0$, dan krijgen we als keuze:

$$\frac{p_{n,v}}{\overline{p}_{n,v}} = \frac{p_n \gamma_{n-1,v-c_n}}{\overline{p}_n \gamma_{n-1,v}}. \tag{3.64}$$

Dit omschrijven naar een vergelijking voor $p_{n,v}$ geeft:

$$p_{n,v} = \frac{\frac{p_{n,v}}{\overline{p}_{n,v}}}{1 + \frac{p_{n,v}}{\overline{p}_{n,v}}} \tag{3.65}$$

$$= \frac{p_n \gamma_{n-1,v-c_n}}{p_n \gamma_{n-1,v-c_n} + \overline{p}_n \gamma_{n-1,v}} \tag{3.66}$$

$$= p_n \frac{\gamma_{n-1,v-c_n}}{\gamma_{n,v}}. \tag{3.67}$$

Dit is hetzelfde als de zero-variance keuze van stelling 3.2.3.

Vergelijking (3.63) kan ook worden gebruikt om een aantal randgevallen te onderzoeken. Stel bijvoorbeeld de situatie voor dat de run zo is gelopen dat het onmogelijk is om in het gebied $A = \{(n, v) \mid n = 0, v \leq 0\}$ te komen als de volgende schuldenaar niet in gebreke blijft. Dit houdt dus in dat $\gamma_{n-1,v} = 0$, en dus ook $0 \leq \sigma^2_{n-1,v} \leq \gamma_{n-1,v}(1 - \gamma_{n-1,v}) = 0$. Vergelijking (3.63) wordt dan $\frac{p_{n,v}}{\overline{p}_{n,v}} = \infty$, dus $p_{n,v} = 1$. Schuldenaar n wordt door de importance sampling als in gebreke verondersteld.

Een andere situatie is als $\gamma_{n-1,v-c_n} = \gamma_{n-1,v} = 1$. Dit komt voor als er niet meer het gewenste gebied A niet meer ontlopen kan worden, bijvoorbeeld als de replicatie zich bevindt in $(n, v) = (5, -1)$. Dan geldt ook dat $\sigma^2_{n-1,v-c_n} = \sigma^2_{n-1,v} = 0$, dus $\frac{p_{n,v}}{\overline{p}_{n,v}} = \frac{p_n}{\overline{p}_n}$. De optimale keuze van $p_{n,v}$ wordt dan $p_{n,v} = p_n$.

3.4 Conclusie

In dit hoofdstuk is een beschrijving van het raamwerk van Rubino en Tuffin [2] gegeven. In dit raamwerk wordt beschreven hoe importance sampling kan worden toegepast in een Markovketen, inclusief de theoretische zero variance verdeling die de gewilde verwachtingswaarde met variantie 0 schat. Daarna is laten zien hoe ons credit risk probleem in dit Markov-raamwerk past. Daarbij is ook een recurrente betrekking afgeleid waarmee de optimale keuze wordt gegeven voor elke toestand (n, v) in het credit risk probleem:

$$\frac{p_{n,v}}{\overline{p}_{n,v}} = \frac{p_n \sqrt{\sigma^2_{n-1,v-c_n} + \gamma^2_{n-1,v-c_n}}}{\overline{p}_n \sqrt{\sigma^2_{n-1,v} + \gamma^2_{n-1,v}}}. \tag{3.68}$$

Uit deze vergelijking kunnen drie keuzes worden afgeleid:

1. De zero-variance keuze ligt ook dichtbij de optimale keuze als de variantie klein is:

$$\frac{p_{n,v}}{\overline{p}_{n,v}} = \frac{p_n \gamma_{n-1,v-c_n}}{\overline{p}_n \gamma_{n-1,v}}. \tag{3.69}$$

2. Het is de optimale keuze om paden die niet leiden tot het gewenste gebied af te sluiten door $p_{n,v} = 1$ te kiezen.

3. Als het gewenste gebied niet meer ontlopen kan worden, is het optimaal om $p_{n,v} = p_n$ te kiezen.

In de volgende hoofdstukken worden twee simulatie-experimenten van importance sampling bij het credit risk model beschreven. Het eerste simulatie-experiment neemt aan dat de schuldenaren gelijke eigenschappen hebben: de schuld en kans op niet-terugbetaling zijn gelijk voor alle schuldenaren. Het tweede experiment varieert de kans op niet-terugbetaling en de schuld van de schuldenaren.

Hoofdstuk 4

Simulatie met gelijke schuldenaren

4.1 Doel experiment

Dit experiment heeft als opzet om te onderzoeken in hoeverre het toepassen van importance sampling in een credit risk probleem leidt tot performancewinst. In het bijzonder wordt getest hoe gevoelig het resultaat is voor accuratere benaderingen van de zero-variance verdeling: is het de moeite waard om de zero-variance overgangskansen goed te benaderen of werken minder accurate benaderingen ook goed?

Allereerst nemen we $p_i = p$ en $c_i = c$ voor alle schuldenaren i. We kunnen hier willekeurige c nemen aangezien $P(c\sum_{i=1}^{m} Y_i \geq x) = P(\sum_{i=1}^{m} Y_i \geq x/c)$ en we dezelfde resultaten voor willekeurige c kunnen vinden door de x te schalen. In dit hoofdstuk nemen we $c = 1$. Merk op dat de Y_i nu allemaal Bernoulli(p) verdeeld zijn, en dus $V = \sum_{i=1}^{m} Y_i$ binomiaal verdeeld is met parameters m en p. Deze verdeling valt vrij gemakkelijk uit te rekenen en dus is het mogelijk de simulatieresultaten te vergelijken met de exacte waarden. In sectie 3.3.1 is aangegeven hoe het credit-risk model is gegoten in een Markovketen. We gebruiken hier de notatie uit dat hoofdstuk.

In de volgende sectie worden de verschillende keuzen van overgangskansen besproken die worden toegepast als importance-sampling-overgangskansen. In de sectie daarna worden de resultaten van het experiment besproken.

4.2 Keuzen overgangskansen

In deze sectie worden de verschillende gekozen IS-overgangskansen besproken. De basis van de IS-overgangskansen is de zero-variance-verdeling. Deze zero-variance verdeling wordt benaderd, aangezien de zero-variance verdeling niet precies gekozen kan worden aangezien deze afhangt van de te vinden staartkans. De benaderingen zijn gesorteerd van grof tot preciezer.

4.2.1 Standaard Monte Carlo

De standaard Monte Carlo aanpak is de aanpak zonder gebruik te maken van importance sampling. De variatiecoefficiënt kan in dit geval exact worden bepaald door $\frac{\sqrt{\gamma(1-\gamma)}}{\gamma} = \frac{\sqrt{1-\gamma}}{\sqrt{\gamma}}$, aangezien de waardes γ rechtstreeks kunnen worden berekend.

$$p_{n,v} = p,$$

voor alle n, v.

4.2.2 Vaste overgangskans

Deze simulatiewijze neemt een vaste keuze van de IS-overgangskansen $p_{n,v}$, en laat deze keuze dus niet afhangen van de toestand waarin het Markovproces zich begeeft. Als keuze voor de $p_{n,v}$ wordt $\frac{x}{m}$ gebruikt, aangezien deze keuze in het voorbeeld van paragraaf 2.1 in de buurt van de optimale keuze leek te liggen.

$$p_{n,v} = \frac{x}{m}$$

voor alle n, v.

4.2.3 Vaste overgangskans, inclusief afsluiten paden

Deze simulatiewijze is ongeveer hetzelfde als de vorige simulatiewijze, met twee toevoegingen. Ten eerste worden nu paden die niet meer kunnen leiden tot het gewenste gebied afgesloten. Concreet betekent dit dat $p_{n,v} = 1$ als $\gamma_{n-1,v} = 0$. Bovendien wordt de $p_{n,v} = p$ gesteld wanneer het gewenste gebied onontkoombaar is, dat wil zeggen $\gamma_{n-1,v} = 1$.

$$p_{n,v} = \begin{cases} 1 & \text{als } \gamma_{n-1,v} = 0 \\ p & \text{als } \gamma_{n-1,v} = 1 \\ \frac{x}{m} & \text{anders} \end{cases}$$

4.2.4 ZV-benadering 1: een-termsbenadering

De optimale overgangskansen worden gegeven door:

$$\frac{p_{n,v}}{\overline{p}_{n,v}} = \frac{p\gamma_{n-1,v-1}}{\overline{p}\gamma_{n-1,v}}.$$

Als zero-variance-benadering 1 nemen we nu een eenterms-benadering van de zero-variance verdeling. Deze benadering berust op $P(Z \geq v) = \sum_{i=v}^{n} P(Z = i) \approx P(Z = v)$ als $v \geq np$, aangezien $P(Z = v)$ de grootste term is in de som. Er volgt dan:

$$\frac{p_{n,v}}{\overline{p}_{n,v}} \approx \frac{p\binom{n-1}{v-1}p^{v-1}(1-p)^{n-v}}{(1-p)\binom{n-1}{v}p^{v}(1-p)^{n-v-1}} \tag{4.1}$$

$$= \frac{\binom{n-1}{v-1}p^{v}(1-p)^{n-v}}{\binom{n-1}{v}p^{v}(1-p)^{n-v}} \tag{4.2}$$

$$= \frac{v!(n-v-1)!}{(v-1)!(n-v)!} \tag{4.3}$$

$$= \frac{v}{n-v}. \tag{4.4}$$

Dit leidt dus voor $v \geq np$ tot:

$$p_{n,v} = \frac{\frac{v}{n-v}}{1+\frac{v}{n-v}} = \frac{v}{n}.$$

Resumerend, ZV-benadering 1 luidt:

$$p_{n,v} = \max(\frac{v}{n}, p).$$

4.2.5 ZV-benadering 2: tweetermsbenadering

De tweede ZV-benadering is een benadering van de zero-variance verdeling door middel van twee termen in de som $\sum_{i=v}^{n} P(Z=i)$. We benaderen de staartkans door de eerste twee termen van de som:

$$P(Z \geq v) \approx P(Z=v) + P(Z=v+1)$$

voor $Z \sim \text{Bin}(n,p)$ als $v \geq np$. De benadering van de zero variance keuze gaat dan als volgt:

$$\frac{p_{n,v}}{\overline{p}_{n,v}} \approx \frac{p}{\overline{p}} \frac{P(S_{n-1}=v-1)+P(S_{n-1}=v)}{P(S_{n-1}=v)+P(S_{n-1}=v+1)} \tag{4.5}$$

$$= \frac{p}{\overline{p}} \frac{\binom{n-1}{v-1}p^{v-1}(1-p)^{n-v} + \binom{n-1}{v}p^{v}(1-p)^{n-v-1}}{\binom{n-1}{v}p^{v}(1-p)^{n-v-1} + \binom{n-1}{v+1}p^{v+1}(1-p)^{n-v-2}} \tag{4.6}$$

$$= \frac{\binom{n-1}{v-1}p^{v}(1-p)^{n-v} + \binom{n-1}{v}p^{v+1}(1-p)^{n-v-1}}{\binom{n-1}{v}p^{v}(1-p)^{n-v} + \binom{n-1}{v+1}p^{v+1}(1-p)^{n-v-1}} \tag{4.7}$$

$$= \frac{\binom{n-1}{v-1}(1-p) + \binom{n-1}{v}p}{\binom{n-1}{v}(1-p) + \binom{n-1}{v+1}p} \tag{4.8}$$

$$= \frac{\frac{v}{n-v}(1-p)+p}{(1-p)+\frac{n-v-1}{v+1}p} \tag{4.9}$$

$$= \frac{v+1}{n-v} \cdot \frac{v(1-p)+(n-v)p}{(v+1)(1-p)+(n-v-1)p}. \tag{4.10}$$

Dit leidt tot

$$p_{n,v} = \frac{v+1}{n} \frac{v(1-p)+(n-v)p}{(v+1)(1-p)+(n-v)p}$$

4.3 Resultaten

In tabel 4.1 zijn de verschillende waarden weergegeven die zijn gekozen voor de parameters m, x en p. De vijf verschillende keuzen van de overgangskansen zijn dus onderworpen alle $3 \cdot 4 \cdot 2 = 24$ combinaties van deze parameters.

Met $x = +a\sigma$ wordt bedoeld hoeveel standaarddeviaties de parameter x wordt gekozen vanaf de verwachtingswaarde. Een x van $+4\sigma$ met $m = 100$ en $p = 0.4$ wordt dus $x = mp + 4 \cdot \sqrt{mp(1-p)} = 40 + 4 \cdot \sqrt{24} \approx 60$, waarbij x wordt afgerond naar een geheel getal.

m	x	p
100	$+4\sigma$	0.4
200	$+6\sigma$	0.1
300	$+8\sigma$	
	$+10\sigma$	

Tabel 4.1: De verschillende gesimuleerde waarden van de parameters m, v en p.

In figuur 4.1 worden de paden van de verschillende replicaties door de toestandsruimte aangegeven. Bij de standaard Monte Carlo (figuur 4.1a) komt geen enkel pad uit in het gewenste gebied, wat leidt tot een schatting van $gam\hat{m}a_{m,x} = 0$. Bij vaste keuze van $p_{n,k}$ (figuur 4.1b) wordt het bereik van de paden steeds groter, en ongeveer de helft van de paden bereikt het gewenste gebied. In figuur 4.1c worden de paden duidelijk gedwongen richting het gewenste gebied. De ZV-benaderingen van figuren 4.1d en 4.1e geven een wat smallere bandbreedte van de paden.

m	Variabele	$+4\sigma$	$+6\sigma$	$+8\sigma$	$+10\sigma$
100	x	22	28	34	40
	$\gamma_{m,x}$	3.12×10^{-4}	3.48×10^{-7}	7.00×10^{-11}	2.95×10^{-15}
200	x	37	45	54	62
	$\gamma_{m,x}$	1.86×10^{-4}	1.76×10^{-7}	8.57×10^{-12}	2.32×10^{-16}
300	x	51	61	72	82
	$\gamma_{m,x}$	1.28×10^{-4}	7.14×10^{-8}	1.91×10^{-12}	2.03×10^{-17}

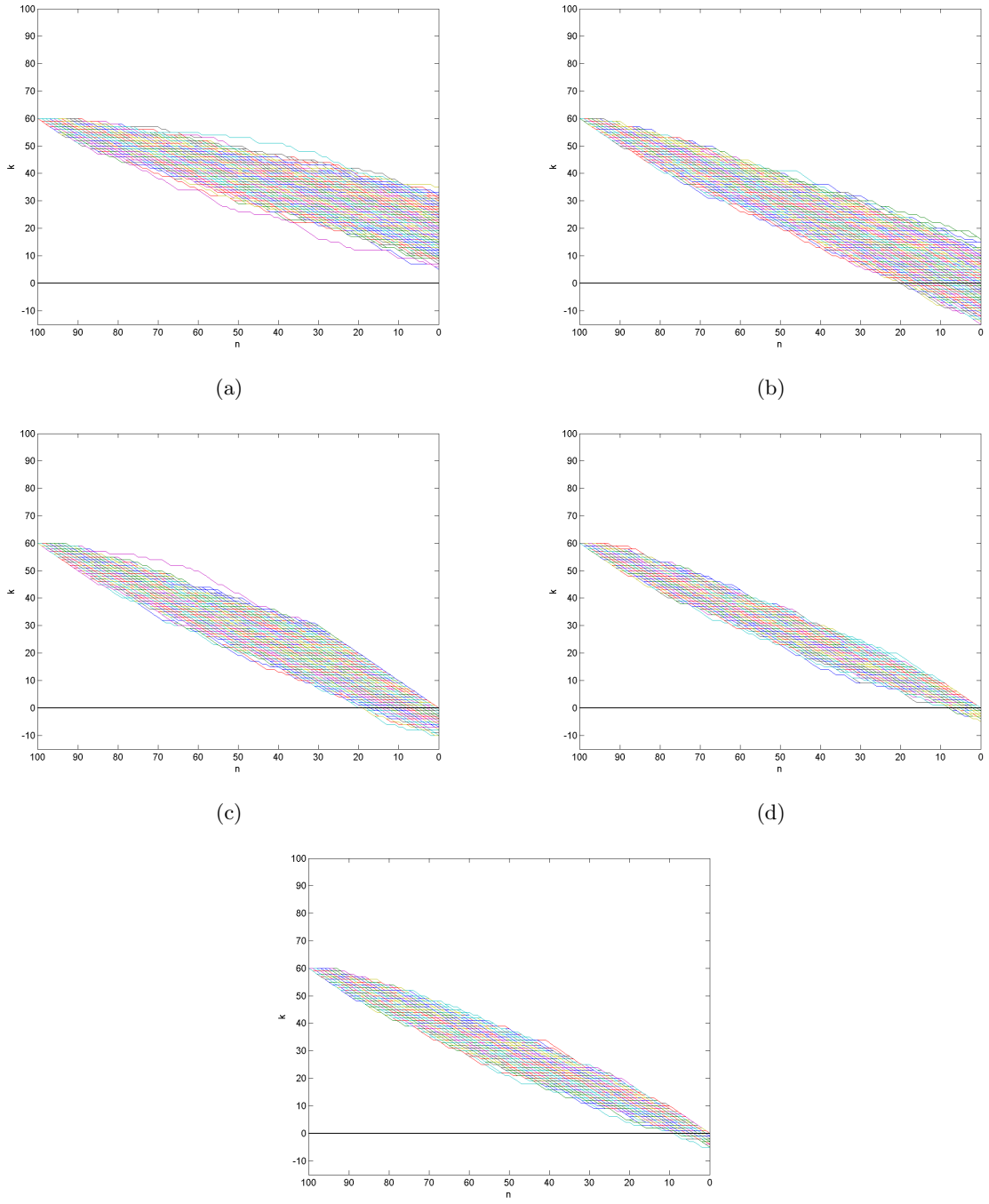
Tabel 4.2: De gebruikte waarden van x en corresponderende waarden van $\gamma_{m,x}$, bij $p = 0.10$.

Ter controle zijn de waarden van $t = \frac{\hat{\gamma}-\gamma}{s/\sqrt{n}}$ weergegeven in tabel 4.3 waarmee een t-toets uitgevoerd kan worden. Hoewel er enkele uitschietende waarden tussen zitten, bijv. 2.68, is dat geen gekke waarde gezien het aantal toetsen dat uitgevoerd is.

m	Benaderingswijze	$+4\sigma$	$+6\sigma$	$+8\sigma$	$+10\sigma$
100	Vaste $p_{n,v}$	0.16	-0.42	1.52	2.68
	Vaste $p_{n,v}$, paden afgesloten	-0.51	1.10	-0.66	0.95
	ZV-benadering 1	-0.84	0.04	-0.62	-0.47
	ZV-benadering 2	0.28	1.35	-0.30	-0.86
100	Vaste $p_{n,v}$	-2.66	-1.22	-0.05	-2.24
	Vaste $p_{n,v}$, paden afgesloten	1.27	0.04	-0.46	0.48
	ZV-benadering 1	-1.56	-1.37	1.32	-0.08
	ZV-benadering 2	0.47	-1.67	0.71	-2.80
100	Vaste $p_{n,v}$	-2.04	-1.54	1.84	1.63
	Vaste $p_{n,v}$, paden afgesloten	-0.80	0.57	0.66	-2.62
	ZV-benadering 1	-0.51	-0.00	0.11	-0.60
	ZV-benadering 2	0.09	0.91	2.60	0.77

Tabel 4.3: t-waarden voor $p = 0.40$.

In tabel 4.4 zijn de geschatte variatiecoefficiënten van de verschillende importance sampling



(a)

(b)

(c)

(d)

(e)

Figuur 4.1: De verschillende paden van de 900 replicaties vanuit begintoestand $(n, v) = (100, 60)$.
Figuur (a) gebruikt een standaard Monte Carlo manier.
Figuur (b) gebruikt een vaste IS-overgangskans.
Figuur (c) gebruikt een vaste IS-overgangskans, inclusief het afsluiten van ongewenste paden.
Figuur (d) gebruikt een eerste orde benadering van de ZV-kansen.
Figuur (e) gebruikt een tweede orde benadering van de ZV-kansen.

m	Benaderingswijze	$+4\sigma$	$+6\sigma$	$+8\sigma$	$+10\sigma$
	Monte Carlo	1.53×10^{2}	1.52×10^{4}	2.25×10^{7}	6.83×10^{11}
	Vaste $p_{n,v}$	2.02	2.34	2.35	2.14
100	Vaste $p_{n,v}$, paden afgesloten	1.35	1.47	1.52	1.45
	ZV-benadering 1	6.05×10^{-1}	5.00×10^{-1}	3.72×10^{-1}	2.27×10^{-1}
	ZV-benadering 2	3.28×10^{-1}	2.17×10^{-1}	9.98×10^{-2}	3.93×10^{-2}
	Monte Carlo	1.52×10^{2}	2.43×10^{4}	1.63×10^{7}	1.53×10^{11}
	Vaste $p_{n,v}$	2.15	2.52	2.82	3.30
200	Vaste $p_{n,v}$, paden afgesloten	1.50	1.74	1.83	1.86
	ZV-benadering 1	7.44×10^{-1}	6.39×10^{-1}	5.32×10^{-1}	4.34×10^{-1}
	ZV-benadering 2	4.72×10^{-1}	3.11×10^{-1}	2.40×10^{-1}	1.31×10^{-1}
	Monte Carlo	1.47×10^{2}	2.18×10^{4}	2.20×10^{7}	1.69×10^{11}
	Vaste $p_{n,v}$	2.23	2.62	2.59	2.99
300	Vaste $p_{n,v}$, paden afgesloten	1.70	1.85	1.95	2.17
	ZV-benadering 1	7.83×10^{-1}	6.85×10^{-1}	6.23×10^{-1}	5.04×10^{-1}
	ZV-benadering 2	5.47×10^{-1}	4.29×10^{-1}	3.53×10^{-1}	2.40×10^{-1}

Tabel 4.4: Geschatte variatiecoefficiënten $\frac{s}{\bar{x}}$ voor de verschillende benaderingswijzen, m's, x's en $p = 0.40$. De variatiecoefficiënten zijn geschat met gebruik van 900 replicaties. De variatiecoefficiënten van de standaard Monte Carlo-schatter zijn de exacte waarden.

schatters gegeven. De variatiecoefficiënten $\frac{\sigma}{\mu}$ zijn geschat door de verhouding tussen de wortel van de steekproefvariantie en steekproefgemiddelde $\frac{s}{\bar{x}}$ over 900 replicaties voor de importance sampling schatters. De variatiecoefficiënt van de standaard Monte Carlo schatter is bepaald door de (theoretische) waarden van $\frac{\sigma}{\mu}$ te gebruiken.

Met behulp van de variantiecoefficiënten zijn de aantallen benodigde simulatiereplicaties N voor een gegeven nauwkeurigheid RSE gemakkelijk te bepalen door de formule $N = \left(\frac{VR}{RSE}\right)^2$. Als voorbeeld kan de situatie van $n = 100$ en op 4σ afstand genomen worden. Als je voor het behalen van 10% RSE standaard Monte Carlo gebruikt, heb je $\left(\frac{1.53 \times 10^2}{0.1}\right)^2 = 2.35 \times 10^6$ replicaties nodig. Met het gebruik van de ZV-benadering 2 zou je slechts 11 replicaties nodig hebben! Voor zeldzamere gebeurtenissen wordt dit verschil alleen maar groter.

Wat opvalt aan de resulaten in tabel 4.4 is dat de importance sampling schatters onnauwkeuriger worden bij grotere m. Waarschijnlijk wordt dit veroorzaakt doordat bij een grotere m de paden langer worden en de benaderingsfouten elke stap optellen.

Voor zeldzamere gebeurtenissen die verder in de staart gezocht worden wordt het verschil tussen IS en de standaard MC groter. Dit is naar verwachting, aangezien er in hoofdstuk 1 was geconcludeerd dat de standaard MC slechter werkt naar mate de gebeurtenis zeldzamer wordt. De vaste IS-methoden worden wat onnauwkeuriger des te verder in de staart van de verdeling gezocht wordt. De adaptieve IS-methoden worden juist nauwkeuriger. Dit komt waarschijnlijk doordat de benaderingen die gebruikt worden in de adaptieve IS-methoden dichterbij de zero-variance overgangskansen komen.

Ter controle zijn de waarden van $t = \frac{\hat{\gamma}-\gamma}{s/\sqrt{n}}$ weergegeven in tabel 4.5 waarmee een t-toets uitgevoerd kan worden. Hoewel er enkele uitschietende waarden tussen zitten, bijv. -2.62, is dat geen gekke waarde gezien het aantal toetsen dat uitgevoerd is.

m	Benaderingswijze	$+4\sigma$	$+6\sigma$	$+8\sigma$	$+10\sigma$
100	Vaste $p_{n,v}$	0.51	0.28	-2.27	-0.13
	Vaste $p_{n,v}$, paden afgesloten	1.18	0.54	0.17	-0.60
	ZV-benadering 1	0.80	-0.16	0.10	-0.69
	ZV-benadering 2	1.02	-0.69	-0.68	0.40
100	Vaste $p_{n,v}$	-1.04	-1.28	0.50	1.57
	Vaste $p_{n,v}$, paden afgesloten	-0.75	0.45	0.43	0.75
	ZV-benadering 1	-0.74	0.45	-2.62	0.22
	ZV-benadering 2	-0.88	1.25	1.00	1.82
100	Vaste $p_{n,v}$	0.84	-0.11	1.18	1.27
	Vaste $p_{n,v}$, paden afgesloten	-0.58	0.17	-0.95	1.22
	ZV-benadering 1	0.93	1.39	0.96	1.07
	ZV-benadering 2	0.74	2.02	0.14	0.60

Tabel 4.5: t-waarden voor $p = 0.10$.

De patronen in de resultaten voor $p = 0.1$ zijn vergelijkbaar met de resultaten voor $p = 0.4$. Een grotere m zorgt voor grotere variantie van de IS-schatters. De verschillen tussen standaard MC en IS worden groter naar mate er verder in de staart gezocht moet worden. De vaste IS-methoden werken slechter naar mate er verder in de staart wordt gezocht. De toestandsafhankelijke IS-methoden worden juist beter verder in de staart.

m	Benaderingswijze	$+4\sigma$	$+6\sigma$	$+8\sigma$	$+10\sigma$
	Monte Carlo	5.66×10^{1}	1.69×10^{3}	1.20×10^{5}	1.84×10^{7}
	Vaste $p_{n,v}$	1.94	2.34	2.88	2.83
100	Vaste $p_{n,v}$, paden afgesloten	1.42	1.52	1.61	1.73
	ZV-benadering 1	5.03×10^{-1}	3.61×10^{-1}	2.87×10^{-1}	2.27×10^{-1}
	ZV-benadering 2	2.32×10^{-1}	1.36×10^{-1}	8.52×10^{-2}	7.00×10^{-2}
	Monte Carlo	7.34×10^{1}	2.38×10^{3}	3.42×10^{5}	6.57×10^{7}
	Vaste $p_{n,v}$	2.09	2.49	2.72	2.86
200	Vaste $p_{n,v}$, paden afgesloten	1.58	1.71	1.85	1.95
	ZV-benadering 1	5.55×10^{-1}	4.89×10^{-1}	3.72×10^{-1}	3.37×10^{-1}
	ZV-benadering 2	3.12×10^{-1}	2.45×10^{-1}	1.87×10^{-1}	1.29×10^{-1}
	Monte Carlo	8.82×10^{1}	3.74×10^{3}	7.24×10^{5}	2.22×10^{8}
	Vaste $p_{n,v}$	2.01	2.56	2.70	2.90
300	Vaste $p_{n,v}$, paden afgesloten	1.69	1.87	2.12	2.08
	ZV-benadering 1	6.56×10^{-1}	5.63×10^{-1}	4.58×10^{-1}	4.19×10^{-1}
	ZV-benadering 2	3.86×10^{-1}	3.08×10^{-1}	2.13×10^{-1}	1.66×10^{-1}

Tabel 4.6: Geschatte variatiecoefficiënten $\frac{s}{\bar{x}}$ voor de verschillende benaderingswijzen, m, x en $p = 0.10$. De variatiecoefficiënten zijn geschat met gebruik van 900 replicaties. De variatiecoefficiënten van de standaard Monte Carlo-schatter zijn de exacte waarden.

Hoofdstuk 5

Simulatie met variatie in schuldenaren

5.1 Opzet experiment

Na het vorige experiment, waar schuldenaren gelijke schuld en gelijke kansen op in gebreke blijven hadden, worden in dit experiment schuldenaren met variërende schuld en kans op in gebreke blijven gebruikt. In sectie 3.3.1 is aangegeven hoe het credit-risk model is gegoten in een Markovketen. We gebruiken hier de notatie uit dat hoofdstuk.

We nemen nu een casus van Glasserman en Li [5]: een portfolio van $m = 1000$ schuldenaren, met schuldenaar i een kans op niet-terugbetaling van $p_i = 0.01(1 + \sin(16 \cdot \pi \cdot i/m))$ en een bedrag van $c_i = (\lceil 5i/m \rceil)^2$, met $i = 1, ..., m$. De c_i worden ook wel de gewichten genoemd. Dit houdt in dat de kansen varieren tussen de 0% en 2%, en dat de bedragen variëren tussen de 1, 4, 9, 16 en 25. Glasserman en Li veronderstellen *afhankelijke* schuldenaren, waar hier *onafhankelijke* schuldenaren gebruikt worden. De resultaten zijn dus niet te vergelijken.

We zoeken weer de staartkansen $\gamma_{m,x} = P(V_m \geq x)$, waarbij $V_m = \sum_{i=1}^{m} c_i Y_i$.

Er geldt voor de verwachting $\mathrm{E}[V_m] = \sum_{i=1}^{m} c_i p_i \approx 104$ en voor de variantie $\mathrm{Var}[V_m] = \sum_{i=1}^{m} c_i^2 p_i(1 - p_i) \approx 1761$. Om een idee te krijgen van de kansverdeling, is in figuur 5.1 een histogram van 200.000 MC-replicaties van V weergegeven.

De kansmassafunctie in het histogram heeft ongeveer dezelfde vorm als de binomiale verdeling, maar is wat schever.

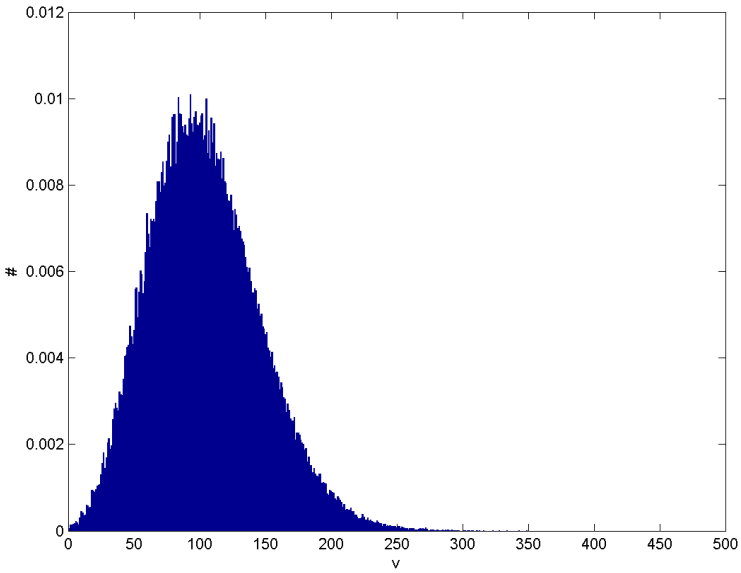
5.2 Alternatieven overgangskansen

Er is een aantal verschillende alternatieven afgeleid voor de keuze van de IS-overgangskansen.

5.2.1 Basis

In het vorige experiment, was te zien dat de ZV-benadering 1:

$$\tilde{p}_{n,v} = \frac{v}{n} = p\frac{v}{np}$$



Figuur 5.1: Histogram van 200.000 Monte Carlo-replicaties van V.

goed functioneerde. We proberen nu een keuze te maken voor de situatie met varierende schuldenaren met dezelfde structuur als de situatie met gelijke schuldenaren.

$$\begin{aligned}\tilde{p}_{n,v} &= p_n \frac{v}{\mathrm{E}\left[L_n\right]} \\ &= p_n \frac{v}{\sum_{i=1}^{n}(c_i p_i)}\end{aligned}$$

Dit is equivalent met

$$\frac{\tilde{p}_{n,v}}{1-\tilde{p}_{n,v}} = \frac{p_n v}{\sum_{i=1}^{n}(c_i p_i) - p_n v}$$

Deze keuze voor $p_{n,v}$ houdt echter geen rekening met c_n, hoewel de gekozen overgangskans groter moet zijn bij een grotere c_n. De benaderde ZV-overgangskans is namelijk:

$$\frac{p_{n,v}}{\overline{p}_{n,v}} = \frac{p\gamma_{n-1,v-c_n}}{\overline{p}\gamma_{n-1,v}},$$

hetgeen leidt tot een hogere keuze van $p_{n,v}$ bij een grotere c_n. Daarom kiezen we

$$\frac{\tilde{p}_{n,v}}{1-\tilde{p}_{n,v}} = \frac{p_n v}{\sum_{i=1}^{n}(c_i p_i) - p_n v}\frac{c_n}{\bar{c}_n},$$

zodat de overgangskans wordt geschaald aan de hand van hoe groot de schuld van schuldenaar n ten opzichte van de gemiddelde schuld van schuldenaren 1 t/m n: $\bar{c}_n = \frac{1}{n-1}\sum_{i=1}^{n-1} c_i$.

Dit leidt tot de keuze van

$$\tilde{p}_{n,v} = \frac{p_n v c_n}{p_n v(c_n - \bar{c}) + \sum_{i=1}^{n}(c_i p_i)\bar{c}}$$

5.2.2 ZV-benadering 1: binomiale benadering, gemiddelde schuldenaar

In de zero-variance benaderingen is de basis voor de keuze van de overgangskansen de vergelijking van de zero-variance overgangskansen:

$$\tilde{p}_{n,v} = p_n \frac{\gamma_{n-1,v-c_n}}{\gamma_{n,v}}.$$

Hierbij zoeken we een benadering van $\gamma_{n,v} = P(L_n \geq v)$, waarbij $L_n = \sum_{i=1}^{n} c_i Y_i$ met $Y_i \sim Ber(p_i)$.

In de eerste ZV-benadering benaderen we L_n door een binomiale verdeling L'_n, waarbij $L'_n = c'B = c' \sum_{i=1}^{n'} Y'_i$ en $Y'_i \sim Ber(p')$. Er geldt dus $B \sim Bin(n', p')$. De verdeling $L'_n = c'B$ wordt gekarakteriseerd door drie parameters: c', n' en p'.

Op het oog is de meest intuitieve keuze:

$$n' = n \tag{5.1}$$

$$c' = \frac{1}{n} \sum_{i=1}^{n} c_i \tag{5.2}$$

$$p' = \frac{1}{n} \sum_{i=1}^{n} p_i \tag{5.3}$$

Dit houdt in dat we het oorspronkelijke portfolioprobleem van n verschillende schuldenaren benaderen door een portfolioprobleem van n identieke schuldenaren met gemiddelde schuld en kans op in gebreke blijven.

5.2.3 ZV-benadering 2: binomiale benadering, gelijke variantie

Als we de centrale momenten om het gemiddelde van de binomiale benadering van de ZV-benadering 1 vergelijken met de waarden van de oorspronkelijke verdeling, valt op dat er verschillen bestaan. In principe kunnen we de parameters n', c' en p' zo kiezen dat de waarden van

	Verwachting	Tweede moment	Derde moment
V_m	$\sum_{i=1}^{m} c_i p_i$	$\sum_{i=1}^{m} c_i^2 p_i(1-p_i)$	$\sum_{i=1}^{m} c_i^3 p_i(1-p_i)(1-2p_i)$
V'_m	ncp	$nc^2p(1-p)$	$nc^3p(1-p)(1-2p)$
Waarde V_m	1.043×10^2	1.7606×10^3	3.4759×10^4
Waarde ZV-benadering 1	1.100×10^2	1.1979×10^3	1.2913×10^4
Waarde ZV-benadering 2	1.043×10^2	1.7606×10^3	2.9781×10^4

verwachting, tweede en derde moment overeenkomen met de waarden van de oorspronkelijke verdeling.

$$\begin{aligned}
\mu &= \sum_{i=1}^{n} c_i p_i && = n'c'p' \\
\mu_2 &= \sum_{i=1}^{n} c_i^2 p_i(1-p_i) && = n'c'^2 p(1-p') \\
\mu_3 &= \sum_{i=1}^{n} c_i^3 p_i(1-p_i)(1-2p_i) && = n'c'^3 p'(1-p')(1-2p')
\end{aligned}$$

Hieruit volgt de keuze voor n', c' en p' :

$$\begin{aligned}
c' &= 2\frac{\mu_2}{\mu} - \frac{\mu_3}{\mu_2} \\
p' &= \frac{\frac{\mu_2}{\mu} - \frac{\mu_3}{\mu_2}}{c'} \\
n' &= \frac{\mu}{c'p'}
\end{aligned}$$

Als we echter de waarden van μ, μ_2 en μ_3 invullen, krijgen we $n' = -36.9$, $c' = 14.1$ en $p' = -0.200$. Deze negatieve waarden voor n' en p' kunnen natuurlijk niet gebruikt worden voor de binomiale verdeling. We kunnen wel n', c' en p' kiezen zodat in ieder geval de verwachte waarde en variantie overeenkomen met de goede waarde. Aangezien n' een geheeltallig moet zijn, kiezen we n' vast. Met vaste n', worden c' en p' gegeven door:

$$c' = \frac{\mu_2}{\mu} + \frac{\mu}{n'} \tag{5.4}$$

$$p' = \frac{\mu}{c'n'} \tag{5.5}$$

Aangezien het derde moment van de binomiale verdeling dichter bij de waarde komt bij grotere n', wordt hier $n' = \sum_{i=1}^{n} c_i$ gekozen.

5.2.4 ZV-benadering 3: binomiale benadering, hybride

De ZV-benadering 2 werkt relatief goed bij variatie in gewichten. Als er geen variatie in de gewichten (meer) is, is de eerste ZV-benadering beter. Dit komt doordat bij gelijke gewichten de discretisatie van de binomiale verdeling relevant wordt. Dit kan inzichtelijk worden gemaakt door een klein voorbeeld.

Neem bijvoorbeeld $m = 10$, $p_i = 0.1$, $c_i = 25$, $x = 130$. Aangezien de kansen en gewichten gelijk zijn, kan de kans $\gamma_{10,130} = P(V_{10} \geq 130) = 1.4690 \times 10^{-4}$ exact worden bepaald door de binomiale verdeling van ZV-benadering 1 ($P(B \geq \frac{130}{25})$ met $B \sim Bin(10, 0.01)$). ZV-benadering 2 benadert de kans $\gamma_{10,130}$ echter door $P(B \geq \frac{130}{22.6}) = 9.569 \times 10^{-4}$ met $B \sim Bin(250, 0.0044)$. De tweede ZV-benadering is dus vrij slecht bij gebrek aan variatie in gewichten.

Als derde ZV-benadering wordt dus een hybride variant van de eerste en tweede ZV-benadering gekozen: de tweede ZV-benadering wordt gebruikt zolang er verschillende gewichten worden gebruikt, anders wordt ZV-benadering 1 gebruikt.

5.3 Sorteringsopties

Naast de benadering van de zero-variancekansen kan je in dit geval nog meer invloed uitoefenen op de simulatie. Je kunt namelijk ook kiezen welke schuldenaren je eerst simuleert: er kan bijvoorbeeld begonnen worden met de schuldenaren met het grootste gewicht of de kleinste kans op failliet.

5.3.1 Voorkeurstoestand voor benaderingen

De beste keuze voor de volgorde van de schuldenaren is erg afhankelijk van de gebruikte benadering. Dit wordt inzichtelijk door te kijken naar de optimale keuze van de overgangskansen:

$$\tilde{p}_{n,v} = p_n \frac{\gamma_{n-1,v-c_n}}{\gamma_{n,v}},$$

of equivalent:

$$\frac{p_{n,v}}{\overline{p}_{n,v}} = \frac{p_n \gamma_{n-1,v-c_n}}{\overline{p}_n \gamma_{n-1,v}}.$$

De verschillende benaderingen proberen zo goed mogelijk $\frac{\gamma_{n-1,v-c_n}}{\gamma_{n,v}}$ (of equivalent $\frac{\gamma_{n-1,v-c_n}}{\gamma_{n-1,v}}$) te benaderen. De schuldenaren kunnen dan zo worden gesorteerd dat deze benaderingen goed kloppen.

De eerste ZV-benadering is voornamelijk goed bij een kleine variatie in gewichten. Bij grote variatie in gewichten zijn de geschatte kansen $\gamma_{n,v}$ ver in de staart onnauwkeurig, doordat de variantie van de benaderende verdeling veel kleiner is dan de oorspronkelijke verdeling. De ratio $\frac{\gamma_{n-1,v-c_n}}{\gamma_{n-1,v}}$ wordt door de eerste ZV-benadering het beste benaderd als c_n klein is: in dit geval worden de fouten in het benaderen van $\gamma_{n-1,v-c_n}$ en $\gamma_{n-1,v}$ zoveel weet je zeker dat de ratio rond de 1 zit.

De tweede ZV-benadering is beter in het schatten van verre staartkansen, hoewel juist de benaderingen minder goed worden bij gebrek aan variatie in gewichten. De derde ZV-benadering combineert het beste van de eerste en tweede ZV-benaderingen.

Er zijn ook twee regio's die onafhankelijk van de benadering goed te schatten zijn. De eerste is bij de "diagonaal" van de (n, v) driehoek. Hierbij is $\gamma_{n-1,v} = 0$, wat betekent dat het pad richting $(n-1, v-1)$ genomen moet worden omdat het pad richting $(n-1, v)$ niet meer in het gewenste gebied kan komen. De optimale keuze voor de IS-overgangskans komt in dit geval uit op $p_{n,v} = 1$.

De tweede regio van de toestandsruimte die goed te benaderen is de grens rond $v = 0$. Als $v - c_n \leq 0$, dan geldt dat $\frac{\gamma_{n-1,v-c_n}}{\gamma_{n-1,v}} = \frac{1}{\gamma_{n-1,v}}$. De benadering van $\gamma_{n-1,v} = P\left(\sum_i^{n-1} (c_i Y_i) \geq v\right)$ is gemakkelijker te maken door te conditioneren naar de Y_i waarbij het gewicht c_i groter is dan v.

Neem bijvoorbeeld de situatie dat $c_{n-1} \geq v$:

$$\gamma_{n-1,v} = P\left(\sum_i^{n-1}(c_i Y_i) \geq v\right) \tag{5.6}$$

$$= 1 - P\left(\sum_i^{n-1}(c_i Y_i) < v\right) \tag{5.7}$$

$$= 1 - P\left(\sum_i^{n-2}(c_i Y_i) < v\right)(1-p_{n-1}) - P\left(\sum_i^{n-2}(c_i Y_i) < v - c_{n-1}\right)(p_{n-1}) \tag{5.8}$$

$$= 1 - P\left(\sum_i^{n-2}(c_i Y_i) < v\right)(1-p_{n-1}) \tag{5.9}$$

$$= 1 - \left(1 - P\left(\sum_i^{n-2}(c_i Y_i) \geq v\right)\right)(1-p_{n-1}) \tag{5.10}$$

Dit proces kan je uitvoeren zodat je uiteindelijk krijgt:

$$P\left(\sum_i (c_i Y_i) \geq v\right) = 1 - \left(1 - P\left(\sum_{i\neq j} (c_i Y_i) \geq v\right)\right)\prod_j (1-p_{j-1}),$$

waarbij j de indices aangeeft waarvoor $c_j \geq v$. $P\left(\sum_{i\neq j}(c_i Y_i)\right)$ kan dan worden benaderd door de verschillende benaderingswijzen.

5.3.2 Mogelijke volgordes schuldenaren

In dit simulatie-experiment worden 8 mogelijke volgordes van de schuldenaren gebruikt. In tabel 5.1 worden de volgordes beschreven.

Sortering	Omschrijving
$p \searrow, c \searrow$	Er wordt begonnen met de schuldenaren met een hoge p_i, waarbij bij gelijke p_i wordt begonnen met schuldenaren met een hoge c_i.
$p \nearrow, c \nearrow$	Er wordt begonnen met de schuldenaren met een lage p_i, waarbij bij gelijke p_i wordt begonnen met schuldenaren met een lage c_i.
$c \searrow, p \searrow$	Er wordt begonnen met de schuldenaren met een hoge c_i, waarbij bij gelijke c_i wordt begonnen met schuldenaren met een hoge p_i.
$c \nearrow, p \nearrow$	Er wordt begonnen met de schuldenaren met een lage c_i, waarbij bij gelijke c_i wordt begonnen met schuldenaren met een lage p_i.
$cp \searrow$	Er wordt begonnen met de schuldenaren met een hoge verwachting.
$cp \nearrow$	Er wordt begonnen met de schuldenaren met een lage verwachting.
$c^2p(1-p) \searrow$	Er wordt begonnen met de schuldenaren met een hoge variantie.
$c^2p(1-p) \nearrow$	Er wordt begonnen met de schuldenaren met een lage variantie.

Tabel 5.1

De verschillende volgordes hebben grote invloed op de paden die doorlopen worden. Als begonnen wordt met de kleine p's zullen de paden richting de diagonaal gaan. Bij grote p's en grote c's worden de paden juist richting de "$v = 0$"-lijn gericht.

De voorkeursvolgorde per benadering is niet meteen duidelijk te benoemen. Er spelen meerdere factoren een rol waarvan de invloed niet direct te kwantificeren zijn:

1. Hoe goed is de benadering precies in de toestanden van de doorlopen paden?
2. Is het voordelig om de paden door een slecht benaderbare regio van de toestandsruimte te sturen naar de regio's waar de overgangskansen optimaal bepaald kunnen worden, of kan je beter langer in een redelijk goed benaderbare regio blijven?

5.4 Resultaten

Deze sectie is gesplitst in twee delen: eerst worden de resultaten vergeleken met een gelijk aantal replicaties per benaderingsmethode. In het tweede deel wordt de rekentijd per benaderingsmethode en sorteringsvolgorde meegenomen.

5.4.1 Resultaten met gelijk aantal replicaties

In totaal zijn er 4 situaties met verschillende begintoestand gesimuleerd, waarbij steeds zeldzamere kansen worden geschat. Per case zijn de 4 verschillende benaderingswijzen gesimuleerd met de 8 verschillende sorteringswijzen.

Schatting $P(L \geq 272)$

Allereerst wordt ter controle de waarden van $t = \frac{\hat{\gamma}-\gamma}{s/\sqrt{n}}$ weergegeven in tabel 5.2 waarmee een t-toets uitgevoerd kan worden. Als waarde voor γ wordt de schatting met de kleinste relatieve standaardfout genomen: $\gamma = 4.94 \times 10^{-4}$.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	-0.07	-2.35	0.69	1.09
$p \nearrow, c \nearrow$	0.24	-9.75	1.62	0.83
$c \searrow, p \searrow$	0.25	0.15	0.01	-0.00
$c \nearrow, p \nearrow$	0.82	1.27	-1.99	0.62
$cp \searrow$	0.09	-3.02	-0.28	-1.70
$cp \nearrow$	-0.11	0.67	0.53	1.84
$c^2p(1-p) \searrow$	0.76	-1.20	0.18	0.08
$c^2p(1-p) \nearrow$	0.60	-0.22	-1.76	0.78

Tabel 5.2: t-waarden van de schattingen van $P(L_m \geq 272)$ (4 standaarddeviaties van $\mathrm{E}[L]$) voor de vier keuzes van overgangskansen en acht sorteringsopties met $N = 500$ replicaties.

In tabel 5.3 zijn de relatieve standaardfouten van de schattingen weergeven. ZV-benadering 1 geeft een uitschieter als t-waarde wanneer eerst wordt begonnen met de schuldenaren met kleine kans op niet-terugbetaling. Deze schatting is dus onbetrouwbaar.

Voor de basiskeuze en ZV-benadering 1 is het voordeling om te beginnen met kleine c_i. Voor ZV-benadering 2 is het juist voordelig om te beginnen met grote c_i. Ditzelfde geldt voor ZV-benadering 3, hoewel daar de verschillen tussen de sorteringsvolgorde kleiner zijn.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	0.1862	0.1087	0.0258	0.0268
$p \nearrow, c \nearrow$	0.2140	*	0.0456	0.0218
$c \searrow, p \searrow$	0.0993	0.1617	0.0163	0.0145
$c \nearrow, p \nearrow$	0.0591	0.0252	0.0317	0.0238
$cp \searrow$	0.1686	0.0830	0.0362	0.0178
$cp \nearrow$	0.0540	0.0275	0.1127	0.0264
$c^2p(1-p) \searrow$	0.0972	0.0961	0.0173	0.0201
$c^2p(1-p) \nearrow$	0.0514	0.0240	0.0339	0.0255

Tabel 5.3: Relatieve standaardfout $\frac{s}{\bar{x}\sqrt{N}}$ van de schatting van $P(L \geq 272)$ voor de vier keuzes van overgangskansen en acht sorteringsopties met $N = 500$ replicaties.

Schatting $P(L \geq 356)$

Allereerst wordt ter controle de waarden van $t = \frac{\hat{\gamma}-\gamma}{s/\sqrt{n}}$ weergegeven in tabel 5.4 waarmee een t-toets uitgevoerd kan worden. Als waarde voor γ wordt de schatting met de kleinste relatieve standaardfout genomen: $\gamma = 2.18 \times 10^{-6}$.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	-0.74	-0.37	-0.68	-0.81
$p \nearrow, c \nearrow$	-0.15	-32.82	-0.35	-0.34
$c \searrow, p \searrow$	-0.43	0.07	-1.18	-0.00
$c \nearrow, p \nearrow$	-0.57	-1.36	-2.16	-0.24
$cp \searrow$	-0.80	-1.83	-1.87	-1.33
$cp \nearrow$	0.00	0.45	0.11	-1.61
$c^2p(1-p) \searrow$	0.18	0.86	-0.87	-0.94
$c^2p(1-p) \nearrow$	-0.06	-2.00	0.46	-0.12

Tabel 5.4: t-waarden van de schattingen van $P(L_m \geq 356)$ (6 standaarddeviaties van $\mathrm{E}[L]$) voor de vier keuzes van overgangskansen en acht sorteringsopties met $N = 500$ replicaties.

In tabel 5.5 zijn de relatieve standaardfouten van de schattingen weergeven. ZV-benadering 1 geeft een uitschieter als t-waarde wanneer eerst wordt begonnen met de schuldenaren met kleine kans op niet-terugbetaling. Deze schatting is dus onbetrouwbaar.

Ook hier geldt dat de performance van de benaderingen erg afhankelijk is van de sorteringen. De hybride ZV-benadering 3 combineert het beste van de andere twee ZV-benaderingen.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	0.1698	0.5434	0.0346	0.0348
$p \nearrow, c \nearrow$	0.2525	*	0.0322	0.0354
$c \searrow, p \searrow$	0.0939	0.1986	0.0238	0.0197
$c \nearrow, p \nearrow$	0.0492	0.0334	0.0599	0.0372
$cp \searrow$	0.1163	0.1910	0.0442	0.0364
$cp \nearrow$	0.0563	0.0399	0.1456	0.0349
$c^2p(1-p) \searrow$	0.1278	0.2945	0.0395	0.0232
$c^2p(1-p) \nearrow$	0.0530	0.0351	0.2556	0.0333

Tabel 5.5: Relatieve standaardfout $\frac{s}{\bar{x}\sqrt{N}}$ van de schatting van $P(L \geq 356)$ voor de vier keuzes van overgangskansen en acht sorteringsopties met $N = 500$ replicaties.

Schatting $P(L \geq 440)$

Allereerst wordt ter controle de waarden van $t = \frac{\hat{\gamma}-\gamma}{s/\sqrt{n}}$ weergegeven in tabel 5.6 waarmee een t-toets uitgevoerd kan worden. Als waarde voor γ wordt de schatting met de kleinste relatieve standaardfout genomen: $\gamma = 3.51 \times 10^{-9}$.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	-0.70	-1.93	-0.37	1.12
$p \nearrow, c \nearrow$	-0.57	-47.94	-2.47	0.35
$c \searrow, p \searrow$	0.49	-2.33	-0.42	-0.00
$c \nearrow, p \nearrow$	0.85	2.03	-0.72	1.96
$cp \searrow$	0.11	-8.01	0.30	-1.62
$cp \nearrow$	0.13	1.02	-1.13	0.13
$c^2p(1-p) \searrow$	0.24	-3.67	-0.99	0.05
$c^2p(1-p) \nearrow$	0.51	-0.32	-0.85	-0.04

Tabel 5.6: t-waarden van de schattingen van $P(L_m \geq 440)$ (8 standaarddeviaties van $\mathrm{E}[L]$) voor de vier keuzes van overgangskansen en acht sorteringsopties met $N = 500$ replicaties.

In tabel 5.7 zijn de relatieve standaardfouten van de schattingen weergeven. ZV-benadering 1 geeft een uitschieter als t-waarde wanneer eerst wordt begonnen met de schuldenaren met kleine kans op niet-terugbetaling en als er wordt begonnen met schuldenaren bij een hoog risico. Deze schatting is dus onbetrouwbaar.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	0.3042	0.5754	0.0454	0.0514
$p \nearrow, c \nearrow$	0.2711	*	0.0353	0.0457
$c \searrow, p \searrow$	0.1051	0.2896	0.0342	0.0343
$c \nearrow, p \nearrow$	0.0696	0.0514	0.1116	0.0526
$cp \searrow$	0.2505	*	0.0699	0.0477
$cp \nearrow$	0.0720	0.0562	0.0758	0.0485
$c^2p(1-p) \searrow$	0.1326	0.3093	0.0559	0.0467
$c^2p(1-p) \nearrow$	0.0674	0.0462	0.0546	0.0450

Tabel 5.7: Relatieve standaardfout $\frac{s}{\bar{x}\sqrt{N}}$ van de schatting van $P(L \geq 440)$ voor de vier keuzes van overgangskansen en acht sorteringsopties met $N = 500$ replicaties.

Schatting $P(L \geq 524)$

Allereerst wordt ter controle de waarden van $t = \frac{\hat{\gamma}-\gamma}{s/\sqrt{n}}$ weergegeven in tabel 5.6 waarmee een t-toets uitgevoerd kan worden. Als waarde voor γ wordt de schatting met de kleinste relatieve standaardfout genomen: $\gamma = 2.56 \times 10^{-12}$.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	-0.82	-1.03×10^{2}	1.11	1.79
$p \nearrow, c \nearrow$	0.24	-9.01×10^{30}	2.36	-0.70
$c \searrow, p \searrow$	0.77	-5.89×10^{2}	0.88	0.02
$c \nearrow, p \nearrow$	0.21	-1.89×10^{54}	-2.87	0.63
$cp \searrow$	-1.60	-1.63×10^{2}	-3.21	0.31
$cp \nearrow$	0.62	-8.14×10^{29}	-0.33	0.82
$c^2p(1-p) \searrow$	0.27	-3.15×10^{2}	-1.13	-0.08
$c^2p(1-p) \nearrow$	0.86	-1.99×10^{43}	0.46	-0.00

Tabel 5.8: t-waarden van de schattingen van $P(L_m \geq 524)$ (10 standaarddeviaties van $\mathrm{E}[L]$) voor de vier keuzes van overgangskansen en acht sorteringsopties met $N = 500$ replicaties.

In tabel 5.9 zijn de relatieve standaardfouten van de schattingen weergeven. Het valt op dat de uitkomsten van ZV-benadering 1 erg afwijken van de rest. Dit komt waarschijnlijk doordat MATLAB de staartkansen benadert door $p_{n,v} = 0$ als de staartkansen heel klein zijn. Hierdoor wordt de voorwaarde dat paden die leiden tot het gewenste gebied niet mogen worden afgesloten overtreden. Dit leidt tot onzuivere schattingen.

Ook hier presteert ZV-benadering 3 het beste, hoewel het verschil met de basiskeuze minder groot is dan minder ver in de staart.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	0.3555	*	0.1172	0.1213
$p \nearrow, c \nearrow$	0.8098	*	0.0797	0.0517
$c \searrow, p \searrow$	0.1610	*	0.0566	0.0600
$c \nearrow, p \nearrow$	0.0831	*	0.0767	0.0567
$cp \searrow$	0.1513	*	0.0812	0.1490
$cp \nearrow$	0.1191	*	0.1075	0.0490
$c^2p(1-p) \searrow$	0.1384	*	0.0621	0.0990
$c^2p(1-p) \nearrow$	0.1031	*	0.2500	0.0485

Tabel 5.9: Relatieve standaardfout $\frac{s}{\bar{x}\sqrt{N}}$ van de schatting van $P(L \geq 524)$ voor de vier keuzes van overgangskansen en acht sorteringsopties met $N = 500$ replicaties.

Wat opvalt bij de resultaten is dat de schattingen onnauwkeuriger worden des te verder in de staart gezocht wordt. Dit is in tegenstelling tot het vorige simulatie-experiment. Waarschijnlijk zijn de benaderingen in dit simulatie-experiment slechter hoe verder in de staart gezocht moet worden.

5.4.2 Rekentijden

De bovenstaande relatieve standaardfouten kunnen met elkaar vergeleken worden bij hetzelfde aantal replicaties N, maar eigenlijk is het eerlijker om de verschillende sorteringen en benaderingen te vergelijken bij dezelfde rekentijd. De ZV-benaderingen zijn namelijk veel complexer en daardoor langzamer uit te rekenen dan de basiskeuze. Bovendien zijn sommige sorteringen sneller, omdat ze de paden sneller richting de diagonaal en nulgrens sturen waar de ZV-benaderingen gemakkelijk zijn uit te rekenen.

De bovenstaande relatieve standaardfouten zijn gegeven door $c_{rs} = \frac{s}{\bar{x}\sqrt{500}}$. De relatieve standaardfouten voor andere N kunnen worden uitgerekend door $RS(N) = \sqrt{500} \cdot \frac{c_{vr}}{\sqrt{N}}$. We nemen aan dat de rekentijd lineair is aan het aantal replicaties: $T(N) = c_{ct}N$. Dan is de relatieve standaardfout als functie van de rekentijd:

$$RS(T) = \frac{\sqrt{500} \cdot c_{rs}}{\sqrt{\frac{T}{c_{ct}}}} = \frac{\sqrt{500} \cdot c_{rs} \cdot \sqrt{c_{ct}}}{\sqrt{T}}$$

De constante $\sqrt{500} \cdot c_{rs} \cdot \sqrt{c_{ct}}$ geeft dus aan hoe efficiënt de schatter is, waarbij een lagere waarde beter is. In de onderstaande tabel worden de waarden van deze constante in alle gevallen gerapporteerd.

Uit de tabellen blijkt dat qua efficiëntie de basiskeuze goed presteert ten opzichte van de ZV-benaderingen. Voor de kansen minder ver in de staart komen de waarden van de basiskeuze en de derde ZV-benadering redelijk overeen, maar op 10 σ afstand presteert de basiskeuze veel beter. Dit wordt veroorzaakt doordat een replicatie met de derde ZV-benadering ongeveer $10\times$ zo lang duurt als een replicatie met de basiskeuze. Daardoor geldt dat de extra rekentijd van de ZV-benadering niet opweegt tegen de lagere variantie van die schatter ten opzichte van de gemakkelijkere basiskeuze.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	1.5942	2.7415	0.6791	0.6923
$p \nearrow, c \nearrow$	1.7836	3.2395	1.4158	0.6756
$c \searrow, p \searrow$	0.8261	3.8085	0.3863	0.3480
$c \nearrow, p \nearrow$	0.4952	0.7621	0.9379	0.7169
$cp \searrow$	1.4051	2.1793	0.8203	0.4082
$cp \nearrow$	0.4498	0.8886	3.5253	0.8358
$c^2p(1-p) \searrow$	0.8080	2.5507	0.3823	0.4525
$c^2p(1-p) \nearrow$	0.4302	0.7578	1.0343	0.7822

Tabel 5.10: Efficiëntie $\sqrt{500} \cdot c_{rs} \cdot \sqrt{c_{ct}}$ van de schatting van $P(L \geq 272)$ voor de vier keuzes van overgangskansen en acht sorteringsopties.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	1.4182	14.0575	0.9346	0.9348
$p \nearrow, c \nearrow$	2.1152	5.1078	1.0162	1.1051
$c \searrow, p \searrow$	0.7837	4.8942	0.5771	0.4784
$c \nearrow, p \nearrow$	0.4124	1.0225	1.7804	1.1105
$cp \searrow$	0.9690	5.2711	1.0172	0.8390
$cp \nearrow$	0.4674	1.3004	4.5638	1.0973
$c^2p(1-p) \searrow$	1.0665	7.7334	0.9049	0.5335
$c^2p(1-p) \nearrow$	0.4473	1.1133	7.8846	1.0215

Tabel 5.11: Efficiëntie $\sqrt{500} \cdot c_{rs} \cdot \sqrt{c_{ct}}$ van de schatting van $P(L \geq 356)$ voor de vier keuzes van overgangskansen en acht sorteringsopties.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	2.5423	14.1590	1.2228	1.3996
$p \nearrow, c \nearrow$	2.2575	10.2506	1.1156	1.4479
$c \searrow, p \searrow$	0.8741	7.2142	0.8565	0.8563
$c \nearrow, p \nearrow$	0.5813	1.5998	3.3976	1.6074
$cp \searrow$	2.0893	5.4590	1.6366	1.1275
$cp \nearrow$	0.5975	1.8359	2.3843	1.5370
$c^2p(1-p) \searrow$	1.1026	8.4927	1.3040	1.0966
$c^2p(1-p) \nearrow$	0.5676	1.4964	1.7049	1.4176

Tabel 5.12: Efficiëntie $\sqrt{500} \cdot c_{rs} \cdot \sqrt{c_{ct}}$ van de schatting van $P(L \geq 440)$ voor de vier keuzes van overgangskansen en acht sorteringsopties.

Sortering	Basis	ZV-1	ZV-2	ZV-3
$p \searrow, c \searrow$	2.9753	10.5253	3.2333	3.3567
$p \nearrow, c \nearrow$	6.7110	10.6455	2.5032	1.6595
$c \searrow, p \searrow$	1.3423	9.1594	1.4084	1.4925
$c \nearrow, p \nearrow$	0.6929	1.7927	2.3034	1.7084
$cp \searrow$	1.2618	15.5319	1.9237	3.5195
$cp \nearrow$	0.9919	2.6311	3.3840	1.5397
$c^2p(1-p) \searrow$	1.1532	12.2411	1.4614	2.2931
$c^2p(1-p) \nearrow$	0.8666	1.6793	7.8034	1.4975

Tabel 5.13: Efficiëntie $\sqrt{500} \cdot c_{rs} \cdot \sqrt{c_{ct}}$ van de schatting van $P(L \geq 524)$ voor de vier keuzes van overgangskansen en acht sorteringsopties.

Hoofdstuk 6

Conclusies en verder onderzoek

6.1 Conclusies en discussie

De hoofdvraag van dit onderzoek is:

Hoe goed werkt importance sampling in een credit risk model met onafhankelijke schuldenaren?

De gebruikte importance sampling technieken werkten zeer goed in vergelijking met de standaard Monte Carlo techniek. Veel minder replicaties waren nodig om tot een redelijke schatting te komen van de kans op een groot verlies voor de bank.

Voor het geval van gelijke schuldenaren in termen van kans op failliet en uitgeleend bedrag leiden de benadering van de zero-variance benaderingen tot goede resultaten. Slechts enkele replicaties (< 30) zijn nodig om de schatting met een 10% relatieve standaardfout te schatten, afhankelijk van het aantal schuldenaren en het gezochte verlies.

In het geval van ongelijke schuldenaren was het moeilijker een goede benadering te vinden die de gezochte kans efficiënt wist te schatten. In tegenstelling tot wat van te voren verwacht was, was voor een hoge totale exposure x de simpele basiskeuze efficiënter dan de complexere zero-variancebenaderingen. Dit wordt waarschijnlijk veroorzaakt door inefficiënte implementatie van de binomiale verdeling.

6.2 Verder onderzoek

1. **Optimaliseren implementatie.** De implementatie van de Zero Variance-benaderingen is niet optimaal. Er zit waarschijnlijk veel rekentijd in de berekening van de staartkansen van de binomiale verdeling die geen significante verschillen in de benadering opleveren. Een verdere analyse van hoe de benaderingen het beste geimplementeerd kunnen worden zal de performance van de ZV-benaderingen verbeteren.

2. **Gebruik andere benaderingsmethoden, bijvoorbeeld via normale verdeling of Poissonverdeling.** In dit onderzoek is gefocust op het benaderen van de ZV-verdeling door middel van een binomiale verdeling. Andere verdelingen zoals de normale verdeling of een Poissonverdeling zullen mogelijk sneller en/of beter zijn.

3. **Uitbreiding naar afhankelijke schuldenaren.** Het credit risk model met onafhankelijke schuldenaren, zoals in dit onderzoek bekeken, is maatschappelijk niet zo relevant. Een uitbreiding naar afhankelijke schuldenaren zal veel complexer, maar ook nuttiger zijn.

4. **Adaptieve methoden.** Elke replicatie van de importance sampling schatter levert niet alleen een schatting van $P(L_m \geq x)$ op, maar ook schattingen van $P(L_{m_1} \geq x - c_n)$ en andere toestanden die op het pad van de replicatie ligt. Deze schattingen kunnen gebruikt worden in de volgende replicaties voor de ZV-benaderingen. Dit leidt tot een zogenaamde adaptieve schatter.

Hoofdstuk 7

Bibliografie

[1] Rice, John: *Mathematical statistics and data analysis.* Cengage Learning, 2006.

[2] Rubino, Gerardo en Bruno Tuffin: *Rare event simulation using Monte Carlo methods.* John Wiley & Sons, 2009.

[3] Awad, Hernan P, Peter W Glynn, en Reuven Y Rubinstein: *Zero-variance importance sampling estimators for markov process expectations.* Mathematics of Operations Research, 38(2):358–388, 2013.

[4] Glasserman, Paul: *Stochastic monotonicity and conditional Monte Carlo for likelihood ratios.* Advances in applied probability, pagina's 103–115, 1993.

[5] Glasserman, Paul en Jingyi Li: *Importance sampling for portfolio credit risk.* Management science, 51(11):1643–1656, 2005.