Concentration of measure and mixing for Markov chains

Malwina Luczak

Fifth Colloquium on Mathematics and Computer Science DMTCS proc. **AI**, 2008, 95–120

Concentration of measure and mixing for Markov chains

Malwina J. Luczak

Department of Mathematics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom
m.j.luczak@lse.ac.uk

We consider Markovian models on graphs with local dynamics. We show that, under suitable conditions, such Markov chains exhibit both rapid convergence to equilibrium and strong concentration of measure in the stationary distribution. We illustrate our results with applications to some known chains from computer science and statistical mechanics.

Keywords: Markov chains, concentration of measure, rapid mixing

1 Introduction

Recent years have witnessed a surge of activity in the mathematics of real-world networks, especially the study of combinatorial and stochastic models. Such networks include, for instance, the Internet, social networks, and biological networks. The techniques used to analyse them draw from a range of mathematical disciplines, such as graph theory, probability, statistical physics, analysis. Strikingly, random processes with rather similar characteristics can occur as models of very different real-world settings.

Random networks can often be regarded as interacting systems of individuals or particles. Under certain conditions, there is a law of large numbers, that is, a large system is close to a deterministic process solving a differential equation derived from the average 'drift', with much simpler dynamics. Further, one may frequently observe *chaoticity*, i.e. asymptotic approximate independence of particles. Unfortunately, it is often difficult to prove the validity of such approximations, especially when the random process has an unbounded number of components in the limit (e.g. the number of vertices or components of size k in a graph of size n, for $k = 1, 2, \ldots$, as $n \to \infty$).

In other instances, it may be difficult to establish good rates of convergence for mean-field approximations, or determine whether the long-term and equilibrium behaviour of the random process also follows that of the deterministic system. Furthermore, some recent attempts at a more accurate representation of real networks still await any kind of mathematically rigorous analysis. We would hope that over the coming years, the intense interest will produce a coherent and widely applicable theory. However, at present, it often appears that each new problem defies the existing theory in an interesting way.

In many complex systems, laws of large numbers and high concentration of measure in equilibrium have been found to co-exist with so-called *rapid mixing* [2; 12; 31], that is mixing in time $O(n \log n)$, where n is a measure of the system size. (Traditionally, such a system was considered to be rapidly mixing

if it converged to equilibrium in a time polynomial in n, but currently the term is more and more restricted to the 'optimal' mixing time $O(n \log n)$, see for example [31; 7].) There are some very notable examples of such behaviour, for instance, the subcritical Ising model, see [4; 15] and references therein, as well as the discussion in Section 3.1 of this paper.

The purpose of this article is to propose a new method to establish concentration of measure in complex systems modelled by Markov chains. We illustrate the technique with an application to a balls-and-bins model analysed in some earlier works by this author and McDiarmid, the *supermarket model* [18; 19]. Strong concentration of measure for this model, over long time intervals starting from a given state, as well as in equilibrium, was established in [18; 19] using the underlying structure of the model that enabled certain functions to be considered as functions of independent random variables so that the bounded differences method could be used.

In Section 4 of the present article we show that such concentration of measure inequalities hold more generally, with fewer assumptions on the structure of the Markov process involved. Our result is somewhat related, in spirit, to results (and arguments) in [16], which establishes transportation cost inequalities for the measure at time t and the stationary measure of a contracting Markov chain, assuming transportation cost inequalities for the kernel. However, the technical approach adopted here is rather different from [16] – discrete and coupling-based rather than functional analytic, and, we think, more 'hands on' and easier to use in practice (though our setting is less general than in [16]). It is striking that our approach, considerably more general than the one taken in [18], enables us to improve on the concentration of measure results proved in [18]. (Accordingly, we could also prove improved versions of results in [19], but we choose not to pursue this here.) The results in Section 4 also significantly extend Lemma 2.6 in [15], which bounds the variance of a real-valued, discrete-time, contracting Markov chain at time t and in equilibrium. We hope many more applications for the ideas presented here will be found in the future.

2 Notation and definitions

Let $X = (X_t)_{t \in \mathbb{Z}^+}$ be a discrete-time Markov chain with a discrete state space S and transition probabilities $P(x, y)$ for $x, y \in S$, where $\sum_{y \in S} P(x, y) = 1$ for each $x \in S$. We assume that, for every pair of states $x, y \in S$, $P(x, y) > 0$ if and only if $P(y, x) > 0$. Then we can form an undirected graph with vertex set S where $\{x, y\}$ is an edge if and only if $P(x, y) > 0$ and $x \neq y$. In general, our chains may be lazy, that is we can have $P(x, x) > 0$ for some $x \in S$. We assume that the graph is locally finite, that is, each vertex is adjacent to only finitely many other vertices. We now endow S with a graph metric d given by $d(x, y) = 1$ if $P(x, y) > 0$ and $x \neq y$, and for all other x, y $d(x, y)$ the length of the shortest path between x and y in the graph, which is assumed to be connected.

This kind of setting is natural and many models in applied probability and combinatorics fit into this framework, including those discussed in Section 3.

For each $t \in \mathbb{Z}^+$, X_t may be viewed as a random variable on a measurable space $(\Omega, \mathcal{F})$, where

$$\Omega = \{\omega = (\omega_0, \omega_1, \ldots) : \omega_i \in S \quad \forall i\},$$

and $\mathcal{F} = \sigma(\cup_{t=0}^{\infty} \mathcal{F}_t)$, with $\mathcal{F}_t = \sigma(X_i : i \leq t)$. Then each X_i is the i-co-ordinate projection, that is $X_i(\omega) = \omega_i$ for $i \in \mathbb{Z}^+$. Then the σ-fields $\mathcal{F}_t$ form the natural filtration for the process.

Let $\mathcal{P}(S)$ be the power set of S. The law of the Markov chain is a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$, and is determined uniquely by the transition matrix P together with a probability measure μ on $(S, \mathcal{P}(S))$

that gives the law of the initial state X_0, according to

$$\mathbb{P}(\{\omega : \omega_j = x_j : j \le i\}) = \mu(\{x_0\}) \prod_{j=0}^{i-1} P(x_j, x_{j+1}),$$

for each $x_0, \ldots, x_i \in S$, for each $i \in \mathbb{Z}^+$. This gives the law of (X_t) conditional on $\mathcal{L}(X_0) = \mu$, and will be denoted by $\mathbb{P}_\mu$ in what follows. Let $P^t(x, y)$ be the t-step transition probability from x to y, given inductively by

$$P^t(x, y) = \sum_{z \in S} P^{t-1}(x, z) P(z, y).$$

Then $\mathbb{P}_\mu(X_t \in A) = (\mu P^t)(A)$ for $A \subseteq S$.

Let $\mathbb{E}_\mu$ denote the expectation operator corresponding to $\mathbb{P}_\mu$. For $t \in \mathbb{Z}^+$ and $f : S \to \mathbb{R}$, define the function $P^t f$ by

$$(P^t f)(x) = \sum_y P^t(x, y) f(y), \quad x \in S.$$

In other words, $(P^t f)(x) = \mathbb{E}_{\delta_x}[f(X_t)] = (\delta_x P^t)(f)$, the expected value of $f(X_t)$ at time t conditional on the Markov process starting at x, i.e. the expectation of the function f with respect to measure $\delta_x P^t$. In general, we write $\mathbb{E}_\mu[f(X_t)] = (\mu P^t)(f)$.

A real-valued function f on S is said to be Lipschitz (or 1-Lipschtitz) if

$$\| f \|_{\text{Lip}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)} \le 1.$$

Here, equivalently, we only need to consider vertices at distance 1, so f is Lipschitz if and only if $\sup_{x,y:d(x,y)=1} |f(x) - f(y)| \le 1$.

Given a probability measure μ on $(S, \mathcal{P}(S))$ and an S-valued random variable X with law $\mathcal{L}(X) = \mu$, we say that μ or X has *normal concentration* if there exist constants $C, c > 0$ such that, for every $u > 0$, uniformly over 1-Lipschitz functions $f : S \to \mathbb{R}$,

$$\mu(|f(X) - \mu(f)| \ge u) \le C e^{-cu^2}. \tag{2.1}$$

We say that μ or X has *exponential concentration* if there exist constants $C, c > 0$ such that, for every $u > 0$, uniformly over 1-Lipschitz functions $f : S \to \mathbb{R}$,

$$\mu(|f(X) - \mu(f)| \ge u) \le C e^{-cu}. \tag{2.2}$$

These definitions are closely related to the notions used by Ledoux [14].

In Section 4 we shall give conditions under which a discrete-time Markov chain (X_t) exhibits normal concentration of measure over long time intervals and in equilibrium.

For probability measures μ_1, μ_2 on $(S, \mathbb{P}(S))$, the *total variation distance* between μ_1 and μ_2 is given by

$$d_{\text{TV}}(\mu_1, \mu_2) = \frac{1}{2} \sum_{x \in S} |\mu_1(x) - \mu_2(x)| = \sup_{A \subseteq S} |\mu_1(A) - \mu_2(A)|.$$

It is well known that the total variation distance satisfies

$$d_{\mathrm{TV}}(\mu_1, \mu_2) = \inf_{\pi} \pi(X \neq Y),$$

where the infimum is over all couplings $\pi = \mathcal{L}(X, Y)$ of S-valued random variables X, Y such that the marginals are $\mathcal{L}(X) = \mu_1$ and $\mathcal{L}(Y) = \mu_2$.

The *Wasserstein distance* between probability measures μ_1 and μ_2 is defined as

$$d_{\mathrm{W}}(\mu_1, \mu_2) = \sup_f \left| \int f d\mu_1 - \int f d\mu_2 \right| = \sup_f |\mu_1(f) - \mu_2(f)|,$$

where the supremum is over all measurable 1-Lipschitz functions $f : S \to \mathbb{R}$. By the Kantorovich – Rubinstein theorem (see [5], Section 11.8),

$$d_{\mathrm{W}} = \inf_{\pi} \{\pi[d(X, Y)] : \mathcal{L}(X) = \mu_1, \mathcal{L}(Y) = \mu_2\},$$

where the infimum is taken over all couplings π on $S \times S$ with marginals μ_1 and μ_2, and we write $\pi[d(X, Y)]$ for the expectation of $d(X, Y)$ under the coupling π. It is well known that the Wasserstein distance metrises weak convergence in spaces of bounded diameter. Also, since the discrete space $(S, \mathcal{P}(S))$ is necessarily complete and separable, so is the space of probability measures on $(S, \mathcal{P}(S))$ metrised by the Wasserstein distance. See [28] for detailed discussions of various metrics on probability measures and relationships between them.

3 Examples of rapid mixing and concentration

In this section we give some examples of known Markov chains exhibiting both concentration of measure in equilibrium and rapid mixing.

3.1 Mean-field Ising model

Let $G = (V, \mathcal{E})$ be a finite graph. Elements of the state space $S := \{-1, 1\}^V$ will be called *configurations*, and for $\sigma \in S$, the value $\sigma(v)$ will be called the *spin* at v. The *nearest-neighbour energy* $H(\sigma)$ of a configuration $\sigma \in \{-1, 1\}^V$ is defined by

$$H(\sigma) := - \sum_{\substack{v,w \in V, \\ v \sim w}} J(v, w)\sigma(v)\sigma(w), \tag{3.1}$$

where $w \sim v$ means that $\{w, v\} \in \mathcal{E}$. The parameters $J(v, w)$ measure the interaction strength between vertices; we will always take $J(v, w) \equiv J$, where J is a positive constant.

For $\beta \geq 0$, the *Ising model* on the graph G with parameter β is the probability measure π on S given by

$$\pi(\sigma) = \frac{e^{-\beta H(\sigma)}}{Z(\beta)}, \tag{3.2}$$

where $Z(\beta) = \sum_{\sigma \in \Omega} e^{-\beta H(\sigma)}$ is a normalising constant.

The parameter β is interpreted physically as the inverse of temperature, and measures the influence of the energy function H on the probability distribution. At *infinite temperature*, corresponding to $\beta = 0$, the measure π is uniform over S and the random variables $\{\sigma(v)\}_{v\in V}$ are independent.

The (single-site) *Glauber dynamics* for π is the Markov chain (X_t) on S with transitions as follows. When at σ, a vertex v is chosen uniformly at random from V, and a new configuration is generated from π conditioned on the set

$$\{\eta \in \Omega \,:\, \eta(w) = \sigma(w),\ w \neq v\}.$$

In other words, if vertex v is selected, the new configuration will agree with σ everywhere except possibly at v, and at v the spin is $+1$ with probability

$$p(\sigma; v) := \frac{e^{\beta M^v(\sigma)}}{e^{\beta M^v(\sigma)} + e^{-\beta M^v(\sigma)}}, \tag{3.3}$$

where $M^v(\sigma) := J \sum_{w\,:\, w\sim v} \sigma(w)$. Evidently, the distribution of the new spin at v depends only on the current spins at the neighbours of v. It is easily seen that (X_t) is reversible with respect to the measure π in (3.2), which is thus its stationary measure.

Given a sequence $G_n = (V_n, E_n)$ of graphs, write π_n for the Ising measure and $(X_t^{(n)})$ for the Glauber dynamics on G_n. For a given configuration $\sigma \in S_n$, let $\mathcal{L}(X_t^{(n)}, \sigma)$ denote the law of $X_t^{(n)}$ starting from σ. The worst-case distance to stationarity of the Glauber dynamics chain after t steps is

$$d_n(t) := \max_{\sigma\in S_n} d_{\mathrm{TV}}(\mathcal{L}(X_t^{(n)}, \sigma), \pi_n). \tag{3.4}$$

The *mixing time* $t_{\mathrm{mix}}(n)$ is defined as

$$t_{\mathrm{mix}}(n) := \min\{t \,:\, d_n(t) \leq 1/4\}. \tag{3.5}$$

Note that $t_{\mathrm{mix}}(n)$ is finite for each fixed n since, by the convergence theorem for ergodic Markov chains, $d_n(t) \to 0$ as $t \to \infty$. Nevertheless, $t_{\mathrm{mix}}(n)$ will in general tend to infinity with n. It is natural to ask about the growth rate of the sequence $t_{\mathrm{mix}}(n)$.

Definition 1 *The Glauber dynamics is said to exhibit a* cut-off *at* $\{t_n\}$ *with* window size $\{w_n\}$ *if* $w_n = o(t_n)$ *and*

$$\lim_{\gamma\to\infty} \liminf_{n\to\infty} d_n(t_n - \gamma w_n) = 1,$$
$$\lim_{\gamma\to\infty} \limsup_{n\to\infty} d_n(t_n + \gamma w_n) = 0.$$

Informally, a cut-off is a sharp threshold for mixing. For background on mixing times and cut-off, see [21].

Here we consider the mean-field case, taking G_n to be K_n, the complete graph on n vertices. That is, the vertex set is $V_n = \{1, 2, \ldots, n\}$, and the edge set $\mathcal{E}_n$ contains all $\binom{n}{2}$ pairs $\{i, j\}$ for $1 \leq i < j \leq n$. We take the interaction parameter J to be $1/n$; in this case, the Ising measure π on $\{-1, 1\}^n$ is given by

$$\pi(\sigma) = \pi_n(\sigma) = \frac{1}{Z(\beta)} \exp\left(\frac{\beta}{n} \sum_{1\leq i<j\leq n} \sigma(i)\sigma(j)\right). \tag{3.6}$$

In the physics literature, this is usually referred to as the *Curie-Weiss* model. To put this into the framework introduced in Section 2, the state space S consists of all n-vectors with components taking values in $\{-1, 1\}$, and two vectors are adjacent if they differ in exactly one co-ordinate.

It is a consequence of the Dobrushin-Shlosman uniqueness criterion that $t_{\rm mix}(n) = O(n \log n)$ when $\beta < 1$; see [1]. (See also [2; 31]). We shall see in Section 4 that, in the same regime, the stationary measure π (the Gibbs measure) exhibits normal concentration of measure for Lipschitz functions in the following sense. Let $X^{(n)}$ be a stationary version of $X_t^{(n)}$. Then, for some constants $c, C > 0$, for all $u > 0$

$$\mathbb{P}_\pi(|f(X^{(n)}) - \mathbb{E}_\pi(f(X^{(n)})) \geq u) \leq Ce^{-u^2/cn}, \tag{3.7}$$

uniformly over all 1-Lipschitz functions on S and over all n. Thinking about (3.7) simply as a statement about the measure π without any mention of the process $X_t^{(n)}$, we can also rewrite it in the form

$$\pi(\{\sigma : |f(\sigma) - \pi(f)| \geq u\}) \leq Ce^{-u^2/cn}.$$

Inequality (3.7) will follow from Theorem 4.1 (i), and is an improvement on Proposition 2.7 in [15].

More precise results about the speed of mixing for $\beta < 1$ can be found in [15], where the occurrence of a cut-off is established. The following is Theorem 1 from [15]:

Theorem 3.1 *Suppose that $\beta < 1$. The Glauber dynamics for the Ising model on K_n has a cut-off at $t_n = [2(1-\beta)]^{-1} n \log n$ with window size n.*

It is also easy to show, using the concentration of the Gibbs measure and the method used to prove Theorem 1.4 in [19], that asymptotically the spin values in a bounded set of vertices become almost independent. (In the language of [31] – see also references therein – this corresponds to the decay of correlations or spatial mixing.)

On the other hand, in the case $\beta \geq 1$, there is no rapid mixing, and no cut-off (see [15; 4] and references therein): $t_{\rm mix}(n)$ is of the order $n^{3/2}$ when $\beta = 1$ and is exponential in n when $\beta > 1$. For the same range of β, the Gibbs measure fails to exhibit normal concentration.

In particular, consider the function $m : S \to \mathbb{R}$ given by $m(\sigma) = \sum_{i=1}^n \sigma(i)$, the *magnetisation*; it is easy to see that $\frac{1}{2}m$ is 1-Lipschitz, and $\mathbb{E}_\pi(m(X)) = \pi(m) = 0$. However, when $\beta > 1$, then there is a constant $c > 0$ such that

$$\pi(\{\sigma : m(\sigma) \geq cn\}) = \pi(\{\sigma : m(\sigma) \leq -cn\}) \geq 1/4,$$

i.e. $m(X)$ is bi-modal for $\beta > 1$. While there is no bi-modality in the case $\beta = 1$, it is easy to calculate directly that $m(X)$ is not concentrated in the sense of (3.7). Further, for $\beta \geq 1$, the spins of vertices are no longer approximately independent for large n.

3.2 Supermarket model

Consider the following well-known queueing model with n separate queues, each with a single server. Customers arrive into the system in a Poisson process at rate λn, where $0 < \lambda < 1$ is a constant. Upon arrival each customer chooses d queues uniformly at random with replacement, and joins a shortest queue amongst those chosen (where she breaks ties by choosing the first of the shortest queues in the list of d). Here d is a fixed positive integer. Customers are served according to the first-come first-served discipline. Service times are independent exponentially distributed random variables with mean 1.

A number of authors have studied this model, as well as its extension to a Jackson network setting [10; 11; 18; 19; 20; 22; 23; 25; 30].

For instance, it is shown by Graham in [10] that the system is *chaotic*, provided that it starts close to a suitable deterministic initial state, or is in equilibrium. This means that the paths of members of any fixed finite subset of queues are asymptotically independent of one another, uniformly on bounded time intervals. This result implies a law of large numbers for the time evolution of the proportion of queues of different lengths, that is, for the empirical measure on path space [10]. In particular, for each fixed positive integer k_0, as n tends to infinity the proportion of queues with length at least k_0 converges weakly (when the infinite-dimensional state space is endowed with the product topology) to a function $v_t(k_0)$, where $v_t(0) = 1$ for all $t \geq 0$ and $(v_t(k) : k \in \mathbb{N})$ is the unique solution to the system of differential equations

$$\frac{dv_t(k)}{dt} = \lambda(v_t(k-1)^d - v_t(k)^d) - (v_t(k) - v_t(k+1)) \tag{3.8}$$

for $k \in \mathbb{N}$. Here one needs to assume appropriate initial conditions $(v_0(k) : k \in \mathbb{N})$ such that $1 \geq v_0(1) \geq v_0(2) \geq \cdots \geq 0$. Further, again for a fixed positive integer k_0, as n tends to infinity, in the equilibrium distribution this proportion converges in probability to $\lambda^{1+d+\cdots+d^{k_0-1}}$, and thus the probability that a given queue has length at least k_0 also converges to $\lambda^{1+d+\cdots+d^{k_0-1}}$.

Although the above results refer only to fixed queue length k_0 and bounded time intervals, they suggest that when $d \geq 2$, in equilibrium the maximum queue length may usually be $O(\log \log n)$. Indeed, one of the contributions of [18] is to show that this is indeed the case, and to give precise results on the behaviour of the maximum queue length. In particular, it turns out that when $d \geq 2$, with probability tending to 1 as $n \to \infty$, in the equilibrium distribution the maximum queue length takes at most two values; and these values are $\log \log n / \log d + O(1)$. Along the way, it is also shown in [18] that the system is rapidly mixing, that is the distribution settles down quickly to the equilibrium distribution. In this context, 'quickly' will mean 'in time $O(\log n)$, as this is a continuous time process with events happening at rate n, and so $O(\log n)$ corresponds to $O(n \log n)$ steps of the discrete-time jump chain. It is further established in [18] that the equilibrium measure is strongly concentrated.

Another natural question concerns fluctuations when in the equilibrium distribution: how long does it take to see large deviations of the maximum queue length from its stationary median? An answer is provided in [18] by establishing strong concentration estimates (for Lipschitz functions of the queue lengths vector) over time intervals of length polynomial in n. The techniques in [18] are partly combinatorial, and are used also in [17] and [19]. In particular, in [19], the concentration estimates obtained in [18] are used to establish quantitative results on the convergence of the distribution of a queue length and on 'propagation of chaos'.

Let us start by discussing the rapid mixing results known for the supermarket model. In [18] two rapid mixing results are established, one in terms of the Wasserstein distance and one in terms of the total variation distance. Unlike for the Ising model in Section 3.1, it turns out to be inappropriate to be looking at the worst-case mixing time, that is the supremum of the mixing times over all possible starting states. In the present case, this quantity is unbounded: the state space is unbounded, and the time to equilibrium from states x with the total number of customers $\| x \|_1 = k \gg n$ is of the order at least k. Then the best one can do is to obtain good upper bounds on the mixing time for copies of the Markov chain starting from nice states – that is, states where the queues are not too 'over-loaded'. This is made more precise below.

Let $X_t^{(n)}$ or X_t be the queue-lengths vector $(X_t^{(n)}(1), \ldots, X_t^{(n)}(n))$ in the supermarket model with n servers. For a positive integer n, $(X_t^{(n)})$ is an ergodic continuous-time Markov chain, with a unique distribution $\pi^{(n)}$ or π.

For any given state x write $\mathcal{L}(X_t^{(n)}, x)$ to denote the law of $X_t^{(n)}$ given $X_0^{(n)} = x$. Also, for $\epsilon > 0$, the mixing time $\tau^{(n)}(\epsilon, x)$ starting from x us defined by

$$\tau^{(n)}(\epsilon, x) = \inf\{t \geq 0 : d_{\mathrm{TV}}(\mathcal{L}(X_t^{(n)}, x), \pi^{(n)}) \leq \epsilon\}.$$

The result below, Theorem 1.1 in [18], shows that starting from an initial state in which the queues are not too long, the mixing time is small. In particular, if $\epsilon > 0$ is fixed and $\mathbf{0}$ denotes the all-zero n-vector, then $\tau^{(n)}(\epsilon, \mathbf{0})$ is $O(\log n)$.

Theorem 3.2 *Let $0 < \lambda < 1$ and let d be a fixed positive integer. For each constant $c > 0$ there exists a constant $\eta > 0$ such that the following holds for each positive integer n. Consider any distribution of the initial queue-lengths vector $X_0^{(n)}$, and for each time $t \geq 0$ let*

$$\delta_{n,t} = \mathbb{P}(|X_0^{(n)}| > cn) + \mathbb{P}(M_0^{(n)} > \eta t).$$

Then

$$d_{\mathrm{TV}}(\mathcal{L}(X_t^{(n)}), \pi^{(n)}) \leq ne^{-\eta t} + 2e^{-\eta n} + \delta_{n,t}.$$

The $O(\log n)$ upper bound on the mixing time τ is of the right order. Indeed, it is also proven in [18] that, for a suitable constant $\theta > 0$, if $t \leq \theta \log n$ then

$$d_{\mathrm{TV}}(\mathcal{L}(X_t^{(n)}), \pi^{(n)}) = 1 - e^{-\Omega(\log^2 n)}. \tag{3.9}$$

Thus $\tau^{(n)}(\epsilon, \mathbf{0})$ is $\Theta(\log n)$ as long as both ϵ^{-1} and $(1-\epsilon)^{-1}$ are bounded polynomially in n.

It would be interesting to consider the mixing times more precisely, to establish whether the supermarket model exhibits a cut-off. Again, here we should not be considering the worst-case mixing time, but rather the worst case over a subset of 'good' initial states, which are states where the total number of customers is not too large and the maximum queue not too long. Also, to bring the supermarket model into the discrete framework of Section 2, let us consider the jump chain of the supermarket model. We shall denote the jump chain by $\hat{X}_t^{(n)}$ or $\hat{X}_t$ in what follows, and its stationary measure by $\hat{\pi}^{(n)}$ or $\hat{\pi}$.

The transition probabilities of the jump chain are as follows. Given the state at time t is x, the next event is an arrival with probability $\lambda/(\lambda+1)$ and is a *potential* departure with probability $1/(\lambda+1)$. Here 'potential' means that it may be a departure or no change of state at all. Given that the next event is an arrival, the queue to which the new customer is sent is determined by selecting a uniformly random d-tuple of queues and directing the customer to a shortest queue among those chosen, in the same way as for the continuous-time process. Given that the next event is a potential departure, the departure queue is chosen uniformly at random from among all n queues. Then a customer will depart if the selected queue is non-empty; otherwise, nothing happens. It is easy to adapt the proofs in [18] (where the arguments are, in fact, based on analysing the jump chain) to show that Theorem 3.2 implies mixing in time of the order $O(n \log n)$ from initial states x such that $\| x \|_1 = O(n)$ and $\| x \|_\infty = O(\log n)$.

Accordingly, we make the following conjecture:

Conjecture 3.3 *Let c be a positive constant, and let $S_0^{(n)}$ be the set of all queue lengths vectors x in the n server supermarket model such that $\| x \|_1 \leq cn$ and $\| x \|_\infty \leq c \log n$. Let $\epsilon > 0$, and let*

$$d_n(\epsilon, t) = \sup_{x \in S_0^{(n)}} d_{\mathrm{TV}}(\mathcal{L}(\hat{X}_t^{(n)}, x), \hat{\pi}^{(n)}).$$

Then $d_n(\epsilon, t)$ has a cut-off in the sense of Definition 1, with window size n.

Our conjecture appears supported by some simulation results. Also it is supported by Conjecture 1 from [15], which states that the Glauber dynamics for the Ising model on transitive graphs G_n has a cutoff if the mixing time is $O(n \log n)$. The jump chain of the supermarket process is of a similar type to Glauber dynamics in that it makes only local transitions, and has mixing time of the order $O(n \log n)$, starting from good initial states. Also, it has a lot of symmetry – its stationary distribution is exchangeable. Thus the supermarket chain appears a good candidate for cut-off, though proving it may not be easy.

More generally, perhaps cut-off can be proven to be a phenomenon that also co-occurs with rapid mixing and concentration of measure in equilibrium much more widely, in the context of Markov chains whose jumps are suitably local.

In [18], the authors upper bound mixing in terms of the total variation distance by first upper bounding the Wasserstein distance between the distribution of the process at time t and the stationary distribution. The following result is Lemma 2.1 in [18].

Theorem 3.4 *Let $0 < \lambda < 1$ and let d be a fixed positive integer. For each constant $c > \frac{\lambda}{1-\lambda}$ there exists a constant $\eta > 0$ such that the following holds for each positive integer n. Let M denote the stationary maximum queue length. Consider any distribution of the initial queue-lengths vector X_0 such that $|X_0|$ has finite mean. For each time $t \geq 0$ let*

$$\delta_{n,t} = 2\,\mathbb{E}[|X_0| \mathbf{1}_{|X_0| > cn}] + 2cn\ \mathbb{P}(M_0 > \eta t).$$

Then

$$d_{\mathrm{W}}(\mathcal{L}(X_t), \pi) \leq n e^{-\eta t} + 2cn\, \mathbb{P}_\pi(M > \eta t) + 2e^{-\eta n} + \delta_{n,t}.$$

The upper bounds on the Wasserstein and total variation distance, and thus on the mixing time, are proven in [18] by means of a monotone coupling. The coupling takes two copies of the queueing process starting in adjacent states (that is, states differing in one customer in one queue) and couples their paths together in such a way that the ℓ_1-distance between them is non-increasing (and so always stays equal to 1 until the processes coalesce). Furthermore, the coupling is such that with high probability the ℓ_1-distance rapidly becomes 0. The coupling is then extended to all pairs of starting states with not too many customers in queues using the fact that the Wasserstein distance is a metric on the space of probability measures, or a *path-coupling* argument [2].

The property that the ℓ_1-distance is non-increasing in the coupling in [18] is very strong and not commonly encountered in path-coupling scenarios. This property is exploited in [18] to prove strong concentration of measure for the supermarket process, starting from a fixed (or highly concentrated state) for a long time interval. The following is Lemma 4.3 in [18].

Lemma 3.5 *There is a constant $c > 0$ such that the following holds. Let $n \geq 2$ be an integer and let f be a 1-Lipschitz function on the state space (set of all queue lengths vectors) S. Let also $x_0 \in S$ and*

assume that the queue-lengths process (X_t) satisfies $X_0 = x_0$ a.s. Let $\mu_t = \mathbb{E}_{\delta_{x_0}}[f(X_t)]$. Then for all times $t > 0$ and all $u \geq 0$,

$$\mathbb{P}_{\delta_{x_0}}(|f(X_t) - \mu_t| \geq u) \leq ne^{-\frac{cu^2}{nt+u}}. \tag{3.10}$$

Lemma 4.3 in [18] is proven by observing that the supermarket process can be 'simulated' by two independent Poisson processes, the arrivals process (with rate λn) and the (potential) departure process (with rate n), together with corresponding independent choices of queues (d independent uniformly random choices for each event in the arrivals process, and one uniformly random choice in the departures process). One then conditions on the number of events in the interval $[0, t]$, and then the state at time t is conditionally determined by a finite family of independent random variables. In other words, the argument is, just like most of the other arguments in [18], based on studying the jump chain $(\hat{X}_t)$, although this is not made explicit therein.

The non-increasing distance coupling property is used to show that a Lipschitz function of the queue lengths vector must satisfy a bounded differences condition, so that the discrete bounded differences inequality can be applied to show concentration of measure for Lipschitz functions in the conditional space. The proof is then completed by deconditioning.

The rapid mixing result can be combined with the long-term concentration of measure result to prove concentration of measure in equilibrium for Lipschitz functions of the queue-lengths vector. The following is Lemma 4.1 in [18].

Lemma 3.6 *There is a constant $c > 0$ such that the following holds. Let $n \geq 2$ be an integer and consider the n-queue system. Let the queue-lengths vector Y have the equilibrium distribution. Let f be a 1-Lipschitz function on S. Then for each $u \geq 0$*

$$\mathbb{P}_\pi\left(|f(Y) - \mathbb{E}_\pi[f(Y)]| \geq u\right) \leq ne^{-cu/n^{\frac{1}{2}}}. \tag{3.11}$$

Lemmas 3.5 and 3.6 prove strong concentration of measure – normal concentration for small deviations and exponential concentration for larger deviations in the case of starting from a fixed state, and exponential concentration in equilibrium. The factor n in the bound on the right-hand sides of both (3.10) and (3.11) is a limitation of the technique and not the right answer. It is natural to expect the truth to be a lot better – that it can be replaced by a constant. In Section 4 we develop concentration inequalities that achieve that. Although we work with the discrete-time jump chain, it is easy to see that our results apply also to the continuous time chain. One further advantage of our inequalities is that they apply to other settings – for instance where rapid mixing is established by a coupling, but the coupling does not have additional useful properties such as the non-increasing Wasserstein distance.

Even so Lemmas 3.5 and 3.6 are quite powerful. We now explore, briefly, some results concerning the queue lengths in the supermarket model in equilibrium that can be obtained using Lemma 3.6. The following is Lemma 4.2 in [18]. (We drop the subscript π to lighten up the notation.)

Lemma 3.7 *Consider the n-queue system, and let the queue-lengths vector Y have the equilibrium distribution. For each non-negative integer k, let $\ell(k, y)$ denote the number of queues of length at least k in state y. Also, for each non-negative integer k, let $\ell(k) = \mathbb{E}[\ell(k, Y)]$. Then for any constant $c > 0$,*

$$\mathbb{P}(\sup_k |\ell(k, Y) - \ell(k)| \geq cn^{\frac{1}{2}} \log^2 n) = e^{-\Omega(\log^2 n)}.$$

Also, there exists a constant $c > 0$ such that

$$\sup_k \mathbb{P}(|\ell(k, Y) - \ell(k)| \geq cn^{\frac{1}{2}} \log n) = o(1).$$

Furthermore, for each integer $r \geq 2$

$$\sup_k |\,\mathbb{E}[\ell(k, Y)^r] - \ell(k)^r| = O(n^{r-1} \log^2 n).$$

Lemma 5.1 in [18], stated below, yields further precise information about the equilibrium behaviour, over long time intervals.

Lemma 3.8 *Let $K > 0$ be an arbitrary constant and let $\tau = n^K$. Let (Y_t) be in equilibrium and let $c > 0$ be a constant. Let B_τ be the event that for all times t with $0 \leq t \leq \tau$*

$$\sup_i |\ell(i, Y_t) - n\lambda^{1+d+\cdots+d^{i-1}}| \leq cn^{1/2} \log^2 n.$$

Then $\mathbb{P}(\overline{B_\tau}) \leq e^{-\Omega(\log^2 n)}$.

In [18], Lemma 5.1 is used to prove two-point concentration for the stationary maximum queue length and its concentration on only a constant number of values over long time intervals. This is Theorem 1.3 in [18]:

Theorem 3.9 *Let $0 < \lambda < 1$ and let $d \geq 2$ be an integer. Then there exists an integer-valued function $m_d = m_d(n) = \log\log n/\log d + O(1)$ such that the following holds. For each positive integer n, suppose that the queue-lengths vector $Y_0^{(n)}$ is in the stationary distribution (and thus so is the maximum queue length $M_t^{(n)}$). Then for each time $t \geq 0$, $M_t^{(n)}$ is $m_d(n)$ or $m_d(n) - 1$ with probability tending to 1 as $n \to \infty$; and further, for any constant $K > 0$ there exists $c = c(K)$ such that, with probability tending to 1 as $n \to \infty$,*

$$\max_{0 \leq t \leq n^K} |M_t^{(n)} - \log\log n/\log d| \leq c. \tag{3.12}$$

The functions $m_2(n)$, $m_3(n)$, ... may be defined as follows. For $d = 2, 3, \ldots$ let $i_d(n)$ be the least integer i such that $\lambda^{\frac{d^i-1}{d-1}} < n^{-\frac{1}{2}} \log^2 n$. Then we let $m_2(n) = i_2(n) + 1$, and for $d \geq 3$ let $m_d(n) = i_d(n)$. (As we have seen, with high probability the proportion of queues of length at least i is close to $\lambda^{\frac{d^i-1}{d-1}}$.)

Also, equation (37) in [18] shows that, for $r = O(\log n)$,

$$\mathbb{P}(M \geq m_d(n) + r) \leq e^{-cr \log n}, \tag{3.13}$$

for a constant $c > 0$.

In [19], strong concentration of measure results from [18] are used to show that in equilibrium the distribution of a typical queue length converges to an explicit limiting distribution and provide explicit convergence rates. Let $Y^{(n)}(1)$ denote the equilibrium length of of queue 1. (Note that the equilibrium distribution is exchangeable.) The following is Theorem 1.1 in [19]. Let $\mathcal{L}_{\lambda,d}$ denote the law of a random variable Y such that $\mathbb{P}(Y \geq k) = \lambda^{(d^k-1)/(d-1)}$ for each $k = 0, 1, \ldots$.

Theorem 3.10 *For each positive integer n let $Y^{(n)}$ be a queue-lengths n-vector in equilibrium, and consider the length $Y^{(n)}(1)$ of queue 1. Then*

$$d_{\mathrm{TV}}(\mathcal{L}(Y^{(n)}(1)), \mathcal{L}_{\lambda,d})$$

is of order n^{-1} up to logarithmic factors.

In fact, it is proven in [19] that the above total variation distance is $o(n^{-1}\log^3 n)$ and is $\Omega(n^{-1})$. Also, the following holds (Corollary 1.2 in [19]).

Corollary 3.11 *For each positive integer k, the difference between the kth moment $\mathbb{E}[Y^{(n)}(1)^k]$ and the kth moment of $\mathcal{L}_{\lambda,d}$ is of order n^{-1} up to logarithmic factors.*

The above results concern the distribution of a single queue length. One may also consider collections of queues and chaoticity. The terms 'chaoticity' and 'propagation of chaos' come from statistical physics [13], and the original motivation was the evolution of particles in physical systems. The subject has since then received considerable attention, especially following the ground-breaking work of Sznitman [29].

The result below (Theorem 1.4 in [19]) establishes chaoticity for the supermarket model in equilibrium. We see that for fixed r the total variation distance between the joint law of r queue lengths and the product law is at most $O(n^{-1})$, up to logarithmic factors. More precisely and more generally we have:

Theorem 3.12 *For each positive integer n, let $Y^{(n)}$ be a queue-lengths n-vector in equilibrium. Then, uniformly over all positive integers $r \le n$, the total variation distance between the joint law of $Y^{(n)}(1), \dots, Y^{(n)}(r)$ and the product law $\mathcal{L}(Y^{(n)}(1))^{\otimes r}$ is at most $O(n^{-1}\log^2 n(2\log\log n)^r)$; and the total variation distance between the joint law of $Y^{(n)}(1), \dots, Y^{(n)}(r)$ and the limiting product law $\mathcal{L}_{\lambda,d}^{\otimes r}$ is at most $O(n^{-1}\log^2 n(2\log\log n)^{r+1})$.*

Analogous time-dependent results (away from equilibrium) are also given in [19] – proven using Lemma 3.5 above (Lemma 4.3 in [18]) but we omit them here for the sake of brevity. Let us mention that the arguments used in [19] to prove Theorems 1.1 and 1.4 (Theorems 3.10 and 3.12 above) are quite generic and would apply in many other settings. The main property needed is concentration of measure for Lipschitz functions of the state vector, the polynomial form of the generator of the Markov process, and, in the case of Theorem 1.1, also the exchangeability of the stationary distribution. The chaoticity result Theorem 3.12 above is a quantitative version of some of the results in [29].

To conclude this section, we mention that analogues of results in [18; 19] are proved in [17] for a related balls-and-bins model, where, instead of queueing up to receive service on a first-come first-served basis, customers (balls) have independent exponentially distributed 'lifetimes' and each departs its queue (bin) as soon as its lifetime has expired.

Current work in progress [9] includes extensions of the results in [18; 19] to the supermarket model where the number of choices $d = d(n)$ and the arrival rate $\lambda = \lambda(n)$ are n-dependent, including the interesting case where $d \to \infty$ and $\lambda \to 1$ with various functional dependencies between λ and d.

4 Coupling and bounded differences method generalised

This section contains our main results and applications. We use the notation introduced in Section 2.

Let us state our first theorem, which gives concentration of measure for Lipschitz functions of a discrete-time Markov chain on state space S and with transition matrix P at time t, under assumptions on the Wasserstein distance between its i step transition measures for $i \le t$.

Theorem 4.1 *Let P be the transition matrix of a discrete-time Markov chain with discrete state space S.*

(i) *Let $(\alpha_i : i \in \mathbb{N})$ be a sequence of positive constants such that, for all i,*

$$\sup_{x,y \in S : d(x,y)=1} d_{\mathrm{W}}(\delta_x P^i, \delta_y P^i) \leq \alpha_i. \tag{4.1}$$

Let f be a 1-Lipschitz function. Then for all $u > 0$, $x_0 \in S$, and $t > 0$,

$$\mathbb{P}_{\delta_{x_0}}(|f(X_t) - \mathbb{E}_{\delta_{x_0}}[f(X_t)]| \geq u) \leq 2e^{-u^2/2(\sum_{i=1}^t \alpha_i^2)}. \tag{4.2}$$

(ii) *More generally, let S_0 be a non-empty subset of S, and let $(\alpha_i : i \in \mathbb{N})$ be a sequence of positive constants such that, for all i,*

$$\sup_{x,y \in S_0 : d(x,y)=1} d_{\mathrm{W}}(\delta_x P^i, \delta_y P^i) \leq \alpha_i. \tag{4.3}$$

Let

$$S_0^0 = \{x \in S_0 : y \in S_0 \text{ whenever } d(x,y) = 1\}.$$

Let f be a 1-Lipschitz function. Then for all $x_0 \in S_0^0$, $u > 0$ and $t > 0$,

$$\mathbb{P}_{\delta_{x_0}}\left(\{|f(X_t) - \mathbb{E}_{\delta_{x_0}}[f(X_t)]| \geq u\} \cap \{X_s \in S_0^0 : \ 0 \leq s \leq t\}\right) \leq 2e^{-u^2/2(\sum_{i=1}^t \alpha_i^2)}. \tag{4.4}$$

If the Markov chain becomes contractive after a finite number of steps, then one can deduce from Theorem 4.1 concentration results for the stationary measure of the Markov chain, as in the following corollary.

Corollary 4.2 (i) *Suppose that there exists $x \in S$ and a sequence $\alpha_i : S \to \mathbb{R}^+$ of functions such that, for all $y \in S$,*

$$d_{\mathrm{W}}(\delta_x P^i, \delta_y P^i) \leq \alpha_i(y), \tag{4.5}$$

where $\alpha_i(y) \to 0$ as $i \to \infty$ for each y, and

$$\sup_k \mathbb{E}_{\delta_x}[\alpha_i(X_k)] = \sup_k (P^k \alpha_i)(x) \to 0 \quad \text{as } i \to \infty. \tag{4.6}$$

Then (X_t) has a unique stationary measure π, and $\delta_y P^t \to \pi$ as $t \to \infty$ for each y.

(ii) *Suppose that (4.1) holds, and the constants α_i in Theorem 4.1 satisfy $\sum_i \alpha_i^2 < \infty$. Suppose further there exists $x \in S$ such that*

$$\sup_k (P^k g)(x) < \infty,$$

where $g(y) = d(x,y)$. Then (X_t) has a unique stationary measure π, $\delta_x P^t \to \pi$ as $t \to \infty$ for each x.

Furthermore, let X be a stationary copy of X_t. Then, for all $u > 0$, and uniformly over all 1-Lipschitz functions f,

$$\mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi[f(X)]| \geq 2u) \leq 2e^{-u^2/2(\sum_{i=1}^\infty \alpha_i^2)}. \tag{4.7}$$

(iii) *Suppose that (X_t) has a unique stationary measure π and condition (4.3) holds, where $\sum_i \alpha_i^2 < \infty$. Let $x \in S_0^0$, and suppose $\delta > 0$ and $t_0 > 0$ are such that $d_{\mathrm{W}}(\delta_x P^{t_0}, \pi) < \delta$ and*

$$\mathbb{P}_{\delta_x}(X_t \in S_0^0 \text{ for } t \leq t_0) \geq 1 - \delta.$$

Let X be a stationary copy of X_t. Then, for all $u \geq \delta$, uniformly over all 1-Lipschitz functions f,

$$\mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi[f(X)]| \geq 2u) \leq 2e^{-u^2/2(\sum_{i=1}^{t_0} \alpha_i^2)} + 2\delta. \tag{4.8}$$

Proof:

(i) Consider the sequence P_i of measures on $(S, \mathcal{P}(S))$ given by $P_i = \delta_x P^i$; we have, using the coupling characterisation of the Wasserstein distance,

$$\begin{aligned} d_{\mathrm{W}}(P_i, P_{i+k}) &= d_{\mathrm{W}}(\delta_x P^i, (\delta_x P^k)P^i) \leq \sum_{y \in S} (\delta_x P^k)(y) d_{\mathrm{W}}(\delta_x P^i, \delta_y P^i) \\ &\leq \sum_{y \in S} (\delta_x P^k)(y)\alpha_i(y) \leq \sup_k \mathbb{E}_{\delta_x}[\alpha_i(X_k)] \to 0 \end{aligned}$$

as $i \to \infty$, by assumption. Thus the sequence (P_i) is a Cauchy sequence and so, since the space of probability measures on $(S, \mathcal{P}(S))$ is complete with respect to the Wasserstein distance, it must converge to a probability measure π on $(S, \mathcal{P}(S))$. It is obvious that this measure must be stationary for P.

Now, take $y \in S$, and let $Q_i = \delta_y P^i$. Then

$$d_{\mathrm{W}}(P_i, Q_i) = d_{\mathrm{W}}(\delta_x P^i, \delta_y P^i) \leq \alpha_i(y) \to 0 \quad \text{as } i \to \infty.$$

It follows that $Q_i \to \pi$ as $i \to \infty$, and so π must be the unique stationary measure.

(ii) The assumption that $\sum_i \alpha_i^2 < \infty$ implies that $\alpha_i \to 0$ as $i \to \infty$. Then it is easily seen (using the fact that the distance $d(y, z)$ between each pair y, z of states in finite) that conditions (4.5) and (4.6) of part (i) hold for x, with $\alpha_i(y) \leq \alpha_i d(x, y)$, and so, as in (i) one can prove that there exists a (necessarily unique) stationary measure π, and that $\delta_x P^t \to \pi$ as $t \to \infty$ for each $x \in S$.

Let us now prove the concentration of measure result, inequality (4.7). Take some $x \in S$. Given $\epsilon > 0$, for t large enough the Wasserstein distance, and hence the total variation distance, between $\delta_x P^t$ and π is at most ϵ. Then, for $u \geq \epsilon$ and all such t, by Theorem 4.1 part (i),

$$\begin{aligned} \mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi[f(X)]| \geq 2u) &\leq \mathbb{P}_{\delta_x}(|f(X_t) - \mathbb{E}_{\delta_x}[f(X_t)]| \geq u) + \varepsilon \\ &\leq 2e^{-u^2/2(\sum_{i=1}^{\infty} \alpha_i^2)} + \varepsilon. \end{aligned}$$

Here we have used the fact that

$$|\,\mathbb{E}_\pi[f(X)] - \mathbb{E}_{\delta_x}[f(X_t)]| \leq \epsilon \leq u.$$

Since ε is arbitrary, the result follows.

(iii) Let

$$A_{t_0} = \{\omega : X_t(\omega) \in S_0 \ \forall t \in [0, t_0]\}.$$

Arguing as in (ii), and using Theorem 4.1 part (ii), we can write, for $u \geq \delta$,

$$\begin{aligned} \mathbb{P}_\pi(|f(X) - \mathbb{E}_\pi[f(X)]| \geq 2u) &\leq \mathbb{P}_{\delta_x}(|f(X_{t_0}) - \mathbb{E}_{\delta_x}[f(X_{t_0})]| \geq u) + \delta \\ &\leq \mathbb{P}_{\delta_x}\left(\{|f(X_{t_0}) - \mathbb{E}_{\delta_x}[f(X_{t_0})]| \geq u\} \cap A_{t_0}\right) \end{aligned}$$

$$\begin{aligned} &+ \quad 2\delta \\ &\leq \quad 2e^{-u^2/2(\sum_{i=1}^{t_0}\alpha_i^2)} + 2\delta, \end{aligned}$$

as required.

□

To prove Theorem 4.1, we shall make use of a concentration inequality from [26]. Let $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ be a probability space, with $\tilde{\Omega}$ finite. Let $\tilde{\mathcal{G}} \subseteq \tilde{\mathcal{F}}$ be a σ-field. Given a bounded random variable Z on $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, the *supremum* of Z in $\tilde{\mathcal{G}}$ is the $\tilde{\mathcal{G}}$-measurable function given by

$$\sup(Z|\tilde{\mathcal{G}})(\omega) = \min_{A \in \tilde{\mathcal{G}}: \omega \in A} \max_{\omega' \in A} Z(\omega'). \tag{4.9}$$

Thus $\sup(Z)$ takes the value at ω equal to the maximum value of Z over the 'smallest' event in $\tilde{\mathcal{G}}$ containing ω. Since $\tilde{\Omega}$ is finite, we are assured that the smallest event containing ω does exist; the arguments used here would work also in many cases where $\tilde{\Omega}$ is countably infinite.

The *conditional range* of Z in $\tilde{\mathcal{G}}$, denoted by $\mathrm{ran}(Z)$, is the $\tilde{\mathcal{G}}$-measurable function

$$\mathrm{ran}(Z \mid \tilde{\mathcal{G}}) = \sup(Z|\tilde{\mathcal{G}}) + \sup(-Z|\tilde{\mathcal{G}}). \tag{4.10}$$

Let $\{\emptyset, \tilde{\Omega}\} = \tilde{\mathcal{F}}_0 \subseteq \tilde{\mathcal{F}}_1 \subseteq \ldots$ be a filtration in $\tilde{\mathcal{F}}$, and let $Z_0, \ldots,$ be the martingale obtained by setting $Z_t = \mathbb{E}(Z|\tilde{\mathcal{F}}_t)$ for each t. For each t let ran_t denote $\mathrm{ran}(Z_t|\tilde{\mathcal{F}}_{t-1})$; by definition, ran_t is an $\tilde{\mathcal{F}}_{t-1}$-measurable function. For each t, let the *sum of squared conditional ranges* R_t^2 be the random variable $\sum_{i=1}^t \mathrm{ran}_i^2$, and let the *maximum sum of squared conditional ranges* $\hat{r}_t^2$ be the supremum of the random variable R_t^2, that is

$$\hat{r}_t^2 = \sup_{\tilde{\omega} \in \tilde{\Omega}} R_t^2(\tilde{\omega}).$$

The following result is Theorem 3.14 in [26].

Lemma 4.3 *Let Z be a bounded random variable on a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ with $\tilde{\mathbb{E}}(Z) = m$. Let $\{\emptyset, \tilde{\Omega}\} = \tilde{\mathcal{F}}_0 \subseteq \tilde{\mathcal{F}}_1 \subseteq \ldots \subseteq \tilde{\mathcal{F}}_t$ be a filtration in $\tilde{\mathcal{F}}$. Then for any $u \geq 0$,*

$$\tilde{\mathbb{P}}(|Z - m| \geq u) \leq 2e^{-2u^2/\hat{r}_t^2}.$$

More generally, for any $u \geq 0$ and any value r_t^2,

$$\tilde{\mathbb{P}}(\{|Z - m| \geq u\} \cap \{R_t^2 \leq r_t^2\}) \leq 2e^{-2u^2/r_t^2}.$$

Proof of Theorem 4.1. Let $f : S \to \mathbb{R}$ be 1-Lipschitz. Fix a time $t \in \mathbb{N}$, $x_0 \in S$ and consider the evolution of X_t conditional on $X_0 = x_0$ for t steps, that is until time t. Since we have assumed that there are only a finite number of possible transitions from any given $x \in S$, we can build this conditional process until time t on a finite probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}_{\delta_{x_0}})$: we can take $\tilde{\Omega}$ to be the finite set of all possible paths of the process starting at time 0 in state x_0 until time t, and $\tilde{\mathcal{F}}$ to be the power set of $\tilde{\Omega}$.

In the conditional space, for each time $j = 0, \ldots, t$, let $\tilde{\mathcal{F}}_j = \sigma(X_0, \ldots, X_j)$, the σ-field generated by $X_0, \ldots, X_j$; so $\tilde{\mathcal{F}}_0 = \{\emptyset, \tilde{\Omega}\}$ and $\tilde{\mathcal{F}}_t = \tilde{\mathcal{F}}$. We write $\mathbb{E}$ instead of $\tilde{\mathbb{E}}$ in what follows to lighten the notation.

Consider the random variable $Z = f(X_t) : \tilde{\Omega} \to \mathbb{R}$. Also, for $j = 0, \ldots, t$ let Z_j be given by

$$Z_j = \mathbb{E}[f(X_t)|\tilde{\mathcal{F}}_j] = \mathbb{E}_{\delta_{x_0}}[f(X_t)|X_0, \ldots, X_j] = (P^{t-j}f)(X_j),$$

where we have used the Markov property in the last equality.

Fix $1 \leq j \leq t$; we want to upper bound $\mathrm{ran}_j = \mathrm{ran}(Z_j \mid \tilde{\mathcal{F}}_{j-1})$. Fix also $x_1, \ldots, x_{j-1} \in S$, and for $x \in S$ consider

$$\begin{aligned} g(x) &= \mathbb{E}[f(X_t)|X_j = x] = \mathbb{E}[f(X_{t-j})|X_0 = x] \\ &= (P^{t-j}f)(x). \end{aligned}$$

Note that $Z_j(\tilde{\omega}) \in \{g(x) : d(x, x_{j-1}) \leq 1\}$ for $\tilde{\omega}$ such that $X_{j-1}(\tilde{\omega}) = x_{j-1}$. It follows that, for such $\tilde{\omega}$,

$$\mathrm{ran}_j(\tilde{\omega}) = \sup_{x,y:d(x,x_{j-1})\leq 1, d(y,x_{j-1})\leq 1} |g(x) - g(y)|.$$

Let us prove part (i) of the theorem. As f is 1-Lipschitz,

$$\begin{aligned} \sup_{x,y:d(x,y)\leq 2} |g(x) - g(y)| &= \sup_{x,y:d(x,y)\leq 2} |(P^{t-j}f)(x) - (P^{t-j}f)(y)| \\ &= \sup_{x,y:d(x,y)\leq 2} |\,\mathbb{E}_{\delta_x P^{t-j}}(f) - \mathbb{E}_{\delta_y P^{t-j}}(f)| \\ &\leq 2 \sup_{x,y:d(x,y)\leq 1} |\,\mathbb{E}_{\delta_x P^{t-j}}(f) - \mathbb{E}_{\delta_y P^{t-j}}(f)| \\ &\leq 2 \sup_{x,y:d(x,y)\leq 1} d_{\mathrm{W}}(\delta_x P^{t-j}, \delta_y P^{t-j}) \\ &\leq 2\alpha_{t-j}, \end{aligned}$$

by assumption. We deduce that $\mathrm{ran}_j(\tilde{\omega}) \leq 2\alpha_{t-j}$ for all $\tilde{\omega} \in \tilde{\Omega}$. It follows that

$$\hat{r}_t^2(\tilde{\omega}) \leq 4 \sum_{r=0}^{t-1} \alpha_{t-r}^2,$$

uniformly over $\tilde{\omega} \in \tilde{\Omega}$. Part (i) of Theorem 4.1 now follows from Lemma 4.3.

To prove (ii), observe that the bound

$$\mathrm{ran}_j(\omega) = \mathrm{ran}(Z_j \mid \tilde{\mathcal{F}}_{j-1})(\omega) \leq 2\alpha_{t-j}$$

still holds on the event $A_t = \{\omega : X_j(\omega) \in S_0^0 \text{ for } j = 0, \ldots, t\}$. ∎

The following special case of model satisfying the hypotheses of Theorem 4.1 is of particular interest and has received considerable attention in computer science literature; see for instance [2; 8; 12]. Suppose (4.1) is satisfied with $\alpha_i = \alpha^i$, where $0 < \alpha < 1$ is a constant. In the language of [2] this corresponds to the following situation. Consider different copies $(X_t), (X_t')$ of the process with initial states x, x' respectively, that is $X_0 = x$ and $X_0' = x'$ almost surely. Suppose that we can couple $(X_t), (X_t')$ so that, uniformly over all pairs of states $x, x' \in S$ with $d(x, x') = 1$,

$$\mathbb{E}[d(X_1, X_1')|X_0 = x, X_0' = x'] \leq \alpha,$$

for a constant $0 < \alpha < 1$. Thus, under the coupling, $(X_t), (X'_t)$ will be getting closer and closer together on average as t gets larger, which implies strong mixing properties [2; 12]. Then, uniformly over $x, x' \in S$ with $d(x, x') = 1$, $d_{\mathrm{W}}(\delta_x P, \delta_{x'} P) \leq \alpha$. By 'path coupling' [2; 12]

$$\mathbb{E}[d(X_1, X'_1)|X_0 = x, X'_0 = x'] \leq \alpha d(x, x'),$$

and hence $d_{\mathrm{W}}(\delta_x P, \delta_{x'} P) \leq \alpha d_{\mathrm{W}}(\delta_x, \delta_{x'})$ for all pairs $x, x' \in S$. By induction on t,

$$d_{\mathrm{W}}(\delta_x P^t, \delta_{x'} P^t) \leq \alpha^t d(x, x')$$

for all $x, x' \in S$ and all $t \in \mathbb{N}$. Then, in the same notation as earlier, we can upper bound

$$\hat{r}^2 \leq 4 \sum_{r=1}^{t} \alpha^{2r} \leq 4\alpha^2 (1 - \alpha^2)^{-1},$$

for all t. Hence we obtain the following corollary.

Corollary 4.4 *Suppose that there is a constant $0 < \alpha < 1$ such that*

$$d_{\mathrm{W}}(\delta_x P, \delta_{x'} P) \leq \alpha \tag{4.11}$$

for all $x, x' \in S$ such that $d(x, x') = 1$. Then for all $t > 0$

$$\mathbb{P}_{\delta_{x_0}}(|f(X_t) - \mathbb{E}_{\delta_{x_0}}[f(X_t)]| \geq u) \leq 2e^{-u^2(1-\alpha^2)/2\alpha^2} \tag{4.12}$$

for all $u > 0$, all $x_0 \in S$, and for every 1-Lipschitz function on S.

Hence, if X has the equilibrium distribution π then, for all $u > 0$ and every 1-Lipschitz function f,

$$\mathbb{P}_{\pi}(|f(X) - \mathbb{E}_{\pi}[f(X)]| \geq u) \leq 2e^{-u^2(1-\alpha^2)/2\alpha^2} \tag{4.13}$$

The particular choice of $\alpha = 1 - c_1/n$ for a constant $c_1 > 0$ corresponds to the 'optimal' mixing time $O(n \log n)$ for a Markov chain in a system with size measure n, and gives concentration of measure in equilibrium of the form

$$\mathbb{P}_{\pi}(|f(X_t) - \mathbb{E}_{\pi}[f(X_t)]| \geq u) \leq 2e^{-u^2/c_2 n}, \tag{4.14}$$

where $c_2 > 0$ is a constant. This is the case, for example, for the subcritical ($\beta < 1$) mean-field Ising model discussed in Section 3 – see for example [21] or [15] for a description of the coupling that implies fast decay of the Wasserstein distance. The same also applies to the Glauber dynamics for colourings on bounded-degree graphs analysed in [7] (see also [8] and [27]). The application is straightforward when the number of colours k is greater than $2D$, where D is the maximum degree of the graph. It is only a little more involved in the case $(2-\eta)D \leq k \leq 2D$, where the proof in [7] relies on *delayed path-coupling* [3], whereby a new Markov chain is used with one step corresponding to cn steps of the original one, n being the size of the graph to colour.

On the other hand $\alpha = 1 - 6/(n^3 - n)$ for the Glauber dynamics on linear extensions of a partial order of size n [2; 12] gives an upper bound $O(n^3 \log n)$ on mixing. The corresponding bound on deviations of a 1-Lipschitz function from its mean of size u is of the form $2e^{-u^2/cn^3}$, which is useless. However, one cannot do much better in general. To see this, consider the partial order on n points consisting of a chain

of length $n-1$ and a single incomparable element. It is not hard to check that in this case the mixing time is of the order n^3 – see [2] for details. It is also easy to see that there is no normal concentration of measure in the sense of (4.14).

We shall now apply Theorem 4.1 and Corollary 4.2 to the supermarket process described in Section 3.2, or rather to the corresponding discrete-time jump chain $\hat{X}_t$. Recall that, when in state x, the next event is an arrival with probability $\lambda/(1+\lambda)$, and is a potential departure with probability $1/(1+\lambda)$. Given that the next event is an arrival, the queue to which the arrival will go is determined by selecting a uniformly random d-tuple of queues and sending the customer to a shortest one among those chosen, ties being split by always going to the first best queue in the list. Given that the next event is a potential departure, the departure queue is chosen uniformly at random among the n possible queues, and departures from empty queues are ignored. In the Markov chain graph, two states are connected by an edge if and only if they differ exactly in one customer in one queue. Then a function f is 1-Lipschitz if and only if it is 1-Lipschitz with respect to the ℓ_1 distance on the state space S.

We focus on the case $d \geq 2$. For $d = 1$, in equilibrium the queue lengths are independent geometric random variables, so normal concentration of measure can be obtained using the standard bounded differences inequality [26].

By Lemma 2.3 in [18], for all $x, y \in S$ such that $d(x,y) = 1$, and all $t \geq 0$,

$$d_{\mathrm{W}}(\delta_x P^t, \delta_y P^t) \leq 1.$$

Let c be a positive constant, and let S_0 be given by

$$S_0 = \{x \in S : \| x \|_1 \leq cn, \| x \|_\infty \leq c \log n\}.$$

It is very easy to modify the proof of Lemma 2.6 in [18] to show that, if $x, y \in S_0$ and $d(x,y) = 1$, then for some constants $\alpha, \beta > 0$,

$$d_{\mathrm{W}}(\delta_x P^t, \delta_y P^t) \leq e^{-\beta t/n} + 2e^{-\beta n} \tag{4.15}$$

for $t \geq \alpha n \log n$.

Take a constant $K > 2$ and let $\tau = n^K$. Then we can put $\alpha_i = 1$ for $t \leq \alpha n \log n$, and $\alpha_i = e^{-\beta t/n} + 2e^{-\beta n}$ for $\alpha n \log n < t \leq \tau$. Then for $t \leq \tau$, we can upper bound

$$\sum_{i=1}^{t} \alpha_i^2 \leq \min\{t, \alpha n \log n + n^{1-\beta/\alpha}\beta^{-1} + 2e^{-\beta n/2}\} \leq \min\{t, 2\alpha n \log n\}.$$

Consider the all-empty state, $\mathbf{0} \in S_0^0$. Then by choosing the constant c in the definition of S_0 sufficiently large, we can ensure that, for $d \geq 2$,

$$\mathbb{P}_{\mathbf{0}}(\hat{X}_t \in S_0^0 \ \forall\, t \leq \tau) \geq 1 - e^{-(\log n)^2/c}.$$

This follows from Lemma 2.3 (monotone coupling for given n and d), Lemma 2.4 (a) and the monotone coupling for given n and different d, d' (see the proof of Lemma 2.4 in [18]) and equation (37) in [18]. (See also the statements of these results in Section 3.2.)

By Theorem 4.1 (i), we can choose c sufficiently large so that, for all $t > 0$, all $u > 0$, and every Lipschitz function f,

$$\mathbb{P}_{\delta_{\mathbf{0}}}(|f(\hat{X}_t) - \mathbb{E}_{\delta_{\mathbf{0}}}[f(\hat{X}_t)]| \geq u) \leq 2e^{-u^2/ct}. \tag{4.16}$$

By Theorem 4.1 (ii), for $\alpha n \log n \leq t \leq \tau$, and all $u > 0$,

$$\mathbb{P}_{\delta_{\mathbf{0}}}(|f(\hat{X}_t) - \mathbb{E}_{\delta_{\mathbf{0}}}[f(\hat{X}_t)]| \geq u) \leq 2e^{-u^2/\alpha n \log n} + e^{-(\log n)^2/c}. \tag{4.17}$$

In particular, for $\alpha n \log n \leq t \leq \tau$, and $u \leq c_0\sqrt{n}\log n$,

$$\mathbb{P}_{\delta_{\mathbf{0}}}(|f(\hat{X}_t) - \mathbb{E}_{\delta_{\mathbf{0}}}[f(\hat{X}_t)]| \geq u) \leq 2e^{-u^2/cn \log n}, \tag{4.18}$$

provided that c is large enough. Inequalities (4.16) – (4.18) improve on what one could obtain for the jump chain from Lemma 3.5 above, for an interesting range of u and t – and it is easy to use them to derive improved concentration of measure inequalities for the continuous chain also. (It is possible to optimise inequality (4.17) by playing with the definition of S_0 to obtain normal concentration for larger u.)

We now want to relate this to concentration of measure in equilibrium, via Corollary 4.2. It is easy to see from earlier work (see [18] and references therein) that the supermarket jump chain has a unique stationary measure. (This could also be proven showing that the hypotheses of Corollary 4.2 (i) are satisfied, via (4.15) above.)

By Lemma 2.1 in [18] and straightforward calculations for the Poisson process, there is a constant $\eta > 0$ such that

$$d_{\mathrm{W}}(\mathcal{L}(\hat{X}_t, \mathbf{0}), \hat{\pi}) \leq ne^{-\eta t/n} + 2cn\,\mathbb{P}_{\hat{\pi}}(M > \eta t/n) + 2e^{-\eta n}, \tag{4.19}$$

where M denotes the maximum queue length in equilibrium, and we may take c the same as in the definition of S_0, assuming that c is sufficiently large. Thus, by (4.19),

$$d_{\mathrm{W}}(\mathcal{L}(\hat{X}_\tau, \mathbf{0}), \hat{\pi}) \leq (n + 2cn + 2)e^{-\eta n}.$$

Let $\hat{Y}$ denote the queue lengths vector in equilibrium. It then follows by Corollary 4.2 (iii), uniformly for all 1-Lipschitz functions f, for $u \geq 1$ and n sufficiently large

$$\mathbb{P}_{\hat{\pi}}(|f(\hat{Y}) - \mathbb{E}_{\hat{\pi}}[f(\hat{Y})]| \geq 2u) \leq 2e^{-u^2/cn \log n} + 2e^{-(\log n)^2/c}. \tag{4.20}$$

So, choosing c to be sufficiently large, for all $u > 0$ and n sufficiently large,

$$\mathbb{P}_{\hat{\pi}}(|f(\hat{Y}) - \mathbb{E}_{\hat{\pi}}[f(\hat{Y})]| \geq 2u) \leq ce^{-u^2/cn \log n} + ce^{-(\log n)^2/c}. \tag{4.21}$$

This improves on Lemma 3.6 above, and gives normal concentration for $u = O(n^{1/2}(\log n)^{3/2})$ (again, it is possible to obtain normal concentration for larger u), but is not the optimal result we are after. In particular, we still cannot show that deviations of size $n^{1/2}\omega(n)$ have probability tending to 0 for $\omega(n)$ tending to infinity arbitrarily slowly. We will now derive another inequality that will enable us to achieve our aim.

Theorem 4.5 *Assume that there exists a set S_0 and numbers $\alpha_i(x, y)$ ($x, y \in S_0$, $i \in \mathbb{N}$) such that, for all i, and all $x, y \in S_0$ with $d(x, y) = 1$,*

$$d_{\mathrm{W}}(\delta_x P^i, \delta_y P^i) \leq \alpha_i(x, y). \tag{4.22}$$

Let

$$S_0^0 = \{x \in S_0 : y \in S_0 \text{ whenever } d(x, y) = 1\}.$$

For $x \in S$, let $g_x(y) = d_{\mathrm{W}}(\delta_y P^i, \delta_x P^i)^2$. Assume that, for some sequence $(\alpha_i : i \in \mathbb{N})$ of positive constants,

$$\sup_{x_0 \in S_0^0} (Pg_{x_0})(x_0) \leq \alpha_i^2. \tag{4.23}$$

Let $t > 0$, let $v = \sum_{i=1}^t \alpha_i^2$, and let

$$\hat{\alpha} = \sup_{1 \leq j \leq t} \sup_{x,y \in S_0 : d(x,y) \leq 2} \alpha_j(x,y). \tag{4.24}$$

Let also $A_t = \{\omega \in \Omega : X_s(\omega) \in S_0^0 \; \forall 0 \leq s \leq t\}$

Then, for all $u > 0$, and uniformly over all 1-Lipschitz functions f,

$$\mathbb{P}_{\delta_{x_0}} \left(|f(X_t) - \mathbb{E}_{\delta_{x_0}}[f(X_t)]| \geq u \cap A_t \right) \leq 2e^{-u^2/(4v(1+(\hat{\alpha}u/6v))}. \tag{4.25}$$

To prove Theorem 4.5, we use another result from [26]. With notation as before, for $j = 1, \ldots, t$, let

$$\mathrm{var}_j = \mathrm{var}(Z_j \mid \tilde{\mathcal{F}}_{j-1}) = \mathbb{E}\left((Z_j - \mathbb{E}(Z_j \mid \tilde{\mathcal{F}}_{j-1}))^2 \mid \tilde{\mathcal{F}}_{j-1} \right);$$

let $V = \sum_{j=1}^t \mathrm{var}_j$. Also, for $j = 1, \ldots, t$, let $\mathrm{dev}_j = \sup(|Z_j - Z_{j-1}| \mid \tilde{\mathcal{F}}_{j-1})$, and let $\mathrm{dev} = \sup_j \mathrm{dev}_j$. The following result is essentially Theorem 3.15 in [26].

Lemma 4.6 *Let Z be a random variable on a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ with $\mathbb{E}(Z) = m$. Let $\{\emptyset, \tilde{\Omega}\} = \tilde{\mathcal{F}}_0 \subseteq \tilde{\mathcal{F}}_1 \subseteq \ldots \subseteq \tilde{\mathcal{F}}_t$ be a filtration in $\tilde{\mathcal{F}}$. Let $\hat{b} = \max \mathrm{dev}$, the maximum conditional deviation (and assume that $\hat{b}$ is finite). Then for any $u \geq 0$,*

$$\mathbb{P}(|Z - m| \geq u) \leq 2e^{-u^2/(2\hat{v}(1+(\hat{b}u/3\hat{v}))},$$

where $\hat{v}$ is the maximum sum of conditional variances (which is assumed to be finite).

More generally, for any $u \geq 0$ and any values $b, v \geq 0$,

$$\mathbb{P}(\{|Z - m| \geq u\} \cap \{V \leq v\} \cap \{\max \mathrm{dev} \leq b\}) \leq 2e^{-u^2/(2v(1+(bu/3v))}.$$

Proof of Theorem 4.5. The proof is similar to the proof of Theorem 4.1. Let $f : S \to \mathbb{R}$ be 1-Lipschitz. Fix a time $t \in \mathbb{N}$, an $x_0 \in S$ and consider the evolution of X_t conditional on $X_0 = x_0$ for t steps, that is until time t. Again this conditional process can be supported by a finite probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}}_{\delta_{x_0}})$.

As before, in the conditional space, for each time $j = 0, \ldots, t$ let $\tilde{\mathcal{F}}_j = \sigma(X_0, \ldots, X_j)$, the σ-field generated by $X_0, \ldots, X_j$; so $\tilde{\mathcal{F}}_0 = \{\emptyset, \tilde{\Omega}\}$ and $\tilde{\mathcal{F}}_t = \tilde{\mathcal{F}}$. Again, we consider the random variable $Z = f(X_t) : \tilde{\Omega} \to \mathbb{R}$. And, for $j = 0, \ldots, t$, Z_j is given by

$$Z_j = \mathbb{E}[f(X_t)|\tilde{\mathcal{F}}_j] = \mathbb{E}_{\delta_{x_0}}[f(X_t)|X_0, \ldots, X_j] = (P^{t-j}f)(X_j).$$

Suppose first for simplicity that $S_0 = S$. We want to apply Lemma 4.6 and for this we need to calculate the conditional variances var_j. To do this, we use the fact that the variance of a random variable Y is equal

to $\frac{1}{2}\,\mathbb{E}(Y-\tilde{Y})^2$, where $\tilde{Y}$ is another random variable with the same distribution as Y and independent of Y.

Fix j and $x_1,\dots,x_{j-1}\in S$, and for $x\in S$ consider

$$\begin{aligned} g(x) &= \mathbb{E}[f(X_t)|X_j=x]=\mathbb{E}[f(X_{t-j})|X_0=x] \\ &= (P^{t-j}f)(x). \end{aligned}$$

Then, for $\tilde{\omega}$ such that $X_{j-1}(\tilde{\omega})=x_{j-1}$, $Z_j(\tilde{\omega})\in\{g(x):d(x,x_{j-1})\le 1\}$, so that

$$\begin{aligned} \mathrm{var}_j(\tilde{\omega}) &= \frac{1}{2}\sum_{x,y}P(x_{j-1},x)P(x_{j-1},y)(g(x)-g(y))^2 \\ &\le \frac{1}{2}\sum_{x,y:d(x_{j-1},x)\le 1,d(x_{j-1},y)\le 1}P(x_{j-1},x)P(x_{j-1},y)d_{\mathrm{W}}(\delta_x P^{t-j},\delta_y P^{t-j})^2 \\ &\le 2\sum_{x:d(x_{j-1},x)\le 1}P(x_{j-1},x)d_{\mathrm{W}}(\delta_x P^{t-j},\delta_{x_{j-1}}P^{t-j})^2 \\ &\le 2\sum_{x}P(x_{j-1},x)\alpha_{t-j}(x_{j-1},x)^2 \\ &\le 2\alpha_{t-j}^2, \end{aligned}$$

by assumption (4.23).

Then we can upper bound the sum

$$\hat{v}\le 2\sum_{j=1}^{t}\alpha_j^2.$$

It remains to bound $\mathrm{dev}=\sup_j \mathrm{dev}_j$. We have, for $\tilde{\omega}$ such that $X_{j-1}(\tilde{\omega})=x_{j-1}$,

$$\begin{aligned} \mathrm{dev}_j(\tilde{\omega}) &\le \sup_{x:d(x,x_{j-1})\le 1}|g(x)-(P^{t-j+1}f)(x_{j-1})| \\ &= \sup_{x:d(x,x_{j-1})\le 1}|(P^{t-j}f)(x)-(P^{t-j+1}f)(x_{j-1})| \\ &\le \sup_{x:d(x,x_{j-1})\le 1}|d_{\mathrm{W}}(\delta_x P^{t-j},\delta_{x_{j-1}}P^{t-j+1}). \end{aligned}$$

It follows that, for each $j=1,\dots,t$,

$$\begin{aligned} \mathrm{dev}_j &\le \sup_{x,y:d(x,y)\le 1}d_{\mathrm{W}}(\delta_x P^{t-j+1},\delta_y P^{t-j}) \\ &\le \sup_{x,y:d(x,y)\le 2}d_{\mathrm{W}}((\delta_x P)P^{t-j},\delta_y P^{t-j}) \\ &\le \hat{\alpha}, \end{aligned}$$

by (4.24) and using the coupling characterisation of the Wasserstein distance. Theorem 4.5 now follows from the first statement in Lemma 4.6 in the case where $S_0=S$. In general, the above bounds on $\hat{v}$ and

dev hold on the event $A_t = \{\omega : X_j(\omega) \in S_0^0 \text{ for } j = 0, \ldots, t\}$, and so Theorem 4.5 also follows from the second statement of Lemma 4.6. ∎

Let us now apply Theorem 4.5 to the supermarket model from [18] discussed above. Again, we focus on the case $d \geq 2$.

Let c be a positive constant, and let S_0 be given by

$$\{x \in S : \ell(k, x) = \sum_{r=1}^{n} \mathbf{1}_{x(r) \geq k} \leq ne^{-k/c} \text{ for } k = 1, \ldots\}.$$

Consider the all-empty state, $\mathbf{0} \in S_0^0$. Let $K > 2$ be a constant. We claim that we can choose c sufficiently large that, if $\tau = n^K$, then

$$\mathbb{P}_{\mathbf{0}}(\hat{X}_t \in S_0^0 : t \leq \tau) \geq 1 - e^{-(\log n)^2/c}.$$

This follows easily from Lemma 3.8 in the present paper, together with equation (3.13).

We now want to calculate the quantity in (4.23). For a state $x_0 \in S_0^0$ and a state x chosen with probability $P(x_0, x)$, these states will only differ in a queue of length greater than k if $P(x_0, x)$ is a probability of an event involving a queue of length at least k – a departure from a queue of length at least k or an arrival into a queue of length at least k. For $x_0 \in S_0^0$ such a transition happens with probability at most $ce^{-k/c}$ (choosing c large enough again).

The proof of Lemma 2.6 in [18] shows that, if $x, y \in S_0$ are adjacent and differ in a queue of length k, then for some constants $\alpha, \beta > 0$ we can upper bound

$$d_{\mathrm{W}}(\delta_x P^t, \delta_y P^t) \leq e^{-\beta t/n} + 2e^{-\beta n}$$

for $t \geq \alpha kn$. Also, by Lemma 2.3 in [18],

$$d_{\mathrm{W}}(\delta_x P^t, \delta_y P^t) \leq 1$$

for all t and hence for $t < \alpha kn$.

Combining the above observations and choosing $\alpha > 1$ large enough, we find that for $t \geq \alpha^2 n$

$$\sup_{x_0 \in S_0^0} \mathbb{E}_{\delta_{x_0}} d_{\mathrm{W}}(\delta_{X_1} P^t, \delta_{x_0} P^t)^2 \leq e^{-t/\alpha n} + e^{-n/\alpha}.$$

Hence, by choosing c large enough, we can upper bound

$$\sum_{i=1}^{\tau} \alpha_i^2 \leq cn.$$

Further, once again using Lemma 2.3 in [18], we can upper bound $\hat{\alpha} \leq 2$.

By Theorem 4.5, there is a constant $c > 0$ such that, uniformly for all 1-Lipschitz functions f, all $t \leq \tau$, and all $u > 0$,

$$\mathbb{P}_{\delta_{\mathbf{0}}}(|f(\hat{X}_t) - \mathbb{E}_{\delta_{\mathbf{0}}}[f(\hat{X}_t)]| \geq u) \leq 2e^{-u^2/4c(n+u)} + e^{-(\log n)^2/c}. \tag{4.26}$$

In particular, we can choose c large enough so that, for $u \leq c_0 \sqrt{n} \log n$,

$$\mathbb{P}_{\delta_{\mathbf{0}}}(|f(\hat{X}_t) - \mathbb{E}_{\delta_{\mathbf{0}}}[f(\hat{X}_t)]| \geq u) \leq 3e^{-u^2/cn}. \tag{4.27}$$

Now, as before, by (4.19),

$$d_{\mathrm{W}}(\delta_{\mathbf{0}}P^{\tau},\hat{\pi}) \le (n+2cn+2)e^{-\eta n}$$

provided c is large enough. It follows that for n large enough, uniformly for all 1-Lipschitz functions f, and all $u \ge 1$,

$$\begin{aligned}\mathbb{P}_{\hat{\pi}}(|f(\hat{Y}) - \mathbb{E}_{\hat{\pi}}[f(\hat{Y})]| \ge 2u) &\le \mathbb{P}_{\delta_{\mathbf{0}}}(|f(\hat{X}_\tau) - \mathbb{E}_{\delta_{\mathbf{0}}}[f(\hat{X}_\tau)]| \ge u)\\ &+ (n+2cn+2)e^{-\eta n}\\ &\le 2e^{-u^2/4c(n+u)} + 2e^{-(\log n)^2/c} \qquad (4.28)\end{aligned}$$

It follows that, for $0 < u \le c_0 n^{1/2}\log n$, we obtain

$$\mathbb{P}_{\hat{\pi}}(|f(\hat{Y}) - \mathbb{E}_{\hat{\pi}}[f(\hat{Y})]| \ge 2u) \le ce^{-u^2/cn}, \qquad (4.29)$$

provided that the constant c is chosen sufficiently large. Choosing $u = \sqrt{n}\omega(n)$, where $\omega(n)$ is a function tending to infinity with n arbitrarily slowly, we obtain

$$\mathbb{P}_{\hat{\pi}}(|f(\hat{Y}) - \mathbb{E}_{\hat{\pi}}[f(\hat{Y})]| \ge u) = o(1)$$

as $n \to \infty$.

Inequalities (4.26) and (4.28) could be optimised (by optimising the choice of set S_0) to obtain normal concentration for larger u.

For a positive integer k, let $\ell(k,\hat{Y})$ be the number of queues of length at least k in the stationary jump chain, and let $\hat{\ell}(k)$ be its expectation. Then for any positive integer s, and any $u > 0$, we can write

$$\mathbb{E}_{\hat{\pi}}[|\ell(k,\hat{Y}) - \hat{\ell}(k)|^s] \le u^s + \sum_{y \ge u} y^{s-1}\,\mathbb{P}_{\hat{\pi}}(|\ell(k,\hat{Y}) - \hat{\ell}(k)| > y).$$

Note that the maximum value that $|\ell(k,\hat{Y}) - \hat{\ell}(k)|^s$ can take is n^s. Then, taking $u = n^{1/2}$, and applying inequality (4.28), we obtain

$$\mathbb{E}_{\hat{\pi}}[|\ell(k,\hat{Y}) - \hat{\ell}(k)|^s] \le cn^{s/2}.$$

assuming the constant c is chosen big enough. Hence, arguing as in Section 4 of [18], it is easy to show that

$$\sup_k |\,\mathbb{E}[\ell(k,\hat{Y})^r - \hat{\ell}(k)^r| = O(n^{r-1}).$$

And hence, arguing as in Section 5 of [18], we obtain that, for some constant c_0,

$$\sup_i |n^{-1}\hat{\ell}(i) - \lambda^{1+d+\ldots+d^{i-1}}| \le c_0 n^{-1}, \qquad (4.30)$$

which improves on equation (27) in [18], implying that

$$\sup_i |n^{-1}\hat{\ell}(i) - \lambda^{1+d+\ldots+d^{i-1}}| \le c_0 n^{-1}(\log n)^2.$$

5 Conclusions

We have derived concentration inequalities for Lipschitz functions of a Markov chain long-term and in equilibrium, depending on contractivity properties of the chain in question. Our results apply to many natural Markov chains in computer science and statistical mechanics.

One open problem is to show that, in a discrete-time Markov chain with 'local' transitions, under suitable conditions, rapid mixing occurs essentially if and only if there is normal concentration of measure long-term and in equilibrium (with non-trivial bounds). Another open question is to explore how these properties relate to the cut-off phenomenon. Is it the case that, again under suitable assumptions, they are necessary and sufficient conditions for a cut-off to occur?

6 Acknowledgement

The author is grateful to Graham Brightwell for reading the paper and making many helpful comments.

References

[1] M. Aizenman and R. Holley (1987). Rapid convergence to equilibrium of stochastic Ising models in the Dobrushin-Shlosman regime. In *Percolation Theory and Ergodic Theory of Infinite Particle Systems* (H. Kesten, ed.). IMA Vol. Math. Applic. Springer-Verlag, Berlin.

[2] R. Bubley and M. Dyer (1997). Path coupling: a technique for proving rapid mixing in Markov chains. *Proc. 38th Ann. Symp. Found. Comp. Sci.* 223 – 231.

[3] A. Czumaj, P. Kanarek, M. Kutylowski and K. Lorys (1999). Delayed path coupling and generating random permutations via distributed stochastic processes. In *10 Ann. Symp. Disc. Alg*, ACM-SIAM, 355–363.

[4] J. Ding, E. Lubetzky and Y. Peres. (2008) The mixing time evolution of Glauber dynamics for the mean-field Ising model. Preprint.

[5] R.M. Dudley (1989). *Real Analysis and Probability*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

[6] M. Dyer, L.A. Goldberg, C. Greenhill, M. Jerrum and M. Mitzenmacher (2000). An extension of path coupling and its application to the Glauber dynamics for graph colourings. In *11 Ann. Symp. Disc. Alg*, ACM-SIAM, 616–624.

[7] M. Dyer, C. Greenhill and M. Molloy (2002). Very rapid mixing of the Glauber dynamics for proper colourings on bounded-degree graphs. *Rand. Struct. Alg* **20** 98–114.

[8] M. Dyer, A. Goldberg, C. Greenhill, M. Jerrum, and M. Mitzenmacher (2001). An extension of path coupling and its application to the Glauber dynamics for graph colourings. *SIAM Journal in Computing* **30** 1962 – 1975.

[9] M. Fairthorne (2009+). PhD Thesis, London School of Economics.

[10] C. Graham (2000). Chaoticity on path space for a queueing network with selection of the shortest queue among several. *J. Appl. Probab.* **37** 198–210.

[11] C. Graham (2004). Functional central limit theorems for a large network in which customers join the shortest of several queues. *Probab. Theor. Relat. Fields* **131** 97–120.

[12] M. Jerrum (1998). Mathematical foundations of MCMC. In *Probabilistic Methods for Algorithmic Discrete Mathematics*. (M. Habib, C. McDiarmid, J. Ramirez and B. Reed, eds.) 116 – 165. Springer – Verlag, Berlin.

[13] M. Kac (1956). Foundation of kinetic theory. *Proc. 3rd Berkeley Symp. on Math. Stat. and Probab.* **3** 171–197. Univ. of Calif. Press.

[14] M. Ledoux (2001). *The Concentration of measure phenomenon*. AMS.

[15] D. Levin, M.J. Luczak and Y. Peres. Glauber dynamics for the mean field Ising model: cut-off, critical power law, and metastability. To appear in *PTRF*.

[16] M.J. Luczak and K. Marton (2008). Transportation cost inequalities for stationary distributions of Markov kernels. Preprint.

[17] M.J. Luczak and C. McDiarmid (2005). On the power of two choices: balls and bins in continuous time. *Ann. Appl. Probab.* **15** 1733–1764.

[18] M.J. Luczak and C. McDiarmid (2006). On the maximum queue length in the supermarket model. *Ann. Probab.* **34** 493–527.

[19] M.J. Luczak and C. McDiarmid (2007). Asymptotic distributions and chaos for the supermarket model. *Elect. J. Probab.* **12** 75 – 99.

[20] M.J. Luczak and J.R. Norris (2005). Strong approximation for the supermarket model. *Ann. Appl. Probab.* **15** 2038–2061.

[21] D. Levin, Y. Peres and E. Wilmer (2009+). *Markov chains and mixing times*. In preparation.

[22] J.B. Martin and Y.M. Suhov (1999). Fast Jackson networks. *Ann. Appl. Probab.* **9** 854–870.

[23] M. Mitzenmacher (1996). The power of two choices in randomized load-balancing. PhD. Thesis, Berkeley.

[24] M. Mitzenmacher, B. Prabhakar, and D. Shah (2002). Load-balancing with memory. *Proc. 43rd Ann. IEEE Symp. Found. Comp. Sci*.

[25] M. Mitzenmacher, A. Richa, and R. Sitaraman (2001). The power of two random choices: a survey of techniques and results. In *Handbook of Randomized Computing* (S. Rajasekaran, P.M. Pardalos, J.H. Reif and J.D.P. Rolim, eds.) **1** 255–312. Kluwer, Dordrecht.

[26] C. McDiarmid (1998). Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*. (M. Habib, C. McDiarmid, J. Ramirez and B. Reed, eds.) 195 – 248. Springer – Verlag, Berlin.

[27] M.Molloy (2001). Very rapidly mixing Markov chains for 2Δ-colourings and for independent sets in a 4-regular graph. *Rand. Struct. Alg* **18** 101–115.

[28] S.T. Rachev (1991). *Probability Metrics and the Stability of Stochastic Models*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Chichester.

[29] A. Sznitman (1989). Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX - 1989* 165 – 251. Springer-Verlag, Berlin.

[30] N.D. Vvedenskaya, R.L. Dobrushin, and F.I. Karpelevich (1996). Queueing system with selection of the shortest of two queues: an asymptotic approach. *Problems Inform. Transmission* **32** 15–27.

[31] D. Weitz (2004). Mixing in time and space for discrete spin systems. PhD Thesis, Berkeley.