Mixing time of an unaligned Gibbs sampler on the square

Balázs Gerencsér*

October 9, 2018

Abstract

The paper concerns a particular example of the Gibbs sampler and its mixing efficiency. Coordinates of a point are rerandomized in the unit square $[0,1]^2$ to approach a stationary distribution with density proportional to $\exp(-A^2(u-v)^2)$ for $(u,v) \in [0,1]^2$ with some large parameter A.

Diaconis conjectured the mixing time of this process to be $O(A^2)$ which we confirm in this paper. This improves on the currently known $O(\exp(A^2))$ estimate.

1 Introduction

A standard use of Markov chains is to sample from a probability distribution that would be otherwise hard to access. This can happen when the distribution is supported on a set implicitly defined by some constraints, e.g., a convex body in a high dimensional space [5], [7], proper colorings of a graph [3], [8], etc. Several frameworks have been designed to achieve this goal including the Metropolis algorithm and the Gibbs sampler and their variants. There is a vast range of applications and studies, we refer the reader to [2], [1] for orientation.

A central and recurring question is the efficiency of these algorithms in the different settings. We highlight two phenomena that can decrease the performance of such algorithms. First, the incremental change the Markov chain allows is usually quite rigid and given by the structure of the state space. However, the desired stationary distribution does not need to be aligned with the directions where the Markov chain mixes fast. Second, some boundary effects might occur if the Markov chain can get trapped in some remote part of the state space.

In this paper we analyze an example of the Gibbs sampling procedure proposed by Diaconis which is surprisingly simple considering it captures both of the two phenomena above. We call the *coordinate Gibbs sampler for the diagonal distribution* the following process. Fix a large positive constant A and on $[0,1]^2$ define the distribution π with density proportional to $\exp(-A^2(u-v)^2)$ for $(u,v) \in [0,1]^2$. At each step randomly choose coordinate u or v and rerandomize it according to the conditional distribution of π. Notice that the distribution of this Markov chain is mostly concentrated near the diagonal of the unit square, while only horizontal and vertical transitions are allowed. Furthermore, near $(0,0)$ and $(1,1)$ we see that both the high density of π and also the boundaries of the square hinder the movement of the chain.

The efficiency of the algorithm is quantified by the mixing time of the Markov chain. For any Markov chain $X(0), X(1), \dots$ on some state space Ω (which is $[0,1]^2$ in our case) let $\mathcal{L}(X(t))$ denote the distribution of the state at time t and η be the stationary distribution assuming it is unique (denoted by π for our case). Using the total variation distance between measures, $\|\rho-\sigma\|_{\mathrm{TV}} := \sup_{S\subseteq\Omega} |\rho(S)-\sigma(S)|$ we define the mixing time as

$$t_{\mathrm{mix}}(X,\varepsilon) := \sup_{X(0)\in\Omega} \min\left\{t : \|\mathcal{L}(X(t)) - \eta\|_{\mathrm{TV}} \le \varepsilon\right\}.$$

Diaconis conjectured that the mixing time of the example proposed is $O(A^2)$, the goal of this paper is to confirm this bound.

*B. Gerencsér is with the Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences and the ELTE Eötvös Loránd University, Department of Probability Theory and Statistics, gerencser.balazs@renyi.mta.hu. His work is supported by NKFIH (National Research, Development and Innovation Office) grant PD 121107.

Theorem 1. *Let $X(t)$ follow the coordinate Gibbs sampler for the diagonal distribution. For any $0 < \alpha < 1$ there exists $\beta > 0$ such that for large enough A,*

$$t_{\mathrm{mix}}(X, \alpha) \leq \beta A^2.$$

Up until now only $O(\exp(A^2))$ was known which easily follows from a minorization condition of the transition kernel.

Observe that the diagonal nature of the distribution plays an important role in the mixing behavior, making the distribution and the randomization steps unaligned. If we took the distribution with density proportional to $\exp(A^2(u - 1/2)^2)$ for $(u, v) \in [0, 1]^2$, then the mixing time would decrease to be $O(1)$. Indeed, this is a product distribution, product of one for u and one (uniform) for v, consequently after a rerandomization is performed along both coordinates, the distribution of the process will exactly match the prescribed one. This will happen with probability arbitrarily close to 1 within a corresponding finite number of steps, not depending on the value of A.

The rest of the paper is organized as follows. In Section 2 a formal definition of the process of interest is provided and further variants are introduced that help the analysis. Section 3 provides the building blocks for the proof, to understand the short-term behavior of the process based on the initialization. Afterwards, the proof of Theorem 1 is aggregated in Section 4. Finally, a complementing lower bound demonstrating that Theorem 1 is essentially sharp is given in Section 5 together with some numerical simulations.

2 Preliminaries, alternative processes

We now formally define the *coordinate Gibbs sampler for the diagonal distribution* which we denote by $X(t)$, then we introduce variants that will be more convenient to handle.

Let $\varphi(x) := \exp(-A^2x^2)$ for some large $A > 0$ and let π be the probability distribution on $[0, 1]^2$ with density $\mathcal{Z}^{-1}\varphi(u - v)$ at $(u, v) \in [0, 1]^2$ (where $\mathcal{Z} = \int_{[0,1]^2} \varphi(u - v)$). We write $\pi(\cdot, v)$ for the conditional distribution of the u coordinate when v is fixed (similarly for $\pi(u, \cdot)$). Denote by π_u the projection of π, that is, the overall distribution of the u coordinate.

When defining the coordinate Gibbs sampler for the diagonal distribution, we separate the decision of the direction of randomization and the randomization itself. For $t = 1, 2, \ldots$ let $r(t)$ be an i.i.d. sequence of variables of characters U, V taking both with probability $1/2$. Given some initial point $X(0) \in [0, 1]^2$ the random variable $X(t) = (X_u(t), X_v(t))$ is generated as a Markov chain from $X(t - 1)$ by randomizing along the axis given by $r(t)$. Formally,

$$X(t) := \begin{cases} \big(u^+, X_v(t-1)\big), & \text{if } r(t) = U, \text{where } u^+ \sim \pi(\cdot, X_v(t-1)), \\ \big(X_u(t-1), v^+\big), & \text{if } r(t) = V, \text{where } v^+ \sim \pi(X_u(t-1), \cdot), \end{cases},$$

where u^+, v^+ are conditionally independent of the past at all steps.

Note that when multiple U's follow each other in the series $r(t)$ (similarly for V), the values u^+ are repeatedly overwritten and forgotten, with no further mixing happening for the overall distribution. Therefore we define an alternative process where this effect does not occur, but rather the directions of randomization are deterministic.

Let $X^*(0) := X(0)$, then the following process is generated:

$$\begin{aligned} X^*(2s) &:= \big(u^+, X^*_v(2s-1)\big), && \text{where } u^+ \sim \pi(\cdot, X^*_v(2s-1)), \\ X^*(2s+1) &:= \big(X^*_u(2s), v^+\big), && \text{where } v^+ \sim \pi(X^*_u(2s), \cdot). \end{aligned}$$

It would be convenient for the analysis if it wasn't necessary to distinguish the steps based on the parity of the time index. For that reason, consider the following modification. At every even step take $X^*(2s)$ as before, at every odd step take $X^*(2s + 1)$ flipped along the diagonal of the square (exchange the two coordinates). Equivalently, flip the process at every step while generating. As a result, the randomization happens in the same direction at every step. Note that the target distribution π is symmetric along the diagonal therefore no adjustment is needed for the flipping. Formally, the process described is the following:

Let $Y(0) := X(0)$, then the random variables $Y(t)$ are generated from $Y(t-1)$ as follows

$$Y(t) := \big(u^+, Y_u(t-1)\big), \qquad \text{where } u^+ \sim \pi(\cdot, Y_u(t-1)).$$

Observe that the scalar process $Y_u(t)$ is a Markov chain by itself simply because $Y(t)$ depends on $Y(t-1)$ only through $Y_u(t-1)$.

3 Dynamics of $Y_u(t)$

In this section we prove two properties of the evolution of $Y_u(t)$, which will be the key elements to compute the mixing time bounds. First, we show that the process cannot stay arbitrarily long at the sides of the unit interval, in $[0, 1/2-\delta)$ or $(1/2+\delta, 1]$, where some small enough parameter $\delta > 0$ will be chosen. Second, we prove that starting from a point in the middle part $[1/2-\delta, 1/2+\delta]$, the distribution of the process quickly approaches the stationary distribution.

3.1 Reaching the middle

We work on the case when the $Y_u(0)$ is away from the middle of $[0, 1]$. We want to ensure that the process does not stay near the boundaries for a long period. To quantify this, the time to reach the middle is defined as follows:

Definition 2. Let $\nu_m := \min\{s : Y_u(s) \in [1/2-\delta, 1/2+\delta]\}$.

Without the loss of generality we may assume that $Y_u(0)$ is on the left part of $[0, 1]$, thanks to the symmetry of π w.r.t. $(1/2, 1/2)$. Therefore we start from $Y_u(0) < 1/2-\delta$. For this period before reaching the middle we introduce a slightly simplified process Y', where both coordinates are allowed to take values in $[0, \infty)$ in principle. This is not supposed to have a substantially different behavior, but will allow more convenient analytic investigation as fewer boundaries are present.

For any $v \in \mathbb{R}$ let σ_v be the measure on $[0, \infty)$ with density proportional to $\varphi(u-v)$ conditioned on $u \in [0, \infty)$. Let $Y'(0) := X(0)$, then define the Markov chain $Y'(t)$ as follows:

$$Y'(t) := \big(u^+, Y'_u(t-1)\big), \qquad \text{where } u^+ \sim \sigma_{Y'_u(t-1)}.$$

We can generate $Y'(t)$ to be coupled to $Y(t)$ as long as possible. For a fixed v, $\pi(u, v)$ is proportional to $\varphi(u-v)$ conditioned on $u \in [0, 1]$. Therefore, when we need to generate u^+ we draw a random sample from $\sigma_{Y'_u(t-1)}$ and use it for both $Y(t)$ and $Y'(t)$ if $u^+ < 1$. Otherwise, we use it for $Y'(t)$ but for $Y(t)$ we draw a new independent sample from $\pi(\cdot, Y'_u(t-1))$. It is easy to verify this is overall a valid method for generating a random variable of distribution $\pi(\cdot, Y'_u(t-1))$.

In the latter case, we also signal decoupling by setting a stopping time $\nu_c^1 = t$. We show this rarely happens, when governed by a variant of ν_m. Let $\tilde{\nu}_m := \min\{s : Y_u(s) \geq 1/2-\delta\}$.

Lemma 3. *For any $\alpha_1 > 0$ there is $\beta_1 > 0$ such that $P(\nu_c^1 < \min(\tilde{\nu}_m, \alpha_1 A^2)) = O(\exp(-\beta_1 A^2))$.*

Proof. We want to bound the probability of decoupling at every point in time.

When u^+ is drawn, $Y'_u(t-1) < 1/2-\delta$ is ensured as $\tilde{\nu}_m$ has not yet occurred. For any $v < 1/2-\delta$ we have

$$\sigma_v(\{u^+ > 1\}) \leq 2P(u > 1, u \sim \mathcal{N}(v, 1/(2A^2))) \leq 2\frac{\exp(-A^2(1/2+\delta)^2)}{2\sqrt{\pi}A(1/2+\delta)}.$$

Here we use that the conditional probability is at most twice the unconditional one (because of $v \geq 0$), use the monotonicity in v, then apply a standard tail probability estimate for the Gaussian distribution.

These exceptional events may occur at most at $\alpha_1 A^2$ different times, therefore by using the union bound the overall probability is $O(\exp(-\beta_1 A^2))$ for any $\beta_1 < (1/2+\delta)^2$. □

Lemma 4. *There exists $\beta_2 > 0$ constant such that $P(\nu_m \neq \tilde{\nu}_m) = O(\exp(-\beta_2 A^2)$.*

Proof. By a similar argument as above this bad event $\{\nu_m \neq \tilde{\nu}_m\}$ happens when $Y_u(t-1) < 1/2 - \delta$ but $Y_u(t) > 1/2 + \delta$ when $\tilde{\nu}_m$ occurs, then a Gaussian tail probability estimate gives an upper bound of

$$2\frac{\exp(-A^2(2\delta)^2)}{2\sqrt{\pi}A(2\delta)}.$$

The lemma holds with $\beta_2 = (2\delta)^2$. □

Handling $Y'(t)$ is still challenging due to the conditional distributions included in the definition. Therefore we introduce the following process that will be both convenient to handle and to relate to $Y'(t)$.

Let $\tilde{Z}(t)$ be a random walk with i.i.d. $\mathcal{N}(0, 1/(2A^2))$ increments, starting from $\tilde{Z}(0) := X_u(0)$. Let $Z(t) := |\tilde{Z}(t)|$.

Let us denote by ϕ the distribution of the centered Gaussian with variance $1/(2A^2)$. During the analysis of $Z(t)$ we will also need to use the distribution of the absolute value of a Gaussian distribution with variance $1/(2A^2)$. We denote it by ϕ_x when the original one is centered at x and it is easy to verify that we can express it for any $A \subset [0, \infty)$ by $\phi_x(A) = \phi(A - x) + \phi(-A - x)$.

Proposition 5. *$Z(t)$ and $Y'_u(t)$ can be coupled such that $Z(t) \leq Y'_u(t)$ for all $t \geq 0$.*

Proof. At 0 we have $Z(0) = Y'_u(0)$. We construct the coupling iteratively, assuming $Z(t-1) \leq Y'_u(t-1)$ we perform the next step of the coupling which will satisfy $Z(t) \leq Y'_u(t)$.

We will use the monotone coupling between the two. For two probability distributions ρ, ρ' the monotone coupling is the one assigning x to x' when $\rho((-\infty, x]) = \rho'((-\infty, x'])$. (We now skip currently irrelevant technical details about continuity, etc.). It is easy to verify that $x \leq x'$ is maintained through this coupling exactly if $\rho((-\infty, y]) \geq \rho'((-\infty, y])$ for all y. In our case we will need the following:

Lemma 6. *For any $v \geq \bar{v} \geq 0$ and $u \geq 0$:*

$$\phi_{\bar{v}}([0, u]) \geq \sigma_v([0, u]).$$

Here $\bar{v}$ corresponds to $Z(t-1)$ and v to $Y'_u(t-1)$ and we compare the distributions for step t.

Proof. We are going to prove the following two inequalities:

$$\phi_{\bar{v}}([0, u]) \geq \phi_v([0, u]), \qquad \phi_v([0, u]) \geq \sigma_v([0, u]).$$

For the first of the two we compute $\partial_v \phi_v([0, u])$:

$$\partial_v \phi_v([0, u]) = \partial_v \left(\phi([-v - u, -v + u])\right)$$

$$= \partial_v \left(\frac{1}{\int_{-\infty}^{\infty} \varphi} \int_{-v-u}^{-v+u} \varphi \right)$$

$$= \frac{1}{\int_{-\infty}^{\infty} \varphi}(-\varphi(-v + u) + \varphi(-v - u)) \leq 0.$$

This last inequality holds because $|-v + u| \leq |-v - u|$ and $\varphi(x)$ is decreasing in $|x|$. Consequently, when $\bar{v}$ is increased to v, the measure of $[0, u]$ decreases confirming the first inequality. This intuitively means that when a Gaussian distribution is shifted to the right then even the reflected Gaussian is shifted (if it was centered at a non-negative point).

The second inequality to confirm is the following:

$$\phi_v([0, u]) = \phi([-v - u, -v + u]) \geq \sigma_v([0, u]).$$

We rearrange and cancel out as much as possible from the domain of integrations.

$$\int_{-v-u}^{-v+u}\varphi \Big/ \int_{-\infty}^{\infty}\varphi \geq \int_{-v}^{-v+u}\varphi \Big/ \int_{-v}^{\infty}\varphi$$
$$\int_{-v-u}^{-v+u}\varphi \cdot \int_{-v}^{\infty}\varphi \geq \int_{-v}^{-v+u}\varphi \cdot \int_{-\infty}^{\infty}\varphi$$
$$\left(\int_{-v-u}^{-v}\varphi + \int_{-v}^{-v+u}\varphi\right)\cdot \int_{-v}^{\infty}\varphi \geq \int_{-v}^{-v+u}\varphi \cdot \left(\int_{-\infty}^{-v}\varphi + \int_{-v}^{\infty}\varphi\right)$$
$$\int_{-v-u}^{-v}\varphi \cdot \int_{-v}^{\infty}\varphi \geq \int_{-v}^{-v+u}\varphi \cdot \int_{-\infty}^{-v}\varphi$$
$$\int_{-v-u}^{-v}\varphi \cdot \left(\int_{-v}^{-v+u}\varphi + \int_{-v+u}^{\infty}\varphi\right) \geq \int_{-v}^{-v+u}\varphi \cdot \left(\int_{-\infty}^{-v-u}\varphi + \int_{-v-u}^{-v}\varphi\right)$$
$$\int_{-v-u}^{-v}\varphi \cdot \int_{-v+u}^{\infty}\varphi \geq \int_{-v}^{-v+u}\varphi \cdot \int_{-\infty}^{-v-u}\varphi$$

We substitute the functions to integrate and transform them to compare them on the same domain.

$$\int_{-v-u}^{-v} e^{-A^2x^2}dx \cdot \int_{-v+u}^{\infty} e^{-A^2y^2}dy \geq \int_{-v}^{-v+u} e^{-A^2x^2}dx \cdot \int_{-\infty}^{-v-u} e^{-A^2y^2}dy$$
$$\int_{0}^{u} e^{-A^2(x+v)^2}dx \cdot \int_{u}^{\infty} e^{-A^2(y-v)^2}dy \geq \int_{0}^{u} e^{-A^2(x-v)^2}dx \cdot \int_{u}^{\infty} e^{-A^2(y+v)^2}dy$$
$$\int_{0}^{u}\int_{u}^{\infty} e^{-A^2(x^2+y^2+2v^2+2v(x-y))}dydx \geq \int_{0}^{u}\int_{u}^{\infty} e^{-A^2(x^2+y^2+2v^2-2v(x-y))}dydx$$

On all the domain of integration we have $x \leq y$. Therefore the exponent is larger at every point for the left hand side, which confirms the second inequality, completing the proof of the lemma. □

Lemma 6 thus ensures that the monotone coupling preserves the ordering, and we can indeed use the recursive coupling scheme while keeping $Z(t) \leq Y'_u(t)$ at every step. □

Proposition 7. *For any $\alpha_3 > 0$ there exists $\beta_3 > 0$ with the following. For large enough A with probability at least $1 - \alpha_3$ we have $\nu_m < \beta_3 A^2$.*

Proof. First we look at the hitting time analogous to $\tilde{\nu}_m$ for Y'_u defined as $\hat{\nu}_m = \min\{s : Y'_u(s) \geq 1/2 - \delta\}$. Without aiming for tight estimates $\hat{\nu}_m \leq t$ can be ensured by $Y'_u(t) \geq 1/2 - \delta$ and by Proposition 5 this holds whenever $Z(t) \geq 1/2 - \delta$. The latter is equivalent to $\tilde{Z}(t) \notin [-1/2 + \delta, 1/2 - \delta]$.

For some $\beta_3 > 0$, the distribution of $\tilde{Z}(\beta_3 A^2)$ is $\mathcal{N}(X_u(0), \beta_3/2)$. Choosing β_3 large enough, the probability of this falling into $[-1/2 + \delta, 1/2 - \delta]$ can be made below $\alpha_3/2$ and this event is a superset of $\hat{\nu}_m > \beta_3 A^2$.

Now apply Lemma 3 with $\alpha_1 = \beta_3$. Note that $\tilde{\nu}_m \neq \hat{\nu}_m$ can only happen if $\nu_c^1 < \tilde{\nu}_m$. Also Lemma 4 ensures that ν_m and $\tilde{\nu}_m$ almost always coincide. Altogether, we have $\nu_m = \tilde{\nu}_m = \hat{\nu}_m < \beta_3 A^2$ with an exceptional probability at most $O(\exp(-\beta_2 A^2)) + O(\exp(-\beta_1 A^2)) + \alpha_3/2$, this stays below α_3 when A is large enough, which completes the proof. □

3.2 Diffusion from the middle

In the previous subsection we have seen that the process $Y_u(t)$ eventually has to reach the middle of the interval $[0, 1]$ as formulated in Proposition 7. Now we complement the analysis and consider the case when the process is initialized from the middle, meaning $Y_u(0) \in [1/2 - \delta, 1/2 + \delta]$. Intuitively, we expect the process to evolve as a random walk with independent Gaussian increments. However, we have to be careful

as boundary effects might alter the behavior of $Y_u(t)$ when it moves near the ends of the interval $[0,1]$. In this subsection we provide the techniques to estimate these boundary effects which will allow to conclude that the mixing of a random walk still translates to comparable mixing of $Y_u(t)$.

Let $W(t)$ be a random walk with i.i.d. $\mathcal{N}(0,1/(2A^2))$ increments, starting from $W(0) := Y_u(0)$. Our goal is to couple $W(t)$ with $Y_u(t)$ which only has a chance as long as $W(t)$ stays within $[0,1]$.

Definition 8. Let $\nu_c^2 := \min\{s : W(s) \notin [0,1]\}$.

Lemma 9. *There exist a coupling of the processes Y_u and W such that $Y_u(t) = W(t)$ whenever $t < \nu_c^2$.*

Proof. Assume the coupling holds until $t-1$, having $Y_u(t-1) = W(t-1)$. Let $\zeta \sim \mathcal{N}(0,1/(2A^2))$ be independent from the past, then define $W(t) = W(t-1) + \zeta$. For $Y_u(t)$, accept $Y_u(t-1)+\zeta$ if it is in $[0,1]$ otherwise redraw it according to $\pi(\cdot, Y_u(t-1))$.

The same values are obtained for the two processes at t except if $W(t)$ is outside $[0,1]$. This is exactly the event we wanted to indicate with ν_c^2 when we allow the two processes to decouple. □

Lemma 10. *For any $\alpha_4 > 0$ there exists $\beta_4 > 0$ with the following property. For A large enough, if $Y_u(0) \in [1/2-\delta, 1/2+\delta]$ there holds $P(\nu_c^2 < \alpha_4 A^2) < \beta_4$. We also have $\beta_4 \to 0$ as we choose $\alpha_4 \to 0$.*

Proof. We need to control the minimum and the maximum of a random walk where we use the following result of Erdős and Kac [4]:

Theorem 11 (Erdős-Kac)**.** *Let $\xi_1, \xi, \ldots$ i.i.d. random variables, $\mathbb{E}\xi_k = 0$, $D^2\xi_k = 1$. Let $S_k = \xi_1 + \xi_2 + \ldots + \xi_k$. Then for any $\alpha \geq 0$*

$$\lim_{n\to\infty} P(\max(S_1, S_2, \ldots, S_n) < \alpha\sqrt{n}) = \sqrt{\frac{2}{\pi}} \int_0^\alpha \exp(-x^2/2)dx.$$

Translating to the current situation, now that we use an initial value $Y_u(0) \in [1/2-\delta, 1/2+\delta]$ as a reference, we want an upper bound on the probability that the partial sums generating $W(t)$ never exceed $1/2-\delta$ (nor they go below $-1/2+\delta$). The increments have variance $1/(2A^2)$ and the number of steps is $\alpha_4 A^2$. Formally,

$$\begin{aligned}
&P\left(\max(0, W(1)-W(0), \ldots, W(\alpha_4 A^2) - W(0)) < 1/2-\delta\right) \\
&= P\left(\max(0, W(1)-W(0), \ldots, W(\alpha_4 A^2) - W(0))\sqrt{2}A < \frac{1-2\delta}{\sqrt{2\alpha_4}}\sqrt{\alpha_4 A^2}\right) \\
&\to \sqrt{\frac{2}{\pi}} \int_0^{\frac{1-2\delta}{\sqrt{2\alpha_4}}} \exp(-x^2/2)dx.
\end{aligned}$$

Now $\nu_c^2 < \alpha_4 A^2$ can only occur if this event fails and the maximum exceeds $1/2-\delta$, meaning $W(t)$ might exceed 1, or alternatively, the minimum of the process goes below $-1/2+\delta$ corresponding to $W(t)$ possibly leaving $[0,1]$ at 0. Consequently, we may fix any small $\varepsilon > 0$, then for any large enough A we get

$$P(\nu_c^2 < \alpha_4 A^2) \leq 2\left(1 - \sqrt{\frac{2}{\pi}} \int_0^{\frac{1-2\delta}{\sqrt{2\alpha_4}}} \exp(-x^2/2)dx\right) + \varepsilon =: \beta_4. \tag{1}$$

Observe that the right hand side of the expression indeed converges to 0 as $\alpha_4 \to 0$. □

Proposition 12. *There exists a constant $\alpha_5 > 0$ such that for A large enough, if $Y_u(0) \in [1/2-\delta, 1/2+\delta]$ we have*

$$\|\mathcal{L}(Y_u(\alpha_5 A^2)) - \pi_u\|_{\mathrm{TV}} < 1/3.$$

Proof. We introduce α_5 as a parameter. We will find sufficient conditions that ensure the claim of the proposition to hold, then pick a α_5 that satisfies the conditions found.

We first compare two simpler distributions, that of $W(\alpha_5 A^2)$ and the uniform μ. By the definition of $W(t)$, the distribution of $W(\alpha_5 A^2)$ is $\mathcal{N}(Y_u(0), \alpha_5/2)$.

$$\|\mathcal{L}(W(\alpha_5 A^2)) - \mu\|_{\mathrm{TV}} = \frac{1}{2}\int_{-\infty}^{\infty} \left| \frac{\exp(-(x - Y_u(0))^2/\alpha_5)}{\sqrt{\alpha_5 \pi}} - \mathbb{1}_{[0,1]}(x) \right| dx$$

The integrand has the form $|a - b|$ which we replace by $a + b - 2\min(a, b)$ (knowing these variables are non-negative). Also, as the probability density functions integrate to 1, we get

$$\begin{aligned}
&\|\mathcal{L}(W(\alpha_5 A^2)) - \mu\|_{\mathrm{TV}} = 1 - \int_{-\infty}^{\infty} \min\left(\frac{\exp(-(x - Y_u(0))^2/\alpha_5)}{\sqrt{\alpha_5 \pi}}, \mathbb{1}_{[0,1]}(x) \right) dx \\
&= 1 - \int_0^1 \min\left(\frac{\exp(-(x - Y_u(0))^2/\alpha_5)}{\sqrt{\alpha_5 \pi}}, 1 \right) dx \qquad (2) \\
&\leq 1 + 2\delta - \int_{-\delta}^{1+\delta} \min\left(\frac{\exp(-(x - 1/2)^2/\alpha_5)}{\sqrt{\alpha_5 \pi}}, 1 \right) dx =: \gamma.
\end{aligned}$$

The last inequality follows because the constant term is increased by 2δ, so is the length of the domain of the integration but the integrand is bounded above by 1. This step also involves an implicit change of variable depending on $Y_u(0)$, and it results in a final expression independent of this starting condition. The γ we get is also independent of A, it does depend on δ but has a limit as $\delta \to 0$.

The claim of the lemma is about two other distributions, now we relate them to the ones just compared. Using Lemma 10 for $\alpha_4 = \alpha_5$ we know that $Y_u(t)$ and $W(t)$ can be coupled well up to $t = \alpha_5 A^2$, which directly implies

$$\|\mathcal{L}(Y_u(\alpha_5 A^2)) - \mathcal{L}(W(\alpha_5 A^2))\|_{\mathrm{TV}} \leq \beta_4, \qquad (3)$$

where β_4 is the constant given by Lemma 10.

To compare π_u with μ we show π_u converges to μ in total variation as $A \to \infty$. For every $x \in [0, 1]$ define

$$p_u(x) = \frac{A}{\sqrt{\pi}} \int_{-x}^{1-x} \varphi(y) dy,$$

this is a function proportional to the density of π_u. By standard Gaussian tail estimates for all $x \in (0, 1)$ we get

$$1 - \frac{\exp(-A^2 x^2)}{2\sqrt{\pi} A x} - \frac{\exp(-A^2 (1 - x)^2)}{2\sqrt{\pi} A (1 - x)} \leq p_u(x) \leq 1.$$

Hence for all $x \in (0, 1)$, $p_u(x) \to 1$ as $A \to \infty$. These are uniformly bounded functions, so $\int_0^1 p_u \to 1$. The expression to consider for the convergence of the distributions is

$$\|\mu - \pi_u\|_{\mathrm{TV}} = \frac{1}{2}\int_0^1 \left| 1 - \frac{p_u(x)}{\int_0^1 p_u} \right| dx.$$

Here $1/\int_0^1 p_u$ is converging to 1 and is therefore bounded after some threshold, so the functions are eventually uniformly bounded and pointwise converging to 0. Thus the integrals also converge, and we get

$$\lim_{A \to \infty} \|\mu - \pi_u\|_{\mathrm{TV}} = 0. \qquad (4)$$

We can now combine our bounds of (2), (3) and (4):

$$\begin{aligned}
\|\mathcal{L}(Y_u(\alpha_5 A^2)) - \pi_u\|_{\mathrm{TV}} \leq \|\mathcal{L}(Y_u(\alpha_5 A^2)) - \mathcal{L}(W(\alpha_5 A^2))\|_{\mathrm{TV}} + \|\mathcal{L}(W(\alpha_5 A^2)) - \mu\|_{\mathrm{TV}} \\
+ \|\mu - \pi_u\|_{\mathrm{TV}} < \beta_4 + \gamma + \varepsilon,
\end{aligned}$$

where $\varepsilon > 0$ can be as small as wanted by setting A large enough. The proposition holds if we can ensure this sum to be small enough.

Note that a strong compromise is present for the choice of the constant α_5. In (3) we want to limit how likely the boundaries of the unit interval are to be reached, at the same time in (2) we want to show that $Y_u(s)$ is already spread out to some extent.

Still, a specific choice is possible. For $\alpha_5 = 0.10$ Lemma 10 provides $\beta_4 \approx 0.051$ when using $\delta = \varepsilon = 0$ and computer calculations for (1). By choosing $\delta, \varepsilon > 0$ but small enough, trusting computers but not too much, we can safely say $\beta_4 < 0.06$. In (2) using the same choice of α_5 we numerically get $\gamma \approx 0.263$ for $\delta = \varepsilon = 0$. Once again we allow a safety margin to only claim $\beta_4 + \gamma + \varepsilon < 1/3$. □

4 Overall mixing

We are now ready to establish mixing time bounds for the process we understand the best, $Y_u(t)$, then we will translate those results to the original process of interest $X(t)$.

Let us define

$$d(t) := \sup_{Y_u(0)\in[0,1]} \|\mathcal{L}(Y_u(t)) - \pi_u\|_{\mathrm{TV}},$$

which measures the distance from the stationary distribution from the worst starting point. We can give good bounds based on the previous sections:

Lemma 13. *There exists a constant $\beta_6 > 0$ such that $d(\beta_6 A^2) < 4/9$.*

Proof. Intuitively, from any starting point we can first wait for the process to reach the middle and then let the diffusion happen from there, as these are components we can already control.

Let us apply Proposition 7 with $\alpha_3 = 1/9$ providing a certain β_3. Once the process is in the middle part $[1/2 - \delta, 1/2 + \delta]$ we know by Proposition 12 that in the subsequent $\alpha_5 A^2$ steps sufficient diffusion occurs. Let $\beta_6 = \beta_3 + \alpha_5$.

Formally, fix $Y_u(0) \in [0, 1]$. We perform our calculations by conditioning on the value of ν_m.

$$\|\mathcal{L}(Y_u((\beta_3 + \alpha_5)A^2)) - \pi_u\|_{\mathrm{TV}} = \left\| \sum_{s=0}^{\infty} P(\nu_m = s)\mathcal{L}(Y_u((\beta_3 + \alpha_5)A^2) \mid \nu_m = s) - \pi_u \right\|_{\mathrm{TV}}.$$

Conditioned on $\nu_m = s$, $Y_u(s) \in [1/2-\delta, 1/2+\delta]$, therefore Proposition 12 provides $\|\mathcal{L}(Y_u(s+\alpha_5 A^2) \mid \nu_m = s) - \pi_u\|_{\mathrm{TV}} < 1/3$. We use this for $s \le \beta_3 A^2$, then performing $\beta_3 A^2 - s$ more steps can only decrease this distance, see [6, Chapter 4] for a detailed discussion about this. For $s > \beta_3 A^2$ we use the trivial bound on the total variation distance. We get

$$\|\mathcal{L}(Y_u((\beta_3 + \alpha_5)A^2)) - \pi_u\|_{\mathrm{TV}} \le \sum_{s=0}^{\beta_3 A^2} P(\nu_m = s) \cdot \frac{1}{3} + P(\nu_m > \beta_3 A^2) \cdot 1 \le \frac{1}{3} + \alpha_3 = \frac{4}{9}.$$

□

A slight variation of $d(t)$ compares the distribution of the process when launched from two different starting points.

$$\bar{d}(t) := \sup_{Y_u^1(0), Y_u^2(0)\in[0,1]} \|\mathcal{L}(Y_u^1(t)) - \mathcal{L}(Y_u^2(t))\|_{\mathrm{TV}},$$

Standard results provide the inequalities $d(t) \le \bar{d}(t) \le 2d(t)$ and the submultiplicativity $\bar{d}(s+t) \le \bar{d}(s)\bar{d}(t)$, see [6, Chapter 4]. The results therein are given for finite state Markov chains but are straightforward to translate to the current case of absolutely continuous distributions and transition kernels.

Proposition 14. *For any $0 < \alpha_7 < 1$ there exists $\beta_7 > 0$ such that*

$$t_{\mathrm{mix}}(Y_u, \alpha_7) \le \beta_7 A^2.$$

Proof. Using Lemma 13 for any $k \geq 1$ we get

$$d(k\beta_6 A^2) \leq \bar{d}(k\beta_6 A^2) \leq (\bar{d}(\beta_6 A^2))^k \leq (2d(\beta_6 A^2))^k \leq \left(\frac{8}{9}\right)^k.$$

For $k = \lceil \log \alpha_7 / \log(8/9) \rceil$ this is less than α_7 thus by setting $\beta_7 = \beta_6 \lceil \log \alpha_7 / \log(8/9) \rceil$ the process will be close enough to the stationary distribution as required at $t = \beta_7 A^2$. □

Lemma 15. *The mixing time of Y_u and Y are nearly the same, for any $0 < \alpha_7 < 1$*

$$t_{\text{mix}}(Y, \alpha_7) = t_{\text{mix}}(Y_u, \alpha_7) + 1.$$

Proof. First, we use that the total variation distance between the marginals is at most the distance between the overall distributions. Consequently, for any t we have $\|\mathcal{L}(Y_u(t-1)) - \pi_u\|_{\text{TV}} \leq \|\mathcal{L}(Y(t)) - \pi\|_{\text{TV}}$. This gives $t_{\text{mix}}(Y, \alpha_7) \geq t_{\text{mix}}(Y_u, \alpha_7) + 1$.

For the other direction, assume $\|\mathcal{L}(Y_u(t)) - \pi_u\|_{\text{TV}} \leq \alpha_7$ for some t. This means there is an optimal coupling with a random variable $\tilde{Y}_u^1$ having distribution π_u such that $P(Y_u(t) \neq \tilde{Y}_u^1) \leq \alpha_7$. As $\tilde{Y}_u^1$ has distribution π_u, it is possible to draw an additional random variable $\tilde{Y}_u^2$ to get $(\tilde{Y}_u^2, \tilde{Y}_u^1)$ with distribution π.

This is the same step when generating $Y_u(t+1)$ from $Y_u(t)$ thus we may keep the above coupling whenever already present. Therefore we have $P\big((Y_u(t+1), Y_u(t)) \neq (\tilde{Y}_u^2, \tilde{Y}_u^1)\big) \leq \alpha_7$ which can also be written as $\|\mathcal{L}(Y(t+1)) - \pi\|_{\text{TV}} \leq \alpha_7$. This implies $t_{\text{mix}}(Y, \alpha_7) \leq t + 1$, completing the proof. □

We are now ready to prove the main theorem of the paper, as stated in the introduction.

Theorem 1. Let $X(t)$ be the coordinate Gibbs sampler for the diagonal distribution. For any $0 < \alpha < 1$ there exists $\beta > 0$ such that for large enough A

$$t_{\text{mix}}(X, \alpha) \leq \beta A^2.$$

Proof. We use Proposition 14 with $\alpha_7 = \alpha/2$ and get a constant β_7 such that $t_{\text{mix}}(Y_u, \alpha/2) \leq \beta_7 A^2$ and by Lemma 15 also $t_{\text{mix}}(Y, \alpha/2) \leq \beta_7 A^2 + 1$. At each step the distribution of X^* and Y might differ only by flipping along the diagonal, which does not change the distance from the (symmetric) π thus also leaves the mixing time the same so we get $t_{\text{mix}}(X^*, \alpha_7/2) \leq \beta_7 A^2 + 1$.

The definition of X^* was based on the observation that when the same coordinate is rerandomized repeatedly, no additional mixing happens and the values at that coordinate simply get overwritten. Let us now quantify this effect, counting how many times did the direction of randomization change:

$$N(t) := |\{s \ : \ 1 \leq s \leq t-1, \ r(s) \neq r(s+1)\}|.$$

With this notation we see that $\mathcal{L}(X(t) \mid N(t) = k, r(1) = V) = \mathcal{L}(X^*(k+1))$ for all $t \geq 1$.

Without the loss of generality we now focus on the case of $r(1) = V$. Let us express the distribution of $X(t)$ conditioning on the value of $N(t)$.

$$\mathcal{L}(X(t) \mid r(1) = V) = \sum_{k=0}^{t-1} P(N(t) = k)\mathcal{L}(X^*(k+1)) = \sum_{k=0}^{t-1} \frac{1}{2^{t-1}} \binom{t-1}{k} \mathcal{L}(X^*(k+1)).$$

We substitute $t = 3\beta_7 A^2$ and evaluate the total variation distance from π.

$$\begin{aligned}
&\|\mathcal{L}(X(3\beta_7 A^2) \mid r(1) = V) - \pi\|_{\text{TV}} \leq \sum_{k=0}^{t-1} \frac{1}{2^{t-1}} \binom{t-1}{k} \|\mathcal{L}(X^*(k+1)) - \pi\|_{\text{TV}} \\
&\leq P(Binom(3\beta_7 A^2 - 1, 1/2) < \beta_7 A^2) \cdot 1 + 1 \cdot \|\mathcal{L}(X^*(\beta_7 A^2 + 1)) - \pi\|_{\text{TV}} \\
&\leq \exp(-\varepsilon \beta_7 A^2) + \frac{\alpha}{2}.
\end{aligned}$$

The last line holds with some positive ε by Hoeffding's inequality for the Binomial distribution and by substituting the upper bound on the total variation distance when we know k is above the mixing time. For large enough A this is below α.

By symmetry, the same bound holds for $\mathcal{L}(X(3\beta_7 A^2) \mid r(1) = U)$ and by convexity it is also true for the mixture of the two, the unconditional distribution of $X(3\beta_7 A^2)$. This concludes the proof with $\beta = 3\beta_7$. □

Finally, let us comment on the multitude of constants α_i, β_i appearing throughout the proofs, verifying that they can be consistently chosen when needed. First, a small enough $\delta > 0$ has to be picked for the proof of Proposition 12 which also relies on Lemma 10. Once it is fixed, observe that in the remaining Sections 3.1 and 4 all the constants only depend on other ones with lower indices, with the last α, β of Theorem 1 also depending on some previous ones. This excludes the issue of circular dependence.

5 Further estimates

In this section we complement the main result Theorem 1 by a lower bound showing that the order of A^2 is exact and by demonstrating the evolution of the distribution via numerical simulations.

Such a lower bound is plausible once having Lemma 9 and Lemma 10, these roughly say that when starting from the middle Y_u behaves like a random walk for order of A^2 steps and reaches only constant distance in order of A^2 steps. Let us proceed by forming a formal argument.

Theorem 16. *Let $X(t)$ be the coordinate Gibbs sampler for the diagonal distribution. There exists constants $\alpha', \beta' > 0$ such that for large enough A*

$$t_{\mathrm{mix}}(X, \alpha') > \beta' A^2.$$

First of all, to bound the mixing time from below it is sufficient to give a lower bound on the number of steps needed for a single starting point. In this spirit, we set $X(0) = (1/2, 1/2)$. With this choice, the arguments in Section 3.2 can be applied.

Set $S = [0, 1/4]^2 \cup [3/4, 1]^2$. Once we prove $\pi(S) - P(X(\beta' A^2) \in S) > \alpha'$ for a proper choice of α', β' that warrants a large total variation distance at the time $\beta' A^2$ and confirms our bound for the mixing time.

Lemma 17. $\pi(S) \geq 1/8$.

Proof. If we divide the unit square to 4-by-4 equal size smaller squares, then S is composed of two of these smaller squares, see Figure 1. It is enough to show that the selected squares forming S have greater or equal

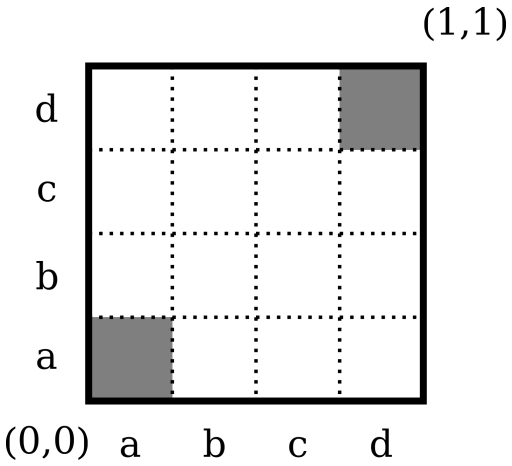


Figure 1: 4-by-4 division of the unit square to smaller squares. Horizontal and vertical intervals are labeled with letters for easier reference. The shaded area represents S.

probability than the other squares w.r.t. π, this directly confirms $\pi(S) \geq 2/16 = 1/8$.

To verify this, we compare the unnormalized density φ on them. We use the simple inequality that for $u, v \in [0, 1/4]$ and any $x \geq 0$ we have

$$\varphi(u - v) \geq \varphi(u - (1/2 - v + x)).$$

Indeed, note that $\varphi(y)$ is monotone decreasing with $|y|$. Then for $u \geq v, x = 0$ an easy comparison of the arguments provides the bound, while the other cases follow similarly. Observe that for $x = 0$ this inequality

compares φ at some point and its reflection to the line $v = 1/4$. Setting $x > 0$ corresponds to a further shift increasing the v coordinate after the reflection.

Using this we see that the points of the small square labeled by (a, a) in Figure 1 correspond to the points of (a, b) after a reflection so φ is pointwise larger on (a, a) by the above inequality. The same comparison holds against (a, c), (a, d) where an additional shift is necessary besides the reflection. Consequently, π is maximal for the (a, a) square compared to the others in its column.

Additionally, note that $\varphi(u - v)$ is invariant under the shift of (u, v) in the direction $(1, 1)$. Therefore π is exactly the same for the four squares on the diagonal. Furthermore, all the other squares are diagonally shifted and/or reflected (w.r.t. the diagonal) copies of the ones considered above, where we have seen that their probability is upper bounded by the probability of the square (a, a). The distribution π is symmetric w.r.t. the diagonal, so we conclude that (a, a) (and therefore also (d, d)) have indeed maximal probability among all squares. □

Lemma 18. *For any $\alpha_1' > 0$ there exists $\beta_1' > 0$ such that for large enough A any $t \le \beta_1' A^2$ satisfies*

$$P(Y(t) \in S) < \alpha_1'.$$

Proof. We want to rely on the previous observations that $Y_u(t)$ behaves like a random walk for a while with certain Gaussian increments. Using Lemma 10 we can choose $\alpha_4 > 0$ so that the corresponding β_4 goes below $\alpha_1'/2$. Let us denote this α_4 by β_2' for convenience.

Also, there exists $\beta_3' > 0$ so that

$$P(\mathcal{N}(1/2, \beta_3'/2) \in [0, 1/4] \cup [3/4, 1]) < \alpha_1'/2,$$

and clearly the same probability bound holds if the variance is decreased. Fixing $Y_u(0) = W(0) = 1/2$, the distribution of $W(\beta_3' A^2)$ is exactly $\mathcal{N}(1/2, \beta_3'/2)$.

To join our estimates we form

$$P(Y(t) \in S) \le P(Y_u(t) \in [0, 1/4] \cup [3/4, 1]) \le P(W(t) \ne Y_u(t)) + P(W(t) \in [0, 1/4] \cup [3/4, 1]).$$

For $t \le \beta_2' A^2$ the first term is below $\alpha_1'/2$ as it is an upper bound for the decoupling of W, Y_u to happen. For $t \le \beta_3' A^2$ the second term is below $\alpha_1'/2$. Altogether, if $t \le \min(\beta_2', \beta_3') A^2$,

$$P(Y(t) \in S) < \alpha_1'.$$

Therefore by choosing $\beta_1' = \min(\beta_2', \beta_3')$ we complete the proof. □

Proof of Theorem 16. Apply Lemma 18 with $\alpha_1' = 1/16$ to get some β_1'. The distribution of $X(\beta_1' A^2)$ is a mixture of the distributions of $Y(t)$ and their diagonally flipped version for $t \le \beta_1' A^2$, where t corresponds to how many times the rerandomization happened in a new direction. The set S is symmetric w.r.t. the diagonal so for $P(X(\beta_1' A^2) \in S)$ we can simply say it is a convex combination of $P(Y(t) \in S)$, $t \le \beta_1' A^2$ without needing any correction for the diagonal flip. Now by Lemma 18 each of these probabilities are below α_1', therefore it follows that

$$P(X(\beta_1' A^2) \in S) < \alpha_1' = \frac{1}{16}.$$

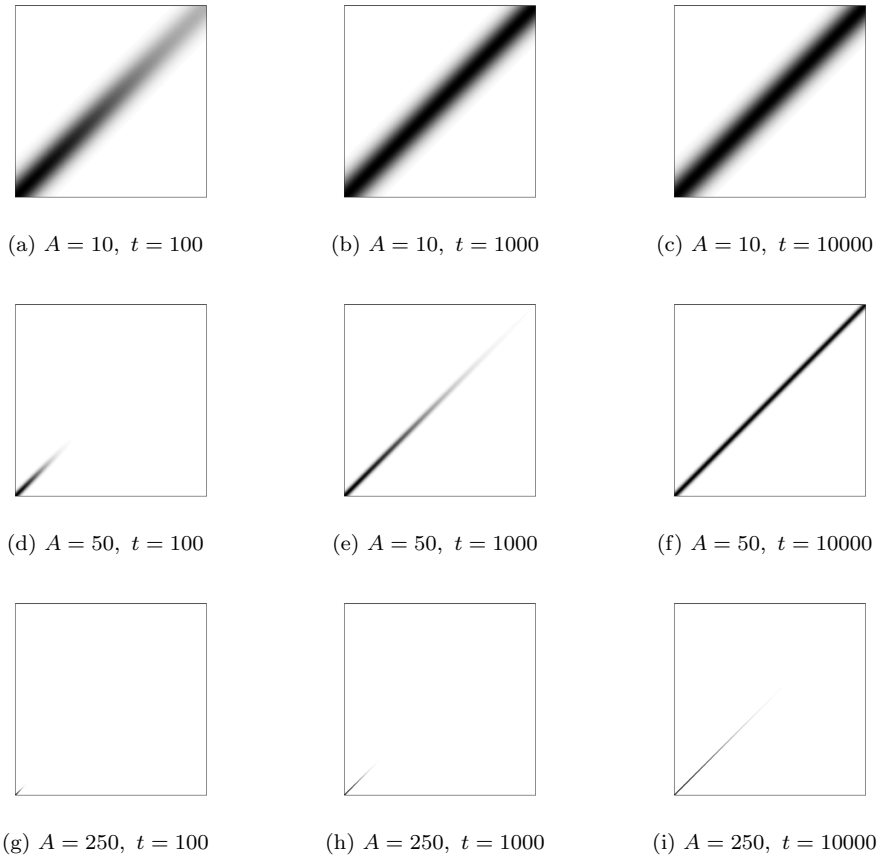
Comparing this with the statement of Lemma 17 we get

$$\pi(S) - P(X(\beta_1' A^2) \in S) > \frac{1}{16}.$$

Consequently $\|\mathcal{L}(X(\beta_1' A^2)) - \pi\|_{\mathrm{TV}} > 1/16$, so $t_{\mathrm{mix}}(X, 1/16) > \beta_1' A^2$. Thus the theorem holds with the choice $\alpha' = 1/16$, $\beta' = \beta_1'$. □

Finally, we present numerical approximations of the evolution of the distribution over time for different values of A. The unit square is discretized with a resolution of 500×500 and the distribution is calculated along these points. The starting point is always $(0, 0)$ at the lower left corner. The results are presented in Figure 2 for different A and different t. Both the convergence to the stationary distribution is visible and

also how this distribution becomes more concentrated along the diagonal for higher values of A. We also computed the time necessary to get within a total variation distance of $1/4$ of the stationary distribution, for $A = 10, t = 71$, for $A = 50, t = 1858$, for $A = 250, t = 47233$ is needed. This is a good proxy for the mixing time, note that only a single (but intuitively bad) starting point was tested and the discretization might have introduced some error. Still, the quadratic growth of t with respect to the increase of A is already apparent.



(a) $A = 10,\ t = 100$ (b) $A = 10,\ t = 1000$ (c) $A = 10,\ t = 10000$

(d) $A = 50,\ t = 100$ (e) $A = 50,\ t = 1000$ (f) $A = 50,\ t = 10000$

(g) $A = 250,\ t = 100$ (h) $A = 250,\ t = 1000$ (i) $A = 250,\ t = 10000$

Figure 2: Density of $\mathcal{L}(X(t))$ for different parameters A and t. Darker colors represent higher values (shades scale individually for each image).

Acknowledgments

The author would like to express his thanks to Persi Diaconis and György Michaletzky for their inspiring comments and to the American Institute of Mathematics for the stimulating workshop they hosted and organized.

References

[1] P. Diaconis, *The Markov chain Monte Carlo revolution*, Bulletin of the American Mathematical Society, 46 (2009), pp. 179–205.

[2] P. Diaconis and L. Saloff-Coste, *What do we know about the Metropolis algorithm?*, Journal of Computer and System Sciences, 57 (1998), pp. 20–36.

[3] M. Dyer, A. D. Flaxman, A. M. Frieze, and E. Vigoda, *Randomly coloring sparse random graphs with fewer colors than the maximum degree*, Random Structures & Algorithms, 29 (2006), pp. 450–465.

[4] P. Erdős and M. Kac, *On certain limit theorems of the theory of probability*, Bulletin of the American Mathematical Society, 52 (1946), pp. 292–302.

[5] R. Kannan, L. Lovász, and M. Simonovits, *Random walks and an $O^*(n^5)$ volume algorithm for convex bodies*, Random Structures & Algorithms, 11 (1997), pp. 1–50.

[6] D. Levin, Y. Peres, and E. Wilmer, *Markov chains and mixing times*, American Mathematical Society, 2009.

[7] L. Lovász and S. Vempala, *Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm*, Journal of Computer and System Sciences, 72 (2006), pp. 392–417.

[8] E. Mossel and A. Sly, *Gibbs rapidly samples colorings of $G(n, d/n)$*, Probability Theory and Related Fields, 148 (2010), pp. 37–69.