Chapter 5

Monte Carlo integration

In numerical integration, we select an appropriate method in accordance with the class of functions to which the integrand in question belongs. The smoother the integrand is, the more efficient numerical integration method we can apply to it. The more complicated the integrand is, the less efficient the applicable method becomes, and the final fort is the Monte Carlo integration.

§ 5.1 deals with integrands defined on $\mathbb{T}^1$ which are $\mathcal{B}_m$-measurable for some $m \in \mathbb{N}^+$. They are univariate functions, but among them there are significant probabilistic examples to which no deterministic numerical integration methods are applicable (Example 5.8 and Example 5.9). In this section, when $2^m \gg 1$, we show that i.i.d.-sampling or pairwise independent sampling is almost optimal in the sense of what is called L^2-robustness. § 5.2 deals with important facts of RWS which we did not mention in § 2.5.2. In § 5.4, we introduce a pairwise independent sampling for all simulatable random variables (§ 1.3), which is called the *dynamical random Weyl sampling*. It is the most reliable Monte Carlo integration method as far as the author knows.

5.1 L^2-robustness

In § 2.5, we mentioned about i.i.d.-sampling and random Weyl sampling (RWS) for numerical integration of functions of m coin tosses. Here we present a special characteristic of them among all kinds of random sampling methods.

Random variables dealt with in § 2.5 are functions on $\{0,1\}^m$, which are regarded as functions on D_m, or $\mathcal{B}_m$-measurable functions on $\mathbb{T}^1$, as was stated in § 1.1. Throughout Chapter 5, integrands are, mainly, functions on $\mathbb{T}^1$.

Theorem 5.1 *(A fundamental inequality about sampling [38]) Let $\{\psi_l\}_{l=1}^{2^m-1}$ be an orthonormal system of $L^2(\mathcal{B}_m)$ such that $\int_{\mathbb{T}^1} \psi_l(x)dx = 0$ holds for each l. Then for any sequence of random variables $\{X_n\}_{n=1}^{2^m} \subset \mathbb{T}^1$, the following inequality holds.*

$$\sum_{l=1}^{2^m-1} \mathbf{E}\left[\left|\frac{1}{N}\sum_{n=1}^{N}\psi_l(X_n)\right|^2\right] \geq \frac{2^m}{N} - 1, \qquad 1 \leq N \leq 2^m. \tag{5.1}$$

Proof. Since $\psi_l \in L^2(\mathcal{B}_m)$, we may assume that $\{X_n\}_{n=1}^{2^m} \subset D_m$. As a special case, each X_n may be deterministic. Conversely, if the theorem holds for any deterministic sequence

$\{x_n\}_{n=1}^{2^m}$, then it holds for any sequence of random variables $\{X_n\}_{n=1}^{2^m}$. So we will prove it for $\{x_n\}_{n=1}^{2^m} \subset D_m$.

First, let

$$g(y) := \frac{2^m}{N} \sum_{n=1}^{N} \mathbf{1}_{[x_n, x_n+2^{-m})}(y). \tag{5.2}$$

Then for any $f \in L^2(\mathcal{B}_m)$, we have

$$\frac{1}{N}\sum_{n=1}^{N} f(x_n) = \langle f, g\rangle_{L^2(\mathcal{B}_m)} := \int_{\mathbb{T}^1} f(x)g(x)dx.$$

Since a set of functions $\{1, \psi_1, \ldots \psi_{2^m-1}\}$ forms a complete orthonormal system of $L^2(\mathcal{B}_m)$, the Parseval identity (or Pythagoras' theorem) implies that

$$\|g\|^2_{L^2(\mathcal{B}_m)} = \langle g, 1\rangle^2_{L^2(\mathcal{B}_m)} + \sum_{l=1}^{2^m-1} \langle g, \psi_l\rangle^2_{L^2(\mathcal{B}_m)}. \tag{5.3}$$

Substitute $\langle g, 1\rangle_{L^2(\mathcal{B}_m)} = 1$ and an inequality

$$\begin{aligned}
\|g\|^2_{L^2(\mathcal{B}_m)} &= \frac{2^{2m}}{N^2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\int_{\mathbb{T}^1} \mathbf{1}_{[x_n,x_n+2^{-m})}\mathbf{1}_{[x_{n'},x_{n'}+2^{-m})}dx \\
&\geq \frac{2^{2m}}{N^2}\sum_{n=n'=1}^{N}\int_{\mathbb{T}^1} \mathbf{1}_{[x_n,x_n+2^{-m})}\mathbf{1}_{[x_{n'},x_{n'}+2^{-m})}dx \qquad (5.4)\\
&= \frac{2^{2m}}{N^2}\sum_{n=n'=1}^{N}\frac{1}{2^m} = \frac{2^m}{N}
\end{aligned}$$

for (5.3), and we see

$$\frac{2^m}{N} \leq 1 + \sum_{l=1}^{2^m-1}\left|\frac{1}{N}\sum_{n=1}^{N}\psi_l(x_n)\right|^2.$$

This completes the proof of (5.1). □

Corollary 5.2 *([38]) Let $2^m \geq N > 1$. For any sequence of random variables $\{X_n\}_{n=1}^{\infty} \subset \mathbb{T}^1$, there exists a non-constant function $f \in L^2(\mathcal{B}_m)$ which satisfies the following inequality.*

$$\mathbf{E}\left[\left|\frac{1}{N}\sum_{n=1}^{N} f(X_n) - \int_{\mathbb{T}^1} f(x)dx\right|^2\right] \geq \left(\frac{1}{N} - 2^{-m}\right)\mathbf{V}[f]. \tag{5.5}$$

Indeed, by Theorem 5.1, at least one ψ_l satisfies (5.5). According to Corollary 5.2, when $2^m \gg N \gg 1$, any sampling method admits an integrand for which the error is almost as large as or larger than i.i.d.-sampling.

If a sampling method — deterministic or random — is very efficient for a certain class of good integrands, we should not use it for integrands which are not in the class, because there must exist a bad integrand for which the error becomes larger than that of i.i.d.-sampling so as to satisfy the inequality (5.1). Thus, such a sampling method may be called a "high risk and high return"-method.

Example 5.3 Let $d_{-i}(n)$ denote the i-th digit of $n \in \mathbb{N}$ in its dyadic expression, i.e., $n = \sum_{i=1}^{\infty} d_{-i}(n)2^i$ (actually a finite sum). The sequence $\{x_n\}_{n=1}^{\infty} \subset \mathbb{T}^1$ defined by

$$x_n := \sum_{i=1}^{\infty} d_{-i}(n)2^{-i}, \quad n \in \mathbb{N}^+,$$

is called the *van der Corput sequence* (cf. [22]). First several terms are

$$\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}, \frac{9}{16}, \frac{5}{16}, \frac{13}{16}, \ldots.$$

About the convergence rate of the numerical integration by means of this sequence, it is known that for any function f of bounded variation,

$$\left| \frac{1}{N} \sum_{n=1}^{N} f(x_n) - \int_0^1 f(t)dt \right| \leq c(N) \|f\|_{BV} \times \frac{\log N}{N}, \quad N \geq 2,$$

where $\|f\|_{BV}$ denotes the total variation of f on $\mathbb{T}^1$, and $c(N)$ is a bounded coefficient, more concretely, $c(N) = \log(N+1)/(\log 2 \cdot \log N)$ will do (cf. [22]). In general, a sequence having this property is called a low discrepancy sequence, or a quasirandom sequence.

The numerical integration by means of van der Corput sequence (a quasi Monte Carlo method) has a much smaller error than i.i.d.-sampling when $\|f\|_{BV}$ is rather small. But when $\|f\|_{BV}$ is huge it may have a much greater error than i.i.d.-sampling. Indeed, for $f(x) = d_{30}(x)$, we have

$$\frac{1}{2^{28}} \sum_{n=1}^{2^{28}} d_{30}(x_n) = 0,$$

which is quite far from the true value $\int_0^1 d_{30}(t)dt = 1/2$.

In contrast to "high risk and high return"-methods, a low risk sampling method, i.e., a numerical integration method that produces stable approximate values for any integrands is said to be *robust*. We here give a quantitative definition of robustness.

Definition 5.4 *A numerical integration method by means of a sequence of random variables* $\{X_n\}_{n=1}^{2^m}$ *is said to be* L^2-robust *(more precisely,* $L^2(\mathcal{B}_m)$-robust*) if for any* $f \in L^2(\mathcal{B}_m)$ *it holds that*

$$\mathbf{E}\left[\left| \frac{1}{N} \sum_{n=1}^{N} f(X_n) - \int_{\mathbb{T}^1} f(x)dx \right|^2\right] \leq \frac{\mathbf{V}[f]}{N}, \quad 1 \leq N \leq 2^m. \tag{5.6}$$

Deterministic sampling is not robust in this sense. (Imagine numerical integration by means of the deterministic sequence $\{x_n\}_{n=1}^{N}$ applied to the function $g(y)$ defined by (5.2).)

Of course, i.i.d.-sampling is L^2-robust, because (5.6) holds as an equality. According to Theorem 5.1 and Corollary 5.2, when $2^m \gg N \gg 1$, the inequality (5.6) can hardly be improved. Thus, i.i.d.-ampling is almost optimal from the viewpoint of L^2-robustness.

RWS is also L^2-robust, because (5.6) holds as an equality. Besides, it works as a secure pseudorandom generator for numerical integration as is mentioned in § 2.5.2.

5.2 Random Weyl sampling (Part 2)

In this section, we look closely at the random Weyl sampling (RWS) introduced in § 2.5.2.

5.2.1 Degeneration of CLT-scaling limit

Since RWS is a numerical integration method by means of pairwise independent random variables, the sequence of its sample means satisfies law of large numbers (cf. [7]). We here show that the central limit theorem scaling (abbreviated as CLT-scaling) of the sample mean converges to 0 in probability (Theorem 5.6).

To this end, we consider RWS on $\mathbb{T}^1$, not on D_m. First, we look at pairwise independence. Theorem 5.5 below is a continuous version of Theorem 2.8.

Theorem 5.5 *([13, 44]) Let $(x, \alpha) \in \mathbb{T}^1 \times \mathbb{T}^1 = \mathbb{T}^2$ be a uniformly distributed random point. Then the sequence $\{x + n\alpha\}_{n\in\mathbb{Z}}$ of $\mathbb{T}^1$-valued random variables has the following properties; if $n \neq n'$, then $(x + n\alpha)$ and $(x + n'\alpha)$ are (pairwise) independent, and each $(x + n\alpha)$ is distributed uniformly in $\mathbb{T}^1$.*

Proof. For any bounded Borel functions F, G defined on $\mathbb{T}^1$, we see

$$\begin{aligned}
\int_{\mathbb{T}^1} d\alpha \int_{\mathbb{T}^1} dx\, F(x+n\alpha)G(x+n'\alpha) &= \int_{\mathbb{T}^1} d\alpha \int_{\mathbb{T}^1} dx\, F(x)G(x+(n'-n)\alpha) \\
&= \int_{\mathbb{T}^1} dx\, F(x) \int_{\mathbb{T}^1} d\alpha\, G(x+(n'-n)\alpha) \\
&= \int_{\mathbb{T}^1} dx\, F(x) \int_{\mathbb{T}^1} d\alpha\, G((n'-n)\alpha) \\
&= \int_{\mathbb{T}^1} dx\, F(x) \int_{\mathbb{T}^1} d\alpha\, G(\alpha). \qquad \square
\end{aligned}$$

As is well-known, Weyl transformation is ergodic on $(\mathbb{T}^1, \mathcal{B}, \mathbb{P})$, and hence the law of large numbers holds for any $F \in L^1(\mathbb{T}^1, \mathcal{B}, \mathbb{P})$. In particular, if F is smooth, then the law of large numbers converges fast ([22]). Indeed, for the function $\exp(2k\pi\sqrt{-1}\,x)$, $0 \neq k \in \mathbb{Z}$, we have

$$\frac{1}{N}\sum_{n=1}^{N} e^{2\sqrt{-1}\,\pi k(x+n\alpha)} = \frac{1}{N} \times \frac{1-e^{2\sqrt{-1}\,\pi N k\alpha}}{1-e^{2\sqrt{-1}\,\pi k\alpha}} \times e^{2\sqrt{-1}\,\pi k(x+\alpha)} = O(N^{-1}), \quad N \to \infty.$$

Since $\int_{\mathbb{T}^1} \exp(2k\pi\sqrt{-1}\,x)dx = 0$, thus the law of large numbers converges at the rate of $O(N^{-1})$. For general functions, we approximate it by finite Fourier series. Since the smoother a function is, the faster its Fourier coefficients converge to 0, in that case, the law of large numbers converges at a rate of nearly $O(N^{-1})$.

In RWS, the parameter $\alpha \in \mathbb{T}^1$ as well as $x \in \mathbb{T}^1$ is chosen at random. The chosen α is irrational with probability 1, and consequently, what we mentioned in the previous paragraph occurs with probability 1. This makes us imagine that the rate of convergence of the law of large numbers about RWS is faster than the rate about i.i.d.-samplimg. In fact, for $1 \leq p < 2$, about the p-th mean error of RWS, we have the following theorem.

Theorem 5.6 *([12, 44]) For any $F \in L^2(\mathbb{T}^1, \mathcal{B}, \mathbb{P})$ and any $1 \le p < 2$, it holds that*

$$\lim_{N\to\infty} \iint_{\mathbb{T}^1\times\mathbb{T}^1} \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \left(F(x+n\alpha) - \int_{\mathbb{T}^1} F(y)dy \right) \right|^p d\alpha\, dx \;=\; 0.$$

Consequently, for any $\varepsilon > 0$, it holds that

$$\mathbb{P}^2\left(\left\{ (x,\alpha) \in \mathbb{T}^2 \;\middle|\; \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \left(F(x+n\alpha) - \int_{\mathbb{T}^1} F(y)dy \right) \right| > \varepsilon \right\} \right) \longrightarrow 0, \quad N \to \infty. \quad (5.7)$$

i.e., the limit distribution of the CLT-scaling of the sample mean degenerates.

Proof. Without loss of generality, we may assume $\int_{\mathbb{T}^1} dx\, F(x) \;=\; 0$. For each $M \in \mathbb{N}^+$, define a function $F_M : \mathbb{T}^1 \to \mathbb{R}$ by

$$F_M(t) \;:=\; \sum_{|l|\le M} \widehat{F}(l) e^{2\sqrt{-1}\pi l t},$$

where $\widehat{F}(l)$ denotes the Fourier coefficient of F ;

$$\widehat{F}(l) \;=\; \int_{\mathbb{T}^1} dt\, F(t) e^{-2\sqrt{-1}\pi l t}.$$

Note that $\int_{\mathbb{T}^1} dtF(t) = 0$ implies $\widehat{F}(0) = 0$. Fix any $1 < p < 2$. By the triangular inequality, Hölder's inequality and Theorem 5.5, we have

$$\begin{aligned}
\left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} F(x+n\alpha) \right\|_p &:= \left(\iint_{\mathbb{T}^1\times\mathbb{T}^1} d\alpha dx \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} F(x+n\alpha) \right|^p \right)^{\frac{1}{p}} \\
&\le \left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} F_M(x+n\alpha) \right\|_p + \left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (F - F_M)(x+n\alpha) \right\|_p \\
&\le \left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} F_M(x+n\alpha) \right\|_p + \left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (F - F_M)(x+n\alpha) \right\|_2 \\
&= \left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} F_M(x+n\alpha) \right\|_p + \sqrt{\mathbf{V}(F - F_M)}. \qquad (5.8)
\end{aligned}$$

Let us compute the first term of the last side of (5.8) in detail. By the definition of F_M,

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} F_M(x+n\alpha) \;=\; \sum_{0<|l|\le M} \left(\widehat{F}(l) e^{2\sqrt{-1}\pi l x} \times \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{2\sqrt{-1}\pi n l \alpha} \right).$$

Taking $L^p(\mathbb{T}^2, d\alpha dx)$-norm, we see

$$\begin{aligned}
\left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} F_M(x+n\alpha) \right\|_p &\le \sum_{0<|l|\le M} \left| \widehat{F}(l) \right| \left(\int_{\mathbb{T}^1} d\alpha \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{2\sqrt{-1}\pi n l \alpha} \right|^p \right)^{1/p} \\
&= \sum_{0<|l|\le M} \left| \widehat{F}(l) \right| \left(\int_{\mathbb{T}^1} d\alpha \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{2\sqrt{-1}\pi n \alpha} \right|^p \right)^{1/p},
\end{aligned}$$

where we used the fact that the transformation $\mathbb{T}^1 \ni \alpha \mapsto l\alpha \in \mathbb{T}^1$ preserves the Lebesgue measure. And then

$$\begin{aligned}
\int_{\mathbb{T}^1} d\alpha \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{2\sqrt{-1}\pi n\alpha} \right|^p &= \int_0^{\frac{1}{2}} d\alpha \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{2\sqrt{-1}\pi n\alpha} \right|^p + \int_{\frac{1}{2}}^{1} d\alpha \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{2\sqrt{-1}\pi n\alpha} \right|^p \\
&= 2\int_0^{\frac{1}{2}} d\alpha \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{2\sqrt{-1}\pi n\alpha} \right|^p \qquad (\alpha \mapsto 1-\alpha) \\
&= 2\int_0^{\frac{1}{2}} d\alpha \left| \frac{1}{\sqrt{N}} \frac{\sin \pi N\alpha}{\sin \pi\alpha} \right|^p \\
&= 2\int_0^{\frac{N}{2}} \frac{dt}{N} \left| \frac{1}{\sqrt{N}} \frac{\sin \pi t}{\sin \pi \frac{t}{N}} \right|^p \qquad (N\alpha \mapsto t) \\
&= 2\left(\frac{1}{N}\right)^{\frac{p}{2}+1} \int_0^{\frac{N}{2}} dt \left| \frac{\pi \frac{t}{N}}{\sin \pi \frac{t}{N}} \right|^p \left| \frac{\sin \pi t}{\pi t} \right|^p N^p \\
&= 2\left(\frac{1}{N}\right)^{1-\frac{p}{2}} \int_0^{\frac{N}{2}} dt \left| \frac{\pi \frac{t}{N}}{\sin \pi \frac{t}{N}} \right|^p \left| \frac{\sin \pi t}{\pi t} \right|^p \\
&< \left(\frac{1}{N}\right)^{1-\frac{p}{2}} 2\left(\frac{\pi}{2}\right)^p \int_0^{\infty} dt \left| \frac{\sin \pi t}{\pi t} \right|^p,
\end{aligned}$$

where we used the fact that $0 < y < \pi/2$ implies $y/\sin y < \pi/2$. Now we see

$$\left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} F_M(x+n\alpha) \right\|_p \leq \sum_{0<|l|\leq M} \left| \widehat{F}(l) \right| \left(\int_{\mathbb{T}^1} d\alpha \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} e^{2\sqrt{-1}\pi n\alpha} \right|^p \right)^{1/p} \underset{N\to\infty}{\longrightarrow} 0,$$

and finally,

$$\overline{\lim_{N\to\infty}} \left\| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} F(x+n\alpha) \right\|_p \leq \sqrt{\mathbf{V}(F-F_M)} \underset{M\to\infty}{\longrightarrow} 0.$$

□

Remark 5.7 Theorem 5.6 can be extended for square integrable functions of several variables. For details, see Theorem 5.20.

From the viewpoint of numerical integration, the convergence in probability (5.7) is much more desirable than the central limit theorem. But for careful readers, we add a remark. Since RWS satisfies

$$\int_{\mathbb{T}^2} \left| \frac{1}{N} \sum_{n=1}^{N} F(x+n\alpha) - \int_{\mathbb{T}^1} F(y)dy \right|^2 dxd\alpha = \frac{\mathbf{V}[F]}{N} \tag{5.9}$$

(cf. Theorem 2.8), if the event of the left hand side of (5.7) should unfortunately occur, the error of the sampling would be very large. Let us consider RWS mentioned in Definition 2.7 and Theorem 2.8. If we should choose $\alpha = 0 \in D_{m+j}$, then $X_n(x,\alpha) = \lfloor x \rfloor_m \in D_m$ for all n, which is the worst sequence for sampling (cf. Remark 2.11). The probability of such an event is $2^{-(m+j)}$, which is much greater than the probability of the same event for i.i.d.-sampling. However, when m is not so small, since the probability of such a very bad event is extremely small, we need not be anxious about it in practice. On the other hand, since (5.9) must be satisfied, when such a bad event does not occur, the error of RWS must be smaller than that of i.i.d.-sampling. As a result, RWS is preferable to i.i.d.-sampling.

Example 5.8 To see the effect of Theorem 5.6 and what we mentioned in the last paragraph, we computed the distribution of $S_{10^6}(g(\omega'))/10^6$ of Example 2.9 by a Monte Carlo method. Using the pseudorandom generator by means of Weyl transformation, we generated 50,000 random seeds $\omega' = (x, \alpha) \in D_{119} \times D_{119}$, and investigated the frequency distribution of $S_{10^6}(g(\omega'))/10^6$.

Table 5.1: Sample mean and sample SD of RWS

Range of data	Sample mean	Sample SD	SD of $\mathcal{N}(0,1)$
Whole	0.546095	0.000434905	1.000000
Central 99.9%	0.546094	0.000307281	0.993631
Central 99%	0.546095	0.000262653	0.956823

The true value of the sample standard deviation (abbreviated as SD) is approximately

$$\sqrt{\frac{0.546095 \cdot (1 - 0.546095)}{10^6}} = 0.000497871,$$

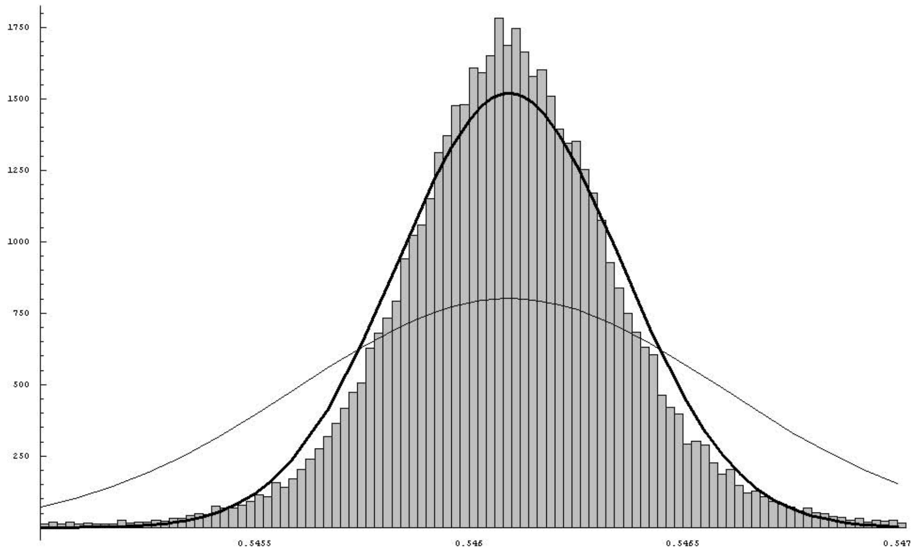
which almost coincides with 0.000434905, the value computed from the 50,000 trials (Table 5.1). The second row of Table 5.1 shows the sample mean and the sample SD of the central 99.9% of the 50,000 samples, i.e., whole data except the smallest 25 data and the largest 25 data. In this case, the sample mean does not change but the sample SD decreases to about 3/4 of the one in the first row. This shows that those exceptional 50 data are far from the mean. The third row shows the sample mean and the sample SD of the central 99% of the 50,000 samples, i.e., whole data except the smallest 250 data and the largest 250 data. In this case, the sample mean does not change, either, but the SD decreases to about 3/5 of the one in the first row. (Similar calculations are done for $\mathcal{N}(0,1)$, which are written in the rightmost column.)

Figure 5.1 shows the frequency distribution of 50,000 samples of $S_{10^6}(g(\omega'))/10^6$. In this figure, the thick curve shows the density function of $\mathcal{N}(0.546095, 0.000262653^2)$, whose mean and SD are same as the third row of Table 5.1. Comparing with this, the distribution of $S_{10^6}(g(\omega'))/10^6$ is more concentrated around the mean and has thicker tails. And the thin curve is the density function of $\mathcal{N}(0.546095, 0.000497871^2)$, which exactly approximates the distribution of $S_{10^6}(\omega)/10^6$ (i.i.d.-sampling). Obviously, RWS is much more preferable to i.i.d.-sampling.

5.2.2 RWS in case $m \gg 1$

When applying RWS to a random variable $\{0,1\}^m \to \mathbb{R}$, Alice chooses an $\omega' \in \{0,1\}^{2m+2j}$. But if m is huge, by the problem of random number again, it would be impossible for her

Figure 5.1: The frequency distribution of 50,000 samples of $S_{10^6}(g(\omega'))/10^6$



to choose ω'. In such a case, Alice chooses ω' with the help of an auxiliary pseudorandom generator $g' : \{0,1\}^n \to \{0,1\}^{2m+2j}$ (cf. Remark 2.12).

In selecting the auxiliary pseudorandom generator g', the drastic reduction of randomness of RWS has the following advantages;

(1) RWS is very insensitive to the quality of pseudorandom generator, i.e., a cheap pseudorandom generator may work well as the auxiliary one.

(2) With almost no slowdown of generating speed of samples, we can use a slow but precise pseudorandom generator, such as a cryptographically secure one, to get most reliable results.

In particular, (2) is important in that it shows the speed of pseudorandom generation is not an important factor in choosing the auxiliary pseudorandom generator g'.

Example 5.9 Let

$$S^{(m)}(x) := \sum_{i=1}^{m} d_i(x), \quad x \in \mathbb{T}^1.$$

In order to see the distribution of $S^{(500)}$ under the Lebesgue probability measure $\mathbb{P}$, we made a histogram of the frequency distribution (Figure 5.2),

$$p_k^{(500)}(N) := \frac{1}{N}\#\left\{1 \le n \le N \,\middle|\, S^{(500)}(x+n\alpha) = k\right\}, \quad k = 0, 1, \ldots, 500,$$

by RWS. Here the seed $(x, \alpha) \in D_{523} \times D_{523}$[†1] is 1046 bit long, which is generated by the pseudorandom generator by means of Weyl transformation (for implementation, see § 6.1.2).

Figure 5.2: The frequency distribution $p_k^{(500)}(N)$, $k = 0, \dots, 500$.

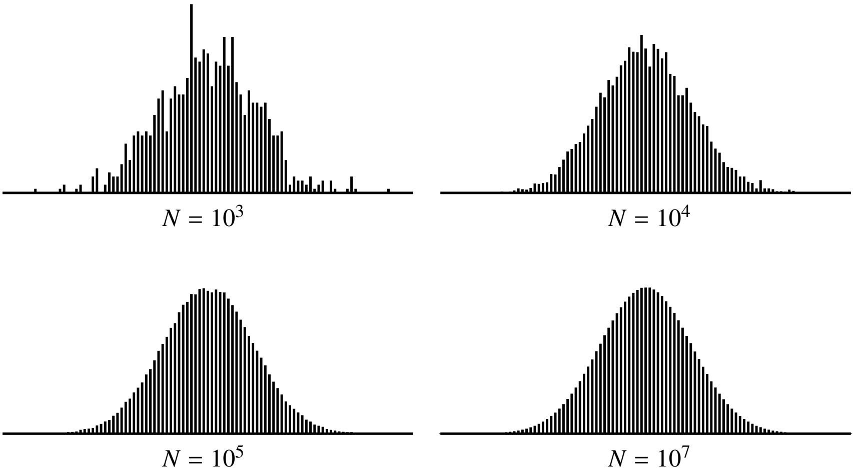
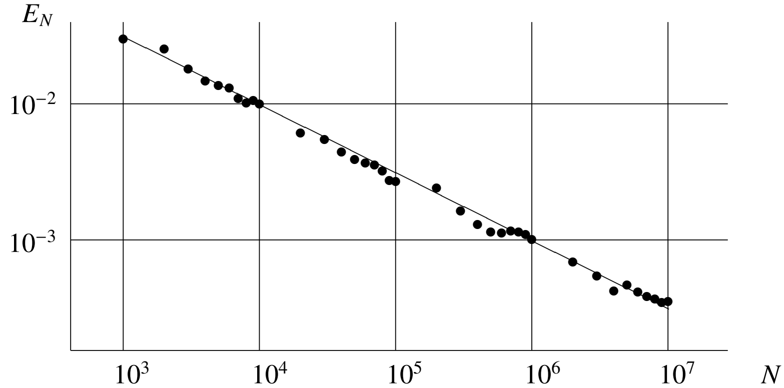


Figure 5.3: Decay of error (log-log scale)



As $N \to \infty$, this frequency distribution approaches to a binary distribution. In fact, as is shown by dots in Figure 5.3, the error

$$E_N := \sqrt{\sum_{k=0}^{500} \left| p_k^{(500)}(N) - q_k^{(500)} \right|^2}, \qquad q_k^{(500)} := \frac{500!}{(500-k)!k!} \times 2^{-500},$$

becomes small as $N \to \infty$. In Figure 5.3, the horizontal axis indicates logarithm of the sample size N (at most 10^7), and the vertical axis indicates logarithm of E_N. The slanting

[†1] $523 = 500 + \lceil \log_2 10^7 \rceil - 1$.

straight line is the graph of the expected values

$$\sqrt{\sum_{k=0}^{500} q_k^{(500)}(1-q_k^{(500)})} \times N^{-1/2} \;=\; 0.03122139 \times N^{-1/2}$$

in the case of i.i.d.-sampling. This figure reads that RWS and i.i.d.-sampling do not much differ for the computation of the distribution of $S^{(500)}$.

Theorem 5.6 asserts that the convergence of RWS is theoretically much faster than that of i.i.d.-sampling. However, in general, it is known that when we apply a quasi-Monte Carlo method to a function of so many independent variables which depends on those variables almost equally, the rate of convergence is observed as slow as i.i.d.-sampling (cf. [34] p.99). Indeed, when we apply RWS to such a function, the advantage of fast convergence does not appear unless the sample size N becomes astronomically huge.

In the case of Example 5.9, $S^{(m)}(x)$ equally depends on m independent variables $d_i(x)$, $i = 1, \ldots, m$. As a matter of fact, when $m \gg 1$, it is shown by the following theorem that the RWS samples of $S^{(m)}(x)$ look like i.i.d. random variables.

Theorem 5.10 *([8]) For almost every irrational α, a stochastic process*

$$\left\{ 2m^{-1/2}\left(S^{(m)}(\,\bullet\, + n\alpha) - \frac{m}{2}\right)\right\}_{n=0}^{\infty}$$

defined on the Lebesgue probability space $(\mathbb{T}^1, \mathcal{B}, \mathbb{P})$ converges to $\mathcal{N}(0,1)$-i.i.d. random variables as $m \to \infty$ in the sense of finite dimensional distribution.

As $m \to \infty$, each one dimensional distribution of this process converges to $\mathcal{N}(0,1)$ by the central limit theorem. But it is not easy to see that the process converges to a Gaussian process. Here instead, we only confirm that the two-term correlation disappears as $m \to \infty$.

Noting properties of Rademacher functions $\{r_i(x) := 1 - 2d_i(x)\}_{i=1}^{\infty}$

$$r_i(x) \;=\; r_1(2^{i-1}x), \qquad \forall c, \quad \int_{\mathbb{T}^1} r_i(x) r_j(x+c)\,dx \;=\; 0, \quad i \neq j,$$

let us calculate the correlation function. Let

$$\varphi(\alpha) \;:=\; \int_{\mathbb{T}^1} r_1(x) r_1(x+\alpha)\,dx \;=\; |2 - 4\alpha| - 1. \quad (\text{cf.}\,(4.83))$$

Since $S^{(m)}(x) - \frac{m}{2} \;=\; -\frac{1}{2}\sum_{i=1}^{m} r_i(x)$,

$$\begin{aligned}
&\int_{\mathbb{T}^1}\left(m^{-1/2}\left(S^{(m)}(x) - \frac{m}{2}\right)\right)\left(m^{-1/2}\left(S^{(m)}(x+n\alpha) - \frac{m}{2}\right)\right)dx \\
&\quad = \frac{1}{4m}\sum_{i=1}^{m}\sum_{j=1}^{m}\int_{\mathbb{T}^1} r_i(x) r_j(x+n\alpha)\,dx \\
&\quad = \frac{1}{4m}\sum_{i=1}^{m}\int_{\mathbb{T}^1} r_i(x) r_i(x+n\alpha)\,dx
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{4m}\sum_{i=1}^{m}\int_{\mathbb{T}^1} r_1(2^{i-1}x)r_1(2^{i-1}x+2^{i-1}n\alpha)dx \\
&= \frac{1}{4m}\sum_{i=1}^{m}\int_{\mathbb{T}^1} r_1(x)r_1(x+2^{i-1}n\alpha)dx \;=\; \frac{1}{4m}\sum_{i=1}^{m}\varphi(2^{i-1}n\alpha). \qquad (5.10)
\end{aligned}
$$

By the ergodicity of the transformation $x \mapsto 2x$, the last term converges to

$$\frac{1}{4}\int_{\mathbb{T}^1}\varphi(x)dx = 0$$

for almost every α as $m \to \infty$ ([15]). Thus the dependency vanishes in the limit.

However, the convergence of (5.10) to 0 is slow, as slow as $O(m^{-1/2})$, which follows from the fact that the central limit theorem and the law of iterated logarithm hold for the sum (5.10) ([8, 15, 25]). Indeed, for $m = 500$, the process $\{S^{(m)}(\bullet + n\alpha)\}_{n=0}^{\infty}$ is not so close to $\mathcal{N}(0,1)$-i.i.d. random variables. Intuitively speaking, when m is large, $S^{(m)}(x)$ and $S^{(m+1)}(x)$ are not so much different for each x, and hence the rate of dependence disappearance cannot be fast. Conversely, we might say that for a function which depends on many independent variables equally and sharply, the rate of dependence disappearance would be fast. In fact, the function $G_m(x) = S^{(m)}(x) \bmod 2$ (cf. (4.7)) depends on each $d_i(x)$ equally, and in addition, its value sensitively changes as soon as one of $d_i(x)$ changes, and hence the dependence disappearance occurs very fast. This observation explains Theorem 4.14 from another point of view.

5.2.3 Another example of pairwise independent random variables

Let GF(2^m) denote the finite field (Galois field) of order 2^m. Let us identify $\{0,1\}^m$, $\{1,2,3,\dots,2^m\}$, and GF(2^m), by any two bijections; ϕ : GF$(2^m) \to \{0,1\}^m$ and ψ : GF$(2^m) \to \{1,2,3,\dots,2^m\}$. For each $\omega' := (x,\alpha) \in \mathrm{GF}(2^m)\times\mathrm{GF}(2^m) \cong \{0,1\}^{2m}$, define

$$\tilde{Z}_n(\omega') := x + n\alpha, \quad n \in \mathrm{GF}(2^m) \cong \{1,2,3,\dots,2^m\}.$$

Then under P_{2m}, the sequence of random variables $\{\tilde{Z}_n\}_{n=1}^{2^m}$ is pairwise independent, and each $\tilde{Z}_n$ is distributed uniformly in $\{0,1\}^m$ ([24] Lecture 5). To see this, let us show that for any $a,b \in \mathrm{GF}(2^m) \cong \{0,1\}^m$, $1 \le n < n' \le 2^m$, we have

$$P_{2m}(\tilde{Z}_n(\omega') = a,\, \tilde{Z}_{n'}(\omega') = b) \;=\; 2^{-2m}.$$

Consider the following system of linear equations with coefficients in GF(2^m)

$$\begin{cases} x + n\alpha &= a, \\ x + n'\alpha &= b, \end{cases}$$

where x, α are unknowns. Let $(x_0,\alpha_0) \in \mathrm{GF}(2^m)\times\mathrm{GF}(2^m)$ be the unique solution to the equations. Then we see

$$P_{2m}(\tilde{Z}_n(\omega') = a,\, \tilde{Z}_{n'}(\omega') = b) \;=\; P_{2m}\left(\{(x_0',\alpha_0')\}\right) \;=\; 2^{-2m},$$

which completes the proof.

As Definition 2.7, define a pseudorandom generator $\tilde{g} : \{0,1\}^{2m} \to \{0,1\}^{Nm}$, $N \leq 2^m$, by

$$\tilde{g}(\omega') := (\tilde{Z}_1(\omega'), \tilde{Z}_2(\omega'), \ldots, \tilde{Z}_N(\omega')) \in \mathrm{GF}(2^m)^N \cong \{0,1\}^{Nm}.$$

Then $\tilde{g}$ is secure for numerical integration of random variables on $(\{0,1\}^m, 2^{\{0,1\}^m}, P_m)$.

In order for two independent D_m-valued uniform random variables to exist, the sample space should be equal to or larger than $D_m \times D_m$. Therefore the size $2m$ of the seed of $\tilde{g}$ is the smallest possible to do pairwise independent sampling for random variables on $(\{0,1\}^m, 2^{\{0,1\}^m}, P_m)$. In particular, it is shorter than $2m + 2j$ of RWS case. But, if m is a little bit large, the multiplication in $\mathrm{GF}(2^m)$ is so complicated that $\tilde{g}$ is not appropriate for practical Monte Carlo integrations.

In [14], a pairwise independent sequence of random variables was constructed on a prime field F_p instead of $\mathrm{GF}(2^m)$.[†2] A synthetic report about pairwise independent sampling methods can be found in [10, 31].

5.3 i.i.d.-sampling for simulatable random variables

We can apply i.i.d.-sampling to simulatable random variables, i.e., functions which is measurable with respect to some stopping time (§ 1.3).

Theorem 5.11 *Let τ be a $\{\mathcal{B}_m\}_m$-stopping time such that $\mathbb{P}(\tau < \infty) = 1$, and let f be a τ-measurable function. Define*

$$y_n(x) := 2^{\sum_{i=1}^{n-1} \tau(y_i(x))} x, \quad x \in \mathbb{T}^1, \quad n \in \mathbb{N}^+.$$

Then the sequence of random variables $\{f(y_n)\}_{n=1}^{\infty}$ on $(\mathbb{T}^1, \mathcal{B}, \mathbb{P})$ is i.i.d., and the common distribution is equal to that of f defined on $(\mathbb{T}^1, \mathcal{B}, \mathbb{P})$.

Proof. Since $y_1(x) = x$, it is clear that the distributions of $f(y_1)$ and f coincide. By (1.7), we have $f(y_n) = f(\lfloor y_n \rfloor_{\tau(y_n)})$, and so it is sufficient to show that $\{\lfloor y_n \rfloor_{\tau(y_n)}\}_{n=1}^{\infty}$ is i.i.d.[†3]

For any $n \in \mathbb{N}^+$ and any $a_1, \ldots, a_n \in D = \cup_{m \in \mathbb{N}^+} D_m$,

$$\begin{aligned}
&\mathbb{P}\left(\lfloor y_i \rfloor_{\tau(y_i)} = a_i,\ 1 \leq i \leq n\right) \\
&= \sum_{m_1,\ldots,m_{n-1} \in \mathbb{N}^+} \mathbb{P}\left(\lfloor y_i \rfloor_{m_i} = a_i, \tau(y_i) = m_i,\ 1 \leq i \leq n-1,\ \lfloor y_n \rfloor_{\tau(y_n)} = a_n,\right) \\
&= \sum_{m_1,\ldots,m_{n-1} \in \mathbb{N}^+} \mathbb{P}\left(\lfloor y_i \rfloor_{m_i} = a_i, \tau(y_i) = m_i,\ 1 \leq i \leq n-1,\ \left\lfloor 2^{\sum_{i=1}^{n-1} m_i} x \right\rfloor_{\tau\left(2^{\sum_{i=1}^{n-1} m_i} x\right)} = a_n,\right) \\
&= \sum_{m_1,\ldots,m_{n-1} \in \mathbb{N}^+} \mathbb{P}\left(\lfloor y_i \rfloor_{m_i} = a_i, \tau(y_i) = m_i,\ 1 \leq i \leq n-1\right) \mathbb{P}\left(\left\lfloor 2^{\sum_{i=1}^{n-1} m_i} x \right\rfloor_{\tau\left(2^{\sum_{i=1}^{n-1} m_i} x\right)} = a_n\right) \\
&= \sum_{m_1,\ldots,m_{n-1} \in \mathbb{N}^+} \mathbb{P}\left(\lfloor y_i \rfloor_{m_i} = a_i, \tau(y_i) = m_i,\ 1 \leq i \leq n-1\right) \mathbb{P}\left(\lfloor x \rfloor_{\tau(x)} = a_n\right) \\
&= \mathbb{P}\left(\lfloor y_i \rfloor_{\tau(y_i)} = a_i,\ 1 \leq i \leq n-1\right) \mathbb{P}\left(\lfloor x \rfloor_{\tau(x)} = a_n\right).
\end{aligned}$$

[†2] Therefore since 1974, the year [14] was published, a secure pseudorandom generator for the Monte Carlo integration has been developed without being noticed so.

[†3] This property is called the strong Markov property.

Repeating this procedure, we see

$$\mathbb{P}\left(\lfloor y_i \rfloor_{\tau(y_i)} = a_i,\ 1 \le i \le n\right) = \prod_{i=1}^{n} \mathbb{P}\left(\lfloor x \rfloor_{\tau(x)} = a_i\right),$$

which completes the proof. □

Under the conditions of Theorem 5.11, we can numerically integrate f by i.i.d.-sampling;

$$\frac{1}{N}\sum_{i=1}^{N} f(y_i(x)). \tag{5.11}$$

Remark 5.12 The time of computation of (5.11) is approximately proportional to the number of required random bits. Therefore its mean is approximately proportional to $N\mathbf{E}[\tau]$. Then if $\mathbf{E}[\tau^2] = \infty$, i.e., $\mathbf{V}[\tau] = \infty$, the time of computation will be possibly extremely long. So, from the practical point of view, $\mathbf{E}[\tau^2] < \infty$ is desirable. For example, if τ is a stopping time associated with von Neumann's rejection method (Example 1.11), its distribution is a geometric distribution, and hence $\mathbf{E}[\tau^2] < \infty$ holds.

The computation of i.i.d.-sampling (5.11) may look complicated, but it can be done by a simple algorithm. Let the integrand f in question satisfy Assumption 1.9. Here we assume that the i.i.d. random variables $\{Z_1, Z_2, \ldots\}$ of Assumption 1.9 are expressed in 2^{-K} precision, i.e., $Z_l^{(K)} \in D_K$. Next let us suppose that a virtual function

- function Random$_m$: D_m-valued;

returns a D_m-valued uniformly distributed random variable which is independent of all the previously generated random variables. (In practice, we use a pseudorandom generator instead.)

Finally, let N be the sample size as in (5.11). Here is the algorithm.

Algorithm of i.i.d.-sampling

- Main routine

```
function Mean_of_f : Real;
begin
   $S$ := 0.0;
   For $i$ := 1 to $N$ do
      begin
         $Z$ :=Random$_K$;
         Try to compute $f$;
         while (another $Z$ is needed to compute $f$ ) do
            begin
               $Z$ :=Random$_K$;
               Try to compute $f$;
            end;      // End of computation of $f$
```

```
        $S := S + f$ ;
      end;
    result:= $S/N$;
  end;
```

The function Mean_of_f returns the value of result, namely, S/N. Here, Z_l's which are needed to compute f are all generated by the random function Random_K.

5.4 Dynamic random Weyl sampling

Let us introduce a pairwise independent sampling method which is applicable to simulatable random variables (§ 1.3), i.e., functions which are measurable with respect to some stopping time (§ 1.3.1). We call it the dynamic random Weyl sampling (abbreviated as DRWS). It can be regarded as a pseudorandom generator which is exclusive for Monte Carlo integrations of all simulatable random variables.

The algorithm of DRWS is so simple that we can write its main program code as easily as i.i.d.-sampling (§ 6.2, [43]), and that the speed of generating pairwise independent samples is sufficiently fast. DRWS is applicable whenever so is i.i.d.-sampling, and it is much more reliable than i.i.d.-sampling (§ 5.4.4). However, since DRWS must keep the random bit-sequence which is needed to generate samples stored in computer memory, we have to note that it spends more memory than i.i.d.-sampling (§ 6.2.4).

5.4.1 Definition and Theorem

Let τ be a $\{\mathcal{B}_m\}_m$-stopping time with $\mathbb{P}(\tau < \infty) = 1$. For $j \in \mathbb{N}^+$, let

$$(x_l, \alpha_l) \in D_{K+j} \times D_{K+j} \subset \mathbb{T}^1 \times \mathbb{T}^1, \quad l \in \mathbb{N}^+, \tag{5.12}$$

be i.i.d. random variables which are uniformly distributed in $D_{K+j} \times D_{K+j}$. Define random variables $\mathbf{x}_n$ by

$$\mathbf{x}_n := \sum_{l=1}^{\infty} 2^{-(l-1)K} \lfloor x_l + \nu_{n,l}\, \alpha_l \rfloor_K \in \mathbb{T}^1, \quad n = 1, \ldots, 2^{j+1}, \tag{5.13}$$

where $\nu_{n,l}$ are random variables defined by

$$\nu_{n,l} := \begin{cases} n & (l = 1), \\ \#\{\, 1 \le u \le n \mid \tau(\mathbf{x}_u) > (l-1)K \,\} & (l > 1). \end{cases} \tag{5.14}$$

Since τ is $\{\mathcal{B}_m\}_m$-stopping time, $\nu_{n,l}$ and $\mathbf{x}_n$ are well-defined.

Now, the following theorem holds.

Theorem 5.13 *([39]) If f is τ-measurable, then random variables $\{f(\mathbf{x}_n)\}_{n=1}^{2^{j+1}}$ are identically distributed and pairwise independent. The common distribution coincides with that of f defined on $(\mathbb{T}^1, \mathcal{B}, \mathbb{P})$.*

Note that as $\{\mathbf{x}_n\}_{n=1}^{2^{j+1}}$ are all uniformly distributed but not pairwise independent. Theorem 5.13 asserts that if they are composed with a τ-measurable function f, then $\{f(\mathbf{x}_n)\}_{n=1}^{2^{j+1}}$ become pairwise independent.

Assume $\mathbf{E}[\tau] < \infty$ and that $f \in L^1(\mathcal{B}_\tau)$. Then the sampling method for the estimation of $\mathbf{E}[f]$ by

$$\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_n), \quad 1 \le N \le 2^{j+1},$$

is called the *dynamic random Weyl sampling* (abbreviated as DRWS). [†4]

Corollary 5.14 *For each $f \in L^2(\mathcal{B}_\tau)$, the mean square error of DRWS is equal to that of i.i.d.-sampling, i.e.,*

$$\mathbf{E}\left[\left|\frac{1}{N}\sum_{n=0}^{N-1} f(\mathbf{x}_n) - \mathbf{E}[f]\right|^2\right] = \frac{\mathbf{V}[f]}{N}, \quad 1 \le N \le 2^{j+1}. \tag{5.15}$$

Remark 5.15 As is mentioned in Remark 5.12, it is desirable that $\mathbf{E}[\tau^2] < \infty$.

5.4.2 Proof of Theorem 5.13

In what follows, we take any $1 \le n < n' \le 2^{j+1}$ and fix them. Set

$$m(n,n') := \max\{\, l \,|\, \nu_{n,l} < \nu_{n',l} \,\}.$$

From $\nu_{n,l} < \nu_{n',l}$, it follows that $\nu_{n,i} < \nu_{n',i}$ for $i = 1, \ldots, l$, and hence

$$m(n,n') = \max\{\, l \,|\, \nu_{n,i} < \nu_{n',i},\ i = 1, \ldots, l \,\}. \tag{5.16}$$

Now define

$$\widetilde{\mathbf{x}}_{n'} := \sum_{l=1}^{\infty} 2^{-(l-1)K} \lfloor x_l + \widetilde{\nu}_{n',l}\,\alpha_l \rfloor_K, \quad \widetilde{\nu}_{n',l} := \begin{cases} \nu_{n',l} & (\, l \le m(n,n')\,), \\ n' & (\, l > m(n,n')\,). \end{cases} \tag{5.17}$$

Lemma 5.16 *(i) $\mathbf{x}_n$ is distributed uniformly in $\mathbb{T}^1$.*
(ii) $\mathbf{x}_n$ and $\widetilde{\mathbf{x}}_{n'}$ are independent.

Proof. In order to prove (i) and (ii), we show that for any $M \in \mathbb{N}^+$, the following $2M$ random variables

$$\lfloor x_l + \nu_{n,l}\,\alpha_l \rfloor_K, \quad \lfloor x_l + \widetilde{\nu}_{n',l}\,\alpha_l \rfloor_K, \qquad l = 1, \ldots, M, \tag{5.18}$$

are all distributed uniformly in D_K and they are independent.

[†4] The term “dynamic” indicates that the sampling points $\{\mathbf{x}_n\}_{n=1}^{2^{j+1}}$ vary in accordance with the integrand, more precisely, with the stopping time τ.

Note that if $l \geq 2$, then both $\nu_{n,l}$ and $\widetilde{\nu}_{n',l}$ depend on $(x_1, \alpha_1), \dots, (x_{l-1}, \alpha_{l-1})$, but they are independent of (x_l, α_l). Note also that we always have $\nu_{n,l} < \widetilde{\nu}_{n',l}$. Now, for any $s_1, t_1, \dots, s_M, t_M \in D_K$,

$$
\begin{aligned}
&\Pr(\lfloor x_l + \nu_{n,l}\,\alpha_l \rfloor_K < s_l,\ \lfloor x_l + \widetilde{\nu}_{n',l}\,\alpha_l \rfloor_K < t_l,\ l = 1, \dots, M)\\
&= \sum_{p<p'} \Pr\left(\begin{array}{l} \lfloor x_l + \nu_{n,l}\alpha_l \rfloor_K < s_l, \\ \lfloor x_l + \widetilde{\nu}_{n',l}\alpha_l \rfloor_K < t_l, \end{array} l = 1, \dots, M-1, \begin{array}{ll} \nu_{n,M} = p, & \lfloor x_M + p\alpha_M \rfloor_K < s_M \\ \widetilde{\nu}_{n,M} = p', & \lfloor x_M + p'\alpha_M \rfloor_K < t_M \end{array} \right)\\
&= \sum_{p<p'} \Pr\left(\begin{array}{l} \lfloor x_l + \nu_{n,l}\alpha_l \rfloor_K < s_l, \\ \lfloor x_l + \widetilde{\nu}_{n',l}\alpha_l \rfloor_K < t_l, \end{array} l = 1, \dots, M-1, \begin{array}{l} \nu_{n,M} = p \\ \widetilde{\nu}_{n',M} = p' \end{array} \right)\\
&\qquad \times \Pr\left(\begin{array}{l} \lfloor x_M + p\alpha_M \rfloor_K < s_M \\ \lfloor x_M + p'\alpha_M \rfloor_K < t_M \end{array} \right).
\end{aligned}
$$

Since $p \neq p'$, Theorem 2.8 implies that two events $\{\lfloor x_M + p\alpha_M \rfloor_K < s_M\}$ and $\{\lfloor x_M + p'\alpha_M \rfloor_K < t_M\}$ are independent. Therefore

$$
\begin{aligned}
&\Pr(\lfloor x_l + \nu_{n,l}\,\alpha_l \rfloor_K < s_l,\ \lfloor x_l + \widetilde{\nu}_{n',l}\,\alpha_l \rfloor_K < t_l,\ l = 1, \dots, M)\\
&= \sum_{p<p'} \Pr\left(\begin{array}{l} \lfloor x_l + \nu_{n,l}\alpha_l \rfloor_K < s_l, \\ \lfloor x_l + \widetilde{\nu}_{n',l}\alpha_l \rfloor_K < t_l, \end{array} l = 1, \dots, M-1, \begin{array}{l} \nu_{n,M} = p \\ \widetilde{\nu}_{n',M} = p' \end{array} \right)\\
&\qquad \times \Pr\left(\lfloor x_M + p\alpha_M \rfloor_K < s_M\right) \Pr\left(\lfloor x_M + p'\alpha_M \rfloor_K < t_M\right)\\
&= \sum_{p<p'} \Pr\left(\begin{array}{l} \lfloor x_l + \nu_{n,l}\alpha_l \rfloor_K < s_l, \\ \lfloor x_l + \widetilde{\nu}_{n',l}\alpha_l \rfloor_K < t_l, \end{array} l = 1, \dots, M-1, \begin{array}{l} \nu_{n,M} = p \\ \widetilde{\nu}_{n',M} = p' \end{array} \right) \times s_M t_M\\
&= \Pr\left(\lfloor x_l + \nu_{n,l}\,\alpha_l \rfloor_K < s_l,\ \lfloor x_l + \widetilde{\nu}_{n',l}\,\alpha_l \rfloor_K < t_l,\ l = 1, \dots, M-1\right) \times s_M t_M.
\end{aligned}
$$

Repeating this procedure, we eventually have

$$
\Pr\left(\lfloor x_l + \nu_{n,l}\,\alpha_l \rfloor_K < s_l,\ \lfloor x_l + \widetilde{\nu}_{n',l}\,\alpha_l \rfloor_K < t_l,\ l = 1, \dots, M\right) = \prod_{i=1}^{M} s_i t_i,
$$

which completes the proof. □

Proof of Theorem 5.13. By Lemma 5.16(i), $f(\mathbf{x}_n)$ and f are identically distributed. Next, by (5.13) with n' substituted for n and (5.17), we have

$$
\lfloor \mathbf{x}_{n'} \rfloor_{m(n,n')K} = \lfloor \widetilde{\mathbf{x}}_{n'} \rfloor_{m(n,n')K}. \tag{5.19}
$$

Let $s := \lceil \tau(\mathbf{x}_{n'})/K \rceil$. Then we have $\tau(\mathbf{x}_{n'}) > (s-1)K$ and hence $\nu_{n,s} < \nu_{n',s}$. Consequently, it follows from (5.16) that

$$
s \leq m(n, n'). \tag{5.20}
$$

(5.19) and (5.20) imply

$$
\lfloor \mathbf{x}_{n'} \rfloor_{sK} = \lfloor \widetilde{\mathbf{x}}_{n'} \rfloor_{sK}. \tag{5.21}
$$

On the other hand, since $\tau(\mathbf{x}_{n'}) \leq sK$ and τ is a $\{\mathcal{B}_m\}_m$-stopping time, the value of $\tau(\mathbf{x}_{n'})$ is determined by $\lfloor \mathbf{x}_{n'} \rfloor_{sK}$. Namely, we see $\tau(\mathbf{x}_{n'}) = \tau(\lfloor \mathbf{x}_{n'} \rfloor_{sK})$. Then by (5.21), we must have

$$
\tau(\widetilde{\mathbf{x}}_{n'}) = \tau(\mathbf{x}_{n'}) \leq sK. \tag{5.22}
$$

(5.21) and (5.22) imply

$$
\lfloor \mathbf{x}_{n'} \rfloor_{\tau(\mathbf{x}_{n'})} = \lfloor \widetilde{\mathbf{x}}_{n'} \rfloor_{\tau(\widetilde{\mathbf{x}}_{n'})}. \tag{5.23}
$$

Since f is $\mathcal{B}_\tau$-measurable, (1.7) and (5.23) imply that $f(\mathbf{x}_{n'}) = f(\widetilde{\mathbf{x}}_{n'})$. Finally, it follows from Lemma 5.16(ii) that $f(\mathbf{x}_n)$ and $f(\mathbf{x}_{n'})$ are independent. This completes the proof. □

5.4.3 Algorithm

Let us show how to implement DRWS. We use the setting of § 5.3. Assume that

$$1 \le N \le 2^{j+1}. \tag{5.24}$$

Algorithm of DRWS

- Global variables

```
$l$                 :  integer;
$\{x_l, \alpha_l\}_l$   :  array (variable length) of $(D_{K+j})^2$-valued vectors;
```

- Procedure and function

```
procedure Set_First_Location;
begin
   $l := 0$;
end;
```

```
function Drws : $D_K$-valued;
begin
   $l := l + 1$;
   if $(x_l, \alpha_l)$ has not been generated;
      then
         begin
            $x_l$ :=Random$_{K+j}$;
            $\alpha_l$ :=Random$_{K+j}$;
         end;
   $x_l := x_l + \alpha_l$;
   result:= $\lfloor x_l \rfloor_K$;
end;
```

- Main routine

```
function Mean_of_f : Real;
begin
   $S := 0.0$;
   For $i := 1$ to $N$ do
      begin
         Set_First_Location;
         Try to compute $f$;
         while ( another $Z$ is needed to compute $f$ ) do
            begin
               $Z$ :=Drws;
               Try to compute $f$;
            end;      // End of computation of $f$
         $S := S + f$;
      end;
   result:= $S/N$;
end;
```

The main routine of DRWS is very similar to that of i.i.d.-sampling (§ 5.3). The only differences are the following; in DRWS, Z_l's are not generated by the direct calls of Random_K, but they are generated by Drws, and we must call Set_First_Location before generating each sample of f.

The random function Random_{K+j} is called only by Drws, only when (x_l, α_l) has not yet been generated, to generate them. Thus, DRWS requires much less randomness than i.i.d.-sampling.

Remark 5.17 DRWS is a Monte Carlo integration, and when we consider it as gambling, the seeds (x_l, α_l) should be chosen by the player, Alice, of her own will. Of course, it is possible for her to input $(K + j)$ bit seed whenever Random_{K+j} is called in the above algorithm. However, in practice, it would be a tiresome task, and hence, an auxiliary pseudorandom generator is usually used for Random_{K+j}. In choosing such an auxiliary pseudorandom generator, what we mentioned in § 5.2.2 are valid for DRWS as well.[†5]

Remark 5.18 In some large scale computations, i.e., when the probability that f requires too many Z_l's is not negligible, DRWS may exhaust computer memory to keep all of (x_l, α_l)'s that have been currently generated. A practical solution of such memory exhaustion can be found in § 6.2.4.

5.4.4 Comparison between i.i.d.-sampling and DRWS

We computed the mean of the random variable f of Example 1.12 by DRWS. Namely, we generated pairwise independent copies of f by DRWS, and computed sample means with sample size being changed from 10^3 to 10^8.[†6] The mean and the variance of f are both 10.[†7]

Table 5.2: Comparison of Errors

Sample size	rand-i.i.d.	MT-i.i.d.	m90-i.i.d.	DRWS
10^3	0.15200000	-0.12500000	0.18600000	0.01700000
10^4	-0.05570000	0.02960000	0.03980000	-0.00030000
10^5	0.00650000	-0.01372000	-0.00170000	0.00076000
10^6	0.00470000	-0.00061300	-0.00382000	0.00007300
10^7	-0.00170760	0.00125260	0.00076940	0.00000560
10^8	-0.00095602	-0.00003483	0.00026567	-0.00000030
Final result	9.99904398	9.99996517	10.00026567	9.99999970
Time (sec.)	13	27	87	35

[†5] In the implementation of DRWS in § 6.2, the pseudorandom generator by means of Weyl transformation is used for this purpose.

[†6] More precisely, we executed the C program `drws.c` of § 6.2.3, with `#define SAMPLE_SIZE` being changed from 10^3 to 10^8.

[†7] This follows from Wald's identity ([7] (1.6) Wald's equation, p.179 and Exercise 1.15, P.182). Or, use a negative binary distribution.

Table 5.2 shows the errors of this computation. For comparison, it also shows the errors of i.i.d.-sampling with different pseudorandom generators, i.e., a standard C function `rand()`, MT (Mersenne twister), and the pseudorandom generator by means of Weyl transformation (`m90randombit()` of § 6.2).[†8] Comparing the errors, DRWS has a decided advantage over i.i.d.-sampling. The error of DRWS with sample size 10^7, which computation spent only 3 seconds, is much smaller than those of i.i.d.-sampling with sample size 10^8.

Table 5.3: Sample mean and sample SD of DRWS

Range of values	Sample mean	Sample SD
Whole	9.99993	0.003129874
Central 99.9%	9.99999	0.000601414
Central 99%	10.00000	0.000231709

After Example 5.8, we computed the frequency distribution of the samples of DRWS applied to f with sample size 10^6 by choosing 10,000 seeds at random (Table 5.3). The true value of the sample SD is

$$\sqrt{\frac{10}{10^6}} = \sqrt{10^{-5}} = 0.00316228,$$

which almost coincides with the first row (Whole) of Table 5.3. The second row (Central 99.9%) of Table 5.3 shows the sample mean and the sample SD of all the samples except the largest 5 ones and the smallest 5 ones. In this case, the sample mean did not change, but the sample SD decreased to about $1/5$ of that of all samples. This means that the excluded 0.1% samples are very far from the sample mean. The third row (Central 99%) of Table 5.3 shows the sample mean and the sample SD of all the samples except the largest 50 ones and the smallest 50 ones. In this case, the sample mean did not change, either, but the sample SD decreased to about $1/13.6$ of that of all samples. Thus a similar phenomenon as Example 5.8 takes place in the case of DRWS, too.

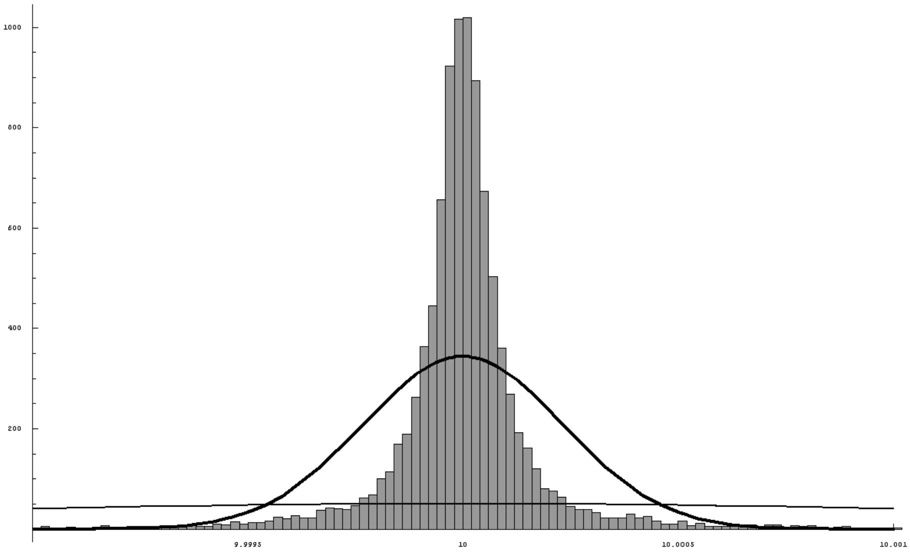
In Figure 5.4, the histogram shows the frequency distribution of the above 10,000 DRWS samples. In this figure, the thick curve is the probability density function of the normal distribution with mean 10 and SD 0.000231709 (same as the third row of Table 5.3). Comparing them, we know that the distribution of DRWS samples is more concentrated around the mean and has a thicker tails than the normal distribution. The thin line which looks almost flat located in very low position in the figure is the density function of the normal distribution with the same mean and the same SD as i.i.d.-sampling, i.e., of $\mathcal{N}(10, 10^{-5})$.

5.4.5 Limit theorem for convergence of DRWS

The error of DRWS seen in the last section was very little with high probability. We conjecture that a limit theorem like Theorem 5.6 should hold for DRWS, too. Indeed, in

[†8] Computer: Panasonic Let's Note CF-Y5 (CPU1.66GHz, RAM 1.49GB, HD55.8GB). Complier: BORLAND C++ COMPILER 5.5, COMMAND LINE TOOLS without any options.

Figure 5.4: Frequency distribution of samples of DRWS



special cases, such as when the stopping time τ is constant, or when DRWS is applied to von Neumann's rejection method, such a theorem holds as we will show below.

In order to state it, let us modify the formulation of DRWS a little. Let $(\mathbb{T}^\infty, \mathcal{B}^\infty, \mathbb{P}^\infty)$ denote the countable direct product of the Lebesgue probability space $(\mathbb{T}^1, \mathcal{B}, \mathbb{P})$. Define an increasing sequence of sub σ-fields $\{\mathcal{B}^m\}_{m=1}^\infty$ of $\mathcal{B}^\infty$ by

$$\mathcal{B}^m \ := \ \sigma(Z_1, Z_2, \ldots, Z_m), \quad m = 1, 2, \ldots,$$

where $Z_i : \mathbb{T}^\infty \to \mathbb{T}^1$ is the coordinate function (projection)

$$Z_i(x) \ := \ x_i, \quad x = (x_1, x_2, \ldots) \in \mathbb{T}^\infty.$$

We call a random variable $\tau : \mathbb{T}^\infty \to \mathbb{N}^+ \cup \{\infty\}$ a $\{\mathcal{B}^m\}_m$-stopping time if

$$\forall m \in \mathbb{N}^+, \quad \{\tau \le m\} \ \in \ \mathcal{B}^m.$$

We assume that $\mathbb{P}^\infty(\tau < \infty) = 1$. For a $\{\mathcal{B}^m\}_m$-stopping time τ, we define a sub σ-algebra

$$\mathcal{B}^\tau \ := \ \{A \in \mathcal{B}^\infty \,;\, \forall m \in \mathbb{N}^+, \, A \cap \{\tau \le m\} \in \mathcal{B}^m\}.$$

A $\mathcal{B}^\tau$-measurable function is simply called a τ-measurable function.

Now, we define a sequence of random variables $\{\mathbf{x}_n\}_{n=1}^\infty$ on the product probability space $(\mathbb{T}^\infty \times \mathbb{T}^\infty, \mathcal{B}^\infty \otimes \mathcal{B}^\infty, \mathbb{P}^\infty \otimes \mathbb{P}^\infty)$ by

$$\begin{aligned} \mathbf{x}_n(x, \alpha) \ &:= \ (x_1 + \nu_{n,1}\alpha_1, x_2 + \nu_{n,2}\alpha_2, \ldots) \in \mathbb{T}^\infty, \\ & \qquad x = (x_1, x_2, \ldots), \ \alpha = (\alpha_1, \alpha_2, \ldots) \in \mathbb{T}^\infty, \end{aligned} \tag{5.25}$$

where

$$\nu_{n,l} := \begin{cases} n & (l = 1), \\ \#\{1 \le u \le n \,|\, \tau(\mathbf{x}_u) > l - 1\} & (l > 1). \end{cases} \tag{5.26}$$

Since τ is a $\{\mathcal{B}^m\}_m$-stopping time, $\nu_{n,l}$ and hence $\mathbf{x}_n(x, \alpha)$ are well-defined.

Theorem 5.19 *Let τ be a $\{\mathcal{B}^m\}_m$-stopping time and $f : \mathbb{T}^\infty \to \mathbb{R}$ be a τ-measurable function. Then the random variables $\{f(\mathbf{x}_n)\}_{n=1}^\infty$ are identically distributed and pairwise independent. The common distribution is equal to that of f defined on $(\mathbb{T}^\infty, \mathcal{B}^\infty, \mathbb{P}^\infty)$.*

This theorem can be proved in a similar way as Theorem 5.13. Of course, a similar assertion as Corollary 5.14 holds in this case, too.

First, let us consider the case $\tau = k$ (constant). In this case, we have

$$\nu_{n,l} = n, \quad n = 1, 2, \ldots, \quad l = 1, 2, \ldots, k,$$

and f is τ-measurable, if and only if it is $\mathcal{B}^k$-measurable, i.e., it is substantially a function on $\mathbb{T}^k$. The following theorem holds.

Theorem 5.20 *(cf. Remark 5.7) For any $F \in L^2(\mathbb{T}^k, \mathcal{B}^k, \mathbb{P}^k)$ and any $1 \le p < 2$, it holds that*

$$\lim_{N\to\infty} \iint_{\mathbb{T}^k \times \mathbb{T}^k} \left| \frac{1}{\sqrt{N}} \sum_{n=1}^N \left(F(x + n\alpha) - \int_{\mathbb{T}^k} F(y)\mathbb{P}^k(dy) \right) \right|^p \mathbb{P}^k(d\alpha)\mathbb{P}^k(dx) = 0.$$

Consequently, for any $\varepsilon > 0$,

$$\lim_{N\to\infty} \mathbb{P}^{2k} \left(\left\{ (x, \alpha) \in \mathbb{T}^{2k} \,\middle|\, \left| \frac{1}{\sqrt{N}} \sum_{n=1}^N \left(F(x + n\alpha) - \int_{\mathbb{T}^k} F(y)\mathbb{P}^k(dy) \right) \right| > \varepsilon \right\} \right) = 0.$$

Proof. As the proof of Theorem 5.6, by using the k-dimensional Fourier series expansion of F, Theorem 5.20 is proved by showing that

$$\int_{\mathbb{T}^k} \left| \frac{1}{\sqrt{N}} \sum_{n=1}^N e^{2\sqrt{-1}\pi n(l_1\alpha_1 + \cdots + l_k\alpha_k)} \right|^p d\alpha_1 \cdots d\alpha_k \to 0, \quad N \to \infty.$$

Here at least one of $l_1, \ldots, l_k \in \mathbb{Z}$ is not 0. If some $l_i = 0$, then the integration in α_i can be removed and it becomes a $(k-1)$-dimensional integral. Thus we may assume that none of l_i's is 0. Then the transformation

$$\mathbb{T}^k \ni (\alpha_1, \ldots, \alpha_k) \mapsto (l_1\alpha, \ldots, l_k\alpha_k) \in \mathbb{T}^k$$

preserves the Lebesgue measure $\mathbb{P}^k$. Therefore we have

$$\int_{\mathbb{T}^k} \left| \frac{1}{\sqrt{N}} \sum_{n=1}^N e^{2\sqrt{-1}\pi n(l_1\alpha_1 + \cdots + l_k\alpha_k)} \right|^p d\alpha_1 \cdots d\alpha_k = \int_{\mathbb{T}^k} \left| \frac{1}{\sqrt{N}} \sum_{n=1}^N e^{2\sqrt{-1}\pi n(\alpha_1 + \cdots + \alpha_k)} \right|^p d\alpha_1 \cdots d\alpha_k.$$

Note that for any bounded measurable function $h : \mathbb{T}^1 \to \mathbb{R}$, we have

$$\int_{\mathbb{T}^1}\int_{\mathbb{T}^1} h(y_1 + y_2)dy_1 dy_2 = \int_{\mathbb{T}^1} h(y)dy.$$

Therefore

$$\begin{aligned}
&\int_{\mathbb{T}^k}\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N} e^{2\sqrt{-1}\pi n(\alpha_1+\cdots+\alpha_k)}\right|^p d\alpha_1\cdots d\alpha_k \\
&= \int_{\mathbb{T}^{k-2}} d\alpha_1\cdots d\alpha_{k-2}\int_{\mathbb{T}^2} d\alpha_{k-1}d\alpha_k\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N} e^{2\sqrt{-1}\pi n(\alpha_1+\cdots+\alpha_{k-2}+\alpha_{k-1}+\alpha_k)}\right|^p \\
&= \int_{\mathbb{T}^{k-1}}\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N} e^{2\sqrt{-1}\pi n(\alpha_1+\cdots+\alpha_{k-1})}\right|^p d\alpha_1\cdots d\alpha_{k-1} \\
&= \cdots \\
&= \int_{\mathbb{T}^1}\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N} e^{2\sqrt{-1}\pi n\alpha_1}\right|^p d\alpha_1.
\end{aligned}$$

As we saw in the proof of Theorem 5.6, the last integral converges to 0 as $N \to \infty$. □

Next, let us consider the case of von Neumann's rejection method (Example 1.11). We assume the following conditions for integrand f.

Assumption 5.21 For $f : \Omega \to \mathbb{R}$, there exist a $\{\mathcal{B}^m\}_m$-stopping time τ and an $r \in \mathbb{N}^+$ such that
(i) f is τ-measurable,
(ii) $\mathbb{P}^\infty(\,\tau \in r\,\mathbb{N}^+\,) = 1$ (then, $\nu_{n,(k-1)r+1} = \cdots = \nu_{n,kr-1} = \nu_{n,kr},\ k = 1,2,\ldots$),
(iii) for any $k \in \mathbb{N}^+$, conditional on an event $\{\,\tau \geq kr\,\}$, $\{\,\tau = kr\,\}$ is independent of $\mathcal{B}^{(k-1)r}$,
(iv) for any $k \in \mathbb{N}^+$ and any $s \in \mathbb{R}$, conditional on $\{\,\tau \geq kr\,\}$, $\{\,f \leq s\,\}$ is independent of $\mathcal{B}^{(k-1)r}$.

Example 5.22 (cf. Example 1.11) Let $0 \leq p(t) \leq M$ ($M > 0$ being a constant) be a probability density function on a bounded interval $[a,b]$. Define a $\{\mathcal{B}^m\}_m$-stopping time τ by

$$\tau := \inf\{\,2l \in 2\mathbb{N}^+ \mid p\,((b-a)Z_{2l-1} + a) \geq MZ_{2l}\,\},$$

and a function f by

$$f(x) := (b-a)Z_{\tau(x)-1}(x) + b, \quad x \in \mathbb{T}^\infty.$$

Then f satisfies Assumption 5.21 with the above τ and $r = 2$, and the distribution of f has the density $p(t)$.

Theorem 5.23 *If f is square integrable and satisfies Assumption 5.21, and if $\{\mathbf{x}_n\}_n$ is the sequence of random variables defined by (5.25) and (5.26), then for any $\varepsilon > 0$, we see*

$$\lim_{N\to\infty} \mathbb{P}^\infty \otimes \mathbb{P}^\infty\left(\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(f(\mathbf{x}_n) - \mathbf{E}[f])\right| > \varepsilon\right) = 0.$$

To prove Theorem 5.23, we prepare a lemma. In what follows, for the sake of simplicity, we assume Assumption 5.21 with $r = 2$. For a general r, the proof is similar.

Lemma 5.24 *For any $l \in \mathbb{N}^+$ and any square integrable function $g : \mathbb{T}^2 \to \mathbb{R}$, define*

$$h_l(x) := \mathbf{1}_{\{\tau \geq 2l\}}(x)\, g(Z_{2l-1}(x), Z_{2l}(x)), \quad x \in \mathbb{T}^\infty.$$

Then for any $\varepsilon > 0$, we see

$$\lim_{N\to\infty} \mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (h_l(\mathbf{x}_n) - \mathbf{E}[h_l]) \right| > \varepsilon \right) = 0.$$

Proof. By induction. First, for $l = 1$, we have $h_1(x) = g(Z_1(x), Z_2(x))$ and hence

$$h_1(\mathbf{x}_n) = g(x_1 + n\alpha_1, x_2 + n\alpha_2), \quad n = 1, 2, \ldots,$$

now the assertion of the lemma is shown by Theorem 5.20.

Next, let us assume $l \geq 2$. Since $\mathbf{1}_{\{\tau \geq 2l\}}(x) = \mathbf{1}_{\{\tau \leq 2l-2\}^c}(x)$ is $\mathcal{B}^{2l-2}$-measurable and $g(Z_{2l-1}(x), Z_{2l}(x))$ is $\sigma(Z_{2l-1}, Z_{2l})$-measurable, they are independent and hence

$$\mathbf{E}[h_l] = q_l\, \mathbf{E}[g], \quad q_l := \mathbb{P}^\infty(\tau \geq 2l).$$

For $l = 2$, noting that $\nu_{n,3} = \nu_{n,4}$, since

$$h_2(\mathbf{x}_n) = \mathbf{1}_{\{\tau \geq 4\}}(\mathbf{x}_n) g(x_3 + \nu_{n,3}\alpha_3, x_4 + \nu_{n,3}\alpha_4), \quad n = 1, 2, \ldots,$$

we have

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} h_2(\mathbf{x}_n) = \frac{1}{\sqrt{N}} \sum_{m=1}^{\nu_{N,3}} g(x_3 + m\alpha_3, x_4 + m\alpha_4).$$

From this, we get the following inequality.

$$\begin{aligned}
\left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (h_2(\mathbf{x}_n) - \mathbf{E}[h_2]) \right| &= \left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (h_2(\mathbf{x}_n) - q_1 \mathbf{E}[g]) \right| \\
&\leq \left| \sqrt{\frac{\nu_{N,3}}{N}} \cdot \frac{1}{\sqrt{\nu_{N,3}}} \sum_{m=1}^{\nu_{N,3}} (g(x_3 + m\alpha_3, x_4 + m\alpha_4) - \mathbf{E}[g]) \right| \\
&\quad + \left| \frac{\nu_{N,3}}{\sqrt{N}} \mathbf{E}[g] - \sqrt{N} q_1 \mathbf{E}[g] \right| \\
&\leq \left| \frac{1}{\sqrt{\nu_{N,3}}} \sum_{m=1}^{\nu_{N,3}} (g(x_3 + m\alpha_3, x_4 + m\alpha_4) - \mathbf{E}[g]) \right| \\
&\quad + \frac{\mathbf{E}[|g|]}{\sqrt{N}} \left| \nu_{N,3} - N q_1 \right|.
\end{aligned}$$

Hence

$$\mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (h_2(\mathbf{x}_n) - \mathbf{E}[h_2]) \right| > \varepsilon \right)$$

$$
\begin{aligned}
&\le \mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\left| \frac{1}{\sqrt{\nu_{N,3}}} \sum_{m=1}^{\nu_{N,3}} (g(x_3 + m\alpha_3, x_4 + m\alpha_4) - \mathbf{E}[g]) \right| > \frac{\varepsilon}{2} \right) \\
&\quad + \mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\frac{\mathbf{E}[|g|]}{\sqrt{N}} \left| \nu_{N,3} - Nq_1 \right| > \frac{\varepsilon}{2} \right) \\
&=: I_1 + I_2. \qquad (5.27)
\end{aligned}
$$

Take any $\delta > 0$. By Theorem 5.20, there exists a $K_0 \in \mathbb{N}^+$ such that for any $K \ge K_0$, it holds that

$$
\mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\left| \frac{1}{\sqrt{K}} \sum_{m=1}^{K} (g(x_3 + m\alpha_3, x_4 + m\alpha_4) - \mathbf{E}[g]) \right| > \frac{\varepsilon}{2} \right) < \frac{\delta}{3}.
$$

Therefore

$$
\mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\left| \frac{1}{\sqrt{\nu_{N,3}}} \sum_{m=1}^{\nu_{N,3}} (g(x_3 + m\alpha_3, x_4 + m\alpha_4) - \mathbf{E}[g]) \right| > \frac{\varepsilon}{2} \,\middle|\, \nu_{N,3} \ge K_0 \right) < \frac{\delta}{3}. \qquad (5.28)
$$

On the other hand, noting that $\nu_{N,3} = \sum_{n=1}^{N} \mathbf{1}_{\{\tau \le 2\}^c}(x_1 + n\alpha_1, x_2 + n\alpha_2)$[†9], by Theorem 5.20, there exists an $N_0 \in \mathbb{N}^+$ such that for any $N \ge N_0$,

$$
I_2 = \mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\frac{\mathbf{E}[|g|]}{\sqrt{N}} \left| \sum_{n=1}^{N} (\mathbf{1}_{\{\tau \le 2\}^c}(x_1 + n\alpha_1, x_2 + n\alpha_2) - q_1) \right| > \frac{\varepsilon}{2} \right) < \frac{\delta}{3}. \qquad (5.29)
$$

Let us estimate I_1. Of course, we may assume $\mathbf{E}[|g|] > 0$. Take an N_1 so that

$$
\forall N \ge N_1, \quad Nq_1 - \frac{\varepsilon \sqrt{N}}{2\mathbf{E}[|g|]} \ge K_0.
$$

Then for any $N \ge N_1$, we see

$$
\begin{aligned}
\mathbb{P}^\infty \otimes \mathbb{P}^\infty (\nu_{N,3} < K_0) &\le \mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\nu_{N,3} - Nq_1 < -\frac{\varepsilon \sqrt{N}}{2\mathbf{E}[|g|]} \right) \\
&= \mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\frac{\mathbf{E}[|g|]}{\sqrt{N}} \left(\sum_{n=1}^{N} (\mathbf{1}_{\{\tau \le 2\}^c}(x_1 + n\alpha_1, x_2 + n\alpha_2) - q_1) \right) < -\frac{\varepsilon}{2} \right) \\
&< \frac{\delta}{3}.
\end{aligned}
$$

Putting

$$
B := \left\{ \left| \frac{1}{\sqrt{\nu_{N,3}}} \sum_{m=1}^{\nu_{N,3}} (g(x_3 + m\alpha_3, x_4 + m\alpha_4) - \mathbf{E}[g]) \right| > \frac{\varepsilon}{2} \right\},
$$

we have that for any $N \ge \max(N_0, N_1)$,

$$
\begin{aligned}
I_1 &= \mathbb{P}^\infty \otimes \mathbb{P}^\infty (B) \\
&= \mathbb{P}^\infty \otimes \mathbb{P}^\infty (B \cap \{\nu_{N,3} \ge K_0\}) + \mathbb{P}^\infty \otimes \mathbb{P}^\infty (B \cap \{\nu_{N,3} < K_0\}) \\
&\le \mathbb{P}^\infty \otimes \mathbb{P}^\infty (B \,|\, \nu_{N,3} \ge K_0) + \mathbb{P}^\infty \otimes \mathbb{P}^\infty (\nu_{N,3} < K_0) \\
&< \frac{\delta}{3} + \frac{\delta}{3} = \frac{2\delta}{3}.
\end{aligned}
$$

[†9] We use this notation because $\mathbf{1}_{\{\tau \le 2\}^c}$ is $\mathcal{B}^2$-measurable.

This and (5.29) imply $I_1 + I_2 < \delta$, and so by (5.27), the proof in the case $l = 2$ is complete.

Now, let us show the assertion of the lemma for $l + 1$ assuming that it is valid for l. In the same way as (5.27), we get

$$\begin{aligned}
&\mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (h_{l+1}(\mathbf{x}_n) - \mathbf{E}[h_{l+1}]) \right| > \varepsilon \right) \\
\le\ & \mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\left| \frac{1}{\sqrt{\nu_{N,2l+1}}} \sum_{m=1}^{\nu_{N,2l+1}} (g(x_{2l+1} + m\alpha_{2l+1}, x_{2l+2} + m\alpha_{2l+2}) - \mathbf{E}[g]) \right| > \frac{\varepsilon}{2} \right) \\
& + \mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\frac{\mathbf{E}[|g|]}{\sqrt{N}} \left| \nu_{N,2l+1} - Nq_l \right| > \frac{\varepsilon}{2} \right) \\
=:\ & I_3 + I_4. \qquad (5.30)
\end{aligned}$$

Putting

$$B' := \left\{ \left| \frac{1}{\sqrt{\nu_{N,2l+1}}} \sum_{m=1}^{\nu_{N,2l+1}} (g(x_{2l+1} + m\alpha_{2l+1}, x_{2l+2} + m\alpha_{2l+2}) - \mathbf{E}[g]) \right| > \frac{\varepsilon}{2} \right\}$$

in the same way as (5.28), it holds that

$$\mathbb{P}^\infty \otimes \mathbb{P}^\infty (B \,|\, \nu_{N,2l+1} \ge K_0) < \frac{\delta}{3}.$$

On the other hand, by Assumption 5.21(iii), there exists a function $\tilde{g}_l : \mathbb{T}^2 \to \mathbb{R}$ such that

$$\mathbf{1}_{\{\tau \ge 2l+2\}}(x) = \mathbf{1}_{\{\tau \ge 2l\}}(x)\,(1 - \mathbf{1}_{\{\tau = 2l\}}(x)) = \mathbf{1}_{\{\tau \ge 2l\}}(x)\,\tilde{g}_l(Z_{2l-1}(x), Z_{2l}(x)), \quad x \in \mathbb{T}^\infty.$$

Consequently, by the assumption of the induction, as $N \to \infty$, we see

$$\mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (\mathbf{1}_{\{\tau \ge 2l+2\}}(\mathbf{x}_n) - q_l) \right| > \varepsilon \right) \to 0.$$

Note that $\sum_{n=1}^{N} \mathbf{1}_{\{\tau \ge 2l+2\}}(\mathbf{x}_n) = \nu_{N,2l+1}$. Then there exists an $N_2 \in \mathbb{N}^+$ such that if $N \ge N_2$, it holds that

$$I_4 = \mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\frac{\mathbf{E}[|g|]}{\sqrt{N}} \left| \nu_{N,2l+1} - Nq_l \right| > \frac{\varepsilon}{2} \right) < \frac{\delta}{3}. \qquad (5.31)$$

Let us estimate I_3. Take $N_3 \in \mathbb{N}^+$ so that

$$\forall N \ge N_3, \quad Nq_l - \frac{\varepsilon\sqrt{N}}{2\mathbf{E}[|g|]} \ge K_0.$$

Then for $N \ge N_3$, we have

$$\mathbb{P}^\infty \otimes \mathbb{P}^\infty(\nu_{N,2l+1} < K_0) < \frac{\delta}{3}.$$

Finally, for any $N \ge \max(N_2, N_3)$, it holds that

$$\begin{aligned}
I_3 &= \mathbb{P}^\infty \otimes \mathbb{P}^\infty(B') \\
&= \mathbb{P}^\infty \otimes \mathbb{P}^\infty(B' \cap \{\nu_{N,2l+1} \ge K_0\}) + \mathbb{P}^\infty \otimes \mathbb{P}^\infty(B' \cap \{\nu_{N,2l+1} < K_0\}) \\
&\le \mathbb{P}^\infty \otimes \mathbb{P}^\infty(B' \,|\, \nu_{N,2l+1} \ge K_0) + \mathbb{P}^\infty \otimes \mathbb{P}^\infty(\nu_{N,2l+1} < K_0) \\
&< \frac{\delta}{3} + \frac{\delta}{3} = \frac{2\delta}{3}.
\end{aligned}$$

This and (5.31) imply $I_3 + I_4 < \delta$, and so by (5.30), the proof for $l + 1$ is complete. □

Proof of Theorem 5.23. Suppose that f satisfies Assumption 5.21 with $r = 2$. For a general r, the proof is similar. For each $L \in \mathbb{N}^+$, put

$$f_L(x) ;= f(x)\mathbf{1}_{\{\tau\le 2L\}}(x), \quad f'_L(x) := f(x) - f_L(x), \quad x \in \mathbb{T}^\infty.$$

Take an arbitrary $\varepsilon > 0$.

$$\begin{aligned} &\mathbb{P}^\infty \otimes \mathbb{P}^\infty\left(\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(f(\mathbf{x}_n) - \mathbf{E}[f])\right| > \varepsilon\right) \\ &\quad\le \mathbb{P}^\infty \otimes \mathbb{P}^\infty\left(\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(f_L(\mathbf{x}_n) - \mathbf{E}[f_L])\right| > \frac{\varepsilon}{2}\right) \\ &\quad\quad+\mathbb{P}^\infty \otimes \mathbb{P}^\infty\left(\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(f'_L(\mathbf{x}_n) - \mathbf{E}[f'_L])\right| > \frac{\varepsilon}{2}\right) =: I_5 + I_6. \end{aligned}$$

Since $|f'_L(x)|^2 \le |f(x)|^2$ and $f'_L(x) \to 0$, as $L \to \infty$, $\mathbb{P}^\infty$-a.s., by Lebesgue's convergence theorem, we see

$$\lim_{L\to\infty} \mathbf{E}\left[|f'_L|^2\right] = 0.$$

Therefore for any $\delta > 0$, there exists an $L_0 \in \mathbb{N}^+$ such that if $L \ge L_0$, we have

$$\mathbf{E}\left[|f'_L|^2\right] < \frac{\varepsilon^2\delta}{8}.$$

Applying Chebyshev's inequality, for any $N \in \mathbb{N}^+$,

$$\begin{aligned} I_6 &= \mathbb{P}^\infty \otimes \mathbb{P}^\infty\left(\left|\sum_{n=1}^{N}(f'_L(\mathbf{x}_n) - \mathbf{E}[f'_L])\right| > \frac{\varepsilon\sqrt{N}}{2}\right) \\ &\le \frac{4}{\varepsilon^2 N}\mathbf{V}\left[\sum_{n=1}^{N} f'_L(\mathbf{x}_n)\right] = \frac{4}{\varepsilon^2}\mathbf{V}[f'_L] \\ &\le \frac{4}{\varepsilon^2}\mathbf{E}\left[|f'_L|^2\right] < \frac{\delta}{2}. \end{aligned}$$

Here we computed the variance by using Theorem 5.19.

Fix $L \ge L_0$. Let us estimate I_5. Since we have

$$I_5 \le \sum_{l=1}^{L}\mathbb{P}^\infty \otimes \mathbb{P}^\infty\left(\left|\frac{1}{\sqrt{N}}\sum_{n=1}^{N}\left(f(\mathbf{x}_n)\mathbf{1}_{\{\tau=2l\}}(\mathbf{x}_n) - \mathbf{E}[f\mathbf{1}_{\{\tau=2l\}}]\right)\right| > \frac{\varepsilon}{2L}\right),$$

we have only to estimate each summand of the right hand side. Because f satisfies Assumption 5.21($r = 2$), for each l, there exists a function $g_l : \mathbb{T}^2 \to \mathbb{R}$ such that

$$f(x)\mathbf{1}_{\{\tau=2l\}}(x) = \mathbf{1}_{\{\tau\ge 2l\}}(x) \cdot g_l(Z_{2l-1}(x), Z_{2l}(x)), \quad x \in \mathbb{T}^\infty.$$

Then Lemma 5.24 implies that there exists an $N_0 \in \mathbb{N}^+$ such that for any $N > N_0$, it holds that

$$\mathbb{P}^\infty \otimes \mathbb{P}^\infty \left(\left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (f(\mathbf{x}_n)\mathbf{1}_{\{\tau=2l\}}(\mathbf{x}_n) - \mathbf{E}[f\mathbf{1}_{\{\tau=2l\}}]) \right| > \frac{\varepsilon}{2L} \right) < \frac{\delta}{2L}, \quad l = 1, \ldots, L.$$

From this we know $I_5 < \delta/2$.

From all above, for $N \geq N_0$, it follows that $I_5 + I_6 < \delta$. □

Remark 5.25 In general, if a sequence of random variables is L^2-bounded and converges to 0 in probability, then for any $p \in [1, 2)$, it converges to 0 in L^p. Consequently, under the assumption of Theorem 5.23, it holds that

$$\lim_{N\to\infty} \mathbf{E}\left[\left| \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (f(\mathbf{x}_n) - \mathbf{E}[f]) \right|^p \right] = 0, \quad 1 \leq p < 2.$$