Recent results in planted assignment problems
Homework 1

1. (Large deviation bound) Let X_i's and Y_i's denote two dependent sequences of random varialbles, where X_i's are i.i.d. copies of $\mathrm{d}P/\mathrm{d}Q$ under distribution P and Y_i's are i.i.d. copies of $\mathrm{d}P/\mathrm{d}Q$ under distribution Q.

   (a) Using the Chernoff bound to prove that
   $$\mathbb{P}\left[\sum_{i=1}^{\ell}(Y_i - X_i) \geq 0\right] \leq \exp(-\ell\alpha),$$
   where $\alpha = -2\log B(P,Q)$ and $B(P,Q) = \int \sqrt{\mathrm{d}P\mathrm{d}Q}$.

   (b) Generalize the above proof to show that
   $$\mathbb{P}\left[\sum_{i=1}^{\ell}(Y_i - X_i) \geq x\ell\right] \leq \exp\left(-\ell(\alpha + x/2)\right).$$

   (c) Prove that when $\sqrt{d}B(P,Q) = 1+\epsilon$, with high probability,
   $$\frac{\mu_W(\mathcal{M}_{\text{near}})}{\mu_W(M^*)} \leq e^{O(\epsilon\delta n)},$$
   where $\mathcal{M}_{\text{near}} = \{M : |M \Delta M^*| < 2\delta n\}$ and μ_W is the posterior distribution of M^* conditioned on the observation W. (Hint: use the first moment method with proper truncation based on part (b) result.)

2. (Planted k-factor model) We consider the following generalization of the planted matching model, see https://arxiv.org/pdf/2010.13700.pdf. Let K_n denote the complete graph on vertex set $[n]$. Let M^* denote a k-factor of K_n, i.e., a spanning k-regular subgraph, that is uniformly chosen at random. Then we generate a weighted graph G as follows. First, connect all node pairs that are connected in M^*. For every node pair that is not connected in M^*, connect them independently with probability d/n. For edges in M^*, the edge weights are drawn independently according to P. The weights for remaining edges are drawn independently from Q.

   Prove that when $\sqrt{kd}B(P,Q) \leq 1$, the maximum likelihood estimator achieves almost perfect recovery of M^*, that is, the fraction of misclassfied edges is $o(1)$ with high probability.

3. (Planted matching model - Gaussian) Consider the planted matching model with Gaussian edge weights drawn from either $P = N(\mu, 1)$ or $Q = N(0,1)$ on a complete bipartite graph.

   (a) Prove that $\sqrt{n}B(P,Q) = 1$ simplifies to $\mu^2 = 4\log n$;

   (b) Prove that when $\mu^2 \geq (2+\epsilon)\log n$, a simple algorithm based on thresholding edge weights achieves almost perfect recovery with high probability;

   (c) Prove that almost perfect recovery is impossible if $\mu^2 \leq (2-\epsilon)\log n$ using the mutual information argument outlined on page 15 of https://arxiv.org/pdf/1806.00118.pdf;

   (d) Prove that the maximum likelihood estimator achieves almost prefect recovery when $\mu^2 \geq (2+\epsilon)\log n$ with an improved analysis.

4. (Planted matching on hypergraphs: open-ended) Read the following paper https://arxiv.org/pdf/2209.03423.pdf, which studies an extension of the planted matching model to hypergraphs. Can you identify and prove some phase transtion threshold?