Afstanden tussen kansmaten

Westhoff, G.G.A.

G.G.A. Westhoff

Afstanden tussen kansmaten

Bachelorscriptie, 18 augustus 2008

Scriptiebegeleider: Dr. O.W. van Gaans

Mathematisch Instituut, Universiteit Leiden

Inhoudsopgave

1 Inleiding

Convergentie in verdeling speelt in de kansrekening een belangrijke rol, bijvoorbeeld in de centrale limietstelling. Deze convergentie kan beschreven worden door aan de ruimte met kansmaten een metriek toe te voegen. Bekende metrieken zijn de *bounded Lipschitz metriek*, *de Prokhorov metriek* en de *Wasserstein metriek*.

De definities van deze metrieken zijn zeer abstract en we vragen ons in dit onderzoek af, of het mogelijk is om voor bepaalde stochasten de afstand tussen hun verdelingen uit te rekenen of op zijn minst goed af te schatten.

Verder onderzoeken we of er in bepaalde gevallen in de Centrale Limiet Stelling (CLS) een snelheid van convergentie kan worden aangegeven ten opzichte van één van de genoemde metrieken.

De scriptie is als volgt ingedeeld. Eerst introduceren we een aantal begrippen en eigenschappen. Vervolgens berekenen we de afstanden tussen enkele concrete stochasten—stochasten met dichtheden en stochasten die maar eindig veel waarden in $\mathbb{R}$ aannemen. Op die manier krijgen we wat ervaring met afschattingstechnieken, die ons van pas komen in het laatste en belangrijkste onderdeel: het genoemde onderzoek aan de convergentiesnelheid in het kader van de CLS.

Bij het schrijven van Hoofdstuk 2 heb ik vooral de seminarnotities van Dr. O. van Gaans geraadpleegd. De stof in het slot van sectie 2.1 is o.a. te vinden in de secties (6.12) en (17.1) van het boek van David Williams.

2 De ruimte van kansmaten op $\mathbb{R}$

2.1 Enkele begrippen en eigenschappen

Zij (S, d) een metrische ruimte. De *Borel σ-algebra* $\mathcal{B}(S)$ is de kleinste σ-algebra die alle open deelverzamelingen van S bevat. De elementen van $\mathcal{B}(S)$ heten *Borel-deelverzamelingen* van S.

Een eindige Borelmaat op S is de afbeelding $\mu\colon \mathcal{B}(S) \to [0, \infty)$, zodanig dat

$$\begin{array}{l} \mu(\emptyset) = 0, \text{ en} \\ A_1, A_2, \ldots \in \mathcal{B}(S) \text{ wederzijds disjunct } \implies \mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(B_i). \end{array}$$

μ heet *Borelkansmaat*, kortweg: kansmaat als tevens $\mu(S) = 1$. Neem van nu af aan: $S = \mathbb{R}$. Aangetoond kan worden dat de Borel σ -algebra $\mathcal{B}(\mathbb{R})$ wordt voortgebracht door de collectie $\{\, (-\infty, x] \,:\, x \in \mathbb{R} \,\}$; alle huis-tuin-en-keuken-deelverzamelingen van $\mathbb{R}$ zijn Boreldeelverzamelingen van $\mathbb{R}$, kort: zijn Borel.

We definiëren de ruimte van kansmaten op $\mathbb{R}$:

$$\mathcal{P}(\mathbb{R}) := \{ \text{ alle (Borel-)kansmaten } \mu \text{ op } \mathbb{R} \,\}.$$

Definieer:

$$C_b(\mathbb{R}) := \{\, f\colon \mathbb{R} \to \mathbb{R} : f \text{ is continu en begrensd} \,\}.$$

Elke $f \in C_b(\mathbb{R})$ is integreerbaar met betrekking tot elke kansmaat op $\mathbb{R}$.

Zij $\mu, \mu_1, \mu_2, \ldots\ldots \in \mathcal{P}(\mathbb{R})$. We zeggen dat de rij (μ_i) *zwak* (of ook *narrow*) convergeert naar μ, notatie $\mu_i \rightsquigarrow \mu$, als:

$$\int_{\mathbb{R}} f \, \mathrm{d}\mu_i \to \int_{\mathbb{R}} f \, \mathrm{d}\mu \text{ als } i \to \infty \text{ voor alle } f \in C_b(\mathbb{R}).$$

In dat geval heet zwakke convergentie ook *convergentie in verdeling* en elke kansmaat μ kunnen we identificeren met een verdelingsfunctie F

$$\mu \leftrightarrow F, \text{ met } F(x) = \mu(-\infty, x].$$

Aangetoond kan worden (literatuur) dat convergentie in verdeling zwakke convergentie impliceert en omgekeerd:

$$F_n \to F \iff \mu_n \rightsquigarrow \mu.$$

We zijn geinteresseerd in het geval dat $F = F_X$, dat wil zeggen, als F de verdelingsfunctie is van zekere reële stochast X. We noteren de bijbehorende kansmaat met μ_X. Tenslotte is er nog de *law* van de stochast X:

$$L_X := \mathbb{P} \circ X^{-1} \text{ , } L_X \colon \mathcal{B} \to [0,1]$$

Het verband tussen de diverse grootheden is nu als volgt:

$$\mu_X(-\infty, x] = L_X(-\infty, x] = F_X(x) = \mathbb{P}(X \leq x)$$

We noteren voor willekeurige Borelverzameling $A \in \mathcal{B}$

$$\mathbb{P}_X(A) := \mathbb{P}(X \in A) =: \mu_X(A).$$

Soms willen we een integraal uitdrukken als verwachting of als een integraal van dichtheden. Laat X stochast in $\mathbb{R}$ op $(\Omega, \Sigma, \mathbb{P})$. Dan is:

$$\int_{\mathbb{R}} f \, \mathrm{d}\mu_X = \mathbb{E}f(X). \tag{1}$$

Als X een dichtheidsfunctie h_X heeft, dan is ook:

$$\int_{\mathbb{R}} f \, \mathrm{d}\mu_X = \mathbb{E}f(X) = \int_{\mathbb{R}} h_X(x) f(x) dx, \tag{2}$$

met integratie naar x in de zin van Lebesgue.

2.2 Afstanden meten in de ruimte van kansmaten: de Prokhorov- en bounded-Lipschitzmetriek

Beide metrieken worden algemeen voor separabele metrische ruimten (S, d) gedefinieerd, met d de metriek op S. We beperken ons weer to $S = \mathbb{R}$.

2.2.1 De Prokhorovmetriek

Definieer voor $\mu, \nu \in \mathcal{P}(\mathbb{R})$:

$$d_P(\mu,\nu) := \inf\{\alpha > 0 : \mu(A) \leq \nu(A_\alpha) + \alpha \text{ and } \nu(A) \leq \mu(A_\alpha) + \alpha \ \forall A \in \mathcal{B}(\mathbb{R})\}, \tag{3}$$

waarin:

$$A_\alpha := \{\, x : |x - A| < \alpha \,\} \text{ als } A \neq \emptyset, \quad \emptyset_\alpha := \emptyset \text{ voor alle } \alpha > 0.$$

(Hierin is $|x - A| = \inf\{|x - a| : a \in A\}$.) Aangetoond kan worden dat d_P een metriek op $\mathcal{P}(\mathbb{R})$ definieert, de *Prokhorovmetriek*. Voor willekeurige $\mu, \mu_1, \mu_2, \ldots \in \mathcal{P}(\mathbb{R})$ kan aangetoond worden dat:

$$\mu_i \rightsquigarrow \mu \iff d_P(\mu_i, \mu) \to 0.$$

Dus convergentie in de Prokhorovmetriek induceert zwakke convergentie en omgekeerd.

2.2.2 De bounded-Lipschitzmetriek

De bounded-Lipschitzmetriek (ook wel *Dudley-metriek* genaamd) is iets makkelijker om mee te werken dan de Prokhorovmetriek. Noteer:

$$\mathrm{BL}(\mathbb{R}) := \{f\colon \mathbb{R} \to \mathbb{R} :\ f \text{ is begrensd en Lipschitz}\}. \tag{4}$$

Definieer voor $f \in \mathrm{BL}(\mathbb{R})$

$$\|f\|_{\mathrm{BL}} = \|f\|_\infty + \mathrm{Lip}(f),$$

waarin:

$$\|f\|_\infty := \sup_{x \in \mathbb{R}} |f(x)|,$$

en:

$$\mathrm{Lip}(f) := \sup_{x,y \in \mathbb{R},\ x \neq y} \frac{|f(x) - f(y)|}{|(x - y|} = \inf\{L\colon |f(x) - f(y)| \leq L|x - y| \ \forall x, y \in \mathbb{R}\}.$$

Dan is $\|\cdot\|_{\mathrm{BL}}$ een norm op $\mathrm{BL}(\mathbb{R})$. Definieer voor $\mu, \nu \in \mathcal{P}(\mathbb{R})$:

$$d_{\mathrm{BL}}(\mu,\nu) := \sup\left\{ \left| \int_{\mathbb{R}} f \,\mathrm{d}\mu - \int_{\mathbb{R}} f \,\mathrm{d}\nu \right| :\ f \in \mathrm{BL}(\mathbb{R}),\ \|f\|_{\mathrm{BL}} \leq 1 \right\}. \tag{5}$$

De functie d_{BL} heet de *bounded Lipschitz metriek op* $\mathcal{P}$. Net als bij de Prokhorovmetriek geldt dat voor willekeurige $\mu, \mu_1, \mu_2, \ldots \in \mathcal{P}(\mathbb{R})$:

$$\mu_i \rightsquigarrow \mu \iff d_{\mathrm{BL}}(\mu_i, \mu) \to 0.$$

2.2.3 Alternatieve notaties voor $d_{P,\mathrm{BL}}(\mu_X, \mu_Y)$

Als $X \sim verdeling1$ en $Y \sim verdeling2$ stochasten zijn, dan zijn de bijbehorende kansmaten μ_X resp. μ_Y. De afstanden tussen beide maten moeten we nu eigenlijk correct noteren met $d_P(\mu_X, \mu_Y)$ (evenzo voor d_{BL}). Maar we wijken daar regelmatig van af en schrijven bijvoorbeeld:

$$d_P(\mu_X, \mu_Y) = d_P(X, Y) = d_P(X, Verdeling2) = d_P(Verdeling1, Verdeling2).$$

3 De verdelingen die een rol spelen in dit verslag

We bespreken de stochasten die de "maat" wordt genomen in dit verslag. Het is een selectie van discrete en continue toevalsvariabelen, zó gekozen, dat er relatief makkelijk mee te rekenen valt.

3.1 De tentverdelingen

De dichtheidsfunctie heeft de vorm van een tent. Als de stochast X verdeeld is als $tent(0)$, $X \sim tent(0)$, dan heeft X dichtheid $h(x)$:

$$h(x) = \begin{cases} x+1 & \text{op } [-1,0]; \\ -x+1 & \text{op } (0,1]; \\ 0 & \text{buiten het interval } [-1,1]. \end{cases} \tag{6}$$

$tent(m)$ is de verdeling $tent(0)$, over een interval $m \in \mathbb{R}$ verschoven.

Als $X \sim tent(m)$ dan:

$$\begin{aligned} \mathbb{E}X &= m \\ VarX &= \tfrac{1}{6} \quad \text{(elementaire calculus)} \end{aligned} \tag{7}$$

De verdeling heeft een compacte drager, dus het gebied waar $h \neq 0$, is gesloten en begrensd. Samen met het lineaire verloop geeft dit prettige integratie-eigenschappen.

3.2 De Bernouilliverdelingen

Geen nadere introductie nodig. Ik bespreek slechts notatiekwesties. Als $X \sim Bern(a, b; p)$ dan is:

$$\begin{aligned} p_X(x) &:= \mathbb{P}(X = x) \\ p_X(b) &= p \\ p_X(a) &= 1 - p \\ \mathbb{E}X &= pb + (1-p)a \end{aligned} \tag{8}$$

$Bern(0, 1; p)$ is de gebruikelijke Bernouilli-verdeling. In het verslag is altijd $p = 1/2$.

3.3 De binomiale verdelingen

De binomiale verdeling $X \sim bin(n, p)$ heeft geen nadere uitleg nodig. Maar ik introduceer hier de verdeling $bin2$ die we nodig hebben in sectie 6. Als boven is $p_X(x) := \mathbb{P}(X = x)$.

$X \sim bin2(n,p)$ is een toevalsvariabele die de gewichten/massa van de gewone binomiale verdeling $Y \sim bin(n,p)$ verdeelt, niet over de gebruikelijke uitkomsten $k = 0, 1, 2, \ldots, n$, maar over de waarden:

$$k = -\sqrt{n}, \frac{2-n}{\sqrt{n}}, \ldots, \frac{-1}{\sqrt{n}}, \frac{-1}{\sqrt{n}}, \ldots, \frac{n-2}{\sqrt{n}}, \sqrt{n}$$

$$p_X\left(\frac{2m-n}{\sqrt{n}}\right) = p_Y(m) = \binom{m}{n} p^m (1-p)^{n-m}, \text{ met } m = 0, 1, 2, \ldots, n. \tag{9}$$

In dit verslag is altijd $p = 1/2$.

3.4 De standaardnormale verdeling

Als X standaard-normaal verdeeld is, dan noteren we dit — zoals gebruikelijk — als $X \sim N(0,1)$. De dichtheid is:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

4 Universele boven- en ondergrenzen van de Prokhorovafstand d_P en de bounded Lipschitz metriek d_{BL}

4.1 De kunst van het afschatten

Het is waarschijnlijk onmogelijk om voor willekeurige kansmaten/verdelingen een compacte uitdrukking te vinden voor de afstand. We kunnen alleen proberen het interval voor de afstand zo klein mogelijk te maken, dwz, onder- en bovengrens streng mogelijk af te schatten. Hier volgen wat recepten.

4.1.1 Begrenzing van de Prokhorovafstand d_P

Zij $\mu, \nu \in \mathcal{P}(\mathbb{R})$. Bekijk de definitie van $d_P(\mu, \nu)$.

- Om een ondergrens te bepalen voor d_P, zoekt men naar een $\alpha > 0$ waarvoor een Borel A is te vinden zodat aan tenminste één van de ongelijkheden NIET is voldaan.
- Als een $\alpha < 1$ bestaat zodat aan beide ongelijkheden is voldaan voor alle Borel A, dan is dit een (nieuwe) bovengrens.
- Beide begrenzingen zijn specifiek voor de verdelingen; voor ieder tweetal verdelingen μ en ν moet opnieuw worden afgeschat.

d_P is een afstandsfunctie, dus 0 is de kleinste waarde die kan worden aaangenomen. Bekijk het stelsel ongelijkheden in de definitie van d_P:

$$\begin{array}{lrcl} \text{(I)} & \mu(A) & \leq & \nu(A_\alpha) + \alpha \\ \text{(II)} & \nu(A) & \leq & \mu(A_\alpha) + \alpha. \end{array} \tag{10}$$

Voor $\alpha \geq 1$ wordt aan beide betrekkingen voldaan, immers ν en μ zijn kansmaten, en zijn dus maximaal 1. Dus $d_P \leq 1$. Voor $\alpha < 1$ en bijvoorbeeld $\mu(A) = 1$, is het noodzakelijk

dat in (I) $\nu(A_\alpha) > 0$ dus dan worden er "eisen" gesteld aan Borel A en ν en hoeft deze α geen bovengrens meer te zijn. Dus algemeen geldt voor $\mu, \nu \in \mathcal{P}$

$$0 \leq d_P(\mu, \nu) \leq 1 \tag{11}$$

Een erg breed interval, maar we hebben nu een eerste kwantitatieve uitspraak; de Prokhorovmetriek laat zich temmen!

4.1.2 Begrenzing van de bounded-Lipschitzafstand d_{BL}

- Ondergrens: door het verschil $|\int_{\mathbb{R}} f \,\mathrm{d}\mu - \int_{\mathbb{R}} f \,\mathrm{d}\nu|$ te bepalen voor een functie f die aan de metriek-voorwaarden voldoet, hebben we een ondergrens voor d_{BL}; dit getal is immers het supremum van de verschillen over alle functies f.
- Bovengrens bepalen: geen simpel afschattingsrecept.
- Boven- en ondergrens zijn specifiek voor de verdelingen; voor ieder tweetal verdelingen μ en ν moet opnieuw worden afgeschat.

De ondergrens van de afstandsfunctie d_{BL} is weer nul. Een universele bovengrens vinden we als volgt. Bekijk het verschil van de twee integralen in (3):

$$\begin{aligned}\left|\int_{\mathbb{R}} f \,\mathrm{d}\mu - \int_{\mathbb{R}} f \,\mathrm{d}\nu\right| &\leq \int_{\mathbb{R}} |f| \,\mathrm{d}\mu + \int_{\mathbb{R}} |f| d\nu \\ &\leq 2\|f\|_\infty \\ &\leq 2,\end{aligned}$$

want μ en ν kansmaten en f voldoet aan de eigenschappen voor de bounded-Lipschitzmetriek.

Dus hebben we voor d_{BL} als universele grenzen:

$$0 \leq d_{\mathrm{BL}}(\mu, \nu) \leq 2 \tag{12}$$

Ook voor de bounded-Lipschitzmetriek hebben we nu een eerste afschatting!

We kunnen de bovengrens voor d_{BL} ook schrijven in termen van verwachtingen. Laat X en Y stochasten in $\mathbb{R}$ op $(\Omega, \Sigma, \mathbb{P})$. Bekijk weer het verschil van de twee integralen in (3):

$$\begin{aligned}\left|\int_{\mathbb{R}} f \,\mathrm{d}\mu_X - \int_{\mathbb{R}} f d\mu_Y\right| &= |\mathbb{E}(f(X)) - \mathbb{E}(f(Y))| && \text{(wegens (1))} \\ &= |\mathbb{E}(f(X) - f(Y))| \\ &\leq \mathbb{E}|f(X) - f(Y)| \\ &= \int_\Omega |(f(X))(\omega) - (f(Y))(\omega)| \,\mathrm{d}\mathbb{P}(\omega) \\ &\leq \int_\Omega |X(\omega) - Y(\omega)| d\mathbb{P}(\omega) \\ &= \mathbb{E}|X - Y|,\end{aligned}$$

want f Lipschitz, $\|f\|_{\mathrm{BL}} \leq 1$ en dus $L \leq 1$.

Dus is:

$$(13) \qquad \begin{array}{rcll} d_{\mathrm{BL}}(\mu_X,\mu_Y) \le \mathbb{E}|X-Y| & \le & \sqrt{\mathbb{E}(X-Y)^2} & \text{(Schwartz)} \\ & = & \sqrt{\mathbb{E}X^2+\mathbb{E}Y^2-2\mathbb{E}XY}, & \end{array}$$

Combineren we (12) en (13):

$$(14) \qquad 0 \le d_{\mathrm{BL}}(\mu,\nu) \le min(2, \sqrt{\mathbb{E}X^2+\mathbb{E}Y^2-2\mathbb{E}XY})$$

5 Afstanden tussen enkele concrete verdelingen

5.1 Afstanden tussen tent-functies

5.1.1 De Prokhorov-afstand tussen $tent(0)$ en $tent(3)$

Zie sectie 3.1 voor de definitie van de tentverdeling.

We berekenen de afstand $d_P(\mu_X,\mu_Y) =: d_P(tent(0), tent(3))$. Bekijk het stelsel (10) zoals toegepast op de kansmaten voor X en Y:

$$(15) \qquad \begin{array}{rrcl} \text{(I)} & \mu_X(A) & \le & \mu_Y(A_\alpha)+\alpha \\ \text{(II)} & \mu_Y(A) & \le & \mu_X(A_\alpha)+\alpha, \end{array}$$

met $A \in \mathcal{B}(\mathbb{R})$. We volgen het recept van sectie 4.1.1 en zoeken een $A \in \mathcal{B}(\mathbb{R})$ waarvoor een α kan worden gevonden zodat *niet* aan beide ongelijkheden wordt voldaan. Kies $A=[-1,1]$ en $\alpha = 1$ dus $A_\alpha = [-2,2]$. Dan $\mu_X(A) = \mu_X(A_\alpha) = 1$ en $\mu_Y(A) = \mu_Y(A_\alpha) = 0$. Het stelsel wordt dan:

$$\begin{array}{rrcl} \text{(I)} & 1 & \le & 0+\alpha \\ \text{(II)} & 0 & \le & 1+\alpha \end{array}$$

Hieraan wordt alleen voldaan door alle $\alpha \ge 1$ en blijkbaar is hier de ondergrens gelijk aan universele bovengrens. Dus is:

$$d_P(tent(0), tent(3)) = 1.$$

5.2 De Prokhorovafstand tussen $tent(0)$ en $tent(m)$ met $m \in [0,1]$

$X \sim$ tent(0) en $Y \sim$ tent(m); $\mathbb{E}X = 0$ en $\mathbb{E}Y = m$. De verdelingen overlappen elkaar nu geheel of gedeeltelijk.

Voor $A \subset \mathbb{R}$ nemen we $A = [-1,0]$, een keuze vooral gedicteerd door rekengemak. Zij $A_\alpha = [-1-\alpha, \alpha] \supset A$, met $0 \le \alpha \le 1$. Dan is

$$\begin{array}{rcl} \mu_X(A) & = & \frac{1}{2}, \\ \mu_Y(A) & = & \frac{1}{2}(1-m)^2, \\ \mu_X(A_\alpha) & = & \frac{1}{2}+\alpha-\frac{1}{2}\alpha^2, \\ \mu_Y(A_\alpha) & = & \frac{1}{2}((1-m)+\alpha)^2. \end{array}$$

(15) wordt dan:

$$\begin{array}{rrcl} \text{(I)} & 1 & \le & ((1-m)+\alpha)^2+2\alpha \\ \text{(II)} & (1-m)^2 & \le & 5-(\alpha-2)^2. \end{array}$$

Na enig rekenen volgt:

$$\begin{array}{lccccl} \text{(I)} & & & \alpha & \geq & -(2-m)+\sqrt{2(2-m)} \\ \text{(II)} & 0 & \leq & \alpha & \leq & 2+\sqrt{5-(1-m)^2}. \end{array}$$

Ongelijkheid (II) verschaft geen informatie en kunnen we buiten beschouwing laten. We concluderen:

$$-(2-m)+\sqrt{2(2-m)} \leq d_P(tent(0), tent(m)) \leq 1. \tag{16}$$

Voor $m = 0$ is de afstand nul, dat klopt natuurlijk. Voor $m = 1$ is het linkerlid van de ongelijkheid $\sqrt{2}-1 \approx 0.41$, dus

$$0.41 \leq d_P(tent(0), tent(1)) \leq 1.$$

Vergelijk dit met de afstand tussen $tent(0)$ en $tent(3)$; het kleinere verschil tussen de gemiddelden van de verdelingen correspondeert met een kleinere afstand d_P.

Als $m \ll 1$, dan kunnen in het rechterlid van (16) hogere-ordetermen in $m/2$ vervallen en dan is de ondergrens $-(2-m)+2(1-1/2m/2) = m/2$ en dus:

$$\frac{m}{2} \leq d_P(tent(0), tent(m)) \leq 1. \tag{17}$$

De ondergrens voor de afstand daalt bij deze keuze voor A met orde (m) naar nul. Het "testinterval" $A = [-1, 0]$ is niet slecht gekozen, want de keuze $A = [-1, 1]$ levert direct een ongunstiger (=kleinere) ondergrens op.

5.2.1 De bounded-Lipschitzafstand tussen $tent(0)$ en $tent(3)$

$X \sim \text{tent}(0)$ en $Y \sim \text{tent}(3)$. Zoek een f die aan de metriekvoorwaarden voldoet en waarvoor $|\int_{\mathbb{R}} f \,\mathrm{d}\mu_X - \int_{\mathbb{R}} f \,\mathrm{d}\mu_Y|$ zo groot mogelijk. Kies de volgende functie:

$$f(x) = \begin{cases} \frac{1}{2}x + \frac{1}{2}, & \text{als } -1 \leq x \leq 0; \\ -\frac{1}{2}x + \frac{1}{2}, & \text{als } 0 < x \leq 1; \\ -\frac{1}{2}x + 1, & \text{als } 2 \leq x \leq 3; \\ \frac{1}{2}x - 2, & \text{als } 3 < x \leq 4; \\ 0, & \text{anders.} \end{cases} \tag{18}$$

Dan is:

$$\begin{array}{rcl} \left| \int_{\mathbb{R}} f \,\mathrm{d}\mu_X - \int_{\mathbb{R}} f \,\mathrm{d}\mu_Y \right| & = & \left| \int_{\mathbb{R}} f(x) h_X(x) \,\mathrm{d}x - \int_{\mathbb{R}} f(x) h_Y(x) \,\mathrm{d}x \right| \\ & = & \left| \int_{-1}^{1} f(x) h_X(x) \,\mathrm{d}x - \int_{2}^{4} f(x) h_Y(x) \,\mathrm{d}x \right|. \end{array}$$

waarin:

$$\begin{array}{rcl} \int_{-1}^{1} f(x) h_X(x) \,\mathrm{d}x & = & 2\int_0^1 \frac{1}{2} x x \,\mathrm{d}x = \frac{1}{3} \text{ en} \\ \int_{2}^{4} f(x) h_Y(x) \,\mathrm{d}x & = & -\frac{1}{3}. \end{array}$$

Dus is $d_{\mathrm{BL}}(\text{tent}(0), \text{tent}(3)) \geq \frac{2}{3}$.

De bovengrens is volgens betrekking (14):

$$d_{\mathrm{BL}}(\mu_X, \mu_Y) \leq \min\left(2, \sqrt{\mathbb{E}X^2 + \mathbb{E}Y^2 - 2\mathbb{E}XY}\right).$$

Wegens (7):

$$\mathbb{E}Y^2 = VarY + (\mathbb{E}Y)^2 = \frac{1}{6} + 9,$$

en omdat X en Y onafhankelijk zijn, is $\mathbb{E}XY = 0$ zodat:

$$\sqrt{\mathbb{E}X^2 + \mathbb{E}Y^2 - 2\mathbb{E}XY} = \sqrt{9 + \frac{1}{3}} > 3 > 2.$$

De conclusie is:

$$\frac{2}{3} \leq d_{\mathrm{BL}}(\mathrm{tent}(0), \mathrm{tent}(3)) \leq 2$$

5.2.2 Een bovengrens voor de bounded-Lipschitzafstand tussen $tent(0)$ en $tent(m)$ met $m \in [0, 2]$

$X \sim tent(0)$ en $Y \sim tent(3)$. Het maximum van de dichtheid van $tent(m)$ ligt bij $x = m$. De doorsnede van beide driehoeken is een nieuwe driehoek met de top op $x = m/2$, hoogte $1 - m/2$. Het oppervlak onder de tentfunctie minus dat van genoemde driehoek, noemen we A.

We zoeken een maximum voor $|\int_{\mathbb{R}} f \,\mathrm{d}\mu_X - \int_{\mathbb{R}} f \,\mathrm{d}\mu_Y|$:

$$\begin{aligned}
\left| \int_{\mathbb{R}} f \,\mathrm{d}\mu_X - \int_{\mathbb{R}} f \,\mathrm{d}\mu_Y \right| &= \left| \int_{\mathbb{R}} f(x) h_X(x) \,\mathrm{d}x - \int_{\mathbb{R}} f(x) h_Y(x) \,\mathrm{d}x \right| \\
&\leq \int_{\mathbb{R}} |\, f(x)| \, |h_X(x) - h_Y(x)\,| \,\mathrm{d}x \\
&\leq \|f\|_\infty \int_{\mathbb{R}} |\, h_X(x) - h_Y(x)\,| \,\mathrm{d}x \\
&\leq \int_{-1}^{m+1} |\, h_X(x) - h_Y(x)\,| \,\mathrm{d}x = 2 \int_{-1}^{\frac{m}{2}} |\, h_X(x) - h_Y(x)\,| \,\mathrm{d}x \\
&= 2A = 2\left(1 - \left(1 - \frac{m}{2}\right)^2\right).
\end{aligned}$$

Verder is: $\mathbb{E}X = \mathbb{E}XY = 0$, $\mathbb{E}Y = m$, $\mathbb{E}X^2 = VarX = \frac{1}{6}$ en $\mathbb{E}Y^2 = VarY + (\mathbb{E}Y)^2 = \frac{1}{6} + m^2$, zodat $\sqrt{\mathbb{E}X^2 + \mathbb{E}Y^2 - 2\mathbb{E}XY} = \sqrt{\frac{1}{3} + m^2}$ en dus:

$$d_{\mathrm{BL}}(tent(0), tent(m)) \leq \min\left(2, \sqrt{\frac{1}{3} + m^2}, 2\left(1 - \left(1 - \frac{m}{2}\right)^2\right)\right) \tag{19}$$

Voor enkele waarden van m berekenen we de bovengrens:

$$d_{\mathrm{BL}}(tent(0), tent(0)) \leq \min\left(2, \sqrt{\frac{1}{3} + 0}, 2\left(1 - (1 - 0)^2\right)\right) = 0 \text{ (triviaal)},$$

$$\begin{aligned}
d_{\mathrm{BL}}(tent(0), tent(1)) &\leq \min\left(2, \sqrt{\tfrac{1}{3} + 1}, 2\left(1 - \left(1 - \tfrac{1}{2}\right)^2\right)\right) \\
&= \min\left(2, 1.2, \tfrac{3}{2}\right) = 1.2,
\end{aligned}$$

$$\begin{aligned} d_{\mathrm{BL}}(tent(0), tent(2)) &\leq \min\left(2, \sqrt{\tfrac{1}{3}+4}, 2\left(1-\left(1-\tfrac{2}{2}\right)^2\right)\right) \\ &= \min\left(2, \sqrt{\tfrac{1}{3}+4}, 2\right) = 2. \end{aligned}$$

Bekijken we nu het regime voor $m \ll 1$. Dan is $(1-\frac{m}{2})^2 \approx 1-m$ en $\sqrt{\frac{1}{3}+m^2} \approx \frac{1}{\sqrt{3}}(1+\frac{3}{2}m^2)$ en voor voldoende kleine m is $2m < \frac{1}{\sqrt{3}}(1+\frac{3}{2}m^2)$ en neemt (19) een prettige gedaante aan:

$$d_{\mathrm{BL}}(\mathrm{Tent}(0), \mathrm{Tent}(m)) \leq 2m.$$

Voor voldoende kleine m neemt de bovengrens dus met orde m af!

5.3 De Prokhorovafstand tussen de Bernoulli- en binomiale verdeling

Dit zijn discrete verdelingen. Het aantal Borelverzamelingen A in (3) waarvoor het stelsel (10) gecontroleerd moet worden, loopt exponentieel op met de parameter n van de binomiale verdeling. Ik heb me daarom beperkt tot een aantal eenvoudige situaties. $X \sim Bern(-1, 1; 1/2)$ en $Y \sim bin(n, 1/2)$. We bekijken de afstanden voor $n = 1$ en $n = 2$.

5.3.1 De afstand tussen $Bern(-1, 1; 1/2)$ en $bin(1, 1/2)$

$X \sim Bern(-1, 1; 1/2)$ en $Y \sim bin(1, 1/2)$. $\mathbb{E}X = 0$ en $\mathbb{E}Y = 1/2$. Op welke metrische verzameling S moeten we nu $\mathcal{P}(S)$ baseren? Niet $S = \mathbb{R}$, want dan moeten we alsnog oneindig veel Borel-deelverzamelingen langslopen. X en Y nemen waarden aan in resp. $\{-1, 1\}$ en $\{0, 1\}$ en neem dus $S = \{-1, 1\} \cup \{0, 1\} = \{-1, 0, 1\}$ met de "gewone" metriek van $\mathbb{R}$. De restrictie van deze metriek tot S geeft de discrete topologie. Nu is $\mathcal{B}(S) = \{0, 1\}^S$ en alle acht deelverzamelingen A van S zijn Borel: $\{0, 1\}^S = \{\emptyset, S, \{-1\}, \{0\}, \{1\}, \{-1, 0\}, \{-1, 1\}, \{0, 1\}\}$.

In het stelsel (10) worden de kansmaten nu geschreven als sommen: $\mu_X(A) = \sum_{x \in A} p_X(x)$ en $\mu_Y(A) = \sum_{y \in A} p_Y(y)$ waarin $p_X(x) = \mathbb{P}(X = x)$ en $p_Y(y) = \mathbb{P}(Y = y)$. En er geldt $A = A_\alpha$ voor $\alpha < 1$.

In tabel 1 worden de resultaten samengevat.
Blijkbaar is:

$$d_P(Bern(-1, 1; 1/2), bin(1, 1/2)) = 1/2.$$

5.3.2 De afstand tussen $Bern(-1, 1; 1/2)$ en $bin(2, 1/2)$

$X \sim Bern(-1, 1; 1/2)$ en $Y \sim bin(2, 1/2)$. Nu is $S = \{-1, 0, 1, 2\}$, $\mathbb{E}X = 0$ en $\mathbb{E}Y = 1$ en er zijn zestien deelverzamelingen waarvoor de ongelijkheden in (10) moeten worden getest. Na een berekening als in tabel 1 verkrijgen we:

$$d_P(Bern(-1, 1; 1/2), Bin(2, 1/2)) = 1/2.$$

en deze afstand is ongewijzigd t.o.v. afstand tussen $Bern(-1, 1; 1/2)$ en $bin(1, 1/2)$)! We verwachten een grotere afstand; een mogelijke verklaring is dat de verdelingen elkaar overlappen. Het is interessant de afstanden nog eens na te gaan voor $n > 2$; misschien is er dan een stapsgewijze toename van de Prokhorovafstand te zien.

A	$\mu_X(A)$	$\mu_X(A)$	Ongelijkheid (I)	Ongelijkheid (II)	Aan (I) en (II) voldaan als
$\emptyset$	0	0	$0 \leq 0 + \alpha$	$0 \leq 0 + \alpha$	$\alpha \geq 0$
S	1	1	$1 \leq 1 + \alpha$	$1 \leq 1 + \alpha$	$\alpha \geq 0$
$\{-1\}$	$\frac{1}{2}$	0	$\frac{1}{2} \leq 0 + \alpha$	$0 \leq \frac{1}{2} + \alpha$	$\alpha \geq \frac{1}{2}$
$\{0\}$	0	$\frac{1}{2}$	$0 \leq \frac{1}{2} + \alpha$	$\frac{1}{2} \leq 0 + \alpha$	$\alpha \geq \frac{1}{2}$
$\{1\}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2} \leq \frac{1}{2} + \alpha$	$\frac{1}{2} \leq \frac{1}{2} + \alpha$	$\alpha \geq 0$
$\{-1, 0\}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2} \leq \frac{1}{2} + \alpha$	$\frac{1}{2} \leq \frac{1}{2} + \alpha$	$\alpha \geq 0$
$\{-1, 1\}$	1	$\frac{1}{2}$	$1 \leq \frac{1}{2} + \alpha$	$\frac{1}{2} \leq 1 + \alpha$	$\alpha \geq \frac{1}{2}$
$\{0, 1\}$	$\frac{1}{2}$	1	$\frac{1}{2} \leq 1 + \alpha$	$1 \leq \frac{1}{2} + \alpha$	$\alpha \geq \frac{1}{2}$

Tabel 1: Boekhouding bij de berekening van de afstand tussen $Bern(-1, 1; 1/2)$ en $bin(1, 1/2)$

5.4 Een ondergrens voor de Prokhorovafstand tussen $Bern(-1, 1; 1/2)$ en $N(0, 1)$

$X \sim Bern(-1, 1; 1/2)$ en $Y \sim N(0, 1)$. Beide verdelingen hebben dezelfde verwachting, $\mathbb{E}X = \mathbb{E}Y = 0$.

Het stelsel ongelijkheden (10) wordt:

$$\begin{array}{rrl} & \text{(I)} & \sum_{x \in A} p_X(x) \leq \int_{A_\alpha} \varphi(x)dx + \alpha \\ & \text{(II)} & \int_A \varphi(x)dx \leq \sum_{x \in A} p_X(x) + \alpha \end{array} \tag{20}$$

We kiezen $A, A_\alpha \subset [-1, 1]$. Dan is $\mu_X(A) = \mu_X(A_\alpha) = 0$ (heel prettig!) en wordt (20):

$$\begin{array}{rrl} & \text{(I)} & 0 \leq \int_A \varphi(y)\,\mathrm{d}y + \alpha \\ & \text{(II)} & \int_A \varphi(y)\,\mathrm{d}y \leq 0 + \alpha \end{array}$$

Handig in de keuze van A is dat de eerste ongelijkheid niet langer hoeft te worden betrokken in de bepaling van α, want iedere waarde van α voldoet. Maar we moeten dan wel opletten met de grenzen van het interval $A = [a, b]$: kiezen we A te "klein", dan levert dat een ongunstige afschatting van α, lees een te kleine waarde. Maken we A te groot, dan kan α zo groot worden, dat niet langer $A_\alpha \subset [-1, 1]$ en dus ook niet langer $\mu_X(A) = \mu_X(A_\alpha) = 0$.

Het gaat nog nét goed voor $A = [-1/2, 1/2]$. De massa van de standaard-normale verdeling in dit interval is 0.39, en dan moet dus $\alpha \geq 0.39$ en kunnen we stellen

$$0.39 \leq d_P(\text{Bern}(-1, 1; 1/2), N(0, 1)) \leq 1.$$

Kiezen we $A_\alpha \supset A = [-1, 1]$ dan:

$$\begin{array}{rrl} & \text{(I)} & 1 \leq \int_{A_\alpha} \varphi(y)\,dy + \alpha \\ & \text{(II)} & 0.68 \leq 1 + \alpha \end{array}$$

Aan (II) is voor alle α voldaan, en voor α voldoende klein is:

$$\begin{array}{rl} \int_{A_\alpha} h(y)\,dy & = \int_{-1}^{1} \varphi(y)\,dy + \frac{2\alpha}{\sqrt{2\pi}} e^{-\frac{1}{2}} \\ & = 0.68 + 0.48\alpha \end{array}$$

waaruit $\alpha \geq 0.22$, dus niet zo goed als de eerste afschatting, 0.39. Bovendien is α een aanzienlijke fractie van de intervalbreedte, zodat $\varphi(y)$ niet zonder meer constant verondersteld mag worden op $[-\alpha-1,-1]$ en $[1,1+\alpha]$.

Schuiven we de Bernoulli-verdeling een bedrag $m>0$ op (zodat dus $\mathbb{E}X=m$) en kiezen we $A_\alpha \supset A=[m-1,m+1]$, dan $\int_{A_\alpha}\varphi(y)\,\mathrm{d}y \to 0$ als $m\to\infty$ en dus $\alpha\to 1$. En dat willen we natuurlijk!

5.5 Een ondergrens voor de bounded-Lipschitzafstand tussen $(Bern(-1,1;1/2)$ en $N(0,1))$

$X\sim \mathrm{Bern}(-1,1;1/2)$ en $Y\sim N(0,1)$, $\mathbb{E}X=0$ en $\mathbb{E}Y=0$. We zoeken een functie f die voldoet aan de metriekvoorwaarden en waarvoor $|\int_{\mathbb{R}} f\,\mathrm{d}\mu_X - \int_{\mathbb{R}} f\,\mathrm{d}\mu_Y|$ zo groot mogelijk is. Een handige keuze met $f(-1)=f(1)=0$ is

$$f(x)=\begin{cases} \frac{1}{2}x+\frac{1}{2}, & \text{als } -1\leq x\leq 0;\\ -\frac{1}{2}x+\frac{1}{2}, & \text{als } \ 0<x\leq 1;\\ 0, & \text{anders.}\end{cases}$$

want dan is $\int_{\mathbb{R}} f\,\mathrm{d}\mu_X=(f(-1)+f(1))/2=0$. En

$$\begin{aligned}\int_{\mathbb{R}} f\,\mathrm{d}\mu_Y &= 2\times\tfrac{1}{\sqrt{2\pi}}\textstyle\int_{-1}^{0}(\frac{1}{2}x+\frac{1}{2})e^{-\frac{1}{2}x^2}\,\mathrm{d}x\\ &= \tfrac{1}{\sqrt{2\pi}}\textstyle\int_{-1}^{0}(x+1)e^{-\frac{1}{2}x^2}\,\mathrm{d}x\\ &= \tfrac{1}{\sqrt{2\pi}}\textstyle\int_{-1}^{0}xe^{-\frac{1}{2}x^2}\,\mathrm{d}x+\tfrac{1}{\sqrt{2\pi}}\int_{-1}^{0}e^{-\frac{1}{2}x^2}\,\mathrm{d}x\\ &= \tfrac{e^{-\frac{1}{2}}-1}{\sqrt{2\pi}}+\tfrac{0.68}{2}\\ &= -0.16+0.34=0.18\end{aligned}$$

zodat $|\int_{\mathbb{R}} f\,\mathrm{d}\mu_X-\int_{\mathbb{R}} f\,\mathrm{d}\mu_Y|=|0-0.18|=0.18$ en dus

$$0.18\leq d_{\mathrm{BL}}(\mathrm{Bern}(-1,1;1/2),N(0,1))\leq 2.$$

5.6 Een ondergrens voor de bounded-Lipschitzafstand tussen $Bern(0,2;1/2)$ en $N(0,1)$

Als boven, maar de Bernouilli-verdeling is nu de afstand 1 opgeschoven: $X\sim\mathrm{Bern}(0,2;1/2)$ en $\mathbb{E}X=1$. Voor f nemen we:

$$f(x)=\begin{cases} \frac{1}{2}x+1, & \text{als } -2\leq x\leq -1;\\ -\frac{1}{2}x, & \text{als } -1<x\leq 0;\\ \frac{1}{2}x, & \text{als } \ 0<x\leq 1;\\ -\frac{1}{2}x+1, & \text{als } \ 1<x\leq 2;\\ 0, & \text{anders.}\end{cases}$$

Populair gezegd betreft het twee tent-functies (geen dichtheden) "in serie geschakeld" zodat f even is. Ook hier is $f=0$ in de punten waar X waarden aanneemt. De tweede "tent" is louter bedoeld om $\int_{\mathbb{R}} f\,\mathrm{d}\mu_Y$ te vergroten en dus de ondergrens van d_{BL} te verbeteren. De berekening van deze integraal is weer elementair en we verkrijgen:

$$\left|\int_{\mathbb{R}} f\,\mathrm{d}\mu_X-\int_{\mathbb{R}} f\,\mathrm{d}\mu_Y\right|=|0-0.34|=0.34,$$

zodat

$$0.34 \leq d_{\mathrm{BL}}(\mathrm{Bern}(0,2;1/2), N(0,1)) \leq 2$$

6 De Centrale Limiet Stelling en convergentie in de ruimte $\mathcal{P}(\mathbb{R})$

We komen nu op het hoofdonderdeel van de scriptie.

Ik herhaal nog even wat zaken. De *Centrale Limiet Stelling* (CLS) heeft betrekking op de limieteigenschappen van sommen van stochasten. Als $X_1, X_2, \ldots$ een rij i.i.d ("independent and identically-distributed") toevalsvariabelen is, met gemiddelde $\mathbb{E}X_1$ en variantie σ^2 en als:

$$S_n := \sum_{k=1}^{n} X_k,$$

dan weten we van de wet van de grote aantallen dat S_n/n in kans naar $\mathbb{E}X_1$ convergeert. De CLS is niet zozeer geïnteresseerd in die limiet, maar eerder in de mate waarin S_n/n om dat gemiddelde fluctueert. Voor dat onderzoek standaardiseren we:

$$Z_n := \frac{S_n - n\mathbb{E}X_1}{\sigma\sqrt{n}}.$$

Z_n heeft gemiddelde 0 en variantie 1. Volgens de CLS convergeert Z_n zwak naar $N(0,1)$, $Z_n \rightsquigarrow N(0,1)$. En onder beide metrieken geldt in $\mathcal{P}(\mathbb{R})$:

$$Z_n \rightsquigarrow N(0,1) \quad \Longleftrightarrow \quad d_{P,\mathrm{BL}}(Z_n, N(0,1)) \to 0.$$

Zoals al aangegeven in de Inleiding van deze scriptie willen we nagaan of we een snelheid van convergentie kunnen bepalen ten opzichte van de beide metrieken: is de convergentie van orde zeg $1/n$ of $1/\sqrt{n}$? Hoe verhouden de convergentiesnelheden zich onder de metrieken?

Concreet onderzoeken we dit aan de hand van de rij i.i.d. stochasten $X_1, X_2, \ldots$ met $X_1 \sim Bern(0,1;1/2)$. Voor deze rij geldt:

$$\begin{aligned} Z_n &= \tfrac{S_n - n\mathbb{E}X_1}{\sigma\sqrt{n}} = \tfrac{S_n - \frac{n}{2}}{\frac{\sqrt{n}}{2}} \\ &= \tfrac{2S_n - n}{\sqrt{n}}. \end{aligned}$$

Nu is S_n als som van een rij i.i.d Bernouilli-verdeelde variabelen binomiaal verdeeld, $S_n \sim bin(n,1/2)$, dus $S_n \in \{0,1,2,\ldots,n\}$ en dus $Z_n \in (2 \times \{0,1,2,\ldots,n\} - n)/\sqrt{n}$, dus

$$Z_n \in \left\{ \frac{-n}{\sqrt{n}}, \frac{2-n}{\sqrt{n}}, \frac{4-n}{\sqrt{n}}, \ldots, \frac{2m-n}{\sqrt{n}}, \ldots, \frac{n-2}{\sqrt{n}}, \frac{n}{\sqrt{n}}, \right\} \quad (m = 0,1,2,\ldots,n)\,.$$

Er is nog onderscheid voor n even en n oneven

$$Z_n \in \left\{ -\sqrt{n}, \frac{2-n}{\sqrt{n}}, \ldots, \frac{-2}{\sqrt{n}}, 0, \frac{-2}{\sqrt{n}}, \ldots, \frac{n-2}{\sqrt{n}}, \sqrt{n}, \right\} \quad (\text{n even})$$

$$Z_n \in \left\{ -\sqrt{n}, \frac{2-n}{\sqrt{n}}, \ldots, \frac{-1}{\sqrt{n}}, \frac{-1}{\sqrt{n}}, \ldots, \frac{n-2}{\sqrt{n}}, \sqrt{n}, \right\} \quad (\text{n oneven})\,.$$

Dus Z_n is verdeeld als de in sectie 3.3 geïntroduceerde verdeling $bin2(n,p)$ met $p = 1/2$.

6.1 Convergentiesnelheid van de Prokhorovafstand tussen Z_n en $N(0,1)$

Z_n als boven, $W \sim N(0,1)$. Kies ten behoeve van (10) als testinterval: $A = [-\sqrt{n}, \sqrt{n}]$ en $A_\alpha = [-\sqrt{n}-\alpha, \sqrt{n}+\alpha]$. Dan wordt het tweetal ongelijkheden:

$$\begin{aligned} &\text{(I)} && 1 \leq \tfrac{1}{\sqrt{2\pi}} \textstyle\int_{-\sqrt{n}-\alpha}^{\sqrt{n}+\alpha} e^{-\frac{1}{2}x^2}\,\mathrm{d}x + \alpha \\ &\text{(II)} && \tfrac{1}{\sqrt{2\pi}} \textstyle\int_{-\sqrt{n}}^{\sqrt{n}} e^{-\frac{1}{2}x^2}\,\mathrm{d}x \leq 1 + \alpha \end{aligned}$$

Op de laatste ongelijkheid hoeven we niet te letten; lossen we de eerste op, dan verkrijgen we:

$$\begin{array}{rlll} n=100\text{:} & 1 \leq 1+\alpha & \text{en dus} & 0 \leq d_P(Z_n,W) \leq 1, \\ n=10\text{:} & 1 \leq 0.998+\alpha & \text{en dus} & 0.000157 \leq d_P(Z_n,W) \leq 1, \\ n=5\text{:} & 1 \leq 0.975+\alpha & \text{en dus} & 0.0250 \leq d_P(Z_n,W) \leq 1, \\ n=2\text{:} & 1 \leq 0.842+\alpha & \text{en dus} & 0.157 \leq d_P(Z_n,W) \leq 1. \end{array}$$

De naar 0 dalende ondergrens is consistent met $d_P \to 0$ als $n \to \infty$.

De ondergrens gaar zeer snel naar nul, maar het is vermoedelijk ondoenlijk om een vergelijkbaar goede rij bovengrenzen te berekenen; *we schenken daarom verder geen aandacht aan deze metriek.*

6.2 Convergentiesnelheid van de bounded-Lipschitzafstand tussen Z_n en $N(0,1)$

We zoeken weer een functie f die aan de voorwaarden van de bounded-Lipschitzmetriek voldoet, die zo min rekenwerk vraagt en waarvoor het verschil $|\int_{\mathbb{R}} f\,\mathrm{d}\mu_{Z_n} - \int_{\mathbb{R}} f\,\mathrm{d}\mu_W|$ zo groot mogelijk wordt. Neem daartoe voor f een functie die 0 is op $(-\infty,-\sqrt{n}) \cup (\sqrt{n},\infty)$ en op het interval $[-\sqrt{n},\sqrt{n}]$ bestaat uit een collectie op elkaar aansluitende tentfuncties (geen dichtheden) met maximumwaarde $1/(2\sqrt{n})$ en die 0 zijn in de punten waar Z_n waarden aanneemt. Dan is $\|f\|_{\mathrm{BL}} = \|f\|_\infty + \mathrm{Lip}(f) = 1/(2\sqrt{n}) + 1/2 \leq 1$. Merk op dat $\|f\|_{\mathrm{BL}} \to 1/2$ voor $n \to \infty$.

$$\int_{\mathbb{R}} f\,\mathrm{d}\mu_{Z_n} = \sum_{z \in \{-\sqrt{n},\ldots,\sqrt{n}\}} f(z) p_{Z_n}(z) = 0. \tag{21}$$

We willen eerst iets zeggen over de ondergrens voor d_{BL}. Afhankelijk van n even of oneven heeft f een andere gedaante, dus we moeten tweemaal afschatten en nemen dan het minimum van de verkregen ondergrenzen.

6.2.1 Ondergrens voor n even

Nu is $f(0) = 0$.

$$\begin{aligned} \textstyle\int_{\mathbb{R}} f\,\mathrm{d}\mu_W &= \textstyle\int_{\mathbb{R}} f(w)\varphi(w)\,\mathrm{d}w \\ &= \textstyle\int_{-\sqrt{n}}^{\frac{2-n}{\sqrt{n}}} f(w)\varphi(w)\,\mathrm{d}w + \int_{\frac{2-n}{\sqrt{n}}}^{\frac{4-n}{\sqrt{n}}} f(w)\varphi(w)\,\mathrm{d}w + \cdots \\ &\quad + \textstyle\int_{\frac{-2}{\sqrt{n}}}^{0} f(w)\varphi(w)\,\mathrm{d}w + \int_{0}^{\frac{2}{\sqrt{n}}} f(w)\varphi(w)\,\mathrm{d}w + \cdots + \int_{\frac{n-2}{\sqrt{n}}}^{\sqrt{n}} f(w)\varphi(w)\,\mathrm{d}w \end{aligned} \tag{22}$$

Nu gaan we afschatten, eerst de waarde van $\int_0^{\frac{2}{\sqrt{n}}} f(w)\varphi(w)\,\mathrm{d}w$.

$$\begin{aligned}\int_0^{\frac{2}{\sqrt{n}}} f(w)\varphi(w)\,\mathrm{d}w &\overset{\text{(i)}}{\geq} \varphi(\tfrac{2}{\sqrt{n}})\int_0^{\frac{2}{\sqrt{n}}} f(w)\,\mathrm{d}w \overset{\text{(ii)}}{=} \varphi(\tfrac{2}{\sqrt{n}})\tfrac{1}{2n}\\ &\overset{\text{(iii)}}{=} \tfrac{\sqrt{n}}{2}\left(\varphi(\tfrac{2}{\sqrt{n}})\tfrac{2}{\sqrt{n}}\right)\tfrac{1}{2n}\\ &\overset{\text{(iv)}}{\geq} \tfrac{\sqrt{n}}{2}\tfrac{1}{2n}\int_0^{\frac{2}{\sqrt{n}}} \varphi(\tfrac{2}{\sqrt{n}})e^{-\frac{1}{2}w^2}\,\mathrm{d}w\\ &\overset{\text{(v)}}{=} \tfrac{1}{4\sqrt{n}}e^{-\frac{2}{n}}\int_0^{\frac{2}{\sqrt{n}}} \varphi(w)\,\mathrm{d}w.\end{aligned}$$

Commentaar bij deze afleiding:

(i) Op het interval $[0, 2/\sqrt{n}]$ (basis van een tent) is $\varphi(w) \geq \varphi(2/\sqrt{n})$.

(ii) $\int_0^{\frac{2}{\sqrt{n}}} f(w)\,\mathrm{d}w = 1/2n$ is het oppervlak onder de tent met hoogte $1/(2\sqrt{n})$ en basis $[0, 2/\sqrt{n}]$.

(iii) Delen door en vermenigvuldigen met $\sqrt{n}/2$. Tussen de haken staat het oppervlak van een rechthoek met breedte $2/\sqrt{n}$ en hoogte $\varphi(2/\sqrt{n})$.

(iv) Dit oppervlak is groter dan dat onder de functie $\varphi(2/\sqrt{n})e^{-\frac{1}{2}w^2}$ op het interval $[0, 2/\sqrt{n}]$.

(v) $\varphi(2/\sqrt{n}) = (1/\sqrt{2\pi})e^{-\frac{1}{2}(\frac{2}{\sqrt{n}})^2} = (1/\sqrt{2\pi})e^{-\frac{2}{n}}$. De factor $1/\sqrt{2\pi}$ onder het integraalteken laten.

Zo kunnen we elk van de $n+1$ termen in (22) afschatten. Omdat φ en f even zijn en wegens (21):

$$\left|\int_{\mathbb{R}} f\,\mathrm{d}\mu_{Z_n} - \int_{\mathbb{R}} f\,\mathrm{d}\mu_W\right| = \int_{\mathbb{R}} f(w)\varphi(w)\,\mathrm{d}w \geq \frac{1}{2\sqrt{n}}e^{-\frac{2}{n}}\int_0^{\sqrt{n}} \varphi(w)\,\mathrm{d}w \tag{23}$$

6.2.2 Ondergrens voor n oneven

Nu is $f(0) = 1/(2\sqrt{n})$. De afschattingsprocedure is vrijwel exact als die voor n even, maar we schatten nu eerst de term $\int_{-1/\sqrt{n}}^{1/\sqrt{n}} f(w)\varphi(w)\,\mathrm{d}w$ af. $\varphi(1/\sqrt{n}) = (1/\sqrt{2\pi})e^{-\frac{1}{2n}}$.

$$\begin{aligned}\int_{-\frac{1}{\sqrt{n}}}^{\frac{1}{\sqrt{n}}} f(w)\varphi(w)\,\mathrm{d}w &\geq \varphi(\tfrac{1}{\sqrt{n}})\int_{-\frac{1}{\sqrt{n}}}^{\frac{1}{\sqrt{n}}} f(w)\,\mathrm{d}w = \varphi(\tfrac{1}{\sqrt{n}})\tfrac{1}{2n}\\ &= \tfrac{\sqrt{n}}{2}\left(\varphi(\tfrac{1}{\sqrt{n}})\tfrac{2}{\sqrt{n}}\right)\tfrac{1}{2n}\\ &\geq \tfrac{\sqrt{n}}{2}\tfrac{1}{2n}\int_{-\frac{1}{\sqrt{n}}}^{\frac{1}{\sqrt{n}}} \varphi(\tfrac{1}{\sqrt{n}})e^{-\frac{1}{2}w^2}\,\mathrm{d}w\\ &= \tfrac{1}{4\sqrt{n}}e^{-\frac{1}{2n}}\int_{-\frac{1}{\sqrt{n}}}^{\frac{1}{\sqrt{n}}} \varphi(w)\,\mathrm{d}w.\end{aligned}$$

En voor n oneven:

$$\left|\int_{\mathbb{R}} f\,\mathrm{d}\mu_{Z_n} - \int_{\mathbb{R}} f\,\mathrm{d}\mu_W\right| = \int_{\mathbb{R}} f(w)\varphi(w)\,\mathrm{d}w \geq \frac{1}{2\sqrt{n}}e^{-\frac{1}{2n}}\int_0^{\sqrt{n}} \varphi(w)\,\mathrm{d}w \tag{24}$$

Combineren we de formules (23) en (24) voor n even resp. oneven dan is de gezochte ondergrens:

$$\min\left(e^{-\frac{1}{2n}}, e^{-\frac{2}{n}}\right)\frac{1}{2\sqrt{n}}\int_0^{\sqrt{n}}\varphi(w)\,\mathrm{d}w \le d_{\mathrm{BL}}(Z_n, N(0,1))$$

Als $n \ge 10$ mogen we stellen: $\int_0^{\sqrt{n}}\varphi(w)\,\mathrm{d}w = 1/2$ zodat dan:

$$d_{\mathrm{BL}}(Z_n, N(0,1)) \ge \min\left(e^{-\frac{1}{2n}}, e^{-\frac{2}{n}}\right)\frac{1}{4\sqrt{n}} \tag{25}$$

Voor $n = 2$ moeten we een correctie toepassen; de factor $1/2$ wordt dan 0.42.

We berekenen de ondergrens voor enkele waarden van n:

$n = 100$:

$$d_{\mathrm{BL}} \ge \min(\, e^{-\frac{1}{200}}, e^{-\frac{1}{50}}\,)\frac{1}{40} = 0.98 \times 1/40 = 0.025$$

$n = 10$:

$$d_{\mathrm{BL}} \ge \min(\, e^{-\frac{1}{20}}, e^{-\frac{1}{5}}\,)\frac{1}{4\sqrt{10}} = 0.82 \times 0.079 = 0.065$$

$n = 2$:

$$d_{\mathrm{BL}} \ge \min(\, e^{-\frac{1}{4}}, e^{-1}\,)\frac{1}{2\sqrt{2}}0.42 = 0.37 \times 0.35 \times 0.42 = 0.054$$

Met toenemende n daalt de ondergrens, in overeenstemming met de CLS. Voor grote n is het minimum gelijk te nemen aan 1 en is in goede benadering:

$$d_{\mathrm{BL}}(Z_n, N(0,1)) \ge \frac{1}{4\sqrt{n}}. \tag{26}$$

Dit suggereert, althans voor deze specifieke f, dat d_{BL} met orde $\frac{1}{4\sqrt{n}}$ naar 0 convergeert.

6.2.3 Berekening bovengrens

Wegens (14) is:

$$d_{\mathrm{BL}}(Z_n, W) \le \min\left(2, \sqrt{\mathbb{E}{Z_n}^2 + \mathbb{E}W^2 - 2\mathbb{E}Z_nW}\right),$$

met $W \sim N(0, 1.$

Substitueer $\mathbb{E}Z_n = \mathbb{E}W = 0$ en $\mathbb{E}Z_n^2 = \mathbb{E}W^2 = 1$:

$$d_{\mathrm{BL}}(Z_n, W) \le \sqrt{2(1 - \mathbb{E}Z_nW)}$$

Als Z_n en W onafhankelijk zijn, is:

$$d_{\mathrm{BL}}(Z_n, N(1,0)) \le \sqrt{2}.$$

Kunnen we dit resultaat verbeteren? Ja, we *moeten* de bovengrens zelfs scherper afschatten, willen we iets kunnnen zeggen over de orde waarmee $d_{\mathrm{BL}}(Z_n, N(1,0)) \to 0$; een afschatting van de ondergrens is niet voldoende. We maken gebruik van de prettige eigenschap dat de afstand tussen Z_n en W, *onafhankelijk* is van de mate waarin W en Z_n gecorreleerd zijn.

Dus zoeken we naar simultane verdelingen $(\widetilde{Z_n}, \widetilde{W})$ met Z_n en W als marginalen, waarvoor $\mathbb{E}\widetilde{Z_n}\widetilde{W}$ zo groot mogelijk is — want dat verkleint de bovengrens van de afstand.

Stel dat het lukt om zo'n verdeling te vinden: hoe bepaal je dan de verwachting $\mathbb{E}\widetilde{Z_n}\widetilde{W}$? Het wordt een Fubini-integraal waarin een discrete met een continue verdeling wordt "gemengd", een onding. Om die te omzeilen vervangen we W door een geschikte discrete verdeling, als volgt.

De driehoeksongelijkheid zegt:

$$d_{\mathrm{BL}}(Z_n, W) \leq d_{\mathrm{BL}}(Z_n, Z_m) + d_{\mathrm{BL}}(Z_m, W) \quad (m = 1, 2, \cdots)$$

zodat $d_{\mathrm{BL}}(Z_n, Z_m) \to d_{\mathrm{BL}}(Z_n, W)$ als $m \to \infty$. Voor voldoende grote m is $d_{\mathrm{BL}}(Z_n, Z_m)$ dus een maat voor $d_{\mathrm{BL}}(Z_n, W)$. Dit suggereert om de stochast W te vervangen door de discrete verdeling Z_m. Maar Z_n en Z_m zijn onafhankelijk, dus $\mathbb{E}Z_nZ_m = \mathbb{E}Z_n\mathbb{E}Z_m = 0$. Dus zoeken we naar simultane verdelingen $(\widetilde{Z_n}, \widetilde{Z_m})$ *waarvoor* $\mathbb{E}\widetilde{Z_n}\widetilde{Z_m}$ *zo groot mogelijk is*. Voor voldoende grote m is dan:

$$d_{\mathrm{BL}}(Z_n, W) \leq \sqrt{2(1 - \mathbb{E}\widetilde{Z_n}\widetilde{Z_m})} \tag{27}$$

en $\mathbb{E}\widetilde{Z_n}\widetilde{Z_m}$ is een aanzienlijk makkelijker te berekenen dubbele som! Dan dient zich het probleem aan: hoe *maken* we afhankelijke stochasten $\widetilde{Z_n}$ en $\widetilde{Z_m}$?

6.2.4 Afhankelijk maken van Z_n en Z_m

Als X en Y reële stochasten zijn met $\mathbb{E}X = \mathbb{E}Y = 0$ en $\mathbb{E}X^2 = \mathbb{E}Y^2 = 1$, dan zijn X en Y maximaal afhankelijk als $X = Y$. De waarden van de simulaten verdeling $(X, Y) \in \mathbb{R}^2$ liggen dan op de diagonaal in het X–Y-vlak. Dat doet ons het idee aan de hand om Z_n en Z_m afhankelijk te maken, door in de n bij m-matrix van gewichten van de simultane verdeling, massa "over te scheppen" naar de diagonaal.

De gewichten van de stochasten zijn als volgt verdeeld over hun uitkomsten: Z_n met gewichten/kansen $\{\, p_0, p_1, p_2, \ldots p_n \,\}$ (de frequentieverdeling van $S_n \sim bin(n, 1/2)$, zie secties 3.3 en 6), verdeeld over de uitkomsten $\{\, \frac{-n}{\sqrt{n}}, \frac{2-n}{\sqrt{n}}, \frac{4-n}{\sqrt{n}}, \ldots, \frac{2i-n}{\sqrt{n}}, \ldots, \frac{n-2}{\sqrt{n}}, \frac{n}{\sqrt{n}}, \,\}$ $(i = 0, 1, 2, \cdots, n)$ en Z_m met gewichten $\{\, q_0, q_1, q_2 \ldots q_m \,\}$ (de frequentieverdeling van $S_m \sim \mathrm{bin}(m, 1/2)$), verdeeld over de uitkomsten $\{\, \frac{-m}{\sqrt{m}}, \frac{2-m}{\sqrt{m}}, \frac{4-m}{\sqrt{m}}, \ldots, \frac{2j-m}{\sqrt{m}}, \ldots, \frac{m-2}{\sqrt{m}}, \frac{m}{\sqrt{m}}, \,\}$ $(j = 0, 1, 2, \cdots, m)$.

De gewichten p_iq_j van de simultane verdeling (Z_n, Z_m) zijn verdeeld over het $n \times m$-rooster opgespannen door de uitkomsten van die verdeling. Door massa over te hevelen vanuit "gebieden" ter weerszijde van de diagonaal van het rooster naar de diagonaal zelf, kunnen we Z_n en Z_m afhankelijk maken; er onstaat een nieuwe simultane verdeling met marginalen die we noteren als $\widetilde{Z_n}$ resp. $\widetilde{Z_m}$. De nieuwe gewichten noteren we als $\widetilde{p_{i,j}}$. We dragen er natuurlijk zorg voor dat de (dubbele) som over de nieuwe gewichten gelijk aan 1 is! Is het rooster rechthoekig, dan doet maar een klein, centraal, deel mee aan dit diagonaliseren. We moeten het rooster dus vierkant maken, door beide zijden ervan op te rekken tot het kleinste gemene veelvoud (kgv) van n en m. Daartoe dupliceren we de uitkomsten van de marginalen Z_n en Z_m met een bedrag afhankelijk van n resp. m en delen er vervolgens weer door. Laat ik een concreet voorbeeld geven. Stel we willen de gewichten van Z_2 en Z_3 diagonaliseren. De kgv van 2 en 3 is 6. We dupliceren de gewichten van Z_2 $6/2 = 3$-voudig en die van Z_3 $6/3 = 2$-voudig. Vervolgens moeten we weer delen door de verveelvoudigingsfactor, teneinde te sommeren tot 1.

De nieuwe gewichten worden dan $\{\, p_0/3, p_0/3, p_0/3, p_1/3, p_1/3, p_1/3, p_2/3, p_2/3, p_2/3 \,\}$ resp. $\{\, q_0/2, q_0/2, q_1/2, q_1/2, q_2/2, q_2/2, q_3/2, q_3/2 \,\}$. Merk op dat de verdelingen symmetrisch zijn, want het zijn die van binomiale verdeling met kans $1/2$. Dit zijn de gewichten van de marginalen $\widetilde{Z_2}$ resp. $\widetilde{Z_3}$. De elementen van het *bereik* van Z_n en Z_m worden ook gedupliceerd, maar worden niet geschaald; we noteren ze als $\tilde{i}$ resp. $\tilde{j}$.

Het overbrengen van gewicht naar de diagonaal is niet voldoende; in de "negatieve kwadranten" van het bereik van de simultane verdeling bevindt zich na dit transport nog teveel massa. Deze laatste resten massa moeten worden overgebracht naar de positieve kwadranten.

De verwachting $\mathbb{E}\widetilde{Z_n}\widetilde{Z_m}$ kunnen we nu berekenen:

$$\mathbb{E}\widetilde{Z_n}\widetilde{Z_m} = \sum_{\tilde{i},\tilde{j}} \tilde{i}\tilde{j}\widetilde{p_{\tilde{i},\tilde{j}}}. \tag{28}$$

Vervolgens kan de bovengrens uit formule (27) worden berekend. Ik heb de MATLAB-programma's `verwachting` en `bovengrens` (M-files) geschreven, die respectievelijk voor gegeven n en m de nieuwe verwachting (28) en een n bij n kruistabel van bovengrenzen (27) berekenen; zie hoofdstuk 8 voor de programmacode. De resultaten vindt men in de tabellen 2 en 3.

6.2.5 Interpretatie van de tabellen

De gekozen waarden voor n en m ($n, m = 1, 4, 9, 19, 49, 99, 199, 499, 999, 1999$) lijken in eerste instantie eigenaardig. Maar tel er 1 bij op en men ziet dat dan een gunstig kleinste gemene veelvoud onstaat, zodat men tot relatief grote waarden voor n en m kan gaan, voordat MATLAB tegen een geheugenbeperking aanloopt.

Bekijk in de tabel met de bovengrens bijvoorbeeld de rij $n = 4$ en ga in gedachten naar rechts in de tabel; dat correspondeert met $m \to \infty$ en met $d_{\mathrm{BL}}(Z_n, Z_m) \to d_{\mathrm{BL}}(Z_n, N(0,1))$ (zie sectie 6.2.3). We nemen aan dat $m = 1999$ voldoende groot is, zodat we mogen stellen $d_{\mathrm{BL}}(Z_n, Z_m) = 0.8961 \approx d_{\mathrm{BL}}(Z_n, N(0,1))$, althans we gaan ervan uit dat de bovengrens voldoende dicht bij d_{BL} ligt.

Ga vervolgens naar beneden in de laatste kolom van tabel 3; dat correspondeert met $d_{\mathrm{BL}}(Z_n, N(0,1)) \to 0$ voor $n \to \infty$. Het is deze dalende rij getallen die het eigenlijk doel van de hele exercitie is: kunnen we het verloop vangen in een empirische formule, dan geeft ons dat misschien de mogelijkheid iets te zeggen over de orde waarmee $d_{\mathrm{BL}}(Z_n, N(0,1)) \to 0$.

We zien in de tabel dat de genoemde rij naar 0 nadert, zoals het hoort, maar het is moeilijk om iets te zeggen over de orde van de convergentiesnelheid.

n,m	1	4	9	19	49	99	199	499	999	1999
1	1.0000	0.7500	0.8203	0.8084	0.8020	0.7999	0.7989	0.7983	0.7981	0.7980
4	0.7500	1.0000	0.8490	0.7500	0.6683	0.6226	0.6020	0.5987	0.5986	0.5985
9	0.8203	0.8490	1.0000	0.8378	0.7202	0.6761	0.6578	0.6548	0.6547	0.6546
19	0.8084	0.7500	0.8378	1.0000	0.7961	0.7194	0.6775	0.6513	0.6457	0.6451
49	0.8020	0.6683	0.7202	0.7961	1.0000	0.8325	0.7386	0.6780	0.6570	0.6461
99	0.7999	0.6226	0.6761	0.7194	0.8325	1.0000	0.8318	0.7186	0.6781	0.6573
199	0.7989	0.6020	0.6578	0.6775	0.7386	0.8318	1.0000	0.7949	0.7185	0.6782
499	0.7983	0.5987	0.6548	0.6513	0.6780	0.7186	0.7949	1.0000	0.8313	0.7380
999	0.7981	0.5986	0.6547	0.6457	0.6570	0.6781	0.7185	0.8313	1.0000	0.8313
1999	0.7980	0.5985	0.6546	0.6451	0.6461	0.6573	0.6782	0.7380	0.8313	1.0000

Tabel 2: De verwachting $\mathbb{E}\widetilde{Z_n}\widetilde{Z_m}$ (na massaoverheveling).

n,m	1	4	9	19	49	99	199	499	999	1999
1	0.0000	0.7071	0.5995	0.6190	0.6293	0.6326	0.6342	0.6352	0.6355	0.6356
4	0.7071	0.0000	0.5496	0.7071	0.8145	0.8688	0.8922	0.8959	0.8960	0.8961
9	0.5995	0.5496	0.0000	0.5696	0.7480	0.8049	0.8272	0.8308	0.8310	0.8311
19	0.6190	0.7071	0.5696	0.0000	0.6386	0.7491	0.8031	0.8351	0.8418	0.8425
49	0.6293	0.8145	0.7480	0.6386	0.0000	0.5788	0.7231	0.8025	0.8283	0.8413
99	0.6326	0.8688	0.8049	0.7491	0.5788	0.0000	0.5800	0.7502	0.8023	0.8279
199	0.6342	0.8922	0.8272	0.8031	0.7231	0.5800	0.0000	0.6405	0.7504	0.8023
499	0.6352	0.8959	0.8308	0.8351	0.8025	0.7502	0.6405	0.0000	0.5808	0.7239
999	0.6355	0.8960	0.8310	0.8418	0.8283	0.8023	0.7504	0.5808	0.0000	0.5809
1999	0.6356	0.8961	0.8311	0.8425	0.8413	0.8279	0.8023	0.7239	0.5809	0.0000

Tabel 3: De bovengrens $\sqrt{2(1-\mathbb{E}\widetilde{Z_n}\widetilde{Z_m})}$ voor $d_{\mathrm{BL}}(Z_n, Z_m)$

7 Discussie en suggesties voor verder onderzoek

Het hoofddoel van deze scriptie is het bepalen van de snelheid waarmee afstanden in de ruimte $\mathcal{P}(\mathbb{R})$, in het kader van de CLS convergeren. Ten behoeve daarvan onderzochten we hoe een gestandaardiseerde rij i.i.d. Bernoulli-verdeelde stochasten Z_n naar de standaard-normale verdeling convergeert. Na een korte berekening van de Prokhorov-afstand $d_P(Z_n, N(0,1))$, die weinig informatie verschafte, zijn we verder gegaan met de berekening van de bounded-Lipschitzafstand $d_{\mathrm{BL}}(Z_n, N(0,1))$.

Het is gelukt om analytisch, in (26), een ondergrens te vinden voor deze afstand.

Het bepalen van een bovengrens dient om redenen, uiteengezet in secties 6.2.3 en 6.2.4, numeriek te gebeuren en daartoe hebben we computerprogramma's geschreven.

De berekeningen zijn uitgevoerd voor $n, m \leq 1999$ en resulteren in de tabellen 2 en 3. Bij vaste n en stijgende m lijken $\mathbb{E}\widetilde{Z_n}\widetilde{Z_m}$ en de bovengrens te convergeren. *Maar niet helemaal*: de convergentie "komt niet echt tot rust" en bekijken we een nevendiagonaal in tabel 3, dan is $d_{\mathrm{BL}}(Z_n, Z_m)$ bij benadering constant, maar.... men verwacht (Cauchy!) dat $d_{\mathrm{BL}}(Z_n, Z_m) \to 0$ langs zo'n diagonaal!

Misschien gaat er nog steeds teveel massa verloren in het overhevelingsproces van sectie 6.2.4. Of is er sprake van een nog onbekend effect.

Om dat met grotere zekerheid te kunnen zeggen, en om uitsluitsel te kunnen geven over de convergentiesnelheid van $d_{\mathrm{BL}}(Z_n, N(0,1))$ aan de hand van de getallen in de laatste kolom van tabel 3, had ik graag met nóg grotere waarden voor n en m gewerkt. Maar ook met de krachtige(r) computers van het Snelliusinstituut loop ik voor $n, m \geq 1999$ tegen de out-of-memoryboodschap van MATLAB aan.

Kortom: voor de ondergrens van $d_{\mathrm{BL}}(Z_n, N(0,1))$ hebben we, zoals gezegd, een keurige formule gevonden maar een dergelijk fraai resultaat voor de bovengrens hebben we dus (nog) niet tot onze beschikking.

Toch is er veel bereikt. De doelen zoals gesteld in de onderzoeksopdracht luidden immers:

- Is het mogelijk voor bepaalde stochasten de afstand tussen hun verdelingen uit te rekenen?
- Of goed af te schatten?
- Kan in bepaalde gevallen in de centrale limietstelling een snelheid van convergentie worden aangegeven ten opzichte van de genoemde metrieken?

En die doelen zijn in belangrijke mate gehaald; ik vond het aanvankelijk moeilijk voor te stellen, dat uit de onmogelijk abstracte definities voor de Prokhorov- en bounded-Lipschitzmetriek, getallen konden worden "geperst"!

Suggesties voor verder onderzoek zijn:

- Onderzoek of in het MATLAB-algoritme het massaverlies nog verder beperkt kan worden,
- Is er (op theoretisch niveau) sprake van een onbekend effect?
- Voer de berekeningen met nog krachtiger computers uit.

8 De MATLAB-bestanden

```
function EZt_nZt_m = verwachting(n,m)
%retourneert de verwachting EZt_nZt_m, waarin Zt_n (Ztilde_n) en Zt_m de stochasten...
%...na massaoverheveling naar diagonaal.

i=0:n; j=0:m;%indexvectoren = uitkomsten van de binominale verdelingen Bin(n,p) resp Bin(m,p)

p_i = binopdf(i,n,.5);%freq functions Z_n ~ Bin(n,1/2)
q_j = binopdf(j,m,.5);%freq functions Z_m ~ Bin(m,1/2)

%maak vierkant: versie met verveelvoudiging ("replicating"). Zie tekst scriptie
width=lcm(n+1,m+1); %lengte van vierkant
factor_x=width/(n+1); %Replicatioe factor voor y-axis <--> Z_m
factor_y=width/(m+1); %Replicatie factor voor x-axis <--> Z_n
```

```
%Verveelvoudig de gewichten.
p_i=replicate (p_i,factor_x);
q_j=replicate (q_j,factor_y);

%.....maar de gewichten moeten dan wel gecorrigeerd worden:
p_i=p_i/factor_x;
q_j=q_j/factor_y;

a = p_i' * q_j; %Opgerekt vierkant met geschaalde gewichten.

%De KERN v h algoritme: breng massa over naar de diagonaal.
c=min(a,a') - diag(diag(a));%matrix met c_ij=min(a_ij,a_ji) en nullen op de diagonaal
d=(sum(c,2))';%vector met de rijsommen van c
d=diag(d,0);%matrix met overal nullen, maar vector c op diagonaal
a=a + d - c; %het getransformeerde vierkant.

%Hevel restmassa van negatieve naar positieve kwadranten
a=RuimRestOp(a);

i = (2*i-n)/sqrt(n);%de waarden die Z_n kan aannemen:
j= (2*j-m)/sqrt(m);%Idem Z_m (bereik Z_m)

%Verveelvoudig de waarden in dit bereik:
i=replicate (i,factor_x);
j=replicate (j,factor_y);

ixj = i'*j;
ixjxa = ixj .* a;
EZt_nZt_m = sum(sum(ixjxa));

%===================================================================
function r=replicate (vector,factor)
%subfunctie van functie <verwachting>
%Elk element van de vector wordt verveelvoudigd met factor.
%Bijv. replicate([1 2 3],4) = [1 1 1 1 2 2 2 2 3 3 3 3]
%
r=repmat(vector,factor,1);
r=reshape(r,[1,prod(size(r))]);
%===================================================================
function r=RuimRestOp(a)
%subfunctie van functie <verwachting>
%Zet diagonalen op nul; doen niet mee aan a_{i,j} ^a_{-i,-j}
b = a - diag(diag(a));%veeg hoofddiagonaal schoon
%
% Gaan andere diagonaal op nul zetten:
c = flipud(b);%verklaar die diagonaal tijdelijk tot hoofddiagonaal.
d = c - diag(diag(c));%veeg die diagonaal schoon en....
```

```
d = flipud(d);%....draai weer terug naar zijn plaats:
%d is matrix a met de diagonalen op nul gezet.

%
e = fliplr(flipud(d));%d gespiegeld tov oorsprong
f = min(d,e);%doe de a_{i,j} ^a_{-i,-j}!!
%
%Willen nu de negatieve kwadranten optellen bij de positieve, maar niet andeersom!!
%Neutraliseer daarom de positieve kwadranten door die met 0 te vermenigvuldigen.
%Maak eerst binair array:
L = size(f,1);%Breedte matrix
v=1:L;%
v=mean(v) - v;%helft v is negatief
v=v'*v;%positieve en nagtieve kwadranten.
bin = (v<0);%binaire matrix; is nul op plekken waar a  >0
%
%Flip f om vert as en tel op bij a....
temp = f.*bin;%..... maar maak eerst de positieve kwadranten van f gelijk nul.
a = a +fliplr(temp);%tel nu de bijdragen v d negatieve kwadranten bij oorspr a op...

r=a-temp;%maar verwijder die dan weer van hun oorspronkelijke plek
%=================================================================


function table=bovengrens(n)
%Argument n is vector van getallen.
%Resultaat:
%tabel(:,:,1)= kruistabel van verwachtingen EZt_nZt_m.
%tabel(:,:,2)= idem van bovengrenzen sqrt(2*(1-table(:,:,1)))
%
length=size(n,2);%
table=zeros(length,length,2);

for i=1:length
  for j=1:length
    table(i,j,1)=verwachting(n(i),n(j));
  end
end

table(:,:,2)=sqrt(2.*(1.-table(:,:,1)));
```

Referenties

[1] van Gaans, Onno, *Probability measures on metric spaces, Notes of the seminar 'Stochastic Evolution Equations'*, Delft University of Technology, 2003.

[2] van Gaans, Onno, en Clément, Philippe, *Notes of the seminar Evolution Equations in Probability Spaces and the Continuity Equation*, Leiden University, Spring 2006.

[3] Williams, David, *Probability with Martingales*, Cambridge Mathematical Textbooks, 1991