Advanced Algorithms

Spectral methods and algorithms

尹一通　栗师　**刘景铖**

Recap

<u>Previous lecture</u>:
<u>Random walks on undirected graphs</u>

- Speeding up bipartite matching
- Return time
- Fundamental theorem of Markov chains
- Pagerank

<u>What next?</u>
<u>Random walks on undirected graphs</u>

- Fundamental theorem of Markov chains
- Spectral analysis
- Mixing time
- Random sampling

Back to Markov chain: What is the stationary distribution?

As in Eulerian directed graph, the stationary distribution of undirected graphs is easy to describe

Let $\vec{d} \in \mathbb{R}^{\mathrm{n}}$ be the degree vector and $m = |E|$

Claim. The distribution $\vec{\pi} = \frac{\vec{d}}{2m}$ is a stationary distribution of the random walk on undirected graphs

In the stationary distribution, the probability of going across an edge is the same for every edge

Fundamental Theorem of Markov Chains in undirected graphs

Does $p_t \to \vec{\pi} = \frac{\vec{d}}{2m}$ as $t \to \infty$ regardless of p_0?

Not necessarily: not when the Markov chain is reducible and periodic.

In undirected graphs, being irreducible just means that the graph is connected.

In undirected connected graphs, aperiodic just means that the graph is non-bipartite

(To see why, recall what happens to an undirected 3-cycle)

So the fundamental theorem of Markov chain just becomes the following in undirected graphs.

Theorem. For any finite, connected, non-bipartite graph, $p_t \to \vec{\pi} = \frac{\vec{d}}{2m}$ as $t \to \infty$ regardless of p_0

Lazy Random Walks

We can remove the non-bipartiteness assumption by doing a **lazy random walk**.

In each step, we stay at the same vertex with probability 1/2,

and we move to a uniform random neighbor with probability 1/2.

In matrix form, $p_t = \left(\frac{1}{2}I + \frac{1}{2}AD^{-1}\right)^t p_0$.

Theorem. For any finite and connected graph, $p_t = \left(\frac{1}{2}I + \frac{1}{2}AD^{-1}\right)^t p_0 \to \frac{\vec{d}}{2m}$ as $t \to \infty$ regardless of p_0.

It will be clear in the proof what the lazy random walk does to remove the non-bipartiteness assumption.

Intuitively, we can see it as making the random walk "very" aperiodic.

Spectra Analysis

Let $W = AD^{-1}$ be the random walk matrix and $Z = \frac{1}{2}I + \frac{1}{2}AD^{-1}$ be the lazy random walk matrix.
To understand $p_t = W^t p_0$, it is very useful to understand the spectrum of W.
One problem is that W is not symmetric.

But W is **similar** to a symmetric matrix: $D^{-\frac{1}{2}}WD^{\frac{1}{2}} = D^{-\frac{1}{2}}(AD^{-1})D^{\frac{1}{2}} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = \mathcal{A}$.
Claim. W and $\mathcal{A}$ have the same spectrum

Note that W may not have an orthonormal basis of eigenvectors.

Spectrum of W

Let $W = \frac{1}{d}A = I - \frac{1}{d}L$.
What do we know about the spectrum $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$?

- We know that $1 \geq \alpha_1$ and $\alpha_n \geq -1$
- We know that $1 = \alpha_1$
- We know that $\alpha_1 > \alpha_2$ if and only if the graph is connected
- We know that $\alpha_1 = -\alpha_n$ if and only if the graph is bipartite

Proof of Fundamental Theorem

Theorem. For any finite, connected, non-bipartite graph, $p_t \to \vec{\pi} = \frac{\vec{d}}{2m}$ as $t \to \infty$ regardless of p_0.

Proof sketch. We just need to show $W^t p_0 \to c_1 v_1$ as $t \to \infty$. Then find out what is $c_1 v_1$.

Note that $W^t p_0 = c_1 \alpha_1^t v_1 + c_2 \alpha_2^t v_2 + \cdots + c_n \alpha_n^t v_n$.

- $\alpha_1 = 1$.
- $\alpha_2 < 1$ if and only if the graph is connected.
- $\alpha_n > -1$ if and only if the graph is non-bipartite.

So the spectral conditions correspond exactly to the combinatorial assumptions of the theorem!

This implies that

- $W^t p_0 \to c_1 v_1$ as $t \to \infty$.

- The convergence is faster if $\alpha_2 < 1 - \epsilon$ and $\alpha_n > -1 + \epsilon$ for a larger $\epsilon > 0$.

Proof for Lazy Random Walks

Theorem. For any finite and connected graph,

$$p_t = \left(\frac{1}{2}I + \frac{1}{2}AD^{-1}\right)^t p_0 \to \frac{\vec{d}}{2m} \text{ as } t \to \infty \text{ regardless of } p_0.$$

Proof is similar, but with a different spectrum.

Exercise: What is the new transition matrix, and what is its spectrum?

Theorem. For any finite, connected, non-bipartite graph,

$p_t \to \vec{\pi} = \frac{\vec{d}}{2m}$ as $t \to \infty$ regardless of p_0.

Mixing Time

From the fundamental theorem of Markov chain, we know that $p_t \to \vec{\pi} = \frac{\vec{d}}{2m}$ as $t \to \infty$ regardless of p_0

We would like to understand how fast it converges to $\vec{\pi}$

Recall how we measure closeness: $d_{TV}(p_t, \pi) = \frac{1}{2}\|p_t - \pi\|_1 = \frac{1}{2}\sum_{i=1}^{n}|p_t(i) - \pi(i)|$

$$= \max_{S \subseteq [n]} |p_t(S) - \pi(S)|$$

Definition. The ***ϵ-mixing time*** of the random walk is defined as the smallest t such that

$$\|p_t - \pi\|_1 \le \epsilon \quad \forall\, p_0.$$

We will bound the mixing time using the **spectral gap**, defined as $\lambda = \min\{1 - \alpha_2, 1 - |\alpha_n|\}$

Mixing Time by Spectral Gap

Theorem. The ϵ-mixing time is upper bounded by $\frac{1}{\lambda}\log\left(\frac{n}{\epsilon}\right)$, where $\lambda = \min\{1-\alpha_2, 1-|\alpha_n|\}$.

For simplicity we give the proof only for d-regular graphs:

Let $v_1, v_2, \dots, v_n$ be an orthonormal basis of A. Then $p_0 = c_1 v_1 + c_2 v_2 + \dots + c_n v_n$, and

$$p_t = W^t p_0 = c_1\alpha_1^t v_1 + c_2\alpha_2^t v_2 + \cdots + c_n\alpha_n^t v_n$$

$$\|x\|_2 \coloneqq \sqrt{x^\top x}$$

By Cauchy-Schwarz, $\|p_t - \pi\|_1 \le \sqrt{n}\|p_t - \pi\|_2$

$$\|p_t - \pi\|_2^2 = \|c_2\alpha_2^t v_2 + \cdots + c_n\alpha_n^t v_n\|_2^2 = c_2^2\alpha_2^{2t}\|v_2\|_2^2 + \cdots + c_n^2\alpha_n^{2t}\|v_n\|_2^2$$

$$= c_2^2\alpha_2^{2t} + \cdots + c_n^2\alpha_n^{2t} \le (1-\lambda)^{2t}(c_2^2 + \cdots + c_n^2)$$

Note that p_0 is a distribution, $\|p_0\|_2^2 = \sum_i p_0(i)^2 \le \sum_i p_0(i) = \|p_0\|_1 = 1$

So $\|p_t - \pi\|_2^2 \le (1-\lambda)^{2t} \Rightarrow \|p_t - \pi\|_1 \le \sqrt{n}(1-\lambda)^t \le \sqrt{n}\, e^{-\lambda t}$

When the spectral gap is a constant (i.e. $\lambda = \Omega(1)$), then the random walk converges in $O\left(\log\frac{n}{\epsilon}\right)$ steps.

When the graph is regular with $\lambda = \Omega(1)$, we can sample an almost uniform vertex in $O(\log n)$ steps.

Aside: for general graphs, $v_1, v_2, \dots, v_n$ is an orthonormal basis of $\mathcal{A}$, which incurs an extra $\sqrt{n}$ factor

Mixing Time for Lazy Random Walks

Theorem. The ϵ-mixing time is upper bounded by $\frac{1}{\lambda}\log\left(\frac{n}{\epsilon}\right)$, where λ is the spectral gap.

In lazy random walks, the spectral gap is simply $\frac{\lambda_2}{2}$, where λ_2 is the second eigenvalue of $\mathcal{L}$.

From Cheeger's inequality, we know that $\lambda_2 \geq \frac{\varphi(G)^2}{2}$.

Theorem. The ϵ-mixing time is of lazy random walks is upper bounded by $\frac{2}{\varphi(G)^2}\log\left(\frac{n}{\epsilon}\right)$.

This implies that lazy random walks mix fast in an **expander graph**, a very important result.

$\varphi(G) \approx$ constant

Further applications: Random Sampling

We have seen algorithmic questions that concerns finding a solution, deciding if a solution exists, finding an optimal solution etc

There is an entire area that concerns on a very different task: sampling a solution according to certain distributions

One of the most important applications for random walks is in designing fast sampling algorithms

More often than not, the main question concerns the mixing time of these random walks

For examples,

- Card shuffling
- Sampling a graph coloring
- Sampling a perfect matching in a bipartite graph
    - Approximating 0-1 permanent
- Sampling a spanning tree
    - Generating a maze for fun
- Approximate counting/inference

Card Shuffling

Say we have a deck of 52 cards. How do you get a random permutation using simple operations?

Let's say the simple operation is to choose a random card and put it at the top of the deck.

1. Does it converge to the uniform distribution of all permutations?
2. How many steps are enough to get an almost uniform distribution?

These questions can be understood as questions about random walks on the big "state" graph.

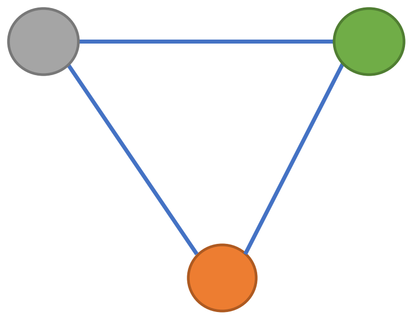
Then the first question is about **stationary distribution**, and the second question is about **mixing time**.

A famous result is 7 "riffle" shuffling will get an almost uniform permutation

- "Trailing the dovetail shuffle to its lair", by Dave Bayer and Persi Diaconis

Cut-off phenomenon

Graph coloring



Given an undirected graph with max. degree Δ and k colors

Goal: generate a k-coloring uniformly at random

This is presumably harder than deciding if there is a k-coloring

Nevertheless, the following random walk has a stationary distribution uniform over all k-colorings:

- Start with any k-coloring σ
- Pick a vertex v and a color c uniformly at random, recolor v with c if it is legal; otherwise do nothing;

This Markov chain is irreducible provided that $k \geq \Delta + 2$, and aperiodic

Conjecture: If $k \geq \Delta + 2$, the above random walk mixes in $\text{poly}(n)$.

We will see a coupling argument assuming $k \geq 4\Delta + 1$

This is known as the Metropolis chain

Other chains: Glauber dynamics, Wang–Swendsen–Kotecký chain, …

Random Combinatorial Objects

We can design a Markov chain to generate a random combinatorial object efficiently.

Another simple example is the **basis-exchange walk** algorithm to generate a random spanning tree.

Sampling algorithms are known for many combinatorial objects (e.g. colorings, perfect matchings, discrepancy minimization)

It is usually easy to construct a Markov chain so that the limiting distribution is uniform

But it is much more difficult to prove that the mixing time is fast

There are books that just focus on mixing time:

- *Markov Chains and Mixing Times*, by Levin and Peres
- *Counting and Markov Chains*, by Jerrum

Many methods are developed, including coupling, conductance, second eigenvalue, etc

Cheeger’s Inequality in Markov chains

It is interesting to see how Cheeger’s inequality can be used.

When we want to bound $\phi(G)$, say in constructing expander graphs,
we can come up with algebraic constructions and bound λ_2 instead

When we want to bound λ_2, say in bounding the mixing time,
we can analyze combinatorial problems and bound $\phi(G)$ instead

An alternative perspective like this is exactly what makes it so powerful

Examples of algorithms from random walk

Finding certain objects faster

- Hitting time / return time
- Ex: Finding bipartite matching, algorithmic Lovász local lemma, 2-SAT, random 3-SAT…

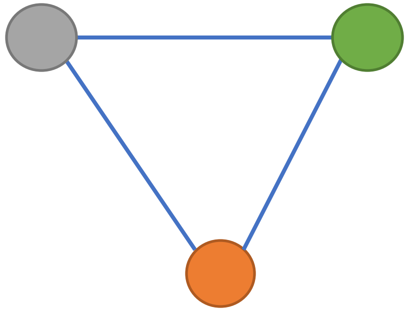
Exploring graphs in space bounded computations

- Cover time
- Ex: checking undirected s-t connectivity, cat and mouse game
- Time-space trade-off (see e.g., Feige’s theorem)

Rapid mixing of random walks: Markov chain Monte Carlo method

- “Local mixing” : local graph partitioning/clustering
- Mixing time
- Ex: Card shuffling, sampling random combinatorial objects, approximate counting
- Exponentially large graph, yet mixes in polynomial time $\approx O(\log N)$ where N is the size of the graph

Recap: Graph coloring



Given an undirected graph with max. degree Δ and k colors
Goal: generate a k-coloring uniformly at random

This is presumably harder than deciding if there is a k-coloring
Nevertheless, the following random walk has a stationary distribution uniform over all k-colorings:

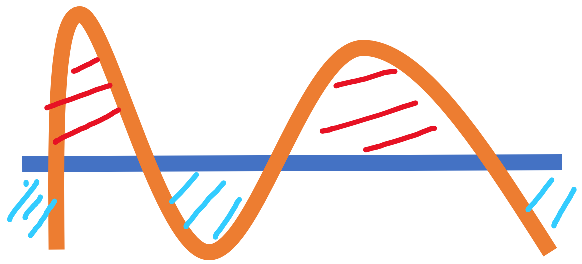
- Start with any k-coloring σ
- Pick a vertex v and a color c uniformly at random, recolor v with c if it is legal; otherwise do nothing;

This Markov chain is irreducible provided that $k \geq \Delta + 2$, and aperiodic

We prove rapid mixing assuming $k \geq 4\Delta + 1$, based on a coupling argument, and explain ideas for $k \geq 2\Delta + 1$
State of the art: $k \geq (\frac{11}{6} - \epsilon)\Delta$ for a small ϵ, or $k \geq \Delta + 3$ for sufficiently large girth graphs

This is known as the Metropolis chain
Other chains: Glauber dynamics, Wang–Swendsen–Kotecký chain, …

Coupling of two distributions



Given distributions p and q over $[n]$, a **coupling** between them is a joint distribution μ over $[n] \times [n]$ such that the marginals are p and q, respectively:

$$\sum_{j \in [n]} \mu(i,j) = p(i)$$
$$\sum_{i \in [n]} \mu(i,j) = q(j)$$

Independently joining p and q is obviously a coupling. More interesting are when they are not independent.

Theorem

For any distributions p and q, and any coupling μ between them, $d_{TV}(p,q) \leq \Pr_{(X,Y)\sim\mu}[X \neq Y]$

Furthermore, there is a coupling μ such that $d_{TV}(p,q) = \Pr_{(X,Y)\sim\mu}[X \neq Y]$

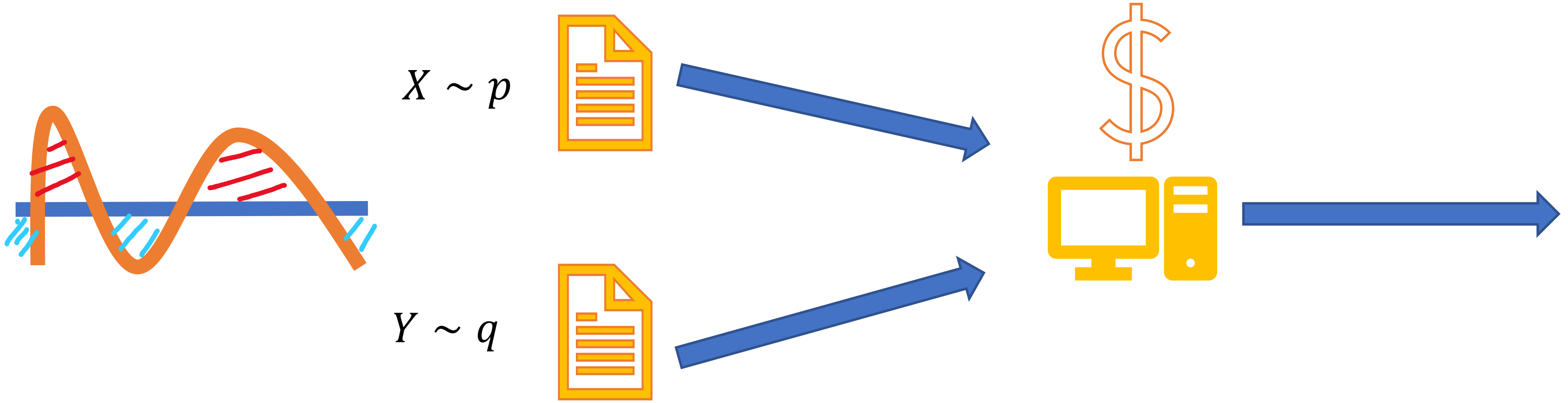
Intuitively, the best we can do is to make the random variables equal in the overlapping regions, that is, $\min\{p_i, q_i\}$; then with the remaining probability, they must be unequal.

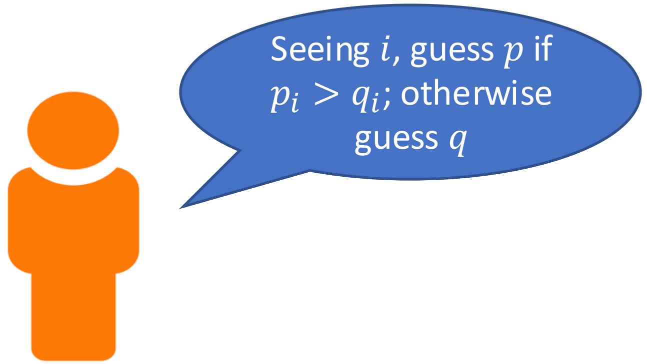
Note that the region in red, and the region in light blue have the same area.

Coupling vs Indistinguishing game

TV distance is also known as statistical distance

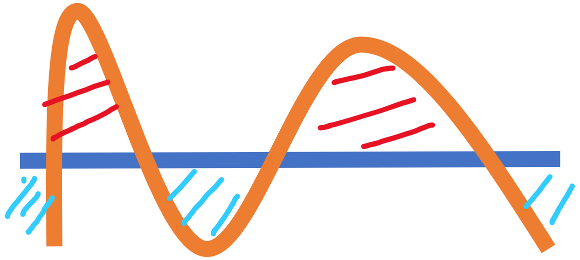
- A game to distinguish two distributions p and q over $[n]$
- Player A draw a sample $X \sim p$ and a sample $Y \sim q$
- Player A flips a fair coin to decide which sample to send to Player B
- Player B now needs to guess which distribution does it came from





Maximum likelihood!
Chance of error = TV distance!

Coupling of two random walks



Let (X_t) and (Y_t) be two copies of a Markov chain over $[n]$. A coupling between them is a joint **process** (X_t, Y_t) over $[n] \times [n]$ such that

1. Marginally, viewed in isolation, (X_t) and (Y_t) are both copies of the original chain
2. $X_t = Y_t \Rightarrow X_{t+1} = Y_{t+1}$

Basically, one can think of two random walkers on the same graph G

In isolation, they each behave faithfully as a random walk on G

But their moves could be dependent

The coupling technique is to design a joint moving process, such that

- The two random walkers meet quickly
- Once they meet, they make identical moves thereafter

Then by the coupling theorem, we know that the time they meet will roughly be an upperbound of mixing time

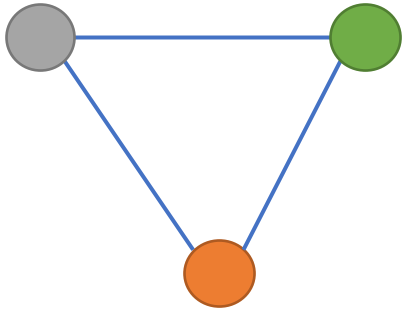
Random walk on the hypercube

- Start with $\sigma \in \{0,1\}^n$
- Pick a coordinate $i \in [n]$ u.a.r., and $b \in \{0, 1\}$ u.a.r.
- Update $\sigma_i = b$

To analyze its mixing time, we consider the following coupling
Say we have two arbitrary copies of the Markov chain, (X_t) and (Y_t)
At each step, we let them choose the same coordinate i and same b

Then, the time that they perfectly couple together is exactly the coupon collecting time!
Note that the probability of not collecting the coupon i after r rounds is at most $\left(1 - \frac{1}{n}\right)^r$
By a union bound, the probability of not collecting all the coupons after $n \ln \frac{n}{\epsilon}$ rounds is at most ϵ
So, the ϵ-mixing time for a random walk on the hypercube is $n \ln \frac{n}{\epsilon}$

Coupling for Graph Coloring



- Start with any k-coloring σ
- Pick a vertex v and a color c uniformly at random, recolor v with c if it is legal; otherwise do nothing

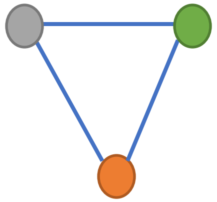
Say we have two arbitrary copies of the Markov chain, (X_t) and (Y_t)

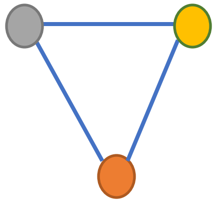
At each step, we let them choose the same vertex v and same color c

Let $d_t = \text{number of vertices } X_t \text{ disagree with } Y_t$

Unlike the previous example, d_t can increase now

We need to consider Good Moves that decrease d_t, and balance them with Bad Moves that increase d_t





Start with any k-coloring σ
Pick a vertex v and a color c u.a.r.,
recolor v with c if legal

Coupling for Graph Coloring

Say we have two arbitrary copies of the Markov chain, (X_t) and (Y_t)

At each step, we let them choose the same vertex v and same color c

Let $d_t = \text{number of vertices } X_t \text{ disagree with } Y_t$

Good Moves that decrease d_t:

If we chose a disagreeing vertex v, and color c does not appear in the neighborhood of v in X_t or Y_t, this is a good move

Because we can safely recolor a disagreeing vertex v with color c, and they agree from then on

Let g_t be the number of good moves (among all possible kn choices)

There are d_t vertices to choose from, and each disagreeing vertex has a neighborhood of at most Δ colors in either process, so each disagreeing vertex has $k - 2\Delta$ "safe colors"

$$g_t \geq d_t(k - 2\Delta)$$

Start with any k-coloring σ
Pick a vertex v and a color c u.a.r.,
recolor v with c if legal

Coupling for Graph Coloring

Say we have two arbitrary copies of the Markov chain, (X_t) and (Y_t)
At each step, we let them choose the same vertex v and same color c
Let $d_t = \text{number of vertices } X_t \text{ disagree with } Y_t$

Bad Moves that increase d_t: a legal move in one process but not the other
This happens when (and only when) the chosen color c is already the color of some neighbor of v in one process but not the other

In other words, v must be a neighbor of some disagreeing vertex u, and c must be the color of u in either X_t or Y_t

Let b_t be the number of bad moves (among all possible kn choices)
There are d_t choices of disagreeing vertex u, then Δ choices for v, then 2 for X_t or Y_t

$$b_t \leq 2\Delta d_t$$

Start with any k-coloring σ
Pick a vertex v and a color c u.a.r.,
recolor v with c if legal

Coupling for Graph Coloring

Say we have two arbitrary copies of the Markov chain, (X_t) and (Y_t)
At each step, we let them choose the same vertex v and same color c
Let $d_t = \text{number of vertices } X_t \text{ disagree with } Y_t$

Combined: $\mathbb{E}[d_{t+1}|d_t] = d_t + \frac{b_t - g_t}{kn} \leq d_t + d_t \frac{4\Delta - k}{kn} \leq d_t \left(1 - \frac{1}{kn}\right)$

Since $d_0 \leq n$, we have $\mathbb{E}[d_t|d_0] \leq 1/e$ for $t = 2k\, n \ln n$. Thus,

$$d_{TV}(p, q) \leq \Pr_{(X_t, Y_t) \sim \mu}[X_t \neq Y_t] \leq \Pr[d_t > 0 | X_0, Y_0] = \Pr[d_t \geq 1 | X_0, Y_0] \leq \mathbb{E}[d_t|d_0] \leq 1/e$$

This concludes that the ϵ-mixing time is $O\left(nk \log \frac{n}{\epsilon}\right)$

To improve this to $k \geq 2\Delta + 1$, one tries to pair bad moves in (X_t) but blocked in (Y_t), with bad moves in (Y_t) but blocked in (X_t)