Westfälische
Wilhelms-Universität
Münster

Seminararbeit

Importance Sampling

Michael Fleermann

12. Februar 2013

Institut für Mathematische Statistik
Prof. Dr. Matthias Löwe, Dr. Andrea Winkler

Inhaltsverzeichnis

1 Einführung

Die vorliegende Seminararbeit hat das Ziel, konventionelle Techniken zum Schätzen von Erwartungswerten (bzw. Integralen) mit der Technik des Importance Sampling zu vergleichen. Die behandelten Schätzmethoden sind die Crude-Monte-Carlo-Methode („i.i.d.-Sampling“) und die Markov-Chain-Monte-Carlo-Methode („dynamisches Sampling“). In beiden Fällen lässt sich Importance Sampling gewinnbringend einsetzen. Hierbei bedeutet „gewinnbringend“, dass eine Varianzreduktion des Schätzers erreicht werden kann.

Da Importance Sampling seinen Ursprung im i.i.d.-Sampling-Kontext hat, werden wir die Technik auch in diesem Kontext einführen und ein einfaches Beispiel kennenlernen, welches sowohl die Macht des Importance Sampling unterstreicht, als auch dessen Vulnerabilität offenlegt. Darauffolgend werden wir das Importance Sampling auf den dynamischen Sampling-Kontext übertragen.

Es gibt bei der Schätzung von Erwartungswerten mittels (MC)MC-Methoden zwei grundlegende Probleme:

1. Es ist oft notwendig, (MC-)Monte-Carlo-Simulationen mit unterschiedlichen Parametern zu wiederholen. Zum Beispiel kann man im Ising-Modell den Parameter β für die inverse Temperatur frei wählen und erhält somit ein von β abhängiges Wahrscheinlichkeitsmaß und darauf basierend verschiedene zu schätzende Erwartungswerte.

2. Man ist unter Umständen mit einer langsamen Konvergenz der Markov-Chain-Monte-Carlo-Verfahren konfrontiert.

Wir werden sehen, dass Importance Sampling beide Probleme unter günstigen Umständen asymptotisch lösen *kann*. Hierbei bedeutet „asymptotisch“, dass man den Stichprobenumfang gegen unendlich streben lässt.

2 Sampling im i.i.d.-Kontext

Sei $(S,\mathcal{S},\mu)$ ein W-Raum, wobei S eine endliche Menge bezeichne und $\mathcal{S} := \mathcal{P}(S)$ die Potenzmenge von S. Es gelte $\mu(\{s\}) > 0 \ \forall\, s \in S$. Es bezeichne ζ das Zählmaß auf $(S,\mathcal{S})$. Ist ν ein beliebiges W-Maß auf $(S,\mathcal{S})$, so bezeichnen wir mit p_ν die Zähldichte von ν, d.h. $p_\nu : S \longrightarrow \mathbb{R}$ wobei $p_\nu(s) = \nu(\{s\})$ für alle $s \in S$. Dann gilt also für jedes W-Maß ν auf $(S,\mathcal{S})$, dass $\nu = p_\nu \zeta$.

Sei nun $f : S \longrightarrow \mathbb{R}$ eine beliebige Funktion. Dann ist f trivialerweise messbar, beschränkt, und somit p-integrierbar für alle $p \geq 1$ bezüglich jeden W-Maßes auf $(S,\mathcal{S})$.

Die zentrale Fragestellung ist: Bestimme möglichst genau $\mathbb{E}_\mu f := \int_S f d\mu$.

2.1 Die Crude-Monte-Carlo-Methode

Simuliere eine Folge $(X_i^\mu)_{i\in\mathbb{N}}$ von i.i.d. Zufallsvariablen $X_i^\mu : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (S, \mathcal{S})$, wobei natürlich $\mathbb{P}^{X_i^\mu} = \mu$ gelte für alle $i \in \mathbb{N}$ und $(\Omega, \mathcal{A}, \mathbb{P})$ einen geeigneten Ausgangs-W-Raum darstelle. Es gilt dann für jedes $i \in \mathbb{N}$:

$$\mathbb{E}f(X_i^\mu) = \int_\Omega f(X_i^\mu) d\mathbb{P} = \int_S f d\mu = \mathbb{E}_\mu f$$

sowie

$$\mathbb{V}f(X_i^\mu) = \int_\Omega \Big(f(X_i^\mu) - \mathbb{E}(f(X_i^\mu))\Big)^2 d\mathbb{P} = \int_S (f - \mathbb{E}_\mu f)^2 d\mu = \mathbb{V}_\mu f$$

Für jedes $n \in \mathbb{N}$ setzen wir dann

$$\hat{f}_n := \frac{1}{n} \sum_{i=1}^n f(X_i^\mu)$$

Dann gilt:

- $\mathbb{E}\hat{f}_n = \frac{1}{n}\sum_{i=1}^n \mathbb{E}f(X_i^\mu) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\mu f = \mathbb{E}_\mu f$. Also ist $\hat{f}_n$ ein erwartungstreuer Schätzer für $\mathbb{E}_\mu f$.
- $\mathbb{V}\hat{f}_n = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}f(X_i^\mu) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}_\mu f = \frac{1}{n}\mathbb{V}_\mu f$.
- SLLN: $\hat{f}_n \longrightarrow \mathbb{E}_\mu f$ $\mathbb{P}$-f.s. für $n \to \infty$, das heißt $\hat{f}_n$ ist ein stark konsistenter Schätzer für $\mathbb{E}_\mu f$.
- CLT: $\sqrt{n}(\hat{f}_n - \mathbb{E}_\mu f) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}_\mu f)$ für $n \to \infty$. Für große n ist $\hat{f}_n$ also approximativ $\mathcal{N}\left(\mathbb{E}_\mu f, \frac{\mathbb{V}_\mu f}{n}\right)$-verteilt.

Um die Crude-Monte-Carlo-Methode anzuwenden, muss man in der Lage sein, die Verteilung μ i.i.d. zu simulieren. Ist dies der Fall, bilden die gerade aufgelisteten Qualitätsmerkmale zusammen mit der einfachen Implemetierung des Verfahrens schlagkräftige Argumente für die Verwendung der Crude-Monte-Carlo-Methode. Lässt sich μ hingegen nicht i.i.d.-simulieren, ist diese Methode nicht anwendbar.

2.2 Importance-Sampling

Das primäre Ziel des Importance Sampling ist die Erhöhung der Genauigkeit der Schätzergebnisse, indem man die Varianz des Schätzers verringert. Letzteres soll erreicht werden, indem man nicht gemäß μ, sondern gemäß eines anderen W-Maßes ν auf $(S, \mathcal{S})$ Stichproben erzeugt, wobei μ absolut stetig bezüglich ν ist (in Zeichen $\mu \ll \nu$).

Sei also ν ein solches W-Maß, dann existiert eine Dichte $\frac{d\mu}{d\nu}$ von μ bzgl. ν, also $\mu = \frac{d\mu}{d\nu}\nu$. Ferner gilt auch $p_\nu(s) > 0$ für alle $s \in S$. Nun gilt:

$$\mathbb{E}_\mu f = \int_S f d\mu = \int_S f \frac{d\mu}{d\nu} d\nu = \mathbb{E}_\nu f \frac{d\mu}{d\nu}$$

Setze $f^\nu := f\frac{d\mu}{d\nu}$, dann gilt $\mathbb{E}_\mu f = \mathbb{E}_\nu f^\nu$. Schätze nun $\mathbb{E}_\nu f^\nu$ mit der Crude-Monte-Carlo-Methode: Erzeuge eine Folge $(X_i^\nu)_{i\in\mathbb{N}}$ von i.i.d. ν-verteilten Zufallsvariablen $X_i^\nu : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (S, \mathcal{S})$. Es gilt also für alle $i \in \mathbb{N} : \mathbb{P}^{X_i^\nu} = \nu$. Setze wieder

$$\hat{f}_n^\nu := \frac{1}{n}\sum_{i=1}^n f^\nu(X_i^\nu) = \frac{1}{n}\sum_{i=1}^n \left(f\frac{d\mu}{d\nu}\right)(X_i^\nu).$$

Dann gilt nach Abschnitt 2.1:

- $\mathbb{E}\hat{f}_n^\nu = \mathbb{E}_\nu f^\nu = \mathbb{E}_\mu f$. Also ist $\hat{f}_n^\nu$ ein erwartungstreuer Schätzer für $\mathbb{E}_\mu f$.
- $\mathbb{V}\hat{f}_n^\nu = \frac{1}{n}\mathbb{V}_\nu f^\nu = \frac{1}{n}\mathbb{V}_\nu f\frac{d\mu}{d\nu}$
- SLLN: $\hat{f}_n^\nu \longrightarrow \mathbb{E}_\mu f$ $\mathbb{P}$-f.s. für $n \to \infty$ (starke Konsistenz)
- CLT: $\sqrt{n}(\hat{f}_n^\nu - \mathbb{E}_\mu f) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}_\nu f\frac{d\mu}{d\nu})$ für $n \to \infty$. Für große n ist $\hat{f}_n^\nu$ also approximativ $\mathcal{N}\left(\mathbb{E}_\mu f, \frac{\mathbb{V}_\nu f\frac{d\mu}{d\nu}}{n}\right)$-verteilt.

Importance Sampling ist also der Crude-Monte-Carlo-Methode vorzuziehen, wenn gilt:

$$\mathbb{V}_\nu f\frac{d\mu}{d\nu} < \mathbb{V}_\mu f.$$

Hierbei werden die Kosten der Simulation außer Acht gelassen. Es könnte zum Beispiel rechenintensiver sein, ν-verteilte Zufallsvariablen zu erzeugen, als μ-verteilte. Wir wollen uns jedoch mit dieser Problematik nicht näher beschäftigen. Stattdessen fragen wir uns, wie man denn ein W-Maß ν wählen kann, sodass eine Varianzverringerung tatsächlich eintritt. Diese Frage wollen wir anhand der folgenden Beispiele beantworten.

2.3 Beispiele

In diesem Abschnitt seien stets $S = \{1, 2, 3\}$ und $f = id_S$. Die W-Maße μ und ν mit $\mu \ll \nu$ besitzen die Zähldichten p_μ und p_ν, wobei per Generalvoraussetzung $p_\mu > 0$ gilt, und dann auch $p_\nu > 0$, da $\mu \ll \nu$. Dann gilt natürlich auch $\nu \ll \mu$. Insgesamt gilt:

- $\mu = p_\mu\zeta,\ \nu = p_\nu\zeta \quad$ sowie $\quad \mu = \frac{d\mu}{d\nu}\nu$
- $\frac{d\mu}{d\nu} = \frac{p_\mu}{p_\nu}$

Unser Ziel ist es, durch passende Wahl von ν die Varianz $\mathbb{V}_\nu f\frac{d\mu}{d\nu}$ klein zu bekommen. Es gilt:

$$\mathbb{V}_\nu f\frac{d\mu}{d\nu} = \sum_{i=1}^{3}\left(f(i)\frac{p_\mu(i)}{p_\nu(i)} - \mathbb{E}_\mu f\right)^2 p_\nu(i) \tag{1}$$

Offenbar wird dieser Term sehr klein, wenn gilt:

$$\forall i = 1,2,3 : f(i)\frac{p_\mu(i)}{p_\nu(i)} \approx E_\mu f \quad \text{bzw.} \quad \forall i = 1,2,3 : p_\nu(i) \approx \frac{f(i)p_\mu(i)}{\mathbb{E}_\mu f}. \tag{2}$$

Will man ν bestimmen, sodass $\mathbb{V}_\nu f\frac{d\mu}{d\nu}$ gering wird, ist dies äquivalent zu einer Bestimmung der Wahrscheinlichkeitsdichte p_ν, sodass $\mathbb{V}_\nu f\frac{d\mu}{d\nu}$ klein wird. Offenbar kann man alle $p_\nu(i)$ gemäß Gleichung (2) optimal wählen, sodass $\mathbb{V}_\nu f\frac{d\mu}{d\nu}$ ganz verschwindet, jedoch bräuchte man hierzu den exakten Wert $\mathbb{E}_\mu f$, welcher ja gerade geschätzt werden soll. (Ist Gleichung (2) erfüllt, erhält man offenbar ein Varianz-minimierendes ν_{opt}, dessen Dichte wir mit $p_{\nu_{opt}}$ bezeichnen.) Trotzdem lässt sich an Gleichung (1) folgende Heuristik ableiten: ν sollte dort viel W-Masse besitzten (d.h. p_ν sollte dort groß sein), wo f und/oder μ groß sind. Dies rechtfertigt die Namensgebung „Importance Sampling“, weil gerade die Bereiche mit großen Werten von f bzw. p_μ am meisten zu dem Erwartungswert $\mathbb{E}_\mu f$ beitragen.

Im Folgenden wird die Notation vereinfacht, indem wir Wahrscheinlichkeitsdichten p_ν bzw. W-Maße ν in kanonischer Weise mit Wahrscheinlichkeitsvektoren im $\mathbb{R}^3$ identifizieren.

2.3.1 Szenario 1

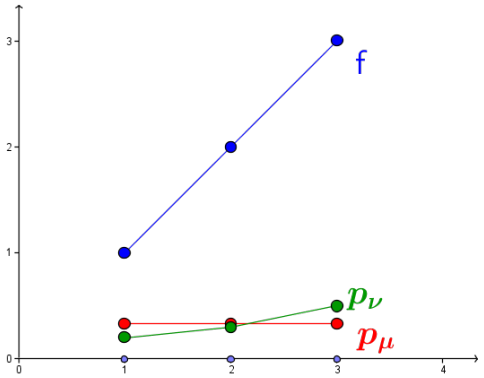
Sei $p_\mu = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, das heißt μ ist die Gleichverteilung auf $(S, \mathcal{S})$. Es gilt:

- $\mathbb{E}_\mu f = \left\langle \begin{pmatrix}\frac{1}{3}\\ \frac{1}{3}\\ \frac{1}{3}\end{pmatrix}, \begin{pmatrix}1\\2\\3\end{pmatrix} \right\rangle = 2 \quad \text{sowie} \quad \mathbb{E}_\mu f^2 = \left\langle \begin{pmatrix}\frac{1}{3}\\ \frac{1}{3}\\ \frac{1}{3}\end{pmatrix}, \begin{pmatrix}1\\4\\9\end{pmatrix} \right\rangle = \frac{14}{3}$
- $\mathbb{V}_\mu f = \frac{14}{3} - 4 = \frac{2}{3} = \frac{18}{27}$

Da p_μ konstant ist, können wir uns bei der Wahl von p_ν ganz auf die Funktion f konzentrieren, also p_ν mit f gemäß den obigen Überlegungen steigen lassen. Dies wird zum Beispiel durch die Wahl $p_\nu = \left(\frac{2}{10}, \frac{3}{10}, \frac{5}{10}\right)$ erreicht.

Es gilt dann

- $\mathbb{E}_\nu f\frac{p_\mu}{p_\nu} = \mathbb{E}_\mu f = 2$
- $\mathbb{E}_\nu (f\frac{p_\mu}{p_\nu})^2 = \left\langle \begin{pmatrix} \frac{2}{10} \\ \frac{3}{10} \\ \frac{5}{10} \end{pmatrix}, \begin{pmatrix} \frac{25}{9} \\ \frac{400}{81} \\ 4 \end{pmatrix} \right\rangle = \frac{109}{27}$
- $V_\nu f\frac{p_\mu}{p_\nu} = \frac{109}{27} - 2^2 = \frac{1}{27}$



Wir erreichen eine drastische Reduktion der Varianz. In der Tat liegt die von uns gewählte Verteilung p_ν sehr nahe an der optimalen Verteilung $p_{\nu_{opt}} = \left(\frac{1}{6}, \frac{2}{6}, \frac{3}{6}\right)$.

2.3.2 Szenario 2

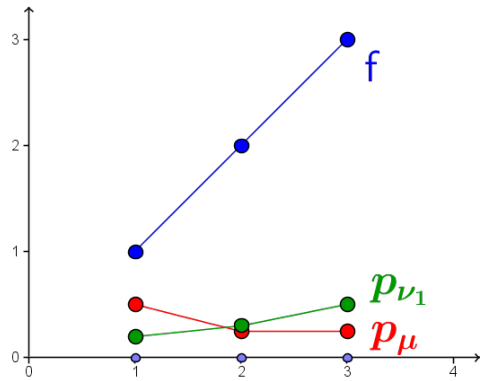
Gehen wir nun von $p_\mu = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$ aus. Nun muss p_ν nicht nur Ausschläge von f beachten, sondern auch von p_μ. Doch zunächst gilt

- $\mathbb{E}_\mu f = \left\langle \begin{pmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \right\rangle = \frac{7}{4}$
- $\mathbb{E}_\mu f^2 = \left\langle \begin{pmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{pmatrix}, \begin{pmatrix} 1 \\ 4 \\ 9 \end{pmatrix} \right\rangle = \frac{14}{4} = \frac{60}{16}$
- $\mathbb{V}_\mu f = \frac{60}{16} - \frac{49}{16} = \frac{11}{16} = \frac{66}{96}$

Wir betrachten zunächst die Verteilung p_{ν_1} aus Szenario 1, also $p_{\nu_1} = \left(\frac{2}{10}, \frac{3}{10}, \frac{5}{10}\right)$.

Dann gilt

- $\mathbb{E}_{\nu_1} f\frac{p_\mu}{p_{\nu_1}} = \frac{7}{4} = \frac{42}{24}$
- $\mathbb{E}_{\nu_1} \left(f\frac{p_\mu}{p_{\nu_1}}\right)^2 = \left\langle \begin{pmatrix} \frac{2}{10} \\ \frac{3}{10} \\ \frac{5}{10} \end{pmatrix}, \begin{pmatrix} \frac{25}{4} \\ \frac{25}{9} \\ \frac{9}{4} \end{pmatrix} \right\rangle = \frac{1848}{576}$
- $\mathbb{V}_{\nu_1} f\frac{p_\mu}{p_{\nu_1}} = \frac{1848}{576} - \frac{1764}{576} = \frac{14}{96}$

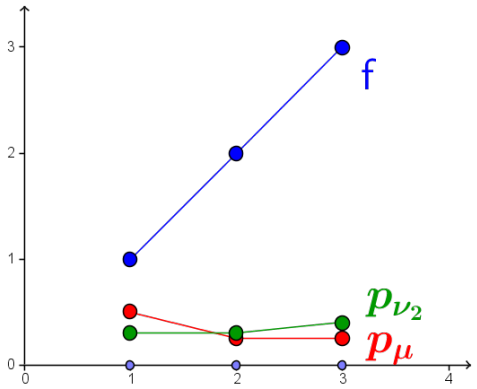


Wir erreichen also eine Varianzreduktion, welche jedoch noch nicht ganz so extrem ausfällt, wie im ersten Szenario.

Nun wollen wir auch berücksichtigen, dass p_μ der $1 \in S$ viel Wahrscheinlichkeitsmasse zuschreibt und setzen $p_{\nu_2} = \left(\frac{3}{10}, \frac{3}{10}, \frac{4}{10}\right)$.

Dann gilt

- $\mathbb{E}_{\nu_2} f \frac{p_\mu}{p_{\nu_2}} = \frac{7}{4}$
- $\mathbb{E}_{\nu_2} \left(f \frac{p_\mu}{p_{\nu_2}}\right)^2 = \left\langle \begin{pmatrix} \frac{3}{10} \\ \frac{3}{10} \\ \frac{4}{10} \end{pmatrix}, \begin{pmatrix} \frac{25}{9} \\ \frac{25}{9} \\ \frac{225}{64} \end{pmatrix} \right\rangle = \frac{295}{96}$
- $\mathbb{V}_{\nu_2} f \frac{p_\mu}{p_{\nu_2}} = \frac{295}{96} - \frac{294}{96} = \frac{1}{96}$



Durch eine Anpassung der dominierenden Dichte sowohl an f, als auch an p_μ erhalten wir wieder eine sehr starke Varianzreduktion. Es gilt hier $p_{\nu_{opt}} = \left(\frac{2}{7}, \frac{2}{7}, \frac{3}{7}\right)$.

Es gibt offenbar viele verschiedene Wahlen für das W-Maß ν, sodass beim Importance Sampling eine Varianzreduktion im Vergleich zur Crude-Monte-Carlo-Methode erreicht werden kann. Wie wir sehen, hängt die Stärke der Varianzminderung von einer geschickten Wahl von ν ab. An dieser Stelle sei angemerkt, dass eine ungünstige Wahl von ν auch zu einer starke Erhöhung der Varianz und damit zu einer deutlichen Verschlechterung des Schätzverfahrens führen kann.

Die Ausführungen in den gerade behandelten Beispielen vermitteln uns sowohl eine Vorstellung davon, wie ein dominierendes W-Maß ν beschaffen sein sollte, damit eine Varianzreduktion erreicht werden kann, als auch mögliche Größenordnungen einer solchen Varianzreduktion. Außerdem zeigen die Beispiele, wie volatil Importance Sampling ist. Es ist kein Verfahren, welches zum Erfolg führen *muss*. Jener steht und fällt mit der nichttrivialen Wahl des dominierenden W-Maßes ν.

3 Sampling im dynamischen Kontext

Nicht jede Verteilung μ oder ν lässt sich i.i.d.-simulieren, auch nicht, wenn $(S, \mathcal{S})$ endlich ist. Endlich bedeutet nicht unbedingt $|S| = 3$ oder $|S| = 1000$, sondern manchmal $|S| = 2^{(10^7)}$. Diese Zahl hat nicht eine oder drei Dezimalziffern, sondern etwas mehr als drei Millionen. Wenn eine Verteilung nur bis auf eine Normierungskonstante bekannt ist und man jene simulieren möchte, kann die direkte Simulation in angemessener Zeit unmöglich

sein, da man zunächst die Normierungskonstante berechnen müsste. Einen Ausweg bildet die Markov-Chain-Monte-Carlo-Methode. Wir wollen sie im Folgenden vorstellen, sowie die Technik des Importance Sampling auf MCMC-Verfahren übertragen.

3.1 Die Markov-Chain-Monte-Carlo-Methode

Lässt sich eine Verteilung μ auf $(S,\mathcal{S})$ nicht mit akzeptablen Kosten i.i.d.-simulieren, lässt sich ggf. die MCMC-Methode anwenden. Hierbei konstruiert man anstelle der i.i.d.-Folge eine irreduzible, aperiodische bzgl. μ reversible Markov Kette $X^\mu = (X_i^\mu)_{i\in\mathbb{N}}$. Diese konvergiert dann in Verteilung gegen μ und besitzt noch andere nützliche Konvergenzeigenschaften. Solche Markov-Ketten lassen sich zum Beispiel mit dem Metropolis-Algorithmus erzeugen, welcher auch ohne die Kenntnis der Normierungskonstanten der μ-Dichte auskommt.

Sei also $X^\mu = (X_i^\mu)_{i\in\mathbb{N}}$ eine solche Kette, die in Verteilung gegen μ konvergiert und in einem beliebigen Punkt startet. Als Schätzfunktion für $\mathbb{E}_\mu f$ wählen wir wieder das arithmetische Mittel der Beobachtungen. Wir setzen

$$\hat{f}_n := \frac{1}{n}\sum_{i=1}^{n} f(X_i^\mu).$$

Doch welche Eigenschaften hat $\hat{f}_n$? Die einzelnen Folgenglieder X_i^μ sind i.A. weder unabhängig, noch identisch verteilt. Man kann also auf die gewöhnlichen Grenzwertsätze (SSLN, CLT) nicht zurückgreifen. Auch ist der Schätzer $\hat{f}_n$ i.A. nicht mehr erwartungstreu, da die einzelnen Summanden nicht mehr μ-verteilt sind. Jedoch gelten andere nützliche Sätze speziell für solche Markov-Ketten wie oben geschildert:

Der (bzw. ein) Ergodensatz liefert ein Pendant zum SLLN:

$$\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n} f(X_i^\mu) \longrightarrow \int_S f d\mu = \mathbb{E}_\mu f \quad \mathbb{P}\text{-f.s. für } \; n\to\infty$$

Außerdem besagt der CLT für reversible Markov-Ketten, dass gilt:

$$\sqrt{n}(\hat{f}_n - \mathbb{E}_\mu f) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}_\mu^a f) \text{ für } n\to\infty$$

wobei gilt:

$$\mathbb{V}_\mu^a f \le \left(\frac{2}{\Delta(X^\mu)}\right)\mathbb{V}_\mu f \tag{3}$$

Hierbei ist $\mathbb{V}_\mu^a f$ die asymptotische Varianz der Zufallsvariablen $\sqrt{n}(\hat{f}_n - \mathbb{E}_\mu f)$. Ferner ist $\Delta(X^\mu)$ die Spektrallücke der Kette X^μ. Eine größere Spektrallücke liefert also genauere Schätzergebnisse. Wir stellen fest, dass die obere Schranke der Ungleichung (3) nicht

unbedingt angenommen werden muss. Wählt man z.B. als Markov-Kette eine i.i.d. Folge $(X_i^\mu)_{i\in\mathbb{N}}$ μ-verteilter Zufallsvariablen, so gilt $\Delta(X^\mu) = 1$ und somit $\mathbb{V}_\mu^a f \leq 2\mathbb{V}_\mu f$. Wir wissen jedoch $\mathbb{V}_\mu^a f = \mathbb{V}_\mu f$, siehe Abschnitt 2.1.

Anmerkung. Sei $h : (S,\mathcal{S}) \longrightarrow \mathbb{R}$ beliebig, ϑ ein beliebiges W-Maß auf $(S,\mathcal{S})$ sowie $X^\vartheta := (X_i^\vartheta)_{i\in\mathbb{N}}$ eine Markov-Kette mit Zustandsraum $(S,\mathcal{S})$, welche aperiodisch, irreduzibel und bzgl. ϑ reversibel ist. Dann bezeichnen wir mit $\mathbb{V}_\vartheta^a h$ die nach dem CLT für Markov-Ketten existente Varianz

$$\mathbb{V}_\vartheta^a h := \mathbb{V}\left(\lim_{n\to\infty}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n h(X_i^\vartheta) - \mathbb{E}_\vartheta h\right)\right) \text{ ,wobei gilt: } \mathbb{V}_\vartheta^a h \leq \left(\frac{2}{\Delta(X^\vartheta)}\right)\mathbb{V}_\vartheta h$$

3.2 Importance Sampling

Wir wollen nun die MCMC-Methode mit Importance Sampling verknüpfen in der Hoffnung, auch unser dynamisches Sampling zu verbessern.

3.2.1 Der einfache Importance-Sampling-Schätzer

Sei ν ein W-Maß auf $(S,\mathcal{S})$ mit $\mu \ll \nu$. Wir setzen wieder $f^\nu := f\frac{d\mu}{d\nu}$ und betrachten eine irreduzible, aperiodische, reversible Markov-Kette $X^\nu := (X_i^\nu)_{i\in\mathbb{N}}$, die in Verteilung gegen ν konvergiert. Wir setzen zunächst

$$\hat{f}_n^\nu := \frac{1}{n}\sum_{i=1}^n f^\nu(X_i^\nu) = \frac{1}{n}\sum_{i=1}^n \left(f\frac{d\mu}{d\nu}\right)(X_i^\nu).$$

und nennen diesen Schätzer den „einfachen Importance-Sampling-Schätzer“. Der Ergodensatz liefert:

$$\hat{f}_n^\nu = \frac{1}{n}\sum_{i=1}^n \left(f\frac{d\mu}{d\nu}\right)(X_i^\nu) \longrightarrow \int_S f\frac{d\mu}{d\nu}d\nu = \mathbb{E}_\mu f \quad \mathbb{P}\text{-f.s. für } n\to\infty$$

Der Schätzer $\hat{f}_n^\nu$ für $\mathbb{E}_\mu f$ ist also stark konsistent.

Bei Betrachtung des einfachen Importance-Sampling Schätzers erkennen wir, dass es bei der Verknüpfung von Importance Sampling mit der MCMC-Methode zu einem Problem kommen kann. Für die Simulation der Schätzfolge $\hat{f}_n^\nu$ ist offenbar die Kenntnis von $\frac{d\mu}{d\nu} = \frac{p_\mu}{p_\nu}$ erforderlich. Dabei ist es gerade einer der Vorteile der MCMC-Methode (bzw. des Metropolis-Algorithmus), die Verteilungen μ bzw. ν (approximativ) simulieren zu können, obwohl man ihre Dichten nur bis auf eine Normierungskonstante kennt. Das heißt die Kette $(X_i^\nu)_{i\in\mathbb{N}}$ ließe sich problemlos simulieren, nicht jedoch $\hat{f}_n^\nu$. Wir wollen dies etwas genauer betrachten und einen Ausweg aus der Misere finden. Dies führt zum Quotientenschätzer.

3.2.2 Der Quotientenschätzer

In der obigen Situation gelte $p_\mu(\cdot) = \frac{\tilde{p}_\mu(\cdot)}{C_\mu}$, $p_\nu(\cdot) = \frac{\tilde{p}_\nu(\cdot)}{C_\nu}$, wobei zwar die Funktionen $\tilde{p}_\mu$ bzw. $\tilde{p}_\nu$ bekannt sind, die Konstanten C_μ bzw. C_ν jedoch unbekannt. Es gilt für den einfachen Importance-Sampling-Schätzer

$$\hat{f}_n^\nu = \frac{1}{n}\sum_{i=1}^n \left(f\frac{d\mu}{d\nu}\right)(X_i^\nu) = \frac{1}{n}\sum_{i=1}^n f(X_i^\nu)\frac{\tilde{p}_\mu(X_i^\nu)C_\nu}{\tilde{p}_\nu(X_i^\nu)C_\mu} = \frac{C_\nu}{C_\mu}\frac{1}{n}\sum_{i=1}^n f(X_i^\nu)\frac{\tilde{p}_\mu(X_i^\nu)}{\tilde{p}_\nu(X_i^\nu)}$$

Es ist also klar, dass die Kenntnis der unbekannten Konstanten C_μ bzw. C_ν für die Berechnung von $\hat{f}_n^\nu$ notwendig ist. Wir möchten diese Konstanten also irgendwie loswerden, und dennoch $\mathbb{E}_\mu f$ möglichst gut schätzen. Wir wenden nun einen Trick an, indem wir im Wesentlichen durch 1 teilen. Dazu setzen wir:

$$\hat{1}_n^\nu := \frac{1}{n}\sum_{i=1}^n \frac{d\mu}{d\nu}(X_i^\nu) = \frac{1}{n}\sum_{i=1}^n \frac{\tilde{p}_\mu(X_i^\nu)C_\nu}{\tilde{p}_\nu(X_i^\nu)C_\mu} = \frac{C_\nu}{C_\mu}\frac{1}{n}\sum_{i=1}^n \frac{\tilde{p}_\mu(X_i^\nu)}{\tilde{p}_\nu(X_i^\nu)}$$

Nach dem Ergodensatz ist dies, also $\hat{1}_n^\nu$, ein stark konsistenter Schätzer für $\int_S \frac{d\mu}{d\nu}d\nu = \int_S d\mu = 1$. Also gilt

- $\hat{1}_n^\nu \longrightarrow 1$ $\mathbb{P}$-f.s.
- $\hat{f}_n^\nu \longrightarrow \mathbb{E}_\mu f$ $\mathbb{P}$-f.s., und somit
- $\frac{\hat{f}_n^\nu}{\hat{1}_n^\nu} \longrightarrow \mathbb{E}_\mu f$ $\mathbb{P}$-f.s.

Das heißt statt $\hat{f}_n^\nu$ lässt sich auch der sogenannte Quotientenschätzer $\frac{\hat{f}_n^\nu}{\hat{1}_n^\nu}$ verwenden, um $\mathbb{E}_\mu f$ stark konsistent zu schätzen. Es ergibt sich dann

$$\frac{\hat{f}_n^\nu}{\hat{1}_n^\nu} = \frac{\sum_{i=1}^n f(X_i^\nu)\frac{\tilde{p}_\mu(X_i^\nu)}{\tilde{p}_\nu(X_i^\nu)}}{\sum_{i=1}^n \frac{\tilde{p}_\mu(X_i^\nu)}{\tilde{p}_\nu(X_i^\nu)}}$$

Diese Schätzfolge lässt sich also ohne die Kenntnis der Konstanten C_ν, C_μ konstruieren. Hierbei lässt sich mit Hilfe des CLT für Markov-Ketten zeigen:

$$\sqrt{n}\left(\frac{\hat{f}_n^\nu}{\hat{1}_n^\nu} - \mathbb{E}_\mu f\right) \xrightarrow{d} \mathcal{N}(0, s_\mu^\nu(f)) \quad \text{für} \quad n \to \infty$$

wobei

$$s_\mu^\nu(f) := \mathbb{V}_\nu^a\left((f-\mathbb{E}_\mu f)\frac{d\mu}{d\nu}\right) \overset{\text{s.o.}}{\leq} \left(\frac{2}{\Delta(X^\nu)}\right)\mathbb{V}_\nu\left((f-\mathbb{E}_\mu f)\frac{d\mu}{d\nu}\right)$$

Die natürliche Frage ist nun, wie sich diese asymptotische Varianz vergleichen lässt mit der asymptotischen Varianz bei Anwendung der Markov-Chain-Monte-Carlo-Methode ohne Importance Sampling.

Anscheinend ist der Quotientenschätzer dem MCMC-Schätzer überlegen, wenn sich $s_\mu^\nu(f)$ besser abschätzen lässt als $\mathbb{V}_\mu^a f$, wenn also gilt:

$$\left(\frac{2}{\Delta(X^\nu)}\right)\mathbb{V}_\nu\left((f-\mathbb{E}_\mu f)\frac{d\mu}{d\nu}\right) < \left(\frac{2}{\Delta(X^\mu)}\right)\mathbb{V}_\mu f \tag{4}$$

Hier spielen nun zwei Faktoren eine Rolle: Sowohl die Varianzen, als auch die Spektrallücken. Es kann zum Beispiel sein, dass $\mathbb{V}_\nu\left((f-\mathbb{E}_\mu f)\frac{d\mu}{d\nu}\right) > \mathbb{V}_\mu f$ gilt, die Ungleichung (4) jedoch trotzdem erfüllt ist, weil die ν-approximierende Markov-Kette schneller mischt bzw. über eine größere Spektrallücke verfügt.

Diese Begebenheit wird unser weiteres Vorgehen bestimmen. Im i.i.d.-Kontext war Importance-Sampling vorteilhaft, wenn das dominierende W-Maß ν sowohl an μ, als auch an den Verlauf von f angepasst wurde. Wir möchten nun jedoch in der Lage sein, die Ungleichung (4) zu überprüfen, ohne f näher zu spezifizieren. Dies führt dazu, dass wir ν nur gemäß μ und der Spektrallücke $\Delta(X^\nu)$ einer ν approximierenden Markov-Kette wählen wollen. Es stellt sich heraus, dass sich Ungleichung (4) besonders gut überprüfen lässt, wenn der Dichtequotient $\frac{d\mu}{d\nu}$ durch ein $A \in \mathbb{R}_+$ beschränkt wird:

Satz 3.1. *Seien μ und ν Verteilungen auf $(S,\mathcal{S})$ mit $\mu \ll \nu$. Sei $A \in \mathbb{R}_+$, sodass gilt:*

$$\frac{d\mu}{d\nu}(s) \leq A \ \forall s \in S$$

Sei $(X_i^\nu)_{i\in\mathbb{N}}$ eine aperiodische, irreduzible, reversible Markov-Kette mit stationärer Verteilung ν. Dann gilt für jede Funktion $f : (S,\mathcal{S}) \longrightarrow (\mathbb{R},\mathcal{B})$:

1. $\mathbb{V}_\nu\left((f-\mathbb{E}_\mu f)\frac{d\mu}{d\nu}\right) \leq A \cdot V_\mu f$
2. $s_\mu^\nu(f) \leq \left(\frac{2A}{\Delta(X^\nu)}\right)\mathbb{V}_\mu f$

Beweis. Es gilt

$$\mathbb{E}_\nu\left((f-\mathbb{E}_\mu f)\frac{d\mu}{d\nu}\right) = \mathbb{E}_\mu(f-\mathbb{E}_\mu f) = 0$$

und somit

$$\begin{aligned}
\mathbb{V}_\nu\left((f-\mathbb{E}_\mu f)\frac{d\mu}{d\nu}\right) =& \mathbb{E}_\nu\left((f-\mathbb{E}_\mu f)\frac{d\mu}{d\nu}\right)^2 \\
=& \mathbb{E}_\mu(f-\mathbb{E}_\mu f)^2\frac{d\mu}{d\nu} \\
\leq& \mathbb{E}_\mu(f-\mathbb{E}_\mu f)^2 A \\
=& A \cdot \mathbb{V}_\mu f
\end{aligned}$$

Dies zeigt den ersten Teil des Satzes. Für den 2. Teil betrachte

$$\begin{aligned} s_\mu^\nu(f) =& \mathbb{V}_\nu^a \left((f - \mathbb{E}_\mu f) \frac{d\mu}{d\nu} \right) \\ &\leq \left(\frac{2}{\Delta(X^\nu)} \right) \mathbb{V}_\nu \left((f - \mathbb{E}_\mu f) \frac{d\mu}{d\nu} \right) \\ &\leq \left(\frac{2}{\Delta(X^\nu)} \right) \cdot A \cdot \mathbb{V}_\mu(f) \end{aligned}$$

□

Mit Hilfe von Satz 3.1 lässt sich nun feststellen: Ist der Dichtequotient $\frac{d\mu}{d\nu}$ beschränkt durch ein $A \in \mathbb{R}_+$, so ist Importance Sampling per Quotientenschätzer der gewöhnlichen MCMC-Methode überlegen, falls

$$\left(\frac{2A}{\Delta(X^\nu)} \right) \mathbb{V}_\mu f < \left(\frac{2}{\Delta(X^\mu)} \right) \mathbb{V}_\mu f \tag{5}$$

(Denn gilt die Ungleichung (5), so folgt mit Satz 3.1, dass sich $s_\mu^\nu(f)$ besser abschätzen lässt als $\mathbb{V}_\mu^a f$.) Dies ist nun eine sehr überschaubare Bedingung, mit der man gut arbeiten kann.

An dieser Stelle bemerken wir, wie uns Importance Sampling bei dem Hauptproblem 2 helfen kann: Wird der Dichtequotient $\frac{d\mu}{d\nu}$ nicht zu groß und ist ν wesentlich schneller mischend als μ, so ist die Ungleichung 5 erfüllt.

Im nächsten Abschnitt werden wir uns dem Hauptproblem 1 zuwenden.

3.2.3 Das Schätzen über Verteilungsfamilien

In vielen Fällen haben wir es nicht nur mit einem zugrunde liegenden W-Maß μ zu tun, sondern mit einer Familie $(\mu_\beta)_{\beta \in B}$, wobei dann $(\mathbb{E}_\beta f)_{\beta \in B}$ die zu schätzenden Größen darstellen. Um Importance Sampling in dieser Situation anzuwenden, muss das dominierende W-Maß ν dann bzgl. der Verteilungsfamilie $(\mu_\beta)_{\beta \in B}$ gewählt werden, sodass gilt: $\mu_\beta \ll \nu \; \forall \beta \in B$. Dieses Vorgehen liefert offenbar eine Lösung zu Hauptproblem 1. Man muss nicht mehr für jedes $\beta \in B$ eine Folge (approximativ) μ_β-verteilter Zufallsvariablen simulieren, sondern simuliert nur einmal eine Folge (approximativ) ν-verteilter Zufallsvariablen.

Um dies genauer zu analysieren, müssen wir zunächst unseren einfachen Importance-Sampling-Schätzer, den Schätzer für die 1 und den sich daraus ergebenden Quotientenschätzer anpassen. Wir definieren:

- $\forall \beta \in B : f_\beta^\nu := f \frac{d\mu_\beta}{d\nu}$
- $\forall \beta \in B : \hat{f}_{\beta,n}^\nu := \frac{1}{n} \sum_{i=1}^n f_\beta^\nu(X_i^\nu)$
- $\forall \beta \in B : \hat{1}_{\beta,n}^\nu := \frac{1}{n} \sum_{i=1}^n \frac{d\mu_\beta}{d\nu}(X_i^\nu)$

Mit den Ausführungen in Abschnitt 3.2.2 ergibt sich für die Quotientenschätzer

$$\forall \beta \in B : \ \frac{\hat{f}_{\beta,n}^\nu}{\hat{1}_{\beta,n}^\nu} = \frac{\sum_{i=1}^n f(X_i^\nu) \frac{\tilde{p}_{\mu_\beta}(X_i^\nu)}{\tilde{p}_\nu(X_i^\nu)}}{\sum_{i=1}^n \frac{\tilde{p}_{\mu_\beta}(X_i^\nu)}{\tilde{p}_\nu(X_i^\nu)}}$$

Dieser Schätzer kann nun für jedes $\beta \in B$ berechnet werden, wobei dazu nur einmal die Kette X^ν simuliert werden muss. Dieses Vorgehen bildet also in der Tat eine Lösung für unser Hauptproblem 1.

Es stellt sich jedoch die Frage, wie bzw. ob sichergestellt werden kann, dass die Schätzer $\hat{f}_{\beta,n}^\nu$ gleichmäßig über alle β gute Werte liefern. Im Ising-Modell ist zum Beispiel die Güte der Verteilungskonvergenz des Metropolis-Algorithmus stark von der inversen Temperatur β abhängig.

Eine intuitive Lösung bietet die folgende Herangehensweise: ν sollte so geschaffen sein, dass es alle Bereiche abdeckt, wo alle W-Maße μ_β konzentriert sind. Dies wird zum Beispiel erreicht, wenn eine Beschränktheitsbedingung eingeführt wird:

$$\frac{d\mu_\beta}{d\nu}(i) \leq A \ \forall i \in S, \ \forall \beta \in B \text{ für ein } A \in \mathbb{R}_+$$

oder äquivalent

$$p_{\mu_\beta}(i) \leq A \cdot p_\nu(i) \ \forall i \in S, \ \forall \beta \in B \text{ für ein } A \in \mathbb{R}_+.$$

Dann lässt sich Satz 3.1 auf diese Situation anwenden und liefert das folgende Ergebnis

Satz 3.2. *Seien $(\mu_\beta)_{\beta \in B}$ eine Verteilungsfamilie und ν eine Verteilung auf $(S,\mathcal{S})$ mit $\forall \beta \in B : \mu_\beta \ll \nu$. Sei $A \in \mathbb{R}_+$, sodass gilt:*

$$\frac{d\mu_\beta}{d\nu}(s) \leq A \ \forall s \in S, \ \forall \beta \in B$$

Sei $(X_i^\nu)_{i \in \mathbb{N}}$ eine aperiodische, irreduzible, reversible Markov-Kette mit stationärer Verteilung ν. Dann gilt für jede Funktion $f : (S,\mathcal{S}) \longrightarrow (\mathbb{R},\mathcal{B})$ und jedes $\beta \in B$:

$$s_{\mu_\beta}^\nu(f) \leq \left(\frac{2A}{\Delta(X^\nu)} \right) \mathbb{V}_{\mu_\beta}(f).$$

Beweis. Die Behauptung folgt sofort mit Satz 3.1. □

Das folgende Beispiel liefert eine interessante Anwendung von Satz 3.2

3.3 Beispiel: Gemischte Verteilung

Seien $D \in \mathbb{N}$ und $\mu_1, \ldots, \mu_D$ W-Maße auf $(S, \mathcal{S})$, f auf diesem Raum reellwertig, sodass uns für $\beta = 1, \ldots, D$ die Erwartungswerte $\mathbb{E}_{\mu_\beta} f$ interessieren. Setze dann

$$\nu = \frac{1}{D}(\mu_1 + \ldots + \mu_D)$$

so folgt

$$p_\nu = \frac{1}{D}(p_{\mu_1} + \ldots + p_{\mu_D})$$

sodass

$$\frac{d\mu_\beta}{d\nu} = \frac{p_{\mu_\beta}}{p_\nu} \leq D \quad \forall \beta = 1, \ldots, D$$

Satz 3.2 liefert

$$s^\nu_{\mu_\beta}(f) \leq \left(\frac{2D}{\Delta(X^\nu)}\right) \mathbb{V}_{\mu_\beta}(f).$$

Dies ist die Abschätzung der asymptotischen Varianz bei der Verwendung des Quotientenschätzers (Importance Sampling). Wird hingegen die gewöhnliche Markov-Chain-Monte-Carlo-Methode für jedes μ_β einzeln angewendet, wobei dann jeweils X^{μ_β} eine μ_β-approximierende Markov-Kette ist, wissen wir, dass die asymptotische Varianz abgeschätzt werden kann durch

$$\mathbb{V}^a_{\mu_\beta}(f) \leq \left(\frac{2}{\Delta(X^{\mu_\beta})}\right) \mathbb{V}_{\mu_\beta}(f)$$

Geht man zunächst von ähnlich großen Spektrallücken aus, das heißt

$$\forall \beta = 1, \ldots, D : \Delta(X^{\mu_\beta}) \approx \Delta(X^\nu),$$

dann ergibt sich, dass folgende Verfahren ähnlich gut funktionieren:

1. Verfahren:

Schätze jeden Wert $\mathbb{E}_{\mu_\beta} f$ mittels MCMC durch je n Samples einer μ_β-approximierenden Markov-Kette. Insgesamt müssen so $n \cdot D$ Samples erstellt werden. Für jedes $\beta \in B$ ist für große n der Schätzer $\hat{f}_{\beta,n} := \frac{1}{n} \sum_{i=1}^n f(X_i^{\mu_\beta})$ für $\mathbb{E}_{\mu_\beta} f$ gemäß $\mathcal{N}\left(E_{\mu_\beta} f, \frac{\mathbb{V}^a_{\mu_\beta}(f)}{n}\right)$-verteilt, wobei

$$\frac{\mathbb{V}^a_{\mu_\beta}(f)}{n} \leq \frac{\left(\frac{2}{\Delta(X^{\mu_\beta})}\right) V_{\mu_\beta}(f)}{n}$$

2. Verfahren:

Schätze jeden Wert $\mathbb{E}_{\mu_\beta} f$ durch $n \cdot D$ Samples einer ν-approximierenden Markov-Kette und wende dynamisches Importance Sampling mittels Quotientenschätzer an. Insgesamt müssen so also ebenfalls $n \cdot D$ Samples erstellt werden. Für jedes $\beta \in B$ ist für große n der Schätzer $\frac{\hat{f}^{\nu}_{\beta,n\cdot D}}{\hat{1}^{\nu}_{\beta,n\cdot D}}$ für $\mathbb{E}_{\mu_\beta} f$ gemäß obiger Schilderung $\mathcal{N}\left(\mathbb{E}_{\mu_\beta} f, \frac{s^{\nu}_{\mu_\beta}(f)}{n\cdot D}\right)$ verteilt, wobei mit Satz 3.2 gilt

$$\frac{s^{\nu}_{\mu_\beta}(f)}{n \cdot D} \leq \frac{\left(\frac{2\cdot D}{\Delta(X^{\nu})}\right) V_{\mu_\beta}(f)}{n \cdot D} = \frac{\left(\frac{2}{\Delta(X^{\nu})}\right) V_{\mu_\beta}(f)}{n}$$

Die Größe der Spektrallücke hängt natürlich sowohl mit dem verwendeten Algorithmus, als auch mit der zu approximierenden Verteilung zusammen. Besitzt die Dichte der zu approximierenden Verteilung mehrere steile Peaks und verwendet man den Metropolis-Algorithmus, dann muss man mit einer langsamen Konvergenz bzw. kleinen Spetrallücke rechnen. Gerade in solch einem Fall ist es also eine plausible Annahme, dass das 2. Verfahren zu einer echten Verbesserung der asymptotischen Varianz führt, da die Dichte p_ν eine Glättung der Dichten p_{μ_β} darstellt. Importance Sampling bietet dann also einen Lösungsansatz für das eingangs angesprochene Hauptproblem 2 bei dem Einsatz von MCMC-Methoden.

4 Fazit und Ausblick

In der vorliegenden Seminarausarbeitung haben wir in das Konzept des Importance Sampling eingeführt mit dem Ziel, die Varianz bzw. asymptotische Varianz der Monte-Carlo- bzw. Markov-Chain-Monte-Carlo-Methoden zu verringern und die eingangs erläuterten Hauptprobleme konventioneller Schätzmethoden zu lösen.

Wir haben gesehen, dass Importance Sampling das Ziel der Varianzreduktion erreichen und beide Probleme (unter bestimmten Umständen) lösen kann. Hierbei hängt der Erfolg des Verfahrens sowohl von der Wahl des dominierenden W-Maßes ν, als auch vom Simulationsverfahren für ν ab. Um solch ein ν geeignet zu wählen, haben wir bestimmte Heuristiken im dynamischen und im i.i.d.-Kontext kennengelernt.

Für folgende Untersuchungen würde es sich zunächst anbieten, die bereits gebotenen Heuristiken zu verfeinern oder neue zu entwickeln. Ferner ist es ratsam, die Wirksamkeit des Importance Sampling in bestimmten Modellen, etwa aus der statistischen Physik, zu untersuchen.