Section 11

1. **(KL Divergence)** Throughout the chapter on coupling and mixing times, we have used the total-variation distance between discrete probability distributions. Here we will show a motivation for the KL-divergence, another widely used way to quantify differences between probability distributions. The KL-divergence is defined for distributions P, Q on the same discrete space $\mathcal{X}$ as:

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log(\frac{P(x)}{Q(x)}) \tag{1}$$

where $P(x), Q(x)$ calculates the probability mass of x under P, Q respectively. The KL-divergence is non-negative and only 0 if $P = Q$.

   (a) What are some disadvantages of this quantity compared to the total-variation distance?

   (b) Now, suppose we have an i.i.d. sample $(X_1, ..., X_T)$ over a finite probability space $\mathcal{X}$. The random variables X_i follow one of two distributions μ, ν, but we do not know which beforehand. We are interested in choosing which distribution is more plausible given these samples.

   A reasonable approach is to compute the probability of the sample under each distribution and choose the distribution with higher probability. Show that this condition is equivalent to picking μ if and only if:

$$F = \frac{1}{T} \sum_{i=1}^{T} \log \frac{\mu(X_i)}{\nu(X_i)} > 0 \tag{2}$$

   (c) What is $\mathbb{E}[F]$ in terms of KL-divergences?

   (d) What is an advantage of KL-divergence compared to total-variation distance? (Hint: consider a setting where $P(x), Q(x)$ can be calculated efficiently but the sample space $\mathcal{X}$ is very large.)

2. **(Coupling) (MU Exercise 12.9)** Consider a Markov chain on n points $[0, n-1]$ lying in order on a circle. At each step, the chain stays at the current point with probability $1/2$ or moves to the next point in the clockwise direction with probability $1/2$. Find the stationary distribution and show that, for any $\varepsilon > 0$, the mixing time $\tau(\varepsilon)$ is $O(n^2 \ln(1/\varepsilon))$.

3. **(Coupling) (MU Theorem 12.8)** Recall that in lecture, we considered a Markov chain approach to sample proper colorings of a graph uniformly. The Markov chain is to start with some proper coloring, then, at each step, pick a vertex and color at random, and change the vertex to that color if the coloring stays proper. Using a coupling argument, we derived a bound on the mixing time when we use a number of colors $c \geq 4\Delta + 1$, where Δ is the maximum degree of the graph. We will improve the coupling argument to reduce the requirement to $c \geq 2\Delta + 1$.

   In the original proof, we defined the set D_t as the set of vertices with different colors in the two copies of the Markov chain at time t. We now additionally define A_t as the set of vertices with matching colors in the two copies of the Markov chain at time t. Also define $|D_t| = d_t$. Our coupling between the chains will differ depending on which set a sampled vertex is in.

(a) Define $d'_t(v)$ to be the the number of vertices adjacent to v that are in the opposite set. Concretely, if $v \in D_t$, $d'_t(v)$ counts the number of vertices adjacent to v that are in A_t, and vice versa if $v \in A_t$.

Denote $m = \sum_{v \in A_t} d'_t(v) = \sum_{v \in D_t} d'_t(v)$. Why are the two sums equal?

(b) Our improved coupling will still involve sampling a single random vertex v for both copies of the Markov chain. If $v \in D_t$, we will sample the same random color for both chains (the same coupling as before). Show the following bound, which is improved over the original proof:

$$P(d_{t+1} = d_t - 1 | d_t > 0) \geq \frac{1}{cn}((c - 2\Delta)d_t + m) \tag{3}$$

(c) Now suppose $v \in A_t$. We will improve the coupling by changing the color correspondence between the two copies of the chain. Give a brief explanation of how this could help.

(d) Using an improved color correspondence, show that when we pick $v \in A_t$:

$$P(d_{t+1} = d_t + 1 | d_t > 0) \leq \frac{m}{cn} \tag{4}$$

(e) Follow the same steps as in the original proof to conclude that the variation distance is at most ε after:

$$t = \lceil \frac{nc}{c - 2\Delta} \ln(\frac{n}{\varepsilon}) \rceil \tag{5}$$