Advanced Algorithms XIII (Fall 2020)

Instructor: Chihao Zhang
Scribed by: Yuan Yao and Zonghan Yang

Last modified on November 30, 2021

In our previous study of discrete time Markov chains, we found that irreducible aperiodic Markov chains converge to a stationary distribution. However, we did not determine how quickly they converge, which is important in a number of algorithmic applications. In this lecture, we introduce the concept of coupling, a powerful method for bounding the rate of convergence of Markov chains.

1 Coupling

1.1 Total variation distance

Definition 1. *The* total variation distance *between two distributions μ and ν on a countable state space Ω is given by*

$$D_{\mathrm{TV}}(\mu, \nu) = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| .$$

We can look to the following figure of two distributions on the sample space. The variation distance is half the area enclosed by the two curves.

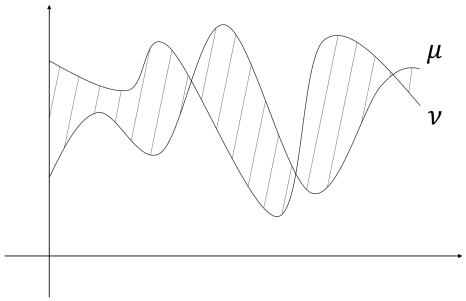

Figure 1: Two distributions on sample space

The total variation distance can be equivalently viewed in the following way.

Lemma 2. *We define $\mu(A) = \sum_{x \in A} \mu(x)$, $\nu(A) = \sum_{x \in A} \nu(x)$, then we have*

$$D_{\mathrm{TV}}(\mu, \nu) = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)| .$$

Proof. Let $\Omega^+ \subseteq \Omega$ be the set of states such that $\mu(x) \geq \nu(x)$, and let $\Omega^- \subseteq \Omega$ be the set of states such that $\nu(x) > \mu(x)$. It can be easily verified that

$$\max_{A\subseteq\Omega} \mu(A) - \nu(A) = \mu(\Omega^+) - \nu(\Omega^+),$$
$$\max_{A\subseteq\Omega} \nu(A) - \mu(A) = \nu(\Omega^-) - \mu(\Omega^-).$$

By $\mu(\Omega) = \nu(\Omega) = 1$,

$$\mu(\Omega^+) + \mu(\Omega^-) = \nu(\Omega^+) + \nu(\Omega^-) = 1,$$

which implies that

$$\mu(\Omega^+) - \nu(\Omega^+) = \nu(\Omega^-) + \mu(\Omega^-).$$

We derive that

$$\max_{A\subseteq\Omega} |\nu(A) - \mu(A)| = \nu(\Omega^-) - \mu(\Omega^-) = \mu(\Omega^+) - \nu(\Omega^+).$$

Therefore,

$$\begin{aligned} D_{\mathrm{TV}}(\mu,\nu) &= \frac{1}{2}\,|\mu(\Omega) - \nu(\Omega)| \\ &= \frac{1}{2}\left(\left|\mu(\Omega^+) - \nu(\Omega^+)\right| + |\mu(\Omega^-) - \nu(\Omega^-)|\right) \\ &= \max_{A\subseteq\Omega} |\nu(A) - \mu(A)|\,. \end{aligned}$$

□

1.2 Coupling lemma

Intuitively, coupling is a process that binds several stochastic processes. Here's the definition.

Definition 3. *Let μ and ν be two distributions on the same space Ω. Let ω be a distribution on the space $\Omega\times\Omega$. If $(x, y) \sim \omega$ satisfies $x \sim \mu$ and $y \sim \nu$, then ω is called a coupling of μ and ν.*

In other words, the marginal probabilities of the disjoint distribution ω are μ and ν respectively. A special case is when x and y are independently. However, in many applications, we want x and y to be correlated while keeping their respect marginal probabilities correct.

We now give a toy example about how to construct different couplings on two fixed distributions. There are two coins: the first coin has probability $\frac{1}{2}$ for head in a toss and $\frac{1}{2}$ for tail, and the second coin has probability $\frac{1}{3}$ and $\frac{2}{3}$ respectively. We now construct two couplings as follows.

prob x \ y	HEAD	TAIL
HEAD	1/3	1/6
TAIL	0	1/2

Table 1: A first coupling

prob x \ y	HEAD	TAIL
HEAD	1/6	1/3
TAIL	1/6	1/3

Table 2: A second coupling

The table defines a joint distribtion and the sum of a certain row/column corresponds to a marginal probability. It is clear that both table are couplings of the two coins. Among all the possible couplings, sometimes we are interested in the one who is "mostly coupled".

Lemma 4 (Coupling Lemma). *Let μ and ν be two distributions on a sample space Ω. Then for any coupling ω of μ and ν it holds that,*

$$\mathbf{Pr}_{(x,y)\sim\omega}[x \neq y] \geq D_{\mathrm{TV}}(\mu,\nu).$$

And furthermore, there exists a coupling ω^* of μ and ν such that

$$\mathbf{Pr}_{(x,y)\sim\omega^*}[x \neq y] = D_{\mathrm{TV}}(\mu,\nu).$$

Proof. For finite Ω, designing a coupling is equivalent to filling a $\Omega \times \Omega$ matrix so that the marginals are correct.

We can learn that

$$\begin{aligned}\mathbf{Pr}[x = y] &= \sum_{t\in\Omega} \mathbf{Pr}[x = y = t] \\ &\leq \sum_{t\in\Omega} \min(\mathbf{Pr}[x = t], \mathbf{Pr}[y = t]).\end{aligned}$$

Thus,

$$\begin{aligned}\mathbf{Pr}[x \neq y] &\geq 1 - \sum_{t\in\Omega} \min(\mathbf{Pr}[x = t], \mathbf{Pr}[y = t]) \\ &= \sum_{t\in\Omega} (\mathbf{Pr}[x = t] - \min(\mathbf{Pr}[x = t], \mathbf{Pr}[y = t])) \\ &= \max_{A\subseteq\Omega} (\mu(A) - \nu(A)) \\ &= D_{\mathrm{TV}}(\mu,\nu).\end{aligned}$$

By taking $\mathbf{Pr}[x = y = t] = \min(\mathbf{Pr}[x = t], \mathbf{Pr}[y = t])$, we can achieve the equality above. □

The coupling lemma provides a way to upper bound the distance between two distributions: For any two distributions μ and ν, we can construct a coupling ω and sample $(x, y) \sim \omega$. The upper bound for $\mathbf{Pr}[x \neq y]$ is an upper bound for $D_{\mathrm{TV}}(\mu,\nu)$. The coupling lemma tells us this upper bound is tight, as long as you are able to find the optimal coupling.

1.3 Coupling of Two Markov Chains

We can also couple two random walks / Markov chains.

Definition 5. *Consider two copies of the chain P that satisfies:*

- *The initial distribution is μ_0 and ν_0.*
- *$\mu_t^T = \mu_0^T P^t$ and $\nu_t^T = \nu_0^T P^t$.*

A coupling of the two chains is joint distribution ω of $\{\mu_t\}_{I\geq 0}$ and $\{\nu_t\}_{t\geq 0}$ that satisfies the following condition. $\{(X_t, Y_t)\}_{t\geq 0} \sim \omega$ is a pair of processes such that

- $\forall a, b \in \Omega$, $\mathbf{Pr}[X_{t+1} = b \mid X_t = a] = P(a, b)$
- $\forall a, b \in \Omega$, $\mathbf{Pr}[Y_{t+1} = b \mid Y_t = a] = P(a, b)$

- $\forall t \geq 0, X_t = Y_t \Longrightarrow X_{t'} = Y_{t'}$ *for all* $t' > t$.

Therefore, marginally $\{X_t\}$ and $\{Y_t\}$ are both chain P. We additionally require that two chains coalesce once they meet.

We can use the coupling of Markov chains to analyze the rate of convergence of Markov chains. If we are able to couple two Markov chains so that they coalesce quickly, then we can apply the coupling lemma to upper bound the distance between the distance between the two chains. Let us first use this tool to prove the fundamental theorem of Markov chains again.

Theorem 6 (Fundamental Theorem via Coupling). *If a finite chain P is irreducible and aperiodic, then it has a unique stationary distribution π. Moreover, for any initial distribution μ, it holds that*

$$\lim_{t\to\infty} \mu^T P^t = \pi^T$$

Proof. We already know that P has a stationary distribution π. What we would like to show is that for all starting distribution μ_0, it holds that

$$\lim_{t\to\infty} D_{\mathrm{TV}}(\mu_t, \pi) = 0\,,$$

where $\mu_t^\top = \mu_0^\top P^t$.

Suppose that $\{X_t\}$ and $\{Y_t\}$ are two identical Markov chains starting from different distribution, where $Y_0 \sim \pi$ while X_0 is generated from an arbitrary distribution μ_0.

Now we have two sequence of random variables:

$$\begin{array}{ccccccccccccc}
\mu_0 & & \mu_1 & & & & & & \mu_t & & & & \\
\wr & & \wr & & & & & & \wr & & & & \\
X_0 & \to & X_1 & \to & X_2 & \to & \cdots & \to & X_t & \to & X_{t+1} & \to & \cdots \\
\\
Y_0 & \to & Y_1 & \to & Y_2 & \to & \cdots & \to & Y_t & \to & Y_{t+1} & \to & \cdots \\
\wr & & \wr & & & & & & \wr & & & & \\
\pi & & \pi & & & & & & \pi & & & &
\end{array}$$

The coupling lemma establishes the connection between the distance of distributions and the discrepancy of random variables. To show that $D_{\mathrm{TV}}(\mu_t, \pi) \to 0$, it is sufficient to construct a coupling ω_t of μ_t and π and then compute $\mathbf{Pr}_{(X_t,Y_t)\sim\omega_t}[X_t \neq Y_t]$.

Here we give a simple coupling. Let $(X_t, Y_t) \sim \omega_t$ and we construct ω_{t+1}. If $X_t = Y_t$ for some $t \geq 0$, then let $X_{t'} = Y_{t'}$ for all $t > t'$, otherwise X_{t+1} and Y_{t+1} are independent. Namely, $\{X_t\}$ and $\{Y_t\}$ are two independent Markov chains until X_t and Y_t reach the same state for some $t \geq 0$, and once they meet together then they move together forever. The coupling lemma tells us that $D_{\mathrm{TV}}(\mu_t, \pi) \leq \mathbf{Pr}_{(X_t,Y_t)\sim\omega_t}[X_t \neq Y_t]$.

The property of irreducibility implies that

$$\forall\, i, j, \quad \exists\, n \quad \text{s.t.} \quad P^n(i,j) > 0\,.$$

We claim that combining with aperiodicity,

$$\exists\, n \quad \text{s.t.} \quad \forall\, i, j, \quad P^n(i,j) > 0\,.$$

Since the state space Ω is finite, it is sufficient to show that

$$\forall\, i, j, \quad \exists\, t_{i,j} \quad \text{s.t.} \quad \forall\, n > t_{i,j}, \quad P^n(i,j) > 0\,.$$

Suppose that there are s loops of length $c_1, c_2, \ldots, c_s$ starting from and ending at state i. Then we have

$$\gcd(c_1, c_2, \ldots, c_s) = 1\,.$$

Thus, by Bézout's identity there exists $x_1, x_2, \ldots, x_s \in \mathbb{Z}$ such that

$$c_1 x_1 + c_2 x_2 + \cdots c_s x_s = 1\,.$$

This implies the following lemma, a proof can be found in e.g. [LP17]

Lemma 7. *For sufficiently large b, there exists $y_1, y_2, \ldots, y_s \in \mathbb{N}$ such that*

$$c_1 y_1 + c_2 y_2 + \cdots c_s y_s = b.$$

The claim then follows from the lemma.

Now we know that $\exists\, n$ s.t. $\forall\, i, j$, $P^n(i,j) > 0$. Then we define

$$\theta \triangleq \min_{x_0, y_0 \in \mathcal{S}} \mathbf{Pr}\left[X_n = Y_n \mid X_0 = x_0, Y_0 = y_0\right]\,.$$

For simplicity, we use $\mathbf{Pr}_{x_0,y_0}[\cdot]$ to denote the conditional probability $\mathbf{Pr}[\,\cdot \mid X_0 = x_0, Y_0 = y_0]$ from now on.

Fix $z \in \Omega$. Let

$$\alpha = \min_{w \in \Omega} \mathbf{P}^n(w, z) > 0\,,$$

and for any $t \geq 0$ and $z' \in \Omega$,

$$\beta_{t,z'} = \mathbf{Pr}_{x_0,y_0}\left[X_t = Y_t = z' \wedge X_{t'} \neq Y_{t'} \text{ for all } t' < t\right]\,.$$

By the Markov property and the independence of $\{X_t\}$ and $\{Y_t\}$ before $X_t = Y_t$, we obtain that

$$\begin{aligned}
&\mathbf{Pr}_{x_0,y_0}[X_n = Y_n] \\
&\geq \mathbf{Pr}_{x_0,y_0}[X_n = Y_n = z] \\
&= \mathbf{Pr}_{x_0,y_0}[X_n = Y_n = z \wedge \forall\, t < n, X_t \neq Y_t] + \mathbf{Pr}_{x_0,y_0}[X_n = Y_n = z \wedge \exists\, t < n, X_t = Y_t] \\
&= \left(P^n(x_0,z)\cdot P^n(y_0,z) - \sum_{t=0}^{n-1}\sum_{z'} \beta_{t,z'} \cdot \left(P^{n-t}(z',z)\right)^2\right) + \sum_{t=0}^{n-1}\sum_{z'} \beta_{t,z'} \cdot P^{n-t}(z',z) \\
&\geq P^n(x_0,z)\cdot P^n(y_0,z) \;\geq \alpha^2\,.
\end{aligned}$$

Hence $\theta > 0$. By the coupling and the Markov property, we have

$$\begin{aligned}
\mathbf{Pr}_{x_0,y_0}[X_{2n} \neq Y_{2n}] &= \sum_{x_n \neq y_n} \mathbf{Pr}_{x_0,y_0}[X_{2n} \neq Y_{2n}, X_n = x_n, Y_n = y_n] \\
&= \sum_{x_n \neq y_n} \mathbf{Pr}_{x_n,y_n}[X_n \neq Y_n] \cdot \mathbf{Pr}_{x_0,y_0}[X_n = x_n, Y_n = y_n] \\
&\leq (1-\theta) \sum_{x_n \neq y_n} \mathbf{Pr}_{x_0,y_0}[X_n = x_n, Y_n = y_n] \;\leq (1-\theta)^2\,,
\end{aligned}$$

and so on ($\mathbf{Pr}_{x_0,y_0}[X_{kn} \neq Y_{kn}] \leq (1-\theta)^k$). It yields directly that

$$\mathbf{Pr}[X_t \neq Y_t] = \sum_{x_0, y_0} \mu_0(x_0)\cdot\pi(y_0)\cdot\mathbf{Pr}_{x_0,y_0}[X_t \neq Y_t] \to 0$$

as $t \to \infty$. □

2 Mixing Time

If the reader carefully examines our proof of the fundamental theorem of Markov chains using coupling above, it can be noticed that the proof provides a way to bound the rates of convergence. The larger the parameter θ, the faster the chain converges. The parameter θ is a lower bound of the probability that two chains meet at time t. Therefore, if we are able to construct a coupling so that the two chains meet fast, we get a good upper bound on the convergence time.

To formally explain the idea, we first define the notion of *mixing time*. For any $0 < \varepsilon < 1$,

$$\tau_{\text{mix}}(\varepsilon) \triangleq \max_{\mu_0} \min_{t} D_{\text{TV}}\left(\mu_0^T P^t, \pi\right) < \varepsilon,$$

which describes the first time t such that the total variation distance between X_t and π is at most ε for any initial μ_0. Mixing time is a measure of the rate of convergence. Sometimes we simply use τ_{mix} to denote $\tau_{\text{mix}}(\frac{1}{4})$.

We show that the coupling lemma can imply a bound for the mixing time. Let $\{X_t\}$ and $\{Y_t\}$ be two Markov chains with same transition rule but different initial distributions. If we can find a coupling of two chains such that for some $t > 0$,

$$\mathbf{Pr}\left[X_t \neq Y_t\right] \leq \varepsilon,$$

then by the coupling lemma we can conclude that $\tau_{\text{mix}}(\varepsilon) \leq t$. This simple fact is a powerful tool and we will see many of its applications.

2.1 Random walk on hypercube

Let's start with a simple example. Consider the random walk on the n-cube. The state space $\Omega = \{0,1\}^n$, and we start from a point $X_0 \in \Omega$. In each step,

1. With probability $\frac{1}{2}$ do nothing.
2. Otherwise, pick $i \in [n]$ uniformly at random and flip $X(i)$.

It's equivalent to the following process:

1. Pick $i \in [n], b \in \{0,1\}$ uniformly at random.
2. Change $X(i)$ to b.

We want to know how many steps should we do to make it ε-far from uniformly random, i.e. $\tau_{\text{mix}}(\varepsilon)$. For two walks X_t, Y_t, we can couple them by choosing the same i, b in every step. Then, the problem for the worst case, $X_0(i) \neq Y_0(i)$ for all i, is exactly the Coupon Collector model. From previous lectures we know that

$$\mathbf{Pr}\left[X_t \neq Y_t\right] \leq e^{-c} \text{ for } t \geq n\log n + cn,$$

so by the coupling lemma it holds that

$$\tau_{\text{mix}}(\varepsilon) \leq n\log n + n\log \varepsilon^{-1}.$$

Let's modify the process a bit by changing $\frac{1}{2}$ into $\frac{1}{n+1}$, i.e. w.p. $\frac{1}{n+1}$ do nothing, to make the 'lazy' walk more active. Curious reader may find it strange to keep a loop with low probability. Actually the main

aim is to keep the Markov chain aperiodic: if we flip one bit w.p. 1 in each step, two walks with different parities can never be mixed up.

Now in this case, we describe another coupling of X_t, Y_t. Without loss of generality, we can reorder the entries of two vectors so that all disagreeing entries come first. Namely there exists an index k such that $X_t(i) \neq Y_t(i)$ if $i \in [k]$, and $X_t(i) = Y_t(i)$ otherwise. Our coupling is as follows:

- If $k = 0$, Y acts the same as X.
- If $k = 1$, Y acts the same as X except when X flips the first entry, Y does nothing and vice versa.
- For $k > 2$, we distinguish between whether X flip indices in $[k]$:
  1. If X did nothing or flipped one of $[n]\backslash[k]$: Y acts the same.
  2. If X flipped $i \in [k]$: Y flips $(i \bmod k) + 1$, i.e. $1 \mapsto 2, 2 \mapsto 3, \cdots, k-1 \mapsto k, k \mapsto 1$.

It's clear that the above is indeed a coupling. In fact, this coupling is like a "doubled speed" Coupon Collector, since in the case $k > 2$ we can always collect two coupons at a time when lady luck is smiling. We therefore state without a rigorous proof that

$$\tau_{\text{mix}} \leq \frac{1}{2} n \log n + O(n).$$

The above two examples are easy to analyze since we can reduce the coalesce time of two chains to problems we are familiar with. To analyze couplings in general, we often require the coupling enjoy the property that the two chains are *expected* closer after every step. Therefore we impose a *distance* $d(\cdot, \cdot)$ between two states and require that

$$\forall t, \mathbf{E}\left[d(X_{t+1}, Y_{t+1}) | (X_t, Y_t)\right] \leq (1 - \alpha) \cdot d(X_t, Y_t).$$

In other words, $\{d(X_t, Y_t)\}_{t \geq 0}$ is a supermartingale.

Without loss of generality, we assume that $\min_{x,y \in \Omega : x \neq y} d(x, y) = 1$ when Ω is finite. By coupling lemma

$$\begin{aligned} D_{\text{TV}}(X_t, Y_t) &\leq \mathbf{Pr}\left[X_t \neq Y_t\right] \\ &= \mathbf{Pr}\left[d(X_t, Y_t) > 0\right] \\ &= \mathbf{Pr}\left[d(X_t, Y_t) \geq 1\right] \\ &\leq \mathbf{E}\left[d(X_t, Y_t)\right] \\ &\leq (1 - \alpha)^t \cdot d(X_0, Y_0) \leq \varepsilon. \end{aligned}$$

This implies

$$\tau_{\text{mix}}(\varepsilon) \leq (\log \varepsilon^{-1} + \log d(X_0, Y_0)) \cdot \log \frac{1}{1 - \alpha}.$$

2.2 Sampling proper colorings

Let's consider the problem of sampling proper colorings. Given a graph $G = (V, E)$, we want to dye the graph using q colors under the condition that no two adjacent vertices share the same color. More formally, a coloring of G is a mapping $c : V \mapsto [q]$, and we call it *proper* iff $\forall (u, v) \in E, c(u) \neq c(v)$. The problem is NP-hard in general. However, for $q > \Delta$ there's always at least one suitable solution and can be easily obtained by a greedy algorithm, where Δ is the maximum degree of the graph.

Consider the following Markov chain to sample proper colorings:

1. Pick $v \in V$ and $c \in [q]$ uniformly at random.

2. Recolor v with c if possible.

The chain is aperiodic since self-loops exist in the walk. For $q \geq \Delta + 2$, the chain is irreducible. The bound $q \geq \Delta + 2$ is tight for irreducibility since when $q = \Delta + 1$, each proper coloring of complete graph is frozen. It is still an open problem if the mixing time of the chain is polynomial in the size of the graph under the condition $q \geq \Delta + 2$. The best bound so far requires that $q \geq (\frac{11}{6} - \varepsilon)\Delta$. Here, we shall give a rapid mixing proof when $q > 4\Delta$ using the method of coupling.

Suppose X_t, Y_t are two proper colorings. We define the distance $d(X_t, Y_t)$ as their Hamming distance, i.e. the number of vertices colored differently in two colorings. Our coupling of two chains is that we always choose the same v, c in each step. The distance between two colorings can change at most 1 since only v is affected. The possible changes can be divided into two kinds:

1. Good move: $X_t(v) \neq Y_t(v)$, and both change into c successfully. It will decrease distance by 1.

2. Bad move: $X_t(v) = Y_t(v)$, one succeeds and one fails in the changing. It will increase distance by 1.

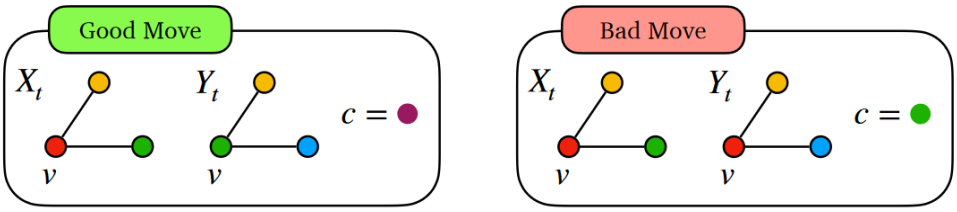

Figure 2: An illustration of moves.

Consider the probabilities of two types of moves. For good moves, w.p. $\frac{d(X_t,Y_t)}{n}$, $X_t(v) \neq Y_t(v)$, and there are at least $q - 2\Delta$ choices of c to make it a good move. So

$$\mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) - 1\right] = \mathbf{Pr}_{(v,c)\in V\times[q]}\left[(v, c) \text{ is a good move}\right] \geq \frac{d(X_t, Y_t)}{n} \cdot \frac{q - 2\Delta}{q}.$$

For bad moves, there exists a neighbor w of v such that its color is different in two colorings, and in one coloring w is of color c. By a counting argument, we have

$$\mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) + 1\right] = \mathbf{Pr}_{(v,c)\in V\times[q]}\left[(v, c) \text{ is a bad move}\right] \leq \frac{\Delta d(X_t, Y_t)}{n} \cdot \frac{2}{q}.$$

Therefore,

$$\begin{aligned}
\mathbf{E}\left[d(X_{t+1}, Y_{t+1})|(X_t, Y_t)\right] &= d(X_t, Y_t) + \mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) + 1\right] - \mathbf{Pr}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) - 1\right] \\
&\leq d(X_t, Y_t) + \frac{\Delta d(X_t, Y_t)}{n} \cdot \frac{2}{q} - \frac{d(X_t, Y_t)}{n} \cdot \frac{q - 2\Delta}{q} \\
&\leq d(X_t, Y_t)(1 - \frac{q - 4\Delta}{nq}).
\end{aligned}$$

In the case $q > 4\Delta$,

$$D_{\mathrm{TV}} \leq \left(1 - \frac{1}{nq}\right)^t n \leq \varepsilon.$$

The mixing time is therefore bounded by

$$\tau_{\mathrm{mix}}(\varepsilon) \leq nq(\log n + \log \varepsilon^{-1}).$$

References