CS 466/666: Algorithm Design and Analysis **University of Waterloo: Fall 2024**

Lecture 24

December 3, 2024

Instructor: Sepehr Assadi

Topics of this Lecture

1 Convergence to Stationary Distribution: Mixing Time

We are going to conclude our course by considering one last topic related to random walks, the *mixing time*: how long does it take for a random walk starting from some arbitrary distribution on the states to "get close enough" to the stationary distribution? Recall that the stationary distribution is a distribution over the states which remains unchanged after taking another step of the random walk.

The first step in addressing this question requires a detour: what does it even mean for two distributions to be close to each other? There are many measures one can use for this purpose, but one of the most standard and useful ones is the *total variation distance* (also called *statistical distance*), which we will define next and use throughout this lecture.

1.1 Detour: Total Variation Distance

The total variation distance between any two distributions μ and ν over the same (discrete) domain $[n]$ is defined as:

$$\Delta_{\text{tvd}}(\mu, \nu) := \frac{1}{2} \cdot ||\mu - \nu||_1 = \frac{1}{2} \cdot \sum_{i=1}^{n} |\mu(i) - \nu(i)|. \tag{1}$$

(Here, $||\cdot||_1$ is the (standard) ℓ_1 norm of the given vector – recall that we can treat a distribution μ as simply a vector in $\mathbb{R}^n$ with non-negative entries and $||\mu||_1 = 1$.)

Total variation distance is a widely used distance metric between two distributions. In the following, we are going to establish some illustrating properties of this distance.

The first property is that total variation distance between two distributions is equal to the maximum gap possible between probability assigned to any event under these two distributions. Formally,

Proposition 1. *The following is true for any pairs of distributions μ and ν on $[n]$:*

$$\Delta_{tvd}(\mu, \nu) = \max_{\Omega \subseteq [n]} \{\mu(\Omega) - \nu(\Omega)\} = \max_{\Omega \subseteq [n]} \{\nu(\Omega) - \mu(\Omega)\}.$$

Proof. Let $\Omega \subseteq [n]$ be $\arg\max\{\mu(\Omega) - \nu(\Omega)\}$. Note that for any $i \in \Omega$, $\mu(i) \geqslant \nu(i)$ as otherwise the event $\Omega \setminus \{i\}$ will have a higher gap in probabilities between μ and ν, contradicting the choice of Ω. Moreover,

$$\sum_{i \in \Omega} \mu(i) - \nu(i) = \sum_{i \notin \Omega} \nu(i) - \mu(i),$$

as $\sum_i \mu(i) = \sum_i \nu(i) = 1$. This way, by Eq (1), we have,

$$\Delta_{\text{tvd}}(\mu, \nu) = \frac{1}{2} \cdot \left(\sum_{i \in \Omega} \mu(i) - \nu(i) + \sum_{i \notin \Omega} \nu(i) - \mu(i) \right) = \sum_{i \in \Omega} \mu(i) - \nu(i) = \mu(\Omega) - \nu(\Omega),$$

as desired. The second part of the equation holds by considering the event $\bar{\Omega}$ (or alternatively noting that TVD is a symmetric function of its inputs). □

Proposition 1 is useful when one wants to bound the probability of an event over a "complicated" probability distribution which is *close* in total variation distance to some "nicer" probability distribution. We can bound the probability of the event under the nicer distribution and get a bound for the original distribution as well which can only be looser compared to the actual bound by the distance between the two distributions. The following proposition extend this reasoning to random variables as well.

Proposition 2. *The following is true for any pairs of distributions μ and ν on $[n]$, and any random variable X with domain $[n]$:*

$$|\mathop{\mathbb{E}}_{\mu}[X] - \mathop{\mathbb{E}}_{\nu}[X]| \leqslant \Delta_{tvd}(\mu, \nu) \cdot \max|X|.$$

Proof. Recall that a random variable X is simply a function from $[n]$ to the range $[-R, R]$ for $R := \max|X|$. We thus have,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mu}[X] &= \sum_{i=1}^{n} \mu(i) \cdot X(i) = \sum_{i:\mu(i) \geqslant \nu(i)} \mu(i) \cdot X(i) + \sum_{i:\mu(i) < \nu(i)} \mu(i) \cdot X(i) \\
&= \sum_{i:\mu(i) \geqslant \nu(i)} (\mu(i) \cdot X(i) - \nu(i) \cdot X(i)) + \sum_{i:\mu(i) \geqslant \nu(i)} \nu(i) \cdot X(i) + \sum_{i:\mu(i) < \nu(i)} \mu(i) \cdot X(i) \\
&\qquad \text{(by adding and subtracting the term } \textstyle\sum_{i:\mu(i) \geqslant \nu(i)} \nu(i) \cdot X(i)\text{)} \\
&\leqslant \sum_{i:\mu(i) \geqslant \nu(i)} (\mu(i) \cdot X(i) - \nu(i) \cdot X(i)) + \sum_{i:\mu(i) \geqslant \nu(i)} \nu(i) \cdot X(i) + \sum_{i:\mu(i) < \nu(i)} \nu(i) \cdot X(i) \\
&\qquad \text{(by the condition on the indices in the last summation)} \\
&= \sum_{i:\mu(i) \geqslant \nu(i)} (\mu(i) \cdot X(i) - \nu(i) \cdot X(i)) + \mathop{\mathbb{E}}_{\nu}[X] \\
&\qquad \text{(as the last two terms sum up to the expectation of } X \text{ under } \nu\text{)} \\
&\leqslant \max|X| \cdot \sum_{i:\mu(i) \geqslant \nu(i)} (\mu(i) - \nu(i)) + \mathop{\mathbb{E}}_{\nu}[X] \qquad \text{(by factoring out maximum value for } X\text{)} \\
&= \max|X| \cdot \Delta_{\text{tvd}}(\mu, \nu) + \mathop{\mathbb{E}}_{\nu}[X]. \qquad \text{(as proven in Proposition 1)}
\end{aligned}$$

A symmetric argument proves the other direction as well, concluding the proof of the proposition. □

Finally, we show that total variation distance is also a good measure for distinguishing two probability distributions given a single sample from them.

Proposition 3. *Let μ and ν be two probability distributions over $[n]$. Suppose one is given a sample s chosen with probability $1/2$ from μ and with probability $1/2$ from ν. Then, the best probability of success in deciding whether s was sampled from μ or ν is exactly $1/2 + 1/2 \cdot \Delta_{tvd}(\mu, \nu)$.*

Proof. Firstly, it is easy to see that the best estimator for the origin of s is the *maximum likelihood estimator* (MLE) that returns μ if $\mu(s) > \nu(s)$, ν if $\nu(s) > \mu(s)$, and either of μ or ν arbitrarily whenever $\mu(s) = \nu(s)$ (for simplicity, we assume in this case also, the algorithm always return ν). We omit the proof.

Let $\Omega \subseteq [n]$ be the set of elements with $\mu(s) > \nu(s)$. Whenever we get $s \in \Omega$, the estimator returns μ and otherwise it returns ν. Thus, the error happens when $s \in \Omega$ is sampled from ν or $s \notin \Omega$ is sampled from μ. The probabilities of these events can be calculated as:

$$\text{for } s \in \Omega: \qquad \Pr(s \text{ chosen from } \nu \mid s) = \Pr(s \mid s \text{ chosen from } \nu) \cdot \frac{\Pr(s \text{ chosen from } \nu)}{\Pr(s)} = \nu(s) \cdot \frac{1}{2 \cdot \Pr(s)}.$$

$$\text{for } s \notin \Omega: \qquad \Pr(s \text{ chosen from } \mu \mid s) = \Pr(s \mid s \text{ chosen from } \mu) \cdot \frac{\Pr(s \text{ chosen from } \mu)}{\Pr(s)} = \mu(s) \cdot \frac{1}{2 \cdot \Pr(s)},$$

where in both case we use the Bayes' rule. As a result, we have

$$\begin{aligned}
\Pr(\text{error}) &= \sum_{s=1}^{n} \Pr(\text{error} \mid s) \cdot \Pr(s) = \frac{1}{2} \cdot \left(\sum_{s \in \Omega} \nu(s) + \sum_{s \in [n] \setminus \Omega} \mu(s) \right) \\
&= \frac{1}{2} \cdot \sum_{s=1}^{n} \min\{\mu(s), \nu(s)\} = \frac{1}{2} \cdot \left(\sum_{s=1}^{n} \frac{\mu(s) + \nu(s) - |\mu(s) - \nu(s)|}{2} \right) \qquad (\text{as } \min\{a,b\} = \tfrac{a+b-|a-b|}{2}) \\
&= \frac{1}{2} \cdot (1 - \Delta_{\text{tvd}}(\mu, \nu)). \qquad (\text{as } \textstyle\sum_{s=1}^{n} \mu(s) = \sum_{s=1}^{n} \nu(s) = 1 \text{ and by Eq (1)})
\end{aligned}$$

This concludes the proof. □

1.2 Mixing Time in Random Walks

We are now ready to go back to our study of random walks. Recall that a Markov chain consists of n states and a probability matrix—called the *transition matrix*—$P \in \mathbb{R}^{n \times n}$, where P_{ij} denotes the probability that the next state of the random walk is j assuming the current state is i. Recall the following definition:

- The **stationary distribution** is a distribution π over the states which remains unchanged after taking another step of the random walk, i.e., $\pi = \pi \cdot P$.

Also recall that if we start from some distribution x over the states and take a t step random walk, the distribution of the state the walker can be computed as $x \cdot P^t$.

We are interested in determining the minimum length t we need to take from any starting state, until the distribution of the state of the walker gets close to the stationary distribution of the random walk (in total variation distance). Formally,

Definition 4. For any parameter $\delta \in (0, 1)$, we define the **δ-mixing time** of a Markov chain $P \in \mathbb{R}^{n \times n}$ with the stationary distribution π as:

$$\arg\min_{t \geqslant 1} \Big(\forall s \in [n] \quad \Delta_{\text{tvd}}(\mathbf{1}_s \cdot P^t, \pi) \leqslant \delta \Big).$$

Now that we defined mixing times formally, we will review a motivating application.

1.3 A Motivating Application: Sampling Colorings

Let $G = (V, E)$ be any undirected graph with maximum degree Δ. Recall that a k-vertex coloring of G is any assignment of k colors to the vertices of G so that no edge receives the same color on both its endpoints. Already in the first lecture of the course, we saw a simple argument that shows as long as $k \geqslant \Delta + 1$, we can always find such a coloring (and very simply, using a greedy algorithm). But, can we also sample a k-coloring of G uniformly at random? In general, a graph can have way too many proper k-colorings (exponentially many), and we cannot simply list all the colorings of G and then sample one chosen randomly. So, what can we do instead?

A simple strategy is to set up a Markov Chain P whose states correspond to proper k-colorings of G and whose stationary distribution π is uniform over the states. Then, we simply start from some state of this Markov chain, namely, any k-coloring of G that we can find, and then run a random walk – as long as the random walk mixes "fast enough", we will be be able to sample a k-coloring of G "almost" uniformly at random (namely, from some distribution with small TVD from uniform).

Let us consider the following Markov chain, whose states are the k-colorings of G, with the following transition at each state:

A Markov chain for sampling k-colorings:

- The states of the Markov chain correspond to k-colorings of G.
- Suppose we are at a fixed k-coloring C of G; the random transition to the next step happens as follows:
  1. Sample a vertex $v \in V$ and color $c \in [k]$ uniformly at random;
  2. if changing the color of v to c still leads to a proper k-coloring of G, then, update the coloring C (namely, move to this new state); otherwise, repeat (namely, stay at this state).

Note that as long as $k \geqslant \Delta + 1$, we will always be able to find a k-coloring and start our Random Walk in the above Markov chain. However, when $k = \Delta + 1$, this Markov chain by not be connected (specifically, when G is a $(\Delta + 1)$-clique, no state has any change to another state). But, it is easy to see that when $k \geqslant \Delta + 2$, the Markov chain is connected and its underlying graph is also not bipartite (more formally, the Markov chain is irreducible and aperiodic). Thus, for $k \geqslant \Delta + 2$, by the Fundamental Theorem of Markov Chains, we know there is a stationary distribution π for our Markov chain and that if we run the random walk long enough, in the limit, our walk is going to mix to its stationary distribution.

There are two questions left though still: what is the stationary distribution π of this Markov chain, and what is its mixing time? It turns out the first one is easy to answer also. Notice that in this Markov chain, for any pair of states i and j,

$$P(i, j) = P(j, i) \in \left\{0, \frac{1}{nk}\right\};$$

either, the two states change in a single vertex with two different colors assigned to the vertex, and both colors are valid for the vertex, in which case the transition probability is $1/nk$; otherwise, the two states are not neighboring and the transition probability is 0. We can prove that stationary distribution of any such Markov chain is uniform distribution.

Lemma 5. *Suppose P is a transition matrix of an irreducible and aperiodic Markov chain such that for every pairs of states i, j, we have $P(i, j) = P(j, i)$; then, the stationary distribution π of P is uniform.*

Proof. Let σ be the uniform distribution over the states. For any state j, if we sample a state from σ and take one step of the random walk, then,

$$\Pr\left(\text{the chain is in state } j\right) = \sum_i \sigma_i \cdot P(i,j) = \sum_i \sigma_i \cdot P(j,i) = \sum_i \sigma_j \cdot P(j,i) = \sigma_j \cdot \sum_i P(j,i) = \sigma_j,$$

where the second equality is by the condition on P in the lemma statement, and the last one is because the probability out of a state is equal to one.

This means that σ satisfies $\sigma = \sigma \cdot P$ and since the stationary distribution is unique in this case, we get that $\pi = \sigma$ and is thus uniform. □

Thus, the last step for solving our problem is to bound the mixing time of this Markov chain (with error, say, $1/\text{poly}(n)$ or even $1/2^{\Omega(n)}$). To do this, we need to introduce a technique for bounding mixing times.

2 Coupling and Bounding Mixing Times

There are various techniques for upper bounding mixing time of Markov chains and we will not be discussing most of them in this course. For this lecture, we present a simple yet quite powerful way for this purpose using the notion of *coupling*, which requires another detour of its own.

2.1 Detour: Coupling

Coupling is a general technique for relating two different distributions/random variables, which among other things, allows us to bound the total variation distance between the variables.

Definition 6. Let X and Y be two different random variables over the same state space Ω. We say a random variable on the state space $\Omega \times \Omega$, denoted by $Z := (Z_1, Z_2)$, is a **coupling** of X and Y iff the following holds: the marginal distribution of Z_1 is X and the marginal distribution of Z_2 is Y (but their joint distribution can be anything). In other words, for any $\omega \in \Omega$:

$$\Pr_X\left(X = \omega\right) = \Pr_Z\left(Z_1 = \omega\right) = \sum_{\omega'} \Pr_Z\left(Z_1 = \omega \wedge Z_2 = \omega'\right);$$
$$\Pr_Y\left(Y = \omega\right) = \Pr_Z\left(Z_2 = \omega\right) = \sum_{\omega'} \Pr_Z\left(Z_2 = \omega' \wedge Z_2 = \omega\right);$$

For example, suppose X, Y are both uniform distribution over $\{0, 1\}$. The following are all couplings of X and Y:

- Let $Z = (X, Y)$ chosen independent of each other.
- Let $Z = (X, 1 - X)$.
- Or, let $Z = (X, X)$.

In each case, it is easy to see that Z is indeed a valid coupling of X and Y.

The following is the key property of couplings we are interested in.

Proposition 7. *For any coupling $Z = (Z_1, Z_2)$ of random variables X and Y,*

$$\Delta_{tvd}(X, Y) \leqslant \Pr_Z\left(Z_1 \neq Z_2\right).$$

Proof. We use Proposition 1. Let E be any event in the space of X and Y. We have,

$$\begin{aligned}
\Pr_X(X \in E) &= \Pr_Z(Z_1 \in E) && \text{(by the definition of coupling)} \\
&= \Pr_Z(Z_1 \in E \wedge Z_1 = Z_2) + \Pr_Z(Z_1 \in E \wedge Z_1 \neq Z_2) \\
&= \Pr_Z(Z_2 \in E \wedge Z_1 = Z_2) + \Pr_Z(Z_1 \in E \wedge Z_1 \neq Z_2) \\
& && \text{(since we already consider } Z_1 = Z_2 \text{ in the first term)} \\
&\leqslant \Pr_Z(Z_2 \in E) + \Pr_Z(Z_1 \neq Z_2) && \text{(by dropping one event from each of the joint probabilities)} \\
&= \Pr_Y(Y \in E) + \Pr_Z(Z_1 \neq Z_2). && \text{(again, by the definition of coupling)}
\end{aligned}$$

Thus, for any event E, we have,

$$\Pr_X(X \in E) - \Pr_Y(Y \in E) \leqslant \Pr_Z(Z_1 \neq Z_2),$$

which, by Proposition 1, implies the same upper bound on $\Delta_{\text{tvd}}(X, Y)$, concluding the proof. □

2.2 Using Coupling for Bounding Mixing Times

We now introduce a general technique for bounding mixing time of a Markov chain using coupling. Suppose we have a Markov chain and we define two random walks

$$(X_0, X_1, X_2, \ldots, X_t, \ldots) \quad \text{and} \quad (Y_0, Y_1, Y_2, \ldots, Y_t, \ldots),$$

on this Markov chain, which can start from some arbitrary distribution X_0 or Y_0, and then X_t (resp. Y_t) is obtained by performing t step of the random walk. Consider the following type of coupling $(Z_0, Z_1, Z_2, \ldots, Z_t, \ldots)$:

- Each Z_t is a tuple $Z_t := (Z_{t,1}, Z_{t,2})$, where $Z_{t,1}$ is marginally distributed as X_t and $Z_{t,2}$ is marginally distributed as Y_t;
- If at some point t, we have, $Z_{t,1} = Z_{t,2}$, then, from thereon, for all $t' \geqslant t$, we also have $Z_{t',1} = Z_{t',2}$.

In words, random variables Z's keep track of *two* random walks on the same Markov chain that may start at different states, each marginally moves according to the transition matrix of the Markov chain, albeit possibly in a correlated way, and moreover, if at any point these two walks collide in the same state, then, from thereon, they will move in sync with each other.

Example. Consider the hypercube $\{0,1\}^n$ and the following Markov chain: each state is a string $x \in \{0,1\}^n$, and whenever we are at a state, with probability $1/2$, we remain at the same state, and with the remaining probability, we uniformly at random go to one of the n strings y which differ from x in exactly one index. Suppose X_0 and Y_0 are two different distributions on states of this Markov chain, and each performing a random walk $X_0, X_1, \ldots, X_t, \ldots$ and $Y_0, Y_1, \ldots, Y_t, \ldots$ according to this Markov chain. Consider defining the coupling Z as follows:

- Sample $i \in [n]$ and $\sigma \in \{0,1\}$ uniformly at random.
- For any state $Z_t = (Z_{t,1}, Z_{t,2})$ of the Markov chain, we move to the state $Z_{t+1} = (Z_{t+1,1}, Z_{t+1,2})$ as follows: the state/string $Z_{t+1,1}$ (resp. $Z_{t+1,2}$) is obtained from the state/string $Z_{t,1}$ (resp. $Z_{t,2}$) by switching its i-th index with the bit σ instead.

We can easily verify that random variables $Z_0, Z_1, \ldots, Z_t, \ldots$ form a coupling of the two random walks.

We can now go back to see how to use this type of coupling for bounding the mixing time. The following proposition is the key to this approach.

Proposition 8. *Suppose* $(X_0, X_1, X_2, \ldots, X_t, \ldots)$ *is a random walk starting from some arbitrary state* s *in a Markov chain,* $(Y_0, Y_1, Y_2, \ldots, Y_t, \ldots)$ *is a random walk according to the stationary distribution* π *on the Markov chain, and* $(Z_0, Z_1, Z_2, \ldots, Z_t, \ldots)$ *is a coupling of these random walks as prescribed above. Then, the* δ*-mixing time of the Markov chain is upper bounded by*

$$\arg\min_{t \geqslant 1} \Big(\Pr(Z_{t,1} \neq Z_{t,2}) \leqslant \delta \Big).$$

Proof. Let t be the smallest index such that $\Pr(Z_{t,1} \neq Z_{t,2}) \leqslant \delta$. By the extra condition of the coupling of the random walks, this implies that the same upper bound on the probability also holds for all $t' \geqslant t$. Moreover, by Proposition 7, for every $t' \geqslant t$, we have $\Delta_{\text{tvd}}(X_{t'}, Y_{t'}) \leqslant \delta$. The distribution of Y_t is the stationary distribution π for all t, and the distribution of X_t is $\mathbf{1}_s \cdot P^t$ where t is the transition matrix of the Markov chain. As such, we have,

$$\Delta_{\text{tvd}}(\mathbf{1}_s \cdot P^t, \pi) \leqslant \delta,$$

proving the upper bound of t on the mixing time. □

Example. Going back to our example from above on the random walk on the hypercube, we can bound the mixing time of the random walk easily as follows. For any two random walks (X_t, Y_t)'s, consider the couplings Z_t's defined in the example. Notice that in the couplings Z_t's, whenever we pick an index $i \in [n]$, the value of $Z_{t',1}$ and $Z_{t',2}$ on the index i will be the same for any time step $t' \geqslant t$. Thus,

$$\Pr(Z_{t,1} \neq Z_{t,2}) \leqslant \sum_{i=1}^{n} \Pr(\text{index } i \text{ is never chosen in transitions } Z_1, Z_2, \ldots, Z_t) = n \cdot \left(1 - \frac{1}{n}\right)^t.$$

Thus, by Proposition 8, the δ-mixing time of this random walk is upper bounded by $n(\ln n + \ln(1/\delta))$.

2.3 Conclusion: Mixing Time for Sampling Colorings

We are now ready to conclude our motivating application of sampling a k-coloring of a given graph $G = (V, E)$ with maximum degree Δ. We will be proving the following result:

Theorem 9. *For any* $k > 4\Delta$*, there is a polynomial time algorithm for sampling a* k*-coloring of any given graph* G *with maximum degree* Δ *wherein the TVD between the distribution of the sampled coloring and uniform distribution is at most* $\exp(-\text{poly}(n))$.

Proof. Consider the two random walks $X_0, X_1, \ldots, X_t, \ldots$ and $Y_0, Y_1, \ldots, Y_t, \ldots$ according to the Markov chain described earlier for this problem. Let $Z_0, Z_1, \ldots, Z_t, \ldots$ be the following coupling of these walks:

- Sample a vertex $v \in V$ and a color $c \in [k]$ uniformly at random.
- For any state $Z_t = (Z_{t,1}, Z_{t,2})$ of the Markov chain, we move to the state $Z_{t+1} = (Z_{t+1,1}, Z_{t+1,2})$ as follows: the state/coloring $Z_{t+1,1}$ (resp. $Z_{t+1,2}$) is obtained from the state/coloring $Z_{t,1}$ (resp. $Z_{t,2}$) by switching the color of vertex v to c if it leads to a proper coloring, and otherwise keeping the coloring unchanged.

We can easily verify that random variables $Z_0, Z_1, \ldots, Z_t, \ldots$ form a coupling of the two random walks.

Our goal is to bound the probability that $\Pr(Z_{t,1} \neq Z_{t,2})$ and then apply Proposition 8. Define D_t as set of vertices with different colors under $Z_{t,1}$ and $Z_{t,2}$ and $d_t := |D_t|$. We have the following main claim.

Claim 10. *For any choice of* $d_t = d$*, we have*

$$\mathbb{E}[d_{t+1} \mid d_t = d] \leqslant \left(1 - \frac{1}{nk}\right) \cdot d.$$

Before proving this claim, let us see how it concludes the proof of the theorem. We have $d_0 \leqslant n$ deterministically and for every $t \geqslant 0$,

$$\mathbb{E}\left[d_{t+1}\right] = \mathop{\mathbb{E}}_{d \sim d_t} \mathbb{E}\left[d_{t+1} \mid d_t = d\right] \leqslant \mathop{\mathbb{E}}_{d \sim d_t}\left[\left(1 - \frac{1}{nk}\right) \cdot d_t\right] = \left(1 - \frac{1}{nk}\right) \cdot \mathbb{E}\left[d_t\right].$$

Thus, by induction, we have that for every $t \geqslant 1$,

$$\mathbb{E}\left[d_t\right] \leqslant \left(1 - \frac{1}{nk}\right)^t \cdot n.$$

Thus, if we set $k = nk \cdot (\ln n + \ln 1/\delta)$, we get that

$$\mathbb{E}\left[d_t\right] \leqslant \exp\left(-\frac{nk \cdot (\ln n + \ln 1/\delta)}{nk}\right) \cdot n = \delta.$$

This in turn implies that

$$\Pr\left(Z_{t,1} \neq Z_{t,2}\right) = \Pr\left(d_t \geqslant 1\right) \leqslant \mathbb{E}\left[d_t\right] \leqslant \delta,$$

where the first inequality is by Markov bound. Hence, for any $\delta = \exp\left(-\text{poly}(n)\right)$, we obtain that the δ-mixing time is polynomial in n, k, which proves the theorem since the steps of the random walk (and its starting state) can be computed in polynomial time as well. It is only left to prove Claim 10.

Proof of Claim 10. Fix the colorings $Z_{t,1}$ and $Z_{t,2}$ which fixes $d_t = d$ as well. We say that a pair of vertex-color (v, c) is *good* if choosing the vertex v and color c randomly in the transition to Z_{t+1} allows for updating both states $Z_{t,1}$ and $Z_{t,2}$ to receive the color c on vertex v, and that previously, v was colored differently between the two colorings. Thus, picking a good (v, c) pair results in d_{t+1} to decrease by one with respect to d_t. We have

$$\begin{aligned}\Pr\left((v, c) \text{ is good}\right) &= \sum_{u \in D_t} \Pr\left(v = u\right) \cdot \sum_{c \in [k]} \Pr\left(c \text{ has not appeared in } N(u) \text{ in either } Z_{t,1} \text{ nor } Z_{t,2}\right) \\ &\geqslant \frac{d_t}{n} \cdot \left(1 - \frac{2\Delta}{k}\right),\end{aligned}$$

since there can be at most 2Δ different colors in the neighborhood of v across both colorings.

On the other hand, we say a pair of vertex-color (v, c) is *bad* if choosing the vertex v and color c randomly in the transition to Z_{t+1} leads to the previously same-colored vertex v to receive a different color in the states $Z_{t+1,1}$ and $Z_{t+1,2}$; this can only happen if v is colored by c in one of the two walks but not the other one. Thus, picking bad (v, c) pair results in d_{t+1} to increase by one with respect to d_t. Note that a vertex v in a bad vertex-color pair (v, c) should always be incident on a vertex $u \in D_t$, and then the color c should be one of colors of u in either $Z_{t,1}$ or $Z_{t,2}$. Thus,

$$\Pr\left((v, c) \text{ is bad}\right) \leqslant \sum_{u \in D_t} \Pr\left(v \in N(u)\right) \cdot \Pr\left(c \text{ is the color of } u \text{ in } Z_{t,1} \text{ or } Z_{t,2}\right) = \frac{d_t \cdot \Delta}{n} \cdot \frac{2}{k}.$$

(by union bound)

Finally, note that there are also vertex-color pairs that are neither good nor bad when sampled by the coupling random walk, but in those cases, value of d_{t+1} remains the same as d_t. As such, we have,

$$\mathbb{E}\left[d_{t+1} \mid d_t = d\right] = d - \frac{d}{n} \cdot \left(1 - \frac{2\Delta}{k}\right) + \frac{d \cdot \Delta}{n} \cdot \frac{2}{k} = d - \frac{d}{n} \cdot \left(1 - \frac{4\Delta}{k}\right) \geqslant d - \frac{d}{nk} = \left(1 - \frac{1}{nk}\right) \cdot d,$$

using the fact that $k \geqslant 4\Delta + 1$. □

This concludes the entire proof of Theorem 9 as well. □